Leveraging passenger mutations to guide cancer treatment and prevention

Carino Gurjao

Université de Paris

Thèse de doctorat de Université de Paris

Ecole doctorale 562 - Bio Sorbonne Paris Cité

Spécialité de doctorat: Génétique

À l'Institut Cochin

INSERM (U1016), CNRS (UMR8104), Université de Paris (UMR-S1016)

Thèse présentée et soutenue à Paris, le 17 décembre 2021:

Leveraging Passenger Mutations to Guide Cancer Treatment and Prevention

par Carino GURJAO

dirigée par Valentina BOEVA et Leonid MIRNY

Composition du jury:

Valentina BOEVA,CR-HDR, Université de Paris, Directrice de thèse

Leonid MIRNY, Professor, Massachusetts Institute of Technology (MIT), Co-Directeur de thèse

Vassili SOUMELIS, Full professor, Université de Paris, Président du jury

Sabine TEJPAR, Full professor, Universitair Ziekenhuis Leuven (UZ Leuven), Rapportrice

VAN BOXTEL, Principal Investigator, Princess Máxima Center, Rapporteur

Massachusetts Institute of Technology

HARVARD MEDICAL SCHOOL

BROAD INSTITUTE

Dana-Farber Cancer Institute

Abstract (in French)

Titre: Potentialité des mutations *passengers* pour le traitement et la prévention du cancer

Résumé:

Le cancer est caractérisé par une prolifération anarchique de cellules, orchestrée par une poignée de mutations. Parallèlement à ces mutations appelées *driver*s, des dizaines de milliers d'autres, appelées *passengers*, s'accumulent au cours de la progression du cancer. La recherche génétique en cancérologie s'est longtemps consacrée aux mutations *drivers* car celles-ci ont un impact majeur sur la trajectoire de progression du cancer et peuvent être exploitées pour certaines thérapies géniques ciblées. En revanche, les mutations *passengers* jouent individuellement un rôle mineur dans la croissance tumorale et leur analyse a été longtemps négligée.

Cependant, deux caractéristiques des *passengers* les rendent particulièrement pertinentes en oncologie. D'une part, les *passengers* peuvent être présentées en grand nombre à la surface des cellules cancéreuses, rendant ainsi ces dernières plus 'visibles' pour le système immunitaire. D'autre part, les *passengers* s'accumulent de par l'effet de différents processus mutagéniques. Ces derniers laissant chacun des empreintes génétiques uniques, l'analyse des *passengers* peut permettre une meilleure caractérisation des mutagènes humains.

Cette thèse étudie comment le paysage mutationnel du génome tumoral, majoritairement composé de *passengers,* peut être exploité pour le traitement et la prévention du cancer. Cette étude porte principalement sur deux caractéristiques des mutations : leur charge mutationnelle, autrement dit

leur nombre, et leur signature mutationnelle, c'est-à-dire leurs combinaisons uniques dans le génome tumoral.

Le premier chapitre interroge l'utilisation clinique de la charge mutationnelle comme biomarqueur de réponse à l'immunothérapie. Malgré l'approbation récente par la *Food and Drug Administration* (FDA) de ce biomarqueur, les analyses de ce premier chapitre en montrent une utilité très limitée. Par conséquent, l'utilisation clinique de la charge mutationnelle pourrait biaiser l'accès aux traitements immunothérapiques pour certains patients qui pourraient en bénéficier.

Dans le même ordre d'idée, le deuxième chapitre de cette thèse approfondit l'analyse de la charge mutationnelle en s'intéressant au cas d'un patient dont le cancer colorectal n'a pas répondu au traitement immunothérapique. L'analyse génomique, transcriptionnelle et pathologique de la tumeur et de son microenvironnement révèlent une évasion immunitaire à stratégie double. La perte de présentation d'antigènes d'une part, et le microenvironnement immunosuppresseur d'autre part, pourraient expliquer la résistance au traitement pour le cancer de ce patient.

Enfin, le troisième chapitre est consacré à l'analyse de signatures mutationnelles de tumeurs colorectales. Grâce à l'analyse *de novo* de 900 échantillons d'ADN, couplée à de nombreuses données sur le régime alimentaire, ce chapitre démontre l'existence d'une signature mutationnelle alkylante associée à une consommation élevée de viande rouge pré-diagnostic. Cette découverte lie pour la première fois une signature mutationnelle à une composante alimentaire, et peut présenter des enjeux majeurs pour la prévention du cancer.

Mots-clés: Cancer, Mutation, Colorectal, Génomique, Prévention, Immunothérapie, Signature mutationnelle, Charge mutationnelle

Abstract (in English)

Title: Leveraging Passenger Mutations to Guide Cancer Treatment and Prevention

Summary:

Cancer is characterized by the uncontrolled proliferation of cells, driven by a handful of genetic alterations. Alongside these ‘driver’ mutations, tens of thousands of other variants, called ‘passenger’ mutations, accumulate during cancer progression. Genomic oncology research has mostly focused on driver mutations as these have the most impact on tumour evolution and offer attractive opportunities for gene-targeted therapies. On the other hand, passenger mutations play individually minor roles in tumour growth and have been widely understudied.

However, two features of passenger mutations make them particularly relevant for cancer research. First, passenger mutations can be presented in large numbers at the cell surface, thus making tumor cells more ‘visible’ to the immune system. This feature has made the total number of mutations an attractive biomarker of response to immune-based therapies. Second, passengers are the products of the different mutagenic processes a tumor underwent. As each mutagenic process leaves a characteristic pattern of mutations, the analysis of passengers can therefore help better understand the underlying mechanisms of carcinogenesis.

This thesis investigates how the overall mutational landscape - overwhelmingly represented by passenger mutations - can be leveraged to guide cancer treatment and prevention. In particular, I examine two features of

mutations: their mutational load (meaning their total number) and their mutational signature (in other words their genetic localization patterns).

The first chapter examines the use of high mutational load as a biomarker of response to immunotherapy. Despite the recent Food and Drug Administration (FDA) approval of this biomarker, there is little evidence showing that higher mutational load predicts patient response to immunotherapy. As a result, using a mutational load threshold for clinical decisions regarding immunotherapy could skew access to treatment for patients who may benefit from it.

In the same vein, the second chapter examines the case of a patient whose colorectal cancer (CRC) did not respond to immunotherapy despite harboring a very high mutational load. This intrinsic resistance was examined through genomic, transcriptional, and pathologic characterizations of the patient's tumour and neighboring immune cells. The analysis suggests a two-pronged immune evasion that combines loss of antigen presentation with an immunosuppressive microenvironment, which could explain the lack of response to immunotherapy.

Last, the third chapter focuses on the mutational signatures of CRC tumours, rather than their overall mutational load. The project involved the *de novo* analysis of sequencing data from 900 CRC cases with extensive dietary information. This allowed the discovery of an alkylating mutational signature associated with high intakes of red meat before diagnosis. These results link for the first time a CRC mutational signature to a component of the diet, which has major implications in cancer prevention.

Key words: Cancer, Mutation, Colorectal, Genomics, Prevention, Immunotherapy, Mutational Signature, Tumor Mutational Burden

Acknowledgements

Little about this thesis would have been possible without the support of many wonderful mentors, friends and family members. I hope they find here the expression of my most sincere thanks for helping me become a better scientist and person.

The very idea of pursuing a PhD would not have been conceivable without my two first academic mentors, Valentina and Leonid. Their fostering of a fun and creative lab environment convinced me to opt for an academic career path. I can only dream of recreating this in my own team in the future.

I owe a greater debt than I could ever express to my parents and grandparents and I owe all my achievements to their selflessness. In addition, my siblings have been role models for many aspects that enabled me to reach this dissertation finish line. Neville for his pragmatism and integrity. Dylena for her determination and creativity. I am also fortunate to have a close bond with my extended family. I thank my ‘cousines préferées’ - Daphnis, Diana and Gracy - for always making time for my impromptu phone calls.

I am particularly grateful for the chosen family I created over the years in Paris, Lyon, and Boston.

Alexiane, David, Marjolaine and Maximien, my primary school/ middle school friends I am lucky to still have in my life.

Laure, without whom I would have not met my most important academic deadlines, including this dissertation. I will forever be grateful for that one 5 a.m. presentation rehearsal during undergrad. Thank you for being *that* friend.

Audrey, Nada, Noelie, Quentin and Rosa, my friends from Lyon who supported me through many life-changing steps.

I would not have stayed in the USA if it wasn't for a friendly first contact. Thank you Anna, Billy, Brittney, Dinesh, Emma, Leaf and Sam for being there for my first PB&J sandwich, my first hamburger and my first New-England hikes.

Aafke, Eric, Johannes, Mariana, Martin, Nezar, Nike, Sam and Sameer, for putting me at ease in the Cambridge scientific mold. I cannot imagine a more diverse and welcoming lab environment.

Along the way in my New-England journey, I was fortunate to form additional close bonds. Guanbo, Mansoureh, Sean and Tracey, whom I had my weekly Friday hangouts with for years until the pandemic. Fernando, Gezel, Hollian and Marc, who somehow turned all my NYC trips into wild and unforgettable life experiences. Daniel and Jeff, for their life mentorship. Baptiste for his patience. Anina and her family for reminding me to appreciate the little things. Tommy for his unwavering support.

I am also thankful for the many connections without whom my career progress would have been considerably slower: Asaf, Beena, Ben, Brendan, Eli, Lionel, Marios, Sebastien and Steven.

This dissertation is dedicated to Mai. I am wearing your ring as I am writing this. I don't quite remember your face or voice anymore, but will never forget how much love we had for each other.

List of abbreviations

ALK: Anaplastic Lymphoma Kinase

APC: Adenomatous Polyposis Coli

AUC: Area Under the Curve

B2M: Beta-2-Microglobulin

BER: Base Excision Repair

BMI: Body mass index

BRAF: B Rapidly Accelerated Fibrosarcoma

CCF: Cancer Cell Fraction

CMS: Consensus Molecular Subtypes

COAD: COlon ADenocarcinoma

COPD: Chronic Obstructive Pulmonary Disease

COSMIC: Catalogue of Somatic Mutations In Cancer

CRC: ColoRectal Cancer

CT: Computed Tomography

CTLA-4: Cytotoxic *T lymphocyte*-associated molecule-*4*

DNA: DeoxyriboNucleic Acid

EBV: Epstein–Barr Virus

EGA: European Genome Archive

EGFR: Epidermal Growth Factor Receptor

FA: Fanconi Anemia

FDA: Food Drug Administration

FDR: False Discovery Rate

FEV1: First Forced Expiratory Volume

FFPE: Formalin-Fixed, Paraffin-Embedded

FFQ: food frequency questionnaires

FOLFOX: drug combination of FOLinic acid, Fluorouracil, and OXaliplatin

FOV: Fields of view

FVC: Forced vital capacity

HBV: Hepatitis B Virus

HCV: Hepatitis C Virus

HIV: Human Immunodeficiency Virus

HLA: Human Leukocyte Antigen

HNSCC: Head and Neck Squamous Cell Carcinoma

HPFS: Health Professionals Follow-Up Study

IARC: International Agency for Research on Cancer

ICB: Immune Checkpoint Blockade

IFNGR: Interferon Gamma Receptor

IGV: Integrative Genomics Viewer

JAK: Janus Kinase

KRAS: Kirsten rat sarcoma virus

KSV: Kaposi sarcoma-associated herpesvirus

LOH: Loss of Heterozygosity

MET: Metabolic Equivalent of Task

MGMT: MethylGuanine MethylTransferase

MHC I: Major Histocompatibility Complex class I

MLH1: MutL Homolog 1

MSI: Microsatellite Instable

MSS: Microsatellite Stable

MUTYH: mutY homolog

NCAM1: Neural Cell Adhesion Molecule 1

NHS: Nurses' Health Study

NHTL1: Nth Like DNA Glycosylase 1

NK cells: Natural Killer cells

NMF: Non-negative matrix factorization

NOC: N-nitroso compounds

NSCLC: Non Small Cell Lung Carcinoma

NTRK: Neurotrophic Tyrosine Receptor Kinase

OS: Overall Survival

TP53: Tumor Protein 53

PCAWG: Pan-Cancer Analysis of Whole Genomes

PCR: Polymerase Chain Reaction

PD-1: Programmed cell Death protein 1

PDCD1: Programmed Cell Death protein 1

PFS: Progression Free Survival

PIK3CA: PhosphatidylInositol-4,5-bisphosphate 3-kinase Catalytic subunit Alpha

PMS2: PostMeiotic Segregation 2

POLE: POLymerase Epsilon

READ: REctal ADenocarcinoma

RNA: Ribonucleic acid

RNF43: RiNg Finger protein 43

ROC: Receiver Operating Curve

RSS: Residual Sum of Squares

SBS: Single Base Substitution

SNV: Single Nucleotide Variant

SNP: Single Nucleotide Polymorphism

STAT: Signal Transducer And Activator Of Transcription

TAP: Transporter associated with Antigen Processing

TCGA: The Cancer Genome Analysis

TLS: Translesion synthesis

TMB: Tumor Mutational Burden

UPR: Unfolded Protein Response

VAF: Variant Allele Fraction

WES: Whole-exome sequencing

ccRCC: clear cell Renal Cell Carcinoma

dMMR: deficient MisMatch Repair

scRNA-seq: single cell RNAseq

t-SNE: t-distributed Stochastic Neighbor Embedding

List of figures

Figure 4.S13: Distribution of overall red meat intake

Figure 4.S14: Variant calling pipeline

Figure 4.S15: MGMT promoter methylation status in TCGA COAD/ READ

Figure 4.S16: Impact of VAF on NMF rank selection

Table of contents

1. Introduction

1.1 Background on cancer genomics

1.1.1 Cancer is a disease of the genome

Cancer is characterized by the uncontrolled proliferation of cells, which can spread to other parts of the body. The subsequent abnormal growths can invade key organs, impair their functions and are often the cause of cancer-related deaths. In 2020, cancer was the second leading cause of death with over 10.0 million deaths and an estimated 19.3 million new cancer cases [1]. While significant progress has been made to treat and prevent cancer, incidence is still predicted to rise by around 50% in 2040 [1] with developing countries anticipated to experience the greatest proportional increases. Understanding the fundamentals of cancer initiation and progression is thus a global health priority.

The commonly accepted etiology of cancer is through DNA mutation(s) in a single normal founder cell, which prime(s) the cell to proliferate abnormally by interfering with the regulation of cell death and cell division. Although a recent analysis suggested that the majority of cancers are the consequence of 'bad luck' [2], the most commonly accepted theory is that less than 10-30% of cancers are due to unavoidable intrinsic risk factors, such as random errors in DNA replication [3,4].

In a minority of cases (~5-10%), cancer mutations are hereditary, due to germline genetic defects [5]. For example, Lynch Syndrome is caused by inherited germline mutation(s) in the DNA mismatch repair machinery and accounts for 3% of all colorectal cancers [6].

However, in most cases (~70-90%) cancer arises from avoidable environmental and lifestyle factors (**Figure 1.1**) [5,7]. These carcinogens often act as direct mutagens by altering DNA.

Carcinogens agents include chemicals, such as the ones from tobacco smoking, which contributes to a third of all cancer deaths [8] in the USA. Dietary factors and obesity are also responsible for almost half of global cancer incidence [9]. In addition, infectious diseases, radiation, autoimmune diseases, and physical agents contribute to the overall cancer death toll [5].

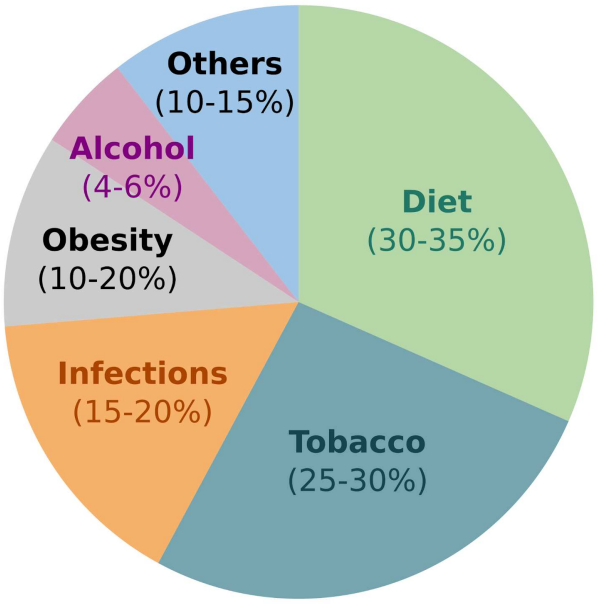


Figure 1.1: Contribution of environmental and lifestyle factors to overall cancer death

Pie chart displaying the percentage of cancer deaths attributable to various environmental and lifestyle factors. Numbers from Anand et al*., Pharm Res 2008* [5]

Although many of these carcinogens have been discovered and characterized [10] through epidemiological studies as well as animal [11] and cell line experiments [12], the mutagenic effects of most mutagens are yet to be observed directly in human tumors.

1.1.2 DNA sequencing and the characterization of the cancer genome

Because of the genetic nature of cancer, oncology research has extensively focused on characterizing the mutational landscape of a wide array of cancer types. In recent years, the study of tumor mutations have stepped up thanks to the emergence of high-throughput sequencing. In particular, Next Generation Sequencing (NGS) of DNA has enabled the study of cancer cells at a single base resolution [13]. Briefly, DNA can be broken into short fragments that are amplified and then sequenced to produce “reads” of the genome. Bioinformatic techniques can then piece together the reads to recapitulate the original sequenced genome [14].

In oncology research, DNA is often sequenced from both normal and tumor tissues. Genetic variants present in both the normal and tumor tissues are called germline mutations and are often hereditary (**Figure 1.2A**). Genetic variants present in the tumor but not in the normal tissues are called somatic mutations (**Figure 1.2B**). These are central during cancer initiation and progression, and are the main focus of this thesis.

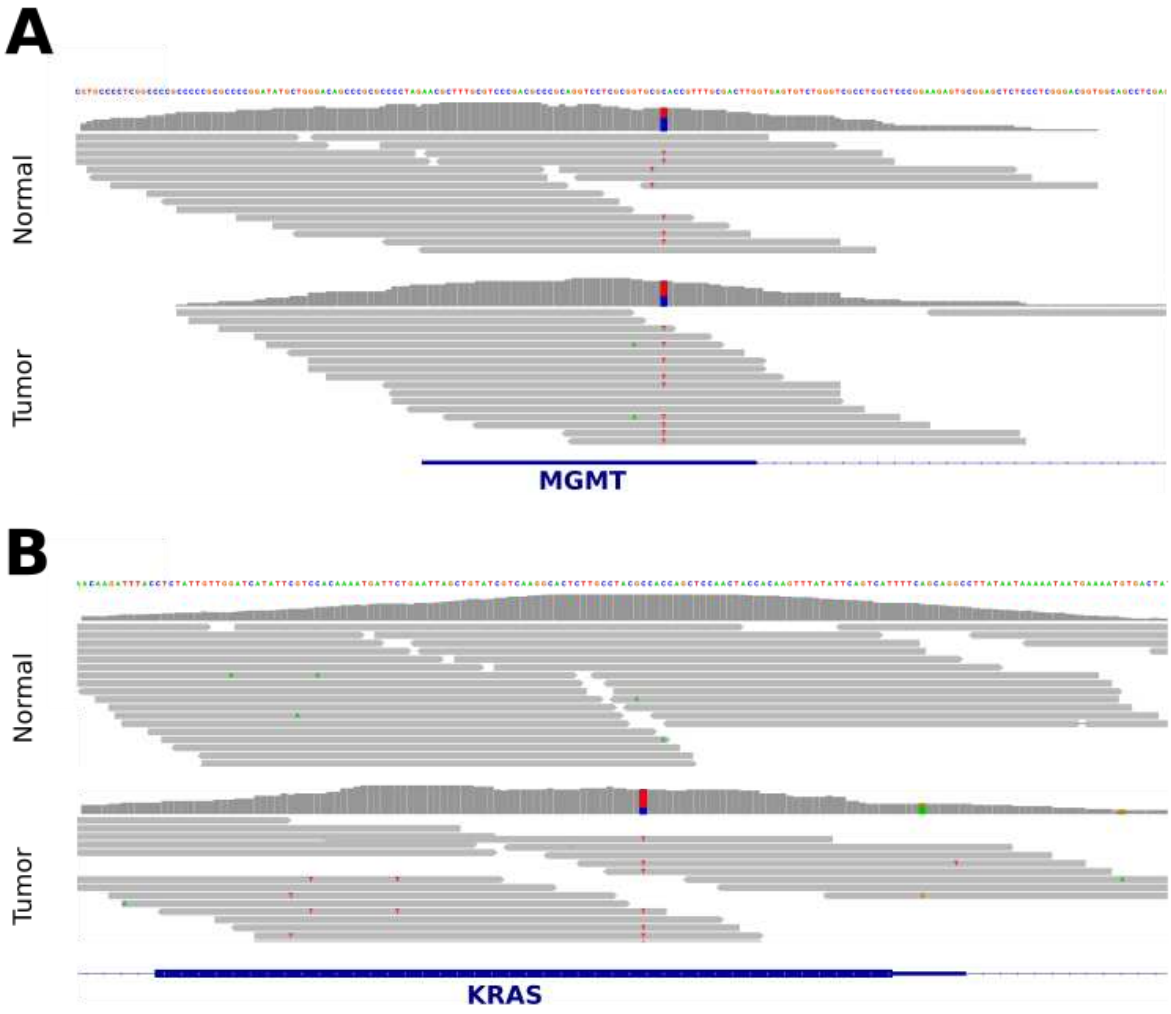


Figure 1.2: Mutations in a matched colorectal tumor/ normal paired sample

Each read (in grey), is aligned to a reference genome (top sequence). Bases that do not match the reference genome (on top) are drawn on the read (A, C, G, or T). (A) Example of a C>T heterozygous germline mutation in the first exon of Methylguanine methyltransferase (MGMT). (B) Example of a C>T somatic mutation in KRAS, a well-characterized driver mutation.

In recent years, technologic and computational advances have enabled the detection of variants with increasing accuracy [15,16]. A wide array of somatic variants can be called, including non-synonymous mutations (which can alter the amino-acid sequence of a protein), synonymous mutations (i.e. point mutations that do not alter the amino-acid sequence) and insertions/ deletions (which often lead to frameshifts). Larger chromosomal events can also be detected, such as copy number changes or translocations.

The lowering cost of whole-genome and whole-exome sequencing provided the means to create large cancer mutational databases such as TCGA (The Cancer Genome Atlas) and PCAWG (Pan-Cancer Analysis of Whole Genomes) [17,18]. These databases have in return enabled better characterization of artefacts and increased mutation calling accuracy. For instance, the filtering of recurrent technical artifacts has drastically improved thanks to the creation of Panel Of Normals [19] and the modelling of a wide array of archival and sequencing-derived artefacts [20].

Ultimately, these advances have allowed so far the characterization of around 50 millions somatic variants making up the mutational landscapes of the most common cancer types [17,18].

1.2 Passenger mutations represent the majority of cancer mutations

1.2.1 Passenger versus driver mutations

Following variant calling, mutations can be classified into two categories according to their effect on overall cell fitness [21,22]. Mutations that drive fast population growth by increasing cell fitness are called *drivers*. In addition to driver mutations, thousands of mutations with individually no role on tumour growth can accumulate alongside the drivers: these mutations are called *passengers*.

Passenger mutations have either neutral or weakly deleterious effects on tumour growth. They can accumulate in the tumoral genome despite negative selections due to two population genetics phenomenons, called Hill-Robertson interference processes: genetic hitchhiking and Muller's ratchet [23,24]. Genetic hitchhiking of passengers happens in association with a driver mutation undergoing a selective sweep. Muller's ratchet is the gradual fixation of passenger mutations in the absence of genetic recombination. Combined, these two evolutionary processes explain why potentially deleterious passenger mutations can survive in large numbers in the tumour's genetic pool.

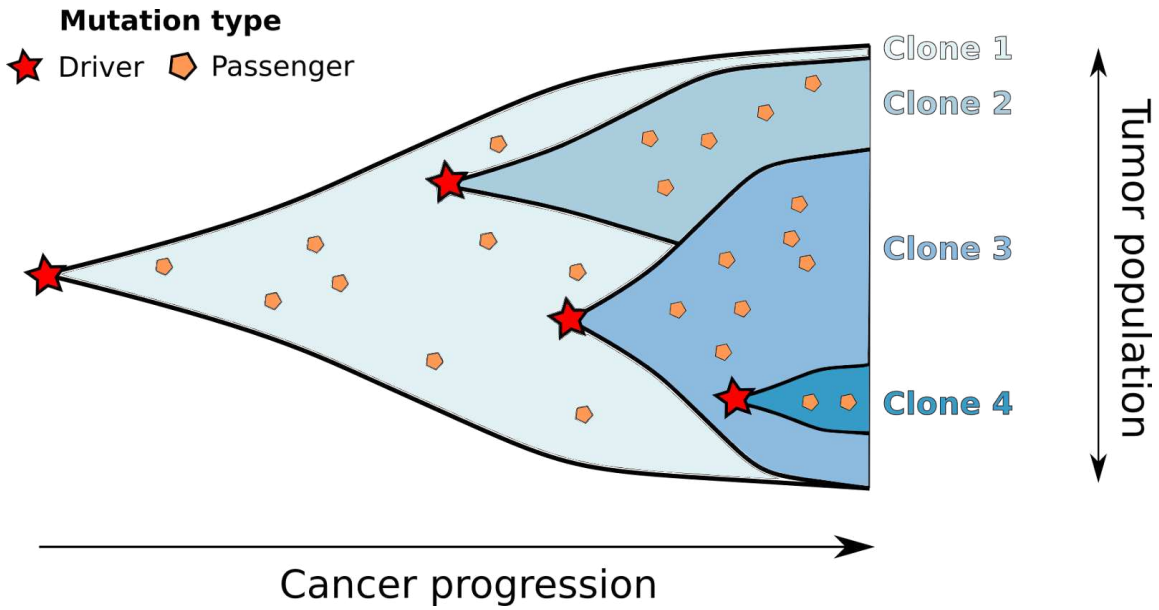


Figure 1.3: Driver and passenger mutations in cancer evolution

Punctual driver mutations (in red) generate phenotypically different clone lineages (in blue) with an elevated fitness. Passenger mutations (in orange) accumulate and fixate during cancer progression but have individually no role on tumour growth.

Driver mutations can stem from the loss of function of a gene regulating cell division and replication, named tumour suppressor gene. They include mutations in *p53* gene (present in almost half of all malignancies [17]) and mutations in *APC* gene, involved in cell adhesion and cell division and common in colorectal cancer [25]. Driver mutations can also arise from the gain of function of a gene, named proto-oncogene, which can then contribute to tumour growth [26]. For instance, mutations in MYC proteins, which code for transcription factors, can activate the expression of pro-proliferative genes leading to cancer[27].

In large sequencing cohorts, the classification of mutations as either a driver or passenger is based on predictive computational methods [17,28]. Briefly, mutations can be assigned a predictive score of molecular functional impact based on various features such as network centrality and inter/intra species conservation [29]. Another approach to detecting driver mutations is by modelling the probability that a base is mutated by chance. Mutations that occur at a higher frequency than expected are interpreted as selected for and subsequently classified as drivers [30–32]. The *in silico* predictions of these approaches can then be experimentally validated.

1.2.2 Mutational landscape of cancers

The studies of driver and passenger mutations in large cancer consortia [17,28] have found that passenger mutations represent the overwhelming majority of tumour mutations: it is estimated that among the tens of thousands of mutations present in a tumour, on average only 5 are drivers [17,22,33].

As drivers are critical for carcinogenesis, the goal of cancer sequencing studies has been historically and primarily to identify them [33,34]. This not only enabled the development of Next Generation Sequencing (NGS) panel tests to prioritize specific therapies based on patients' mutations [35] but also opened novel avenues for gene-targeted therapies such as gene inhibitors [36–38]. For example, FDA-approved treatments for non-small cell lung cancer include various targeted therapies depending on the presence of specific mutations such as those in the *KRAS*, *EGFR*, *ALK*, *ROS1*, *BRAF*, *RET*, *MET*, or *NTRK* genes [39].

On the other hand, little attention has been given to passenger mutations as they have individually no role in tumour progression. However,

due to their large number, they offer exciting avenues for novel mathematical and statistical approaches to understand cancer.

1.3 Passenger mutations offer a window into mutagenic processes

1.3.1 Concept of mutational signature

Passenger mutations do not provide a growth advantage to the tumour. Therefore, they are not strongly selected for and can be viewed as residual molecular fingerprints of the various mutagenic processes [40] that a tumour has undergone before sequencing.

Mutagenic processes include mutations resulting from normal cellular processes (called endogenous mutations) such as errors in replication, depurination, deamination, and damage by oxidative stress [41]. Conversely, exogenous mutations are due to environmental and lifestyle factors including exposure to ultraviolet light, smoking and diet [7], as mentioned previously (**Section 1.1.1**). Exogenous mutations are of particular interest as they are potentially modifiable factors that could be leveraged for cancer prevention in a larger subset of cancer patients. It is thought that 30% to 50% of cancers in the United States could be prevented by avoiding risk factors [42].

Because they are molecular imprints of mutagenic processes, passenger mutations can consequently help better characterize the carcinogenic potency of a putative carcinogen. Notably, recent pioneering work

on the concept of “mutational signatures” [40,43] has enabled a greater understanding of the contribution of different mutational processes to the overall mutational landscape.

The concept of “mutational signature” relies on the observation that mutagenic processes do not affect all loci equally. For example (**Figure 1.4**), ageing has a higher likelihood to cause a C>T mutation. Similarly, microsatellite instability (MSI) can have a T>C bias, and polymerase epsilon instability a C>A one. These mutagenic processes cumulatively shape the final mutational spectrum (**Figure 1.4**, in orange) with different contributions. In addition to the substitution type, mutagenic processes can further present a bias in the trinucleotide context. For example, aging preferentially targets A[C>T]G compared to A[C>T]A [43]. Consequently, mutational signatures are often represented by the 96 possible mutation types (six substitution types and their immediate 5′ and 3′ trinucleotide context) (**Figure 1.5C**). Larger broader contexts (e.g. pentanucleotide and heptanucleotide context) have also been studied [44] although their comprehensive analysis is hindered by the exponential loss of statistical power.

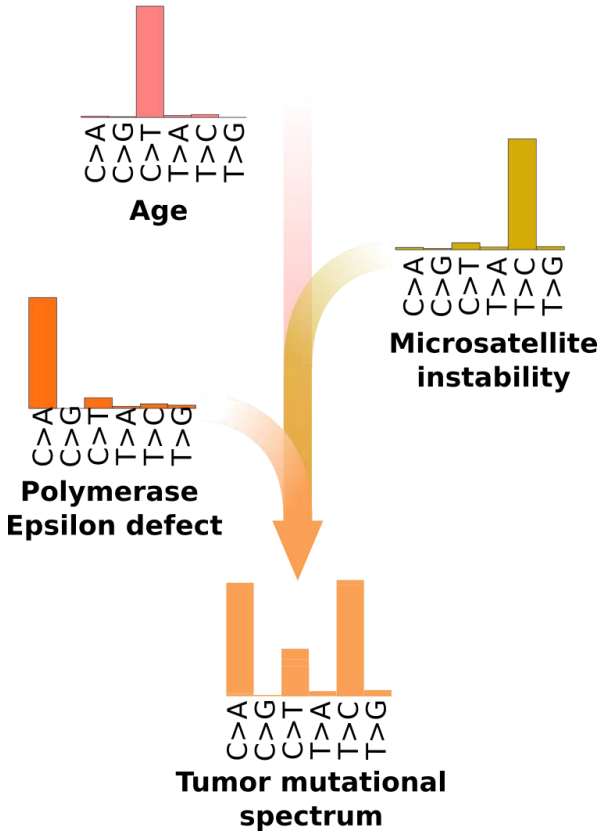


Figure 1.4: Concept of mutational signatures

Throughout the genesis and progression of a tumour (vertical arrow), a variety of mutational processes can affect the tumoral genome. The final tumour mutational landscape is a cumulation of those processes.

1.3.2 Computational deconvolution of mutagenic processes

Dimension reduction methods based on Non-negative Matrix Factorization (NMF) have enabled the deconvolution of mutagenic processes from mutation calls [43]. NMF allows data signal separation and is an efficient method to identify distinct molecular patterns.

Starting from mutation calls, a catalog matrix of the 96 possible trinucleotide contexts can be constructed (**Figure 1.5A**). This catalog matrix can then be used as input for an NMF which results in two factorized matrices (**Figure 1.5B**). The 'Mutational Signature' matrix contains the 96-trinucleotide context imprints left by the different mutagenic processes (example given in **Figure 1.5C**). The 'Signature Activity' matrix consists of the contribution of each mutagenic process for a given patient.

Ideally, an NMF should be run only on passenger mutations as they are not by definition affected by selection (unlike drivers) and are thus molecular fingerprints representative of the operative molecular processes. In practice, an NMF is run on mutations regardless of their fitness advantage as passengers constitute the overwhelming majority of mutations (>99%) and the effect of drivers on NMF results is negligible. For comparison, the rate of false positives when calling variants is usually at least a few percent [45]; hence, driver mutations' effect on NMF is often below noise level.

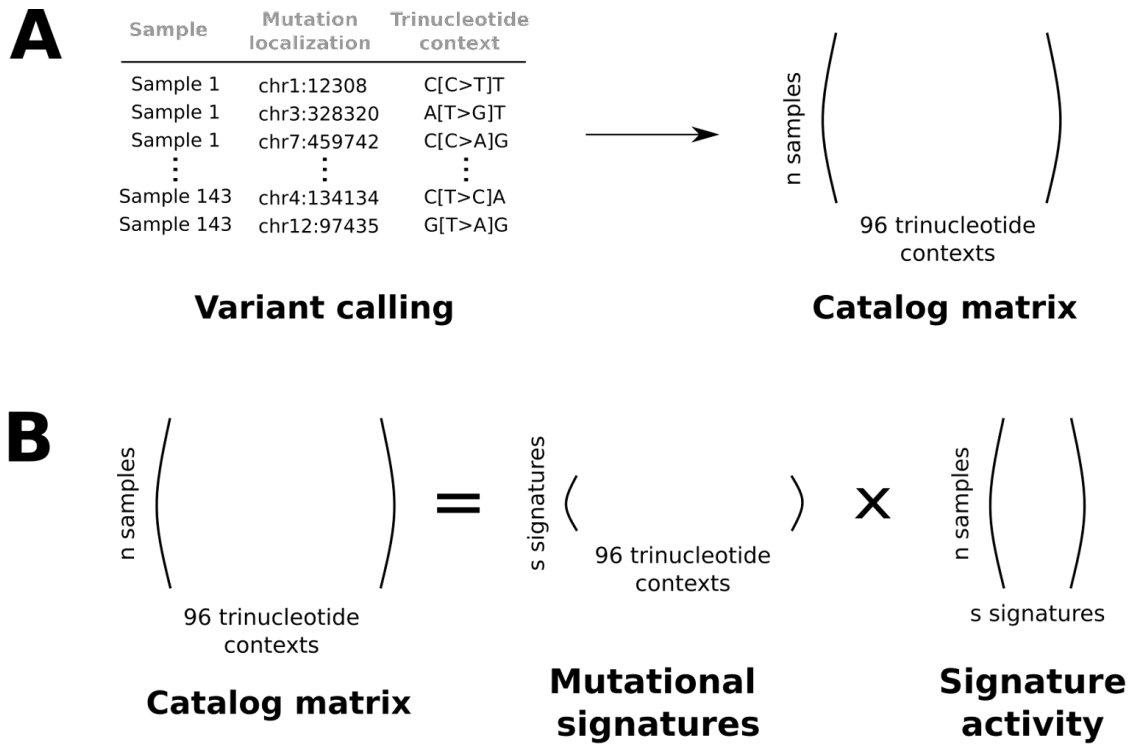


Figure 1.5: Workflow for mutational signature analysis

(A) Variants are called for multiple tumour samples and are used to construct the catalog matrix. In the latter, each row corresponds to a sample and each column contains the mutation counts for a specific trinucleotide context. (B) A Non-Negative Matrix Factorization method can be applied to the catalog matrix for a given rank (i.e. the number of mutagenic processes in the tumours). One of the resulting matrices contains the mutational signatures which are the characteristic combinations of mutations arising from each mutagenic process.

The other matrix contains the signature activities with the relative contributions of each mutagenic process for each tumour. (C) Example of a published mutational signature [46]. SBS4 is a mutational signature associated with tobacco smoking.

Signature analyses have been previously performed on large Whole-Genome Sequencing (WGS) and Whole-Exome Sequencing (WES) cohorts such as TCG and PCAWG and resulted in a compendium of mutational signatures in cancer [46]. In addition to point mutation signatures, signatures of doublet base substitutions and indels [46] have been characterized, as well as differential replication timing and strand asymmetry for various signatures [47].

These mutational signatures were associated with a wide array of biological processes, such as rare cancer predisposition syndromes [48], environmental agents [49], microbiota [50] , chemotherapeutic drugs and even poor outcomes to specific therapies [51]. However, larger and more comprehensively annotated cancer-specific cohorts are needed to fully recapitulate cancer-specific mutagenic processes.

In addition to revealing active mutational processes, mutational signatures can also be leveraged to predict their potential induction of driver mutations [52,53]. Indeed, mutational signatures can be viewed as the probability of a mutation happening at a specific trinucleotide context. For instance, a smoking signature SBS 4 (**Figure 1.5C**) has a higher probability to target C[C>A]A mutations such as *KRAS* G12C. The estimation of how mutagenic processes contribute to driver events can inform prevention efforts, in particular for avoidable risk factors. Previous studies [53] have demonstrated that driver mutations in lung and skin cancers' can largely be attributed to actionable mutational processes, although the role of these processes in the cancer initiation itself is still debated [54].

1.4 Passenger mutations can collectively trigger an immune response

1.4.1 Mutations can be presented at the cell surface

As part of the normal metabolism, intracellular proteins can be ultimately fragmented into smaller peptides that are transported to the endoplasmic reticulum and presented on the cell surface as part of a complex with Major Histocompatibility Complex class I (MHC I) molecules [55] (**Figure 1.6**). These antigens are part of the 'self' and do not trigger an immune response.

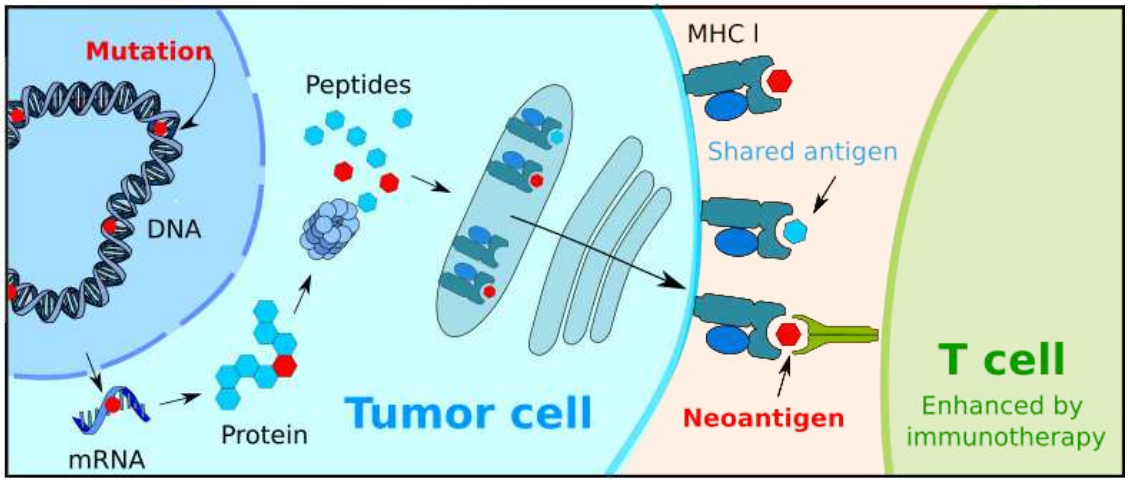


Figure 1.6: Neoantigen theory

Tumour mutations can be expressed and translated into proteins, which can then be degraded into smaller peptides. These peptides can form a complex with MHC class I molecules and be presented at the cell surface as 'neoantigens' that can bind to T-cell receptors and trigger an immune response.

In the case of a cancer cell, a mutated protein can be aberrantly presented on the cell surface (**Figure 1.6**) as part of the 'non-self' [55]. Because these *neoantigens* are new to the immune system, they can be recognized as 'foreign', in particular by killer T-cells, and trigger an anti-tumour response. This might partially explain why tumours often evade immunosurveillance by upregulating the expression of specific immune checkpoints (i.e. regulators of the immune system) [56].

Immune-based therapies have shown clinical benefits for a wide array of malignancies (**Figure 1.7**). In essence, these therapies prompt the immune system to trigger an antitumor response: for example, by blocking specific immune checkpoints and consequently unleashing an immune response against cells detected as 'foreign'.

Such treatments include inhibitors of the negative regulators of T-cells such as cytotoxic T-lymphocyte-associated protein 4 (CTLA-4) and programmed cell death protein 1 (PD-1). These inhibitors can be used alone as front-line therapy, but are often more effective in combination [57](**Figure 1.7**). However, although effective these therapies are encumbered by high variability in patients' response [58–61].

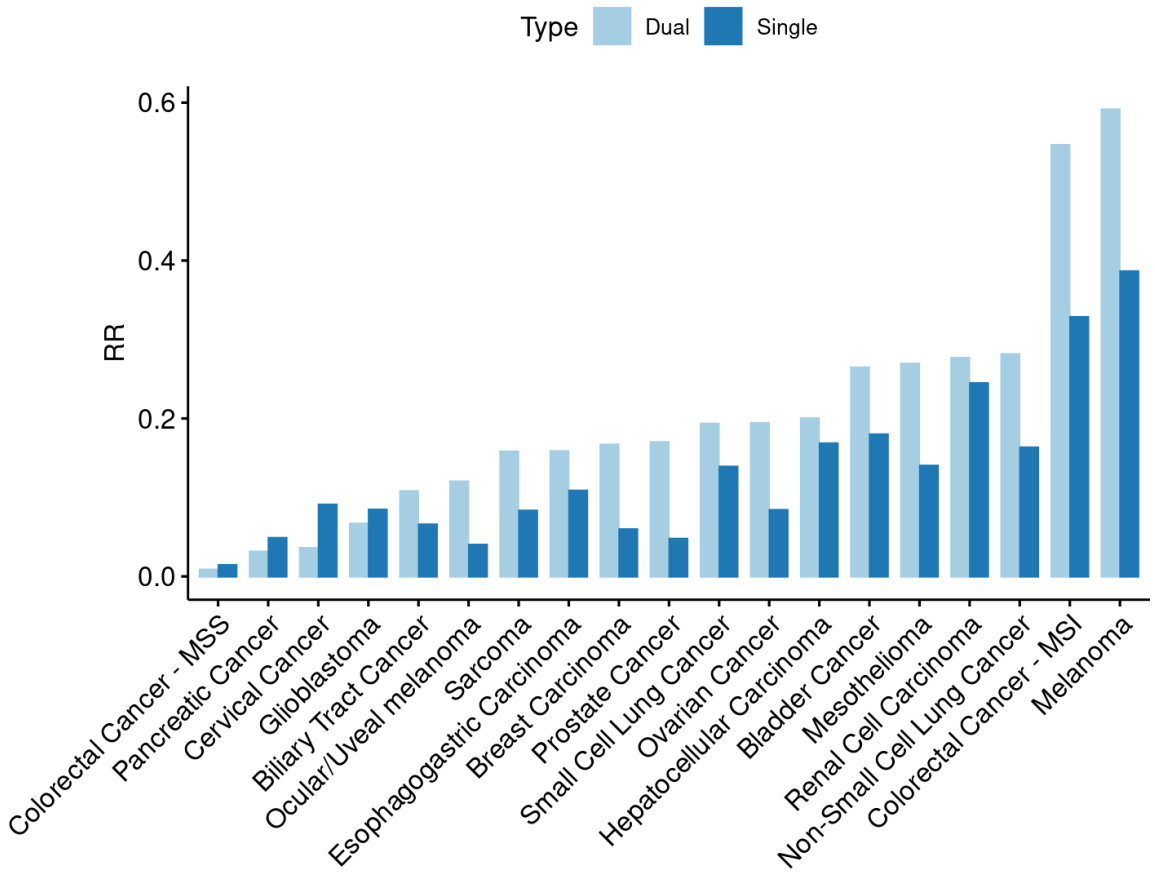


Figure 1.7: Immunotherapy response rate for different cancers

Objective response rate (RR, on the y-axis) for a range of cancers (on the x-axis) for both monotherapy (dark blue) and combination immunotherapy (light blue). Based on numbers from a previously published meta-analysis [62]. MSS: Microsatellite Stable. MSI: Microsatellite Instable

1.4.2 Neoantigen theory

Neoantigens are immunologically active proteins presented at the cell surface, and are the main functional targets of immune checkpoint blockade therapies [63]. Because the total Tumor Mutational Burden (TMB) is mainly composed of passengers, immune reactivity is mainly directed towards

passengers-derived neoantigens [63–65]. Consequently, it is widely thought that passenger load is proportional to the likelihood of an immunologically active neoantigen to be present and targeted by immune-based cancer therapies [66–70].

The United States Food and Drug Administration (FDA) approved in June 2020 the use of TMB as a biomarker of response to pembrolizumab, an immune checkpoint blockade therapy [71]. Pembrolizumab is a highly selective humanized antibody directed against PD-1. This is one of the first tissue-agnostic drug approvals for solid tumours, which allows all adult and pediatric patients with TMB> 10 mutations/Mb to undergo immunotherapy. This FDA approval led to the creation of assays to directly measure TMB [72] and optimized gene panels recapitulating TMB [73].

1.5 Thesis overview

The tumour mutational landscape consists of a few driver mutations alongside thousands of passenger mutations. The latter are viewed as collateral damage with no role in tumour growth and have been historically largely understudied. Throughout this thesis, I explore the use of leveraging passengers in combination with driver mutations to further inform cancer therapy and prevention (**Figure 1.8**).

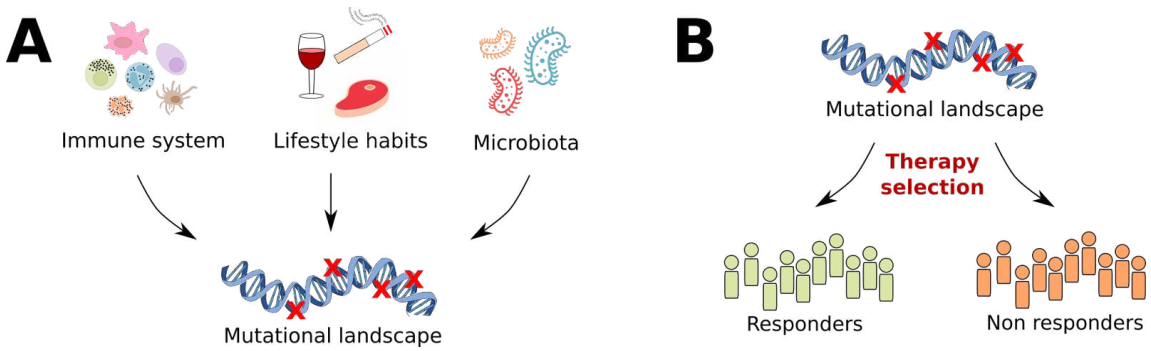


Figure 1.8: Summarizing diagram of the objects of study

Through the analysis of passengers, this thesis explores what factors shape mutational patterns and how these patterns can in return inform therapy selection. (A) Lifestyle habits and the microbiome can be genotoxic and leave an imprint on tumour DNA. In addition, the immune system shapes the mutational landscape by weeding out cells with too many mutations (a theory called "neoantigen theory"). (B) The overall mutational landscape, in addition to specific driver mutations, can be leveraged for therapy selection.

This dissertation focuses on two features of mutations: their total number (called ‘mutational load’ or ‘mutational burden’) and their genetic localization patterns (through mutational signature analysis). The projects developed in this thesis include: (i) a meta-analysis of high mutational load as a biomarker of response to immunotherapy [74]; (ii) a case report of a patient's colorectal tumour who did not respond to immunotherapy despite having a very high mutational load [75]; and (iii) the mutational signature analyses of large colorectal cancer sequencing cohort [76].

1.5.1 High tumour mutational burden as a biomarker of response to immunotherapy

The first chapter of this thesis challenges one of the central paradigms in cancer immunology: the notion that a high number of mutations — and the resulting neoantigens, recognized as ‘foreign’ by T-cells — elicits a better antitumour immune response. Several pioneering papers [66–70] have suggested that tumours with a high load of mutations better respond to immunotherapy; these studies have been the basis of the FDA approval of TMB to prioritize patients who would most likely benefit from immunotherapy.

We revisit this claim by conducting a pan-cancer meta-analysis: we aggregate the largest available dataset of immunotherapy patients, reuniting more than 2,500 individuals, with available TMB data and clinical annotations. To conduct this analysis, we not only leverage standard biomarker metrics (e.g. Receiver Operating Curves analysis), but we also create a novel statistical framework to correct for multiple hypotheses when the tests were non-independent. Finally, we build a mathematical model of the neoantigen

theory reproducing our observations of the association between TMB and treatment response. Overall, this first chapter further assesses the use of directly measuring passenger load in the clinic to predict immunotherapy response.

1.5.2 Exploratory analysis of an immunotherapy-treated patient with high tumour mutational burden

High TMB tumours, in particular those with DNA repair defects such as Microsatellite Instability (MSI), are FDA-approved for immunotherapy; yet not all patients respond. The second chapter of this thesis aims at refining our understanding of the link between TMB and immunotherapy response. To this end, we conduct an exploratory analysis of a patient who did not respond to immunotherapy despite having an MSI colorectal tumour.

To do so, we use whole-exome and bulk whole-transcriptome sequencing (RNA-Seq) to characterize the driver and passenger mutational landscape of the patient. In addition, we leverage cutting-edge technologies such as multiplex immunofluorescence and single-cell RNA-Sequencing to comprehensively profile not only the tumour genome but also the associated tumour-immune microenvironment. Ultimately, this chapter aims at uncovering the mechanisms of intrinsic resistance to immunotherapy and can help understand how to better integrate passenger load with driver mutations to predict treatment response.

1.5.3 Molecular imprints in colorectal cancer

Colorectal cancer (CRC) is currently the third most common cancer in the world. In the last decade, there has been a concerning trend of early-onset CRC: by 2040, it is predicted to become the leading cause of cancer death in individuals between 20 and 40 years old [77]. Although the reasons for this trend are unclear, diet, in particular red meat consumption, has been hypothesized to play a role. However, no mutagenic role of red meat has been observed in human colons yet.

The third chapter of this thesis presents the mutational signature analysis of a comprehensive CRC molecular profiling study: 900 tumours were sequenced from tissue biopsies collected over four decades. Along with molecular data, this study leveraged detailed information, collected every other year, on the lifestyle of CRC patients. This comprehensive cohort allows us to accurately estimate mutagenic processes in CRC and better understand how they are related to lifestyle factors.

2. Limited evidence of tumour mutational burden as a biomarker of response to immunotherapy

Pre-print published on Biorxiv on November 17, 2020.

2.1 Author information

Carino Gurjao [1,2,3], Dina Tsukrov [3], Maxim Imakaev [3], Lovelace J. Luquette [4] and Leonid A. Mirny [3,2,1] *

1. Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
2. Broad Institute of MIT and Harvard, Cambridge, MA, USA
3. Institute for Medical Engineering and Science, and Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA
4. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

2.2 Abstract

Cancer immunotherapy by immune checkpoint blockade (ICB) is effective for several cancer types [61], however, its clinical use is encumbered by a high variability in patient response. Several studies have suggested that Tumour Mutational Burden (TMB) correlates with patient response to ICB treatments [66–70], likely due to immunogenic neoantigens generated by novel mutations accumulated during cancer progression [78]. Association of TMB and response to checkpoint inhibitors has become widespread in the oncoimmunology field, within and across cancer types [62,78–81], and has led to the development of commercial TMB-based biomarker platforms. Furthermore, patient prioritization for ICB based on individual TMB level was recently approved by the FDA [71]. Here we revisit the association of mutational burden with response to checkpoint inhibitors by aggregating the largest pan-cancer dataset with more than 2500 ICB-treated patients with sequencing data and clinical annotation. Surprisingly, we find little evidence that TMB is predictive of patient response to immunotherapy. Our analysis suggests that previously reported associations arise from a combination of confounding disease subtypes and incorrect statistical testing. We show that using a TMB threshold for clinical decisions regarding immunotherapy could skew access to treatment for patients who may benefit from these therapies. Finally, we present a simple mathematical model that extends the neoantigen theory, is consistent with the lack of association between TMB and response to ICB and highlights the role of immunodominance. Our analysis calls for caution in the use of TMB as a biomarker and emphasizes the necessity of continuing the search for other genetic and non-genetic determinants of response to immunotherapy.

2.3 Introduction

Immune checkpoint blockade (ICB) treatments such as anti-CTLA-4 and anti-PD1, which target regulatory pathways in T-lymphocytes to enhance anti-tumour immune responses, have already proven to elicit durable clinical responses for some patients [58–61]. However, the genetic determinants of response to immunotherapy have yet to be found. Several studies [66–70] suggested that Tumour Mutational Burden (TMB), computed as the total number of nonsynonymous somatic mutations, is correlated with response to immunotherapy in cancer. The underlying hypothesis posits that a fraction of nonsynonymous mutations become exposed as epitopes and constitute neoantigens, which can trigger an anticancer response by the immune system. The association between high mutational burden and response to immunotherapy, within and across cancer types [62,78–81] (**Figure 2.1A**), has been widely reported in the scientific literature and the media. As a result, TMB is currently discussed as the most clinically advanced biomarker of response to immune checkpoint blockade [82,83], and the FDA approved the use of TMB to identify patients most likely to derive clinical benefit [71]. These studies also triggered a search for inexpensive assays to evaluate TMB directly [84], as well as TMB-derived measures, such as neoantigens, neoepitopes, and mutation clonality [85], which are all currently under investigation to further stratify patients most likely to respond to immunotherapy. Our analysis focuses on TMB itself, as this is the most widely used and only FDA-approved measure.

2.4 Results

2.4.1 Data aggregation

To evaluate the association of TMB with response to ICB across a broader range of cancer types, we aggregated and analyzed data for 882 immunotherapy patients with publicly available pre-treatment whole-exome sequencing data (referred below as CPI800+, **Table 2.S1** and **Material and Methods**). We included patient-level data from an aggregate of early seminal studies [86] as well as recent clear cell renal cell cancer [87,88], non-small cell lung cancer [70], bladder cancer [89] and melanoma [66,90,91] ICB-treated cohorts. For every dataset examined, we retrieved TMB levels and survival data (Progression-Free Survival (PFS) or Overall Survival (OS)) for each patient, and provided . Cohorts provided response classification for most patients.

We also leveraged a very recent meta-analysis of 1283 patients (termed CPI1000+) who underwent immunotherapy [92] , have unified TMB (n=1083) and response definition, as well as survival measures for some patients (n=545 with OS data). Furthermore, we obtained gene panel data (MSK-IMPACT) for 1662 patients (**Table 2.S1**) who underwent immunotherapy. To the best of our knowledge, together this dataset constitutes the largest pan-cancer aggregate of ICB-treated patients with sequencing and clinical data, which allow a robust unified statistical assessment of TMB as a predictor of ICB response

First, we examined the difference in TMB between responders and non-responders. All datasets showed a considerable overlap in TMB between responders and non-responders, as well as a large range of TMB values for the same cancer type.

Consistent with published studies [70,86,87,90,93], we find (**Figure 2.1A** and **Figure 2.S1**) that only the melanoma datasets (mel1 and mel2) and non-small cell lung cancer datasets (lung1 and lung2) yield a significant difference in TMB between responders and nonresponders ($p=8.3\times10^{-6}$ and $p=7.7\times10^{-3}$ for lung1 and lung2, $p=2.6\times10^{-2}$ and $p=4.1\times10^{-2}$ for mel1 and mel2, Mann-Whitney U test). Two of the three other cancer types analyzed – clear cell Renal Cell Carcinoma (ccRCC) and Head and Neck Squamous Cell Carcinoma (HNSCC) – showed no trend or an unexpected inverse one between TMB and response, although the association was non-significant (**Figure 2.1A**). The lack of associations between TMB and response to ICB in these two cancers was previously observed [88,94,95]. An identical analysis in CPI1000+ revealed similar results: there is no association between TMB and response in HNSCC and ccRCC, as well as breast cancer. Of note, CPI1000+ comprises colorectal cancer (CRC) specimens, and more bladder cancers (n=250) than our CPI800+ aggregate. Both showed an association between TMB and response ($p=0.044$ and $p=5.4\times10^{-7}$, respectively) (**Figure 2.S1**).

Next we examined (i) whether TMB can be used as a biomarker to predict response to ICB in cancers/cohorts with significant association, yet considerable overlap in TMB between responders and non-responders; (ii) whether the association between TMB and response is confounded by cancer subtypes, i.e. due to the different response rates of cancer subtypes with

different TMB ranges, and (iii) a potential association between TMB and survival, rather than a binary response/no-response classification.

2.4.2 Measuring performance of TMB as a biomarker of response to immunotherapy

The key component for validating a biomarker is acceptable classification accuracy, i.e. the biomarker's capacity to correctly classify a patient's response[96]. ROC curves analysis (**Figure 2.1B**) is a standard tool for measuring the quality of a predictor; it provides a comprehensive quantification of specificity and sensitivity over all possible cutoffs, with the Area Under the ROC Curve (AUC) being an aggregate measure of predictor performance (AUC=0.5 for a predictor performing as well as random). Our ROC-curve analysis shows the (i) lack of a clear TMB cutoff that could be used in the clinic; (ii) poor performance of the TMB-based predictor of response to ICB, as evident from the low AUC in most datasets: mel1 and mel2 yielding of 0.62 and 0.59, and lung2 has an AUC of 0.68. Lung1, however, has the highest AUC of 0.85, which, as we show below, is still insufficient to select patients for ICB. ROC curve analysis on CPI1000+ cohort, with unified TMB, also shows a similarly poor AUC of 0.6 (**Figure 2.S8**).

Using a poor predictor for treatment decisions can lead to patient misclassification, i.e. patients who could benefit from the therapy would be deprived of it (responders below the TMB threshold), and patients who get the treatment but don't benefit from it (non-responders above the threshold). To quantify the shortcomings of TMB-based selection of patients for treatment we computed the proportion of misclassified patients based on the recent FDA

approval of 10 mutations/Mb threshold to select patients for ICB (**Figure 2.1C** and **Figure 2.1D).** We find that, on average across the non-small cell lung cancer and melanoma datasets, 62% of responders were below the treatment prioritization threshold and 19% of non-responders were above. While these misclassification rates vary across datasets, fractions of potential responders remain under the TMB threshold. Moreover, the poor predictive power of TMB indicates that current efforts of harmonizing TMB measures would not address the shortcoming TMB as a biomarker of response to ICB. Indeed, our ROC analysis shows that even the optimal cutoff (Youden index associated cutoff) for each dataset would result in approximately 25% of responders below the treatment prioritization threshold and thus discouraged from receiving a potentially efficacious and life-extending treatment (**Figure 2.1E**). As such, the main challenge in using TMB in the clinic does not reside in harmonizing the values but in inherently poor association between TMB and response to treatment.

2.4.3 Cancer subtypes can confound the association of TMB with clinical benefit

We hypothesized that different cancer subtypes, with distinct TMB ranges and response rates, confound the observed increase in TMB in patients with clinical response to ICB. Understanding different responses of cancer subtypes can be also important for unraveling underlying biology of response. In particular, acral and mucosal melanomas are known to yield lower TMB and have a poorer prognosis [97]. Similarly, non-small cell lung tumours from smokers have higher TMB and published studies showed that ICB confers a survival

advantage in smokers compared to never smokers [98]. Consistent with a previous study [90], we find that stratifying melanoma patients based on their disease subtype removes the association observed between TMB and clinical benefit in mel1 and mel2 (**Figure 2.2B**), i.e. among patients of the same subtype there is no association between TMB and response.

However, stratifying non-small cell lung tumours based on the patient smoking status still showed a significantly higher TMB for responders versus non-responders, for smokers (current and former) ($p=1.3\times10^{-4}$ and $p=1.8\times10^{-2}$ for lung1 and lung2, Mann-Whitney U test)), but not among non-smokers. Other factors may contribute to a substantially better response of higher TMB in smoker patients. In particular, the presence of Chronic Obtrusive Pulmonary Disease (COPD) [99] could be a factor underlying the better response of high-TMB patients. While none of the ICB-treated cohort provided COPD status, several recent observations are consistent with the confounding role of COPD in response to high-TMB patients. First, COPD status is associated with increased survival after ICB [98,100]. Second, we find that TMB is significantly increased in COPD patients from TCGA (**Figure 2.S2**). Third, the presence of EGFR mutation, that is infrequent in lung tumours of COPD patients [101,102], has been reported to correlate with poor response to immunotherapy [86]. Consistently, a large lung study that excluded patients with targetable EGFR mutation (KEYNOTE-189 [103], n=293), observed no association of high TMB with survival and clinical response to ICB. Our analysis alongside results of KEYNOTE-189 suggest that COPD status can be a confounder that could explain higher TMB among patients that respond to ICB; this hypothesis and the underlying biology can be tested by future studies. Similarly, the recent KEYNOTE-177 [104] showed no association of TMB with response to ICB in hypermutated (MSI positive) colorectal tumors, suggesting MSI status

confounds the association between TMB and clinical benefit to ICB. Cancer subtypes with higher response rate constitute a rich ground for understanding the biology of ICB response.

2.4.4 TMB and response across cancer types

We also revisit a meta-analysis that reported a positive correlation between response rates and TMB across different cancer types [62,81]. In that study, each cancer type is characterized by a median TMB and a median response rate; and splits melanoma and colorectal cancers – but not other cancers – into subtypes. We find that the correlation of the cancer-median TMB with the response rate reported in this study is driven solely by the TMB-response association of melanoma and colorectal cancers subtypes (**Figure 2.S3**): when three points representing these cancers and their subtypes (melanoma, MSI+, MSI-) were removed, the correlation across all remaining cancer types becomes non-significant ($p=0.10$ for monotherapy, and $p=0.21$ for combination therapy). Thus, beyond subtypes with extreme differential response and a high response in melanoma, no association between TMB and response rate across different cancer types is present in available data.

Overall, the evidence of an association between TMB and response to ICB relies largely on data for two cancer types: melanoma and non-small cell lung cancer. However, melanoma is confounded by subtypes and lung cancer requires more data and COPD stratification to validate the use of TMB. Crucially, an elevated TMB among responders does not imply the suitability of

TMB for patient classification and treatment prioritization neither within nor across cancer types.

2.4.5 TMB association with survival post-immunotherapy

To evaluate the use of TMB for prioritizing patients, and to go beyond the binary response classification, we examined an association between TMB and survival time (OS or PFS). Since survival data is “censored” i.e. only lower bound on survival is known for some patients, standard correlation-based methods cannot be used to evaluate such association. Nevertheless, groups of patients can be compared in their survival. Hence we tested whether it is possible to find a TMB threshold that can separate patients into groups with significantly distinct survival.

Strikingly, plots of survival versus TMB (**Figure 2.3A** and **Figure 2.S4**) do not show a visible correlation or TMB cutoff that could differentiate longer and shorter surviving patients. Nevertheless, several studies established such TMB thresholds [67,69] and reported a seemingly statistically significant difference in survival between patients below and above the threshold. One caveat of this approach is that it suffers from inherent multiple hypothesis testing made when the TMB thresholds have been selected among numerous possible values. This inherent multiple hypothesis testing would require further correction of the p-values; a step that is missing in all of the studies. However, standard approaches (e.g., Bonferroni correction, FDR correction) for multiple hypotheses testing would be too stringent because the hypotheses generated by comparing survival in two groups at multiple TMB thresholds are not independent.

Hence, we used a randomization approach to address this limitation. This approach is similar to known multiple hypothesis testing methods [105,106] and earlier statistical studies that examined associations between dose and response in epidemiological studies [107]. We define the optimal TMB threshold as that which maximizes the difference in survival (i.e. minimizes the logrank p-value, a standard survival analysis test) between groups above and below the threshold. First, the optimal TMB threshold and its p-value (p_{real}) was found for the original data. Next, we randomly shuffled TMB among patients, while keeping survival and censored labels unchanged, and found the optimal TMB threshold and its p-value (p_{shuf}) for each randomized data. Finally, the p-value corrected for multiple hypotheses is derived by repeating the shuffling 1000 times and computing the fraction of shufflings where $p_{shuf} < p_{real}$ (**Figure 2.3B)**.

Applied to the melanoma and lung cancer datasets (**Figure 2.3C** and **Figure 2.3D**), we find that the majority (~60-70%) of randomly shuffled datasets produced p_{shuf} below the standard 0.05 threshold, creating a seemingly significant TMB-survival association and emphasizing the need for multiple hypothesis correction. Overall, our correction for multiple hypothesis testing reveals the lack of a TMB threshold that can classify patients into groups with significantly different survival. In particular for lung cancer, for which we previously observed a significant association between TMB and clinical benefit, we obtain a corrected p-value of 0.06 among smokers for lung1 and 0.23 for lung2. Of note, lung1 cohort that has p-value is close to significance contains 50% more EGFR patients than lung2, further suggesting that the observed weak association might be due to confounding effects to EGFR mutation status and/ or COPD status (see above).

We also ran our analysis using OS (for datasets where both are available: mel1, mel2 and lung1) instead of PFS as an endpoint and showed similar results, suggesting that survival definitions do not drive the results of our analysis (**Figure 2.S5**). A similar analysis for individual cancer types (bladder, melanoma and non-small cell lung cancer; **Figure 2.S6**) from CPI1000+ shows the lack of significant TMB threshold that can differentiate patients with significantly distinct survival rates.

We further obtained consistent results for 1662 patients of MSK-IMPACT cohort treated with ICB but genotyped with gene panels rather than whole-exome sequencing (**Figure 2.S6**). Most of the 10 cancer types tested had a non-significant p-value including colorectal cancer (p=0.088) and melanoma (p=0.093) which have marginally significant p-values, and except for non-small cell lung cancer (p=0.034). This study did not provide additional information such as tumour location for melanoma, Microsatellite Instability (MSI) status for colorectal cancer, or COPD for non-small cell lung tumours, which can confound the association of TMB with response [90,108].

Taken together, our analysis shows the lack of a single TMB threshold that establishes a high-TMB group with a significantly longer survival.

2.4.6 TMB and cancer immunogenicity

Neoantigen theory is widely used to argue that cancers with high TMB are more likely to elicit an immune response after ICB. Although our results show the lack of such dependence, we demonstrate that the effect we observe

can nevertheless be explained by a simple mathematical model of neoantigens and immunogenicity.

Our model (**Materials and Methods**) aims to explain (i) the lack of association between TMB and response; and (ii) the response by cancers with even very low TMB; (iii) the lack of detectable selection against neoantigens [109]. In our model, each mutation has a probability $P_{immunogenic}$ to become immunogenic, i.e. to be expressed and presented as an epitope, to interact with the major histocompatibility complex, and to trigger an immune response (**Figure 2.4**). To include possible limited sensitivity of the immune system, we further require that at least k_{crit} such mutations are present to mount an immune response (for k_{crit}=1, a single mutation that becomes immunogenic triggers a full response). The components of our model are illustrated in **Figure 2.4A** and further explained in **Materials and Methods.**

Figure 2.4B shows the probability of eliciting a response ($P_{immune\ response}$) as a function of TMB for a range of $P_{immunogenic}$ and k_{crit} values. Our model has two regimes: If individual mutations are unlikely to be immunogenic ($P_{immunogenic}$<0.1, **Figure 2.S8**), the response rate increases gradually with TMB, as widely expected by the neoantigen theory, but inconsistent with clinical data where, as we showed above, such increase of response with TMB is absent . On the contrary, if single mutations are likely to be immunogenic $P_{immunogenic}$>0.1, the probability of response saturates for TMB ≳ 10, making tumours respond to ICB irrespective of TMB, as we observed above. For $P_{immunogenic}$ in the range estimated in silico [110] (0.22 for weak binders to T cells, and 0.64 for strong binders) and $k_{crit} \simeq 1\text{-}2$, the probability of eliciting a response quickly approaches 1 for TMB≳10 and stays constant and independent of TMB. (**Materials and Methods).** The model further suggests that for the regime consistent with the

data ($P_{immunogenic}$=0.2-0.6; $k_{crit} \simeq 1\text{-}2$) (i) >90% of tumours with as little as 10 non-synonymous mutations are immunogenic; (ii) when 90% of tumours are immunogenic they have on average as few as 2 immunogenic mutation. Such quick saturation of immunogenicity with TMB in our model suggests that further immunogenic mutations experience not negative selection (i.e. threshold epistasis), as was recently demonstrated [NatGenetics] These results are also consistent with recently observed immunodominance hierarchies of the T cell responses [111]: low TMB tumours can mount responses as robust as high TMB tumours since only a small subset of neoantigens are targeted by T cells.

Taken together, our model and analysis of the available data indicate that cancer with even very few mutations can be immunogenic, suggesting that patients with low TMB might also mount robust immune responses, as has been recently shown for pediatric patients with acute lymphoblastic leukaemia [111].

2.5 Discussion

Tumour Mutational Burden, a measure of the total somatic nonsynonymous mutations in a tumour, recently became a popular biomarker of response to ICB, notably because of its relative simplicity to assess.

However, this paradigm is largely based on a series of early papers that examined response in melanoma and lung cancer that we show here to be potentially problematic statistically and further confounded by tumour subtype. Several recent studies have also reported poor association of TMB with response for specific cancer types [cite:recent anti-TMB], and highlighted TMB

and its expression/presentation-based derivatives as problematic for clinical cohort classification [93]. In particular for melanoma, recent analyses [90] and our results indicate that the site location can explain the observed association between TMB and response to ICB. For lung cancer, our analysis points to the possibility that co-occurrence with COPD may explain the association between TMB and response to ICB among smokers. Overall we demonstrate that while most cohorts and cancer types show the lack of association of TMB and response or survival, the remaining statistical signal in some cohorts can arise due to confounders such as clinical subtypes. Future studies can examine the underlying biology -including TMB and neoantigens- explaining the better responses to ICB in certain clinical and cancer subtypes.

Critically, even if responders show significantly but slightly elevated TMB, such associations do not imply the suitability of TMB as a biomarker of response. In particular, we show that no TMB cutoff can distinguish groups of patients with significantly different survival rates. Besides, we show that TMB has poor accuracy as a classifier of response, even in the best-case scenario (Youden optimal cutpoint). This result challenges a recent FDA approval of TMB for prioritizing patients for ICB. If implemented, such TMB-based clinical decision making would deprive many patients who can benefit from ICB from receiving a life-extending treatment.

A recent ICB clinical trial that used FDA-approved TMB threshold (KEYNOTE-158) [112] has focused on rare cancers, excluding melanoma and lung cancer. While claiming a higher response rate among high-TMB patients, the trial observed little, if any, difference in overall survival of high-TMB and other patients, putting in question the clinical use of TMB-based prioritization.

We also put forward a simple model that reconciles our findings with the neoantigen theory. Our model shows that if each mutation has a high chance of triggering an immune response, then only a few new mutations make a cancer immunogenic, consistent with the observed immunodominance when the immune response is mounted against only a few of the neoantigens. This result is also consistent with the observed lack of association between antigen density and T-cell presence previously reported [113]. Moreover, our model suggests that most cancers are immunogenic, arguing that failures of ICB likely arise due to factors independent of cancer immunogenicity. Quantitative measurements [114] and modelling of neo-antigenic effects can deepen our understanding of cancer development and response to immunotherapy.

Although attractive and scalable, TMB does not consider the effect of specific mutations (missense, frameshift etc), their presentation and clonality [85], nor the state of the tumour, its microenvironment, and interactions with the immune system that can be integrated into potentially better predictors of response to ICB [75,115]. For the biology of oncoimmunity, our analysis suggests that, contrary to the neoantigen theory, cancer immunogenicity does not increase with the growing load of neoantigens, and that clinical subtypes can underlie better response to ICB.

Altogether, our analysis indicates that low TMB should not be used to deprive otherwise eligible patients of immunotherapy treatment, and stimulates further research into other determinants of response to immunotherapy.

2.6 Material and Methods

2.6.1 Immunotherapy study population

CPI800+ was formed of eight independent WES cohorts (n=882, detailed in **Table 2.S1**). The TMB and clinical annotations were not modified from the original studies. Post ICB sequenced samples were excluded from our analysis. In addition, gene panel datasets (n=1662, detailed in **Table 2.S1**) were identified from cbioportal [116].

2.6.2 TCGA data

Lung cancer TCGA data were also retrieved from cbioportal [116], and additional clinical annotations were downloaded from The Cancer 3′ UTR Atlas [117]. COPD status was assessed based on the standard spirometric classification, i.e. post-bronchodilator ratio of forced expiratory volume in one second (FEV1) and forced vital capacity (FVC) below 70%.

2.6.3 Statistical analysis

We used R version 3.6.2 to perform statistical analyses. Two-group comparisons were evaluated by a two-sided Mann–Whitney U test unless otherwise indicated. $P < 0.05$ was considered statistically significant.

2.6.4 Code availability

The R code and data used to reproduce the analysis and figures from the paper are available on GitHub https://github.com/mirnylab/TMB_analysis

2.6.5 Model of cancer immunogenicity

Response to ICB treatment requires that the cancer is immunogenic and that immunotherapy can mount the immune response to this immunogenic cancer: $P_{response} = P_{immune\ response} * P_{therapy}$. The probability that immunotherapy works, given that the cancer is immunogenic, $P_{therapy}$, depends on the specifics of treatment and other physiological variables, so we'll assume it to be constant. The probability of being immunogenic, $P_{immune\ response}$, however, depends on the ability of mutations to trigger the immune response. Assume that every nonsynonymous mutation has the probability $P_{immunogenic}$ (noted below as p) to be expressed and presented as an epitope, to interact with the major histocompatibility complex, and to trigger an immune response.

In the scenario of immunodominance, in which the immune response is mounted against only a few of the neopeptides, only $k \leq k_{crit}$ such mutations are sufficient to mount an immune response. Hence, the probability of being immunogenic is the probability of having at least k_{crit} presented and immunogenic mutations out of TMB:

$$P_{immune\ response} = \sum_{k=k_{crit}}^{TMB} p^k(1-p)^{TMB-k}C^k_{TMB} \approx 1 - \sum_{k=0}^{k_{crit}-1} \text{Poisson}(k, TMB*p),$$

where $p=P_{immunogenic}$.

In the case of k_{crit} =1, even a single mutation, if immunogenic, can trigger a response yielding $P_{immune\ response}$= 1-(1-p)^{TMB}≈1-exp(-TMB*p). It is easy to see that for this case $P_{immune\ response}$ saturates at p*TMB~1. Thus to achieve approximately constant $P_{immune\ response}$ for TMB>10-20, one needs p>0.1 for k_{crit} =1. Achieving a similar effect for k_{crit} >1 (i.e. $P_{immune\ response}$ that doesn't depend on TMB for TMB>10-20) requires even higher p>0.2. Moreover for k_{crit} =1, one can estimate the expected number of immunogenic mu[118]tations (p*TMB) present when 90% of cancers are immunogenic: 0.9=$P_{immune\ response}$≈1-exp(-TMB*p), gives p*TMB=2.3. I.e. irrespective of specific value of p, when 90% of cancers are immunogenic they carry only ~2 immunogenic mutations.

Furthermore, our model also explains a puzzling observation that immunoediting, i.e. negative selection against immunogenic mutations, is inefficient, allowing tumours to accumulate a high TMB [109,118]. Indeed, once a cancer accumulates mutations making it immunogenic, additional mutations incur no additional selective disadvantage i.e. show "the epistasis of diminishing return", and hence accumulate as neutral or weakly damaging passenger mutations [119–121] Moreover, according to this argument, cancer would have to develop means to suppress the immune response early in its development, a prediction that can be tested in future studies of cancer clonal evolution.

2.7 Acknowledgement

We are grateful for Krupa Thakkar, Kevin Litchfield and Charles Swanton for running our analysis on their recent "CPI1000+" meta-aggregate of immunotherapy patients [92].

We are thankful for many productive and exciting discussions of this work to Christopher McFarland, Johannes Berg, Donate Weghorn, Eli van Allen, Kenneth Kehl, Shamil Sunyaev, Martha Luksza, Michael, Lassig, Boris Reizis, Virginia Savova, Gregory Kryukov, Sebastian Amigorena, Sean McGrath, Baptiste Boisson, Toni Choueiri, Paul B Robbins and all members of the Mirny lab. We are grateful to the organizers and participants of “Physicists working on Cancer” workshop at the Weizmann Institute of Science, Schwartz/Reisman Institute for Theoretical Physics, particularly, Eytan Domany, Herbert Levine, Caterina La Porta and Stefano Zapperi. We are also grateful to the organizers and participants of the workshop “Tumors and Immune Systems: From Theory to Therapy” seminar at Institut d'Etudes Scientifique de Cargèse, particularly, Alexandra Walczak, Thierry Mora, Vassili Soumelis, Paul Thomas and Jason George.

LJL was supported by the Training Program in Bioinformatics and Integrative Genomics (NIH T32HG002295, PI: P.Park). This project grew from the qualifying examination problem in Bioinformatics and Integrative Genomics given by LAM to LJL. We acknowledge support of the MIT-France Seed Fund, and The Chicago Region Physical Science Oncology Center (PS-OC, National Cancer Institute U54CA193419).

2.8 Figures

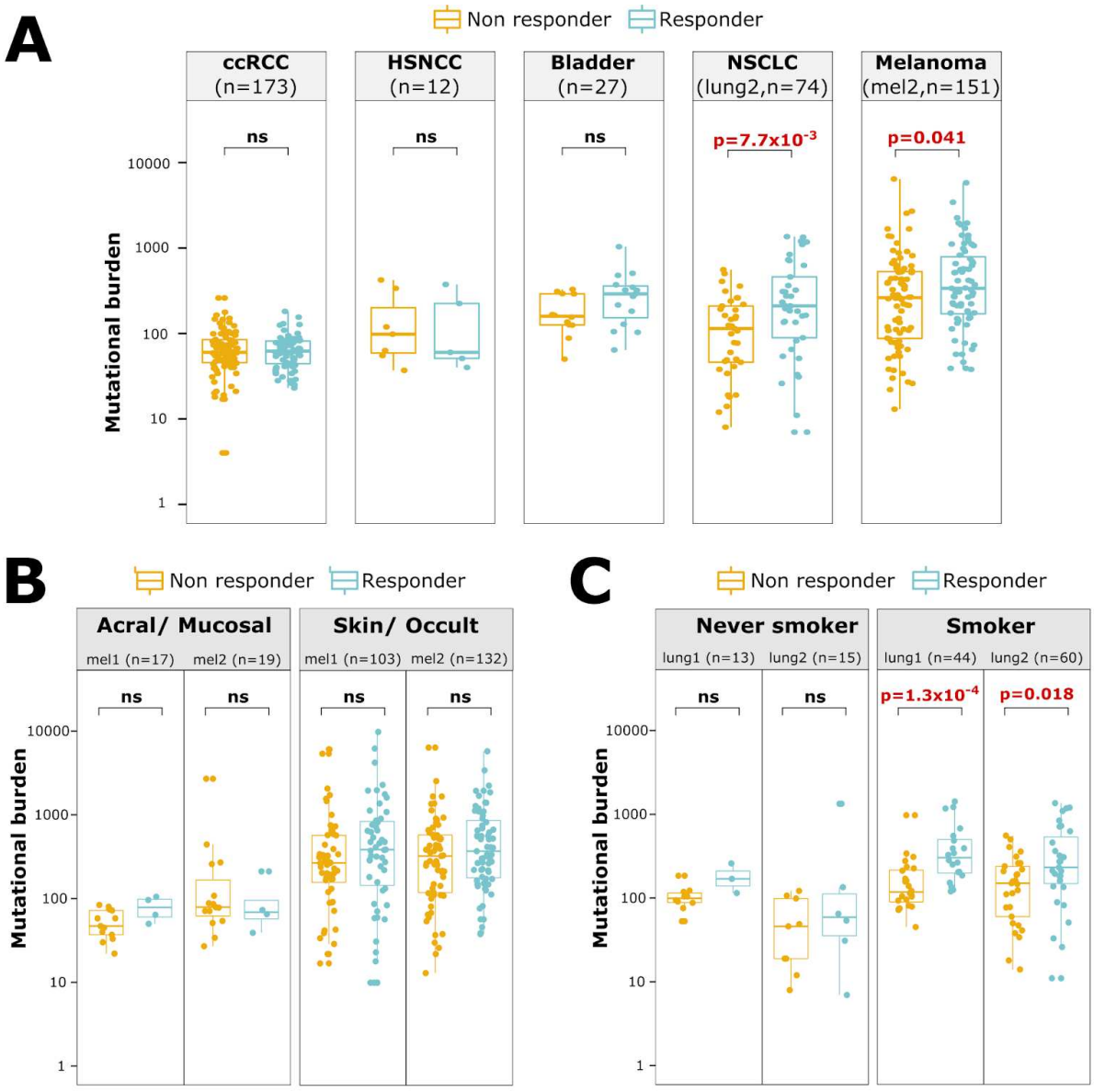


Figure 2.1: TMB association with clinical benefit from ICB across cancers

(A) *Association of TMB with response to ICB across five cancer types from CPI800+ (the largest cohorts of each cancer type are plotted here, the others are shown in Figure S1A). Only melanoma and non-small cell lung cancer have a significantly different TMB between responders and non-responders. ccRCC:*

clear cell Renal Cell Carcinoma, HNSCC: Head and Neck Squamous Cell Carcinoma; NSCLC: Non Small Cell Lung Carcinoma ***(B)*** *Association of TMB with response to ICB for specific melanoma subtypes. When split into subtypes, TMB does not associate with response to ICB.* ***(C)*** *Association of TMB with response to ICB for subtypes of non-small cell lung cancer patients. When split into subtypes, TMB associates with response to ICB only among smokers.*

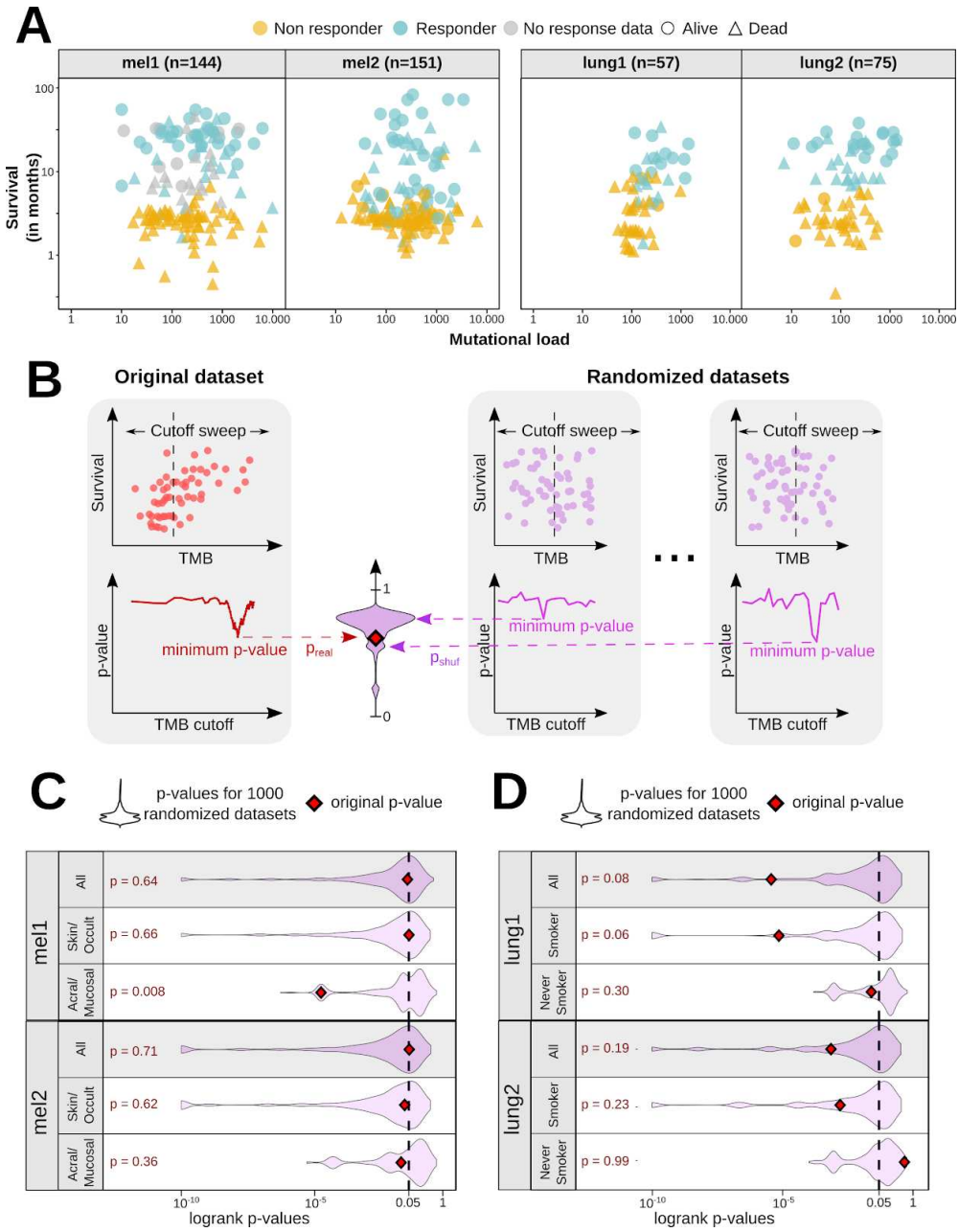


Figure 2.2: TMB association with progression-free survival post-immunotherapy

(A) *Plots of progression-free survival and TMB for melanoma and lung cancer ICB cohorts show the lack of correlation or of an obvious TMB cutoff. (Similar plots by cancer subtypes are shown in* ***Figure 2.S3****)* ***(B)*** *Overview of the randomization analysis. Left: the optimal cutoff is found to maximize the*

difference between survival between groups above and below the cutoff (i.e to minimize the logrank p-value, yielding p_{real}). Right: the same procedure for shuffled data yields p_{shuf}. The fraction of $p_{shuf}<p_{real}$ produces a p-value corrected for multiple hypothesis testing for non-independent tests. ***(C)*** *Results of the randomization analysis in the melanoma cohorts and stratification by subtypes (p-values < 10^{-10} not shown)* ***(D)*** *Randomization analysis results in the lung cancer cohorts and stratification by subtypes (p-values < 10^{-10} not shown).*

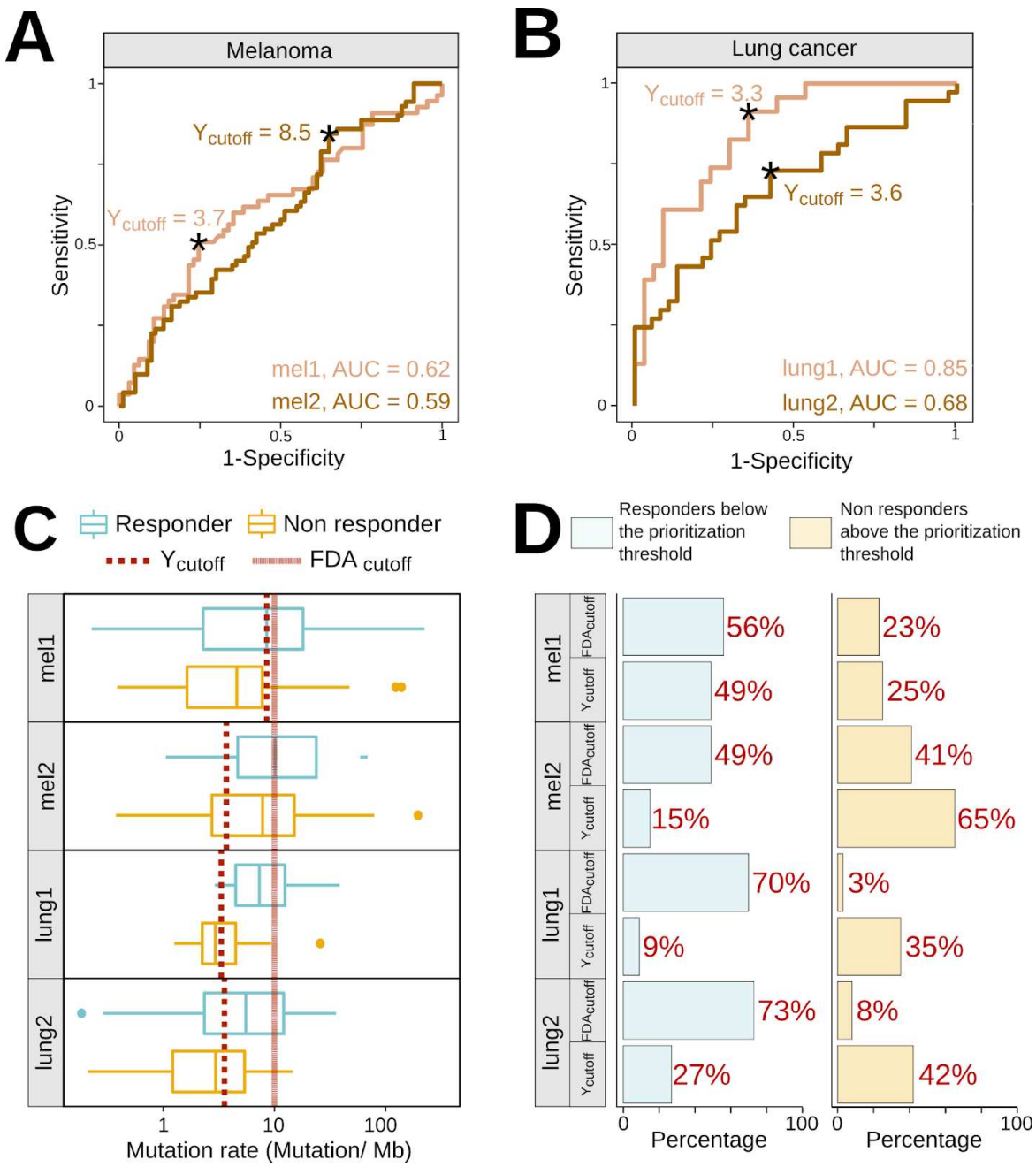


Figure 2.3: TMB as a biomarker of response to immunotherapy

(A)(B) *ROC curves for the melanoma and lung cancer cohorts. Youden index associated cutoffs are also plotted.* ***(C)*** *Boxplots of nonsynonymous mutation rates across responders and non responders in the melanoma and lung cancer cohorts. The FDA-approved cutoff (10 mutations/Mb) and the best cutoff*

(Youden index associated cutoff) are shown by vertical lines. ***(D)*** *Proportion of misclassified patients based on the FDA-approved cutoff, as well as the Youden index cutoff for each dataset. The use of either cutoff leads to substantial fraction of misclassified patients (potential responders below the treatment cutoff, or non-responders above the cutoff).*

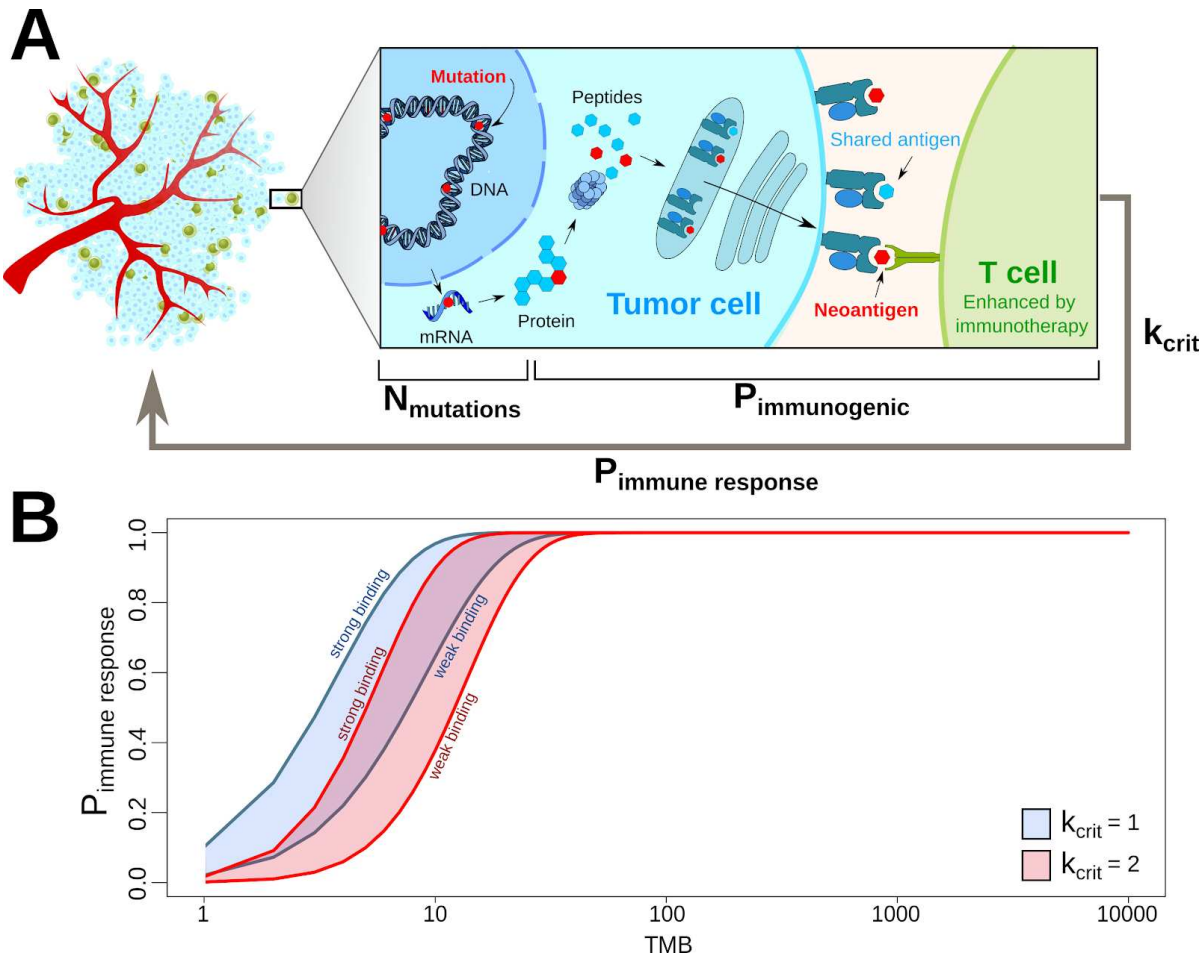


Figure 2.4: TMB and cancer immunogenicity

(A) *Our model of cancer immunogenicity coarse-grains several cellular processes into the probability that a mutation becomes immunogenic ($P_{immunogenic}$). If the number of immunogenic mutations reaches k_{crit}, the cancer triggers an immune response* ***(B)*** *The probability of immune response $P_{immune\ responce}$ as a function of TMB for a range of k_{crit} and $P_{immunogenic}$. Rapid saturation of $P_{immune\ responce}$. TMB requires low k_{crit} and sufficiently high $P_{immunogenic}$>0.1 (see* ***Materials and Methods****).*

2.9 Supplementary Figures

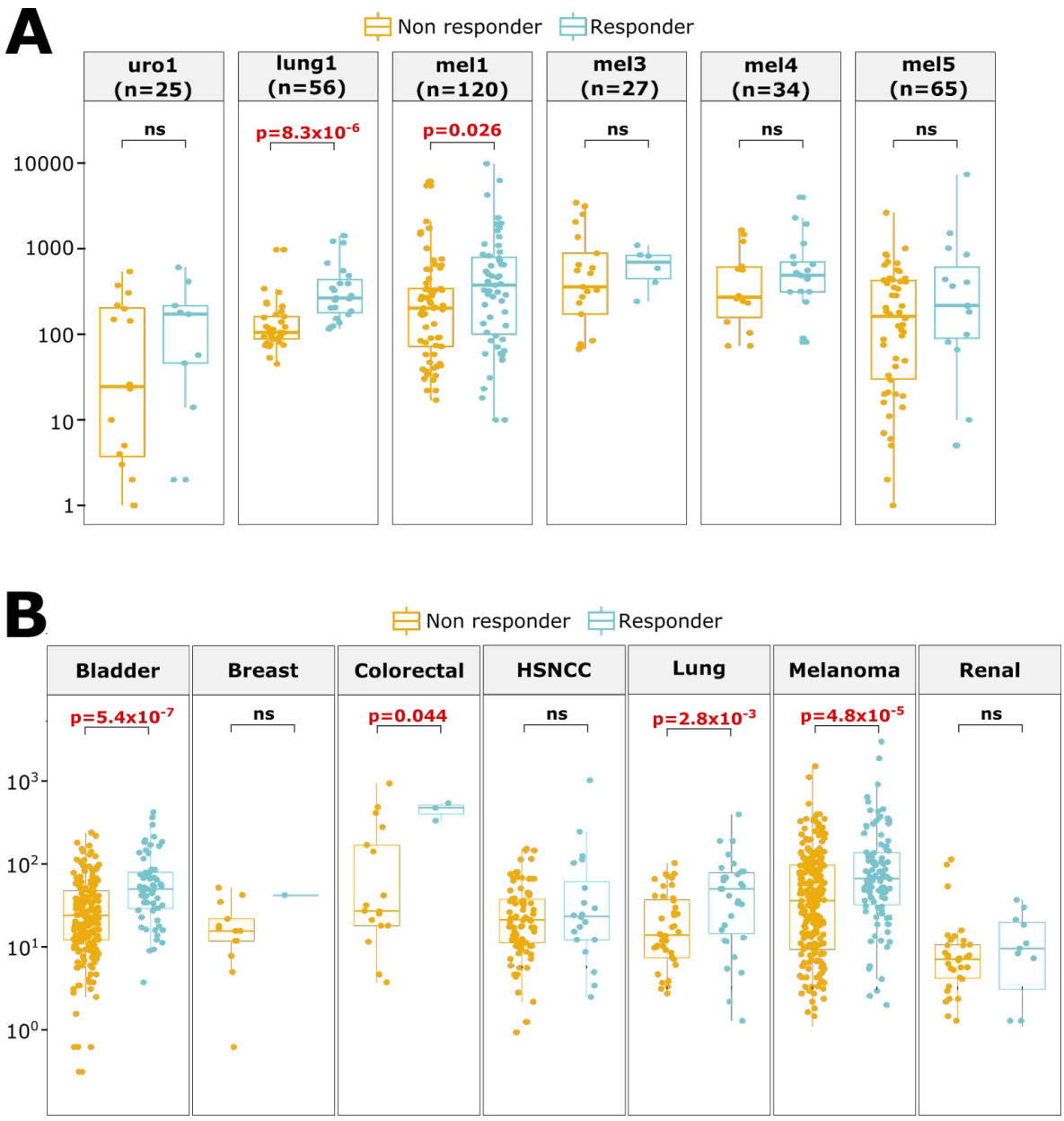


Figure 2.S1: TMB association with clinical benefit from ICB across cancers

(A) *Association of TMB with response to ICB across three cancer types from CPI800+. Only melanoma and non-small cell lung cancer have a significantly different TMB between responders and non-responders.* ***(B)*** *Association of TMB with response to ICB across seven cancer types from CPI1000+. Melanoma, non-small cell lung, bladder and colorectal cancer have a significantly different TMB between responders and non-responders.*

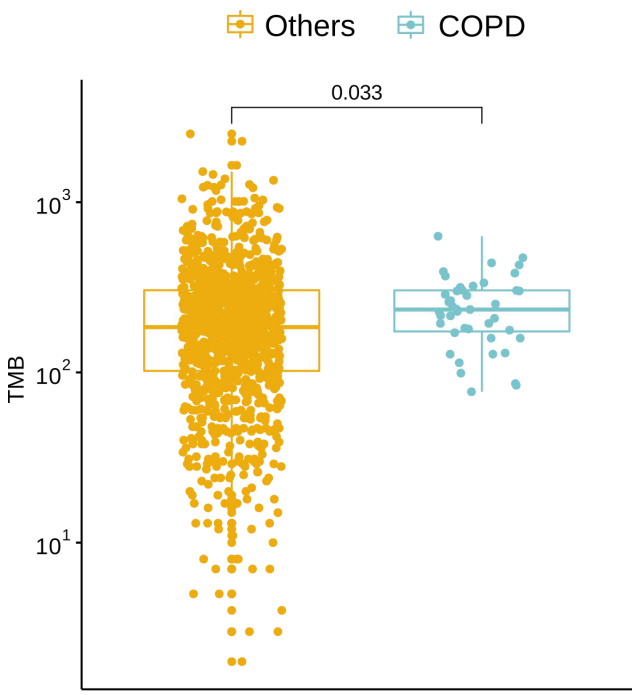


Figure 2.S2: Chronic obtrusive pulmonary disease status and TMB

Association of TMB with COPD in TCGA. Of the 83 patients with COPD data, 43 were diagnosed with COPD. Here we compare COPD patients (n=43) to the rest of the cohort (n=1101).

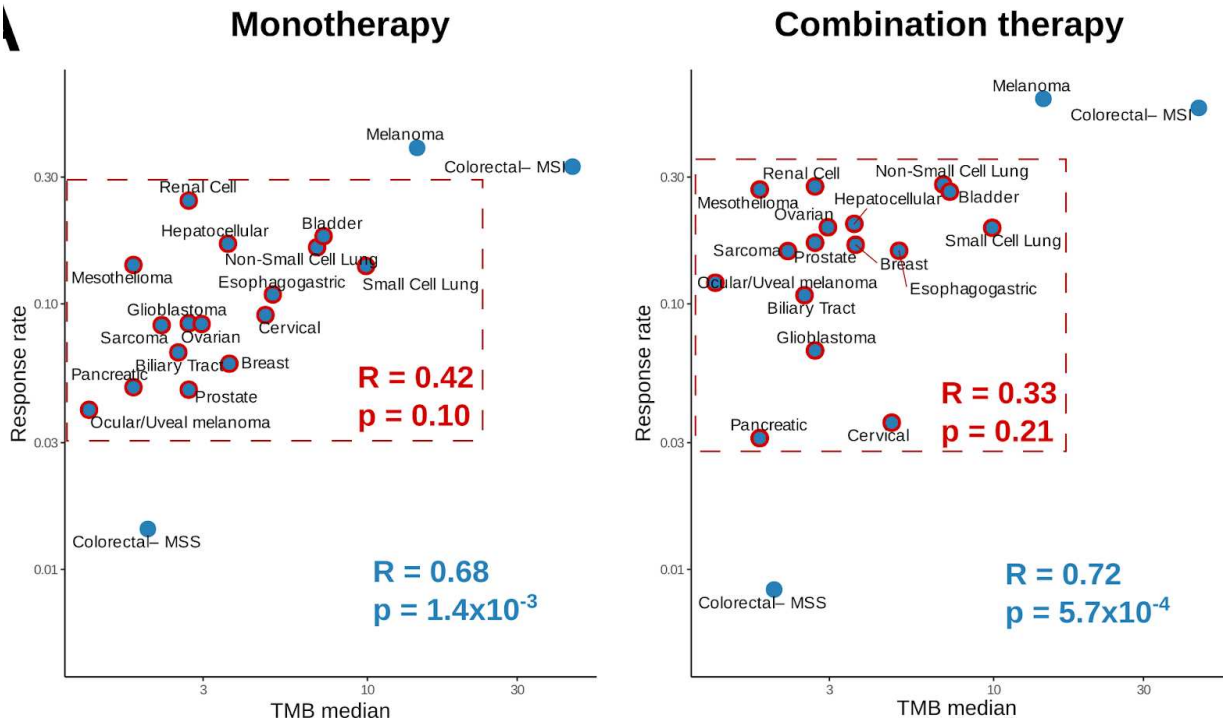


Figure 2.S3: Correlation between response rates and TMB across cancer types

Median TMB in 19 cancer types of patients who underwent immunotherapy treatment (monotherapy or combination therapy). Pearson R correlation coefficient and p-value was calculated for all patients (in blue) and a subset of patients (red box and values) after removing melanoma and colorectal cancers.

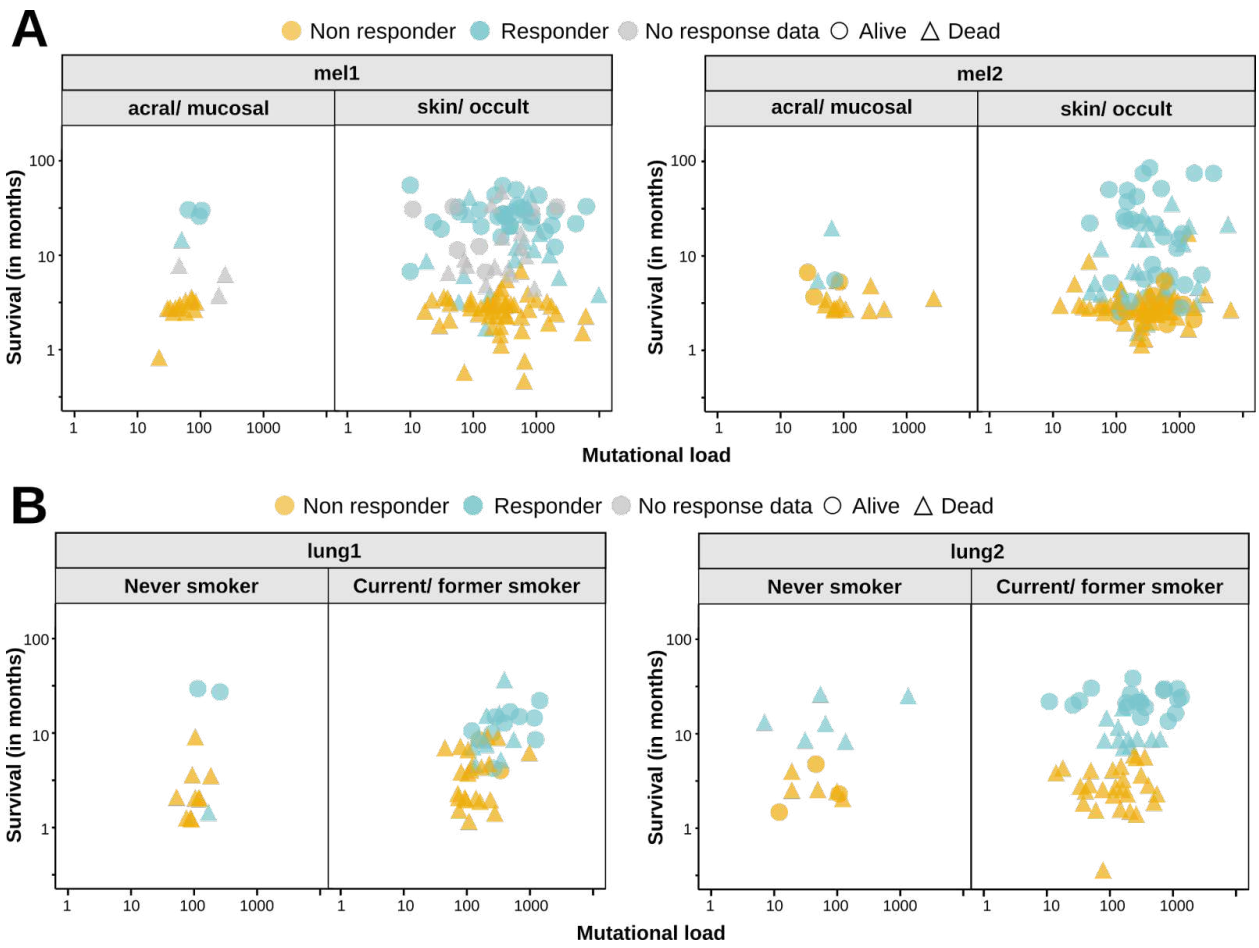


Figure 2.S4: TMB association with progression-free survival post-immunotherapy

(A) (B) Plots of progression-free survival and TMB for melanoma and lung cancer ICB cohorts labeled by cancer subtype, showing the lack of correlation or of an obvious TMB cutoff.

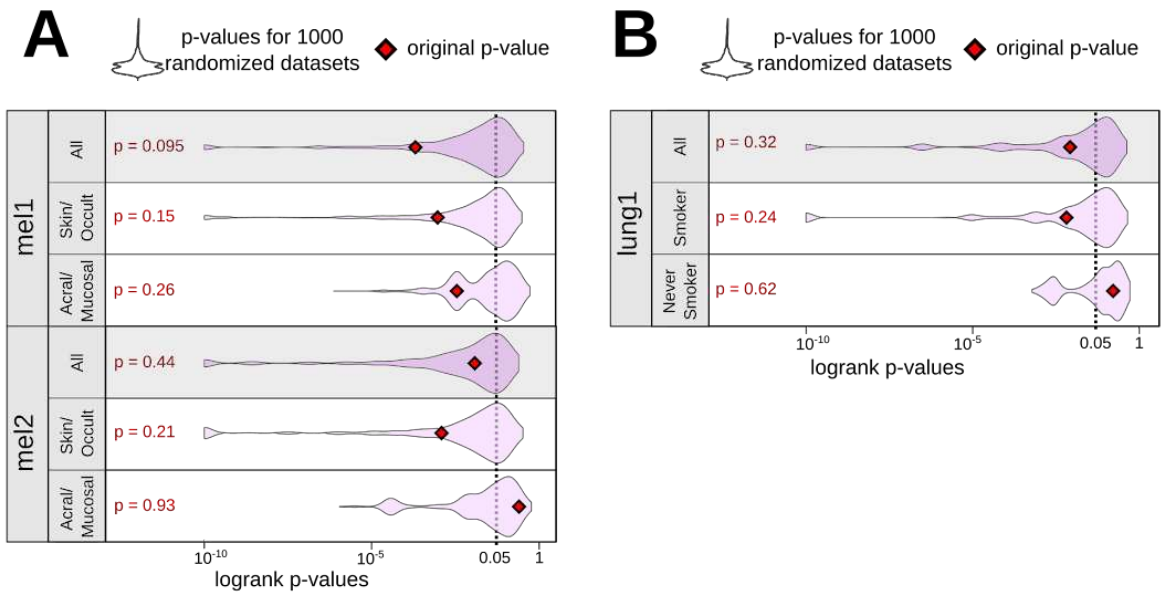


Figure 2.S5: TMB association with overall survival post-immunotherapy

(A) *Randomization analysis results in mel1 and mel2 and stratification by subtypes (p-values < 10^{-10} not shown)* ***(B)*** *Randomization analysis results in lung1 and stratification by subtypes (p-values < 10^{-10} not shown). When corrected for multiple hypotheses all cohorts fail to provide a statistically significant cutoff.*

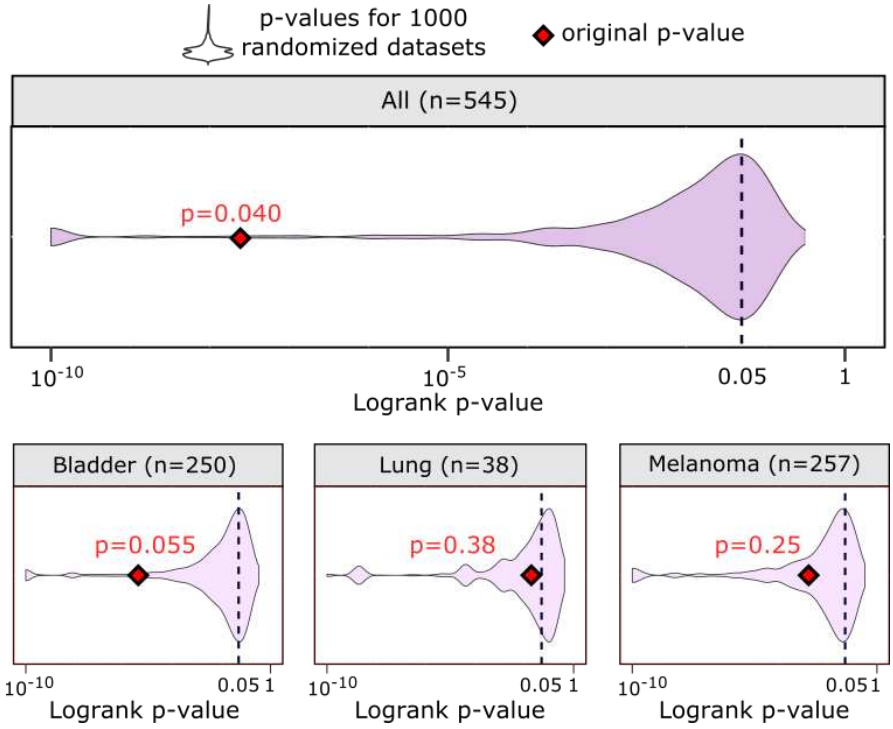


Figure 2.S6: TMB association with overall survival post-immunotherapy

Results of the randomization analysis in CPI1000+ (p-values < 10-10 not shown).When cancer types of CPI1000+ were combined, a nominally significant p-value (p=0.04) arises, likely due to cancer types with different TMB ranges showing significantly different survival rates to ICB.

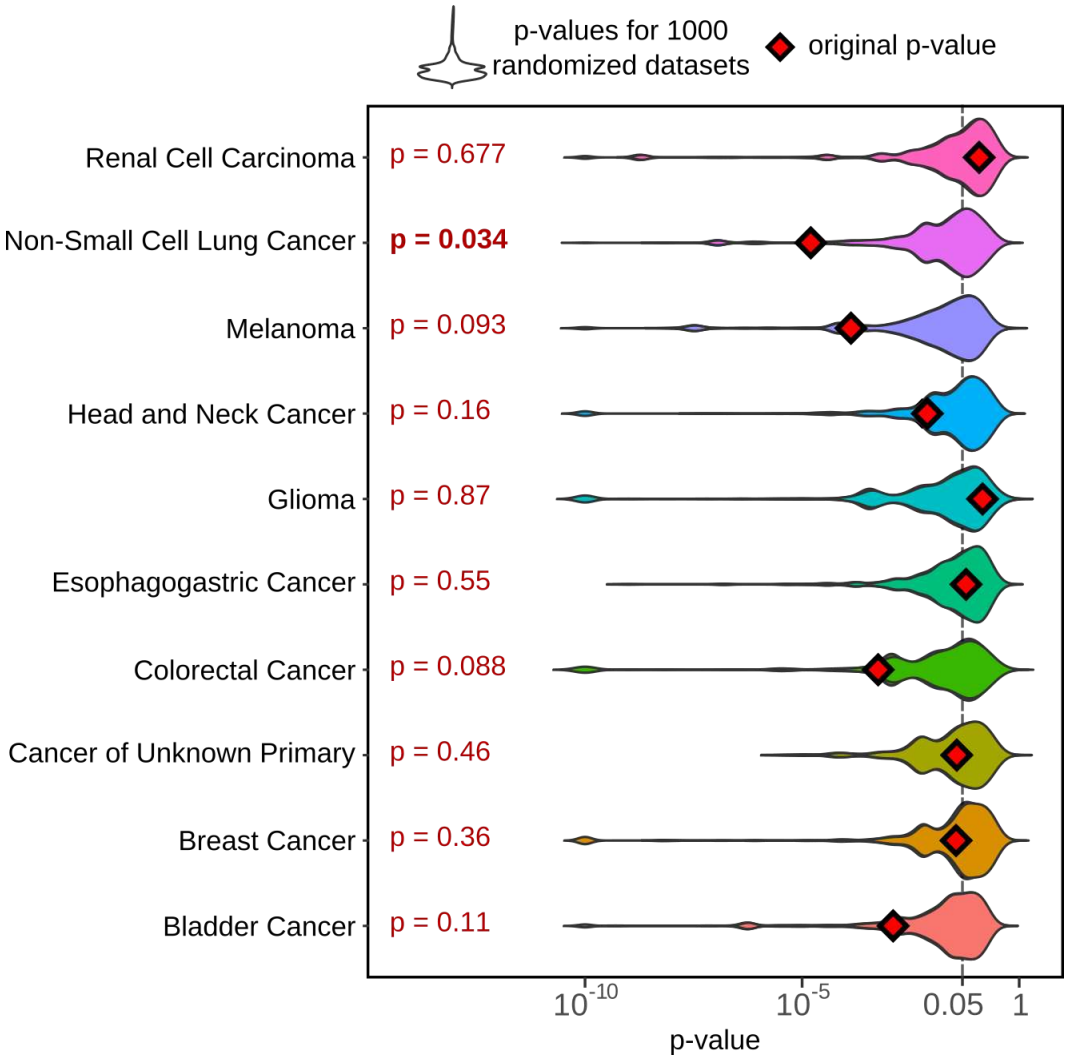


Figure 2.S7: TMB association with overall survival post-immunotherapy

Randomization analysis results in multiple cancer types with MSK-IMPACT targeted next-generation sequencing data (p-values < 10^{-10} not shown)

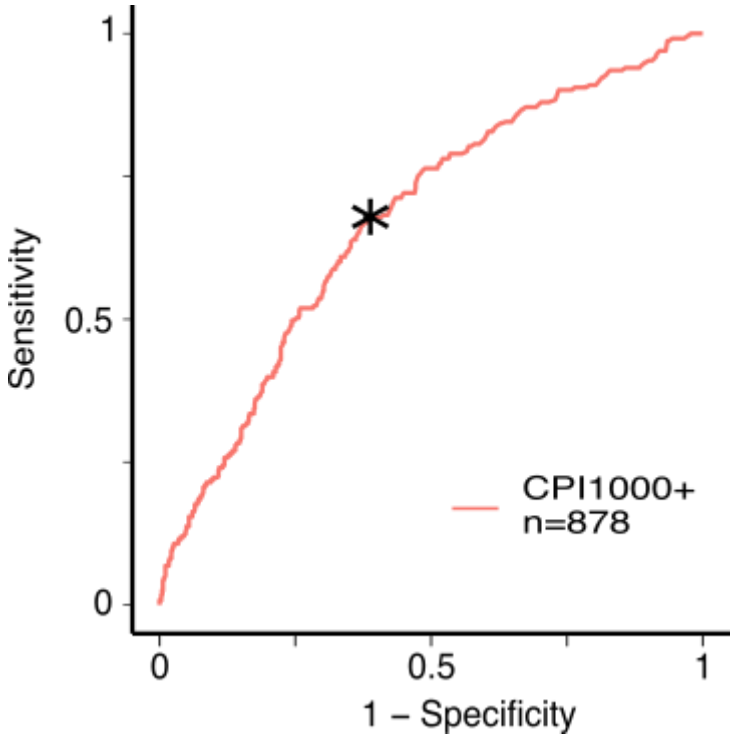


Figure 2.S8: TMB as a biomarker of response to immunotherapy

ROC curve for CPI1000+. The Youden index associated cutoffs is also plotted.

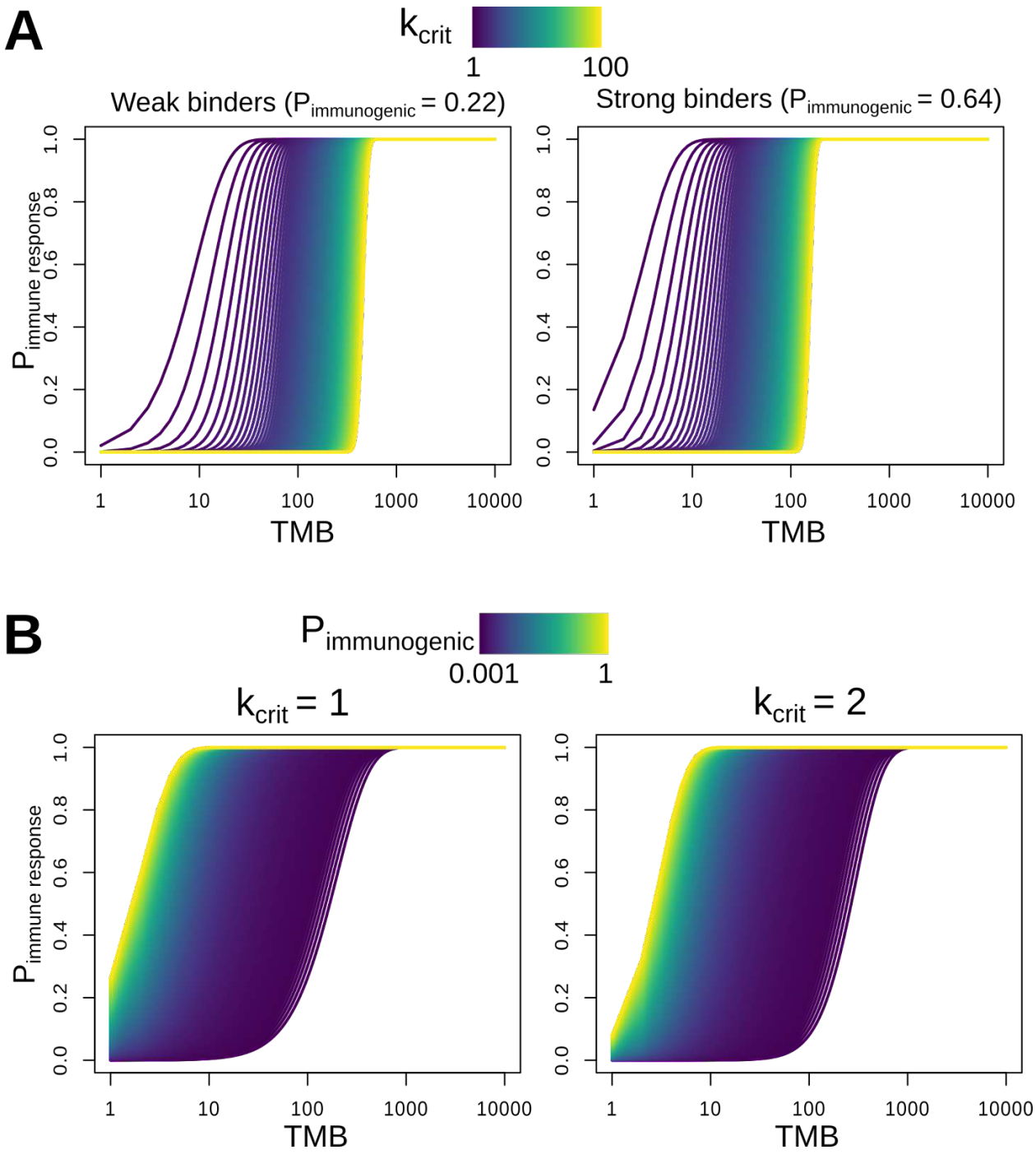


Figure 2.S9: Components of cancer immunogenicity

(A) Probability of eliciting an immune response for a range of k_{crit} values

(B) Probability of eliciting an immune response for a range of $P_{immunogenic}$ values

3. Intrinsic resistance to immune checkpoint blockade in a mismatch repair deficient colorectal cancer

Published in Cancer Immunology Research on June 19, 2019.

3.1 Author information

Carino Gurjao,[1,2] David Liu,[1,2] Matan Hofree,[2] Saud H. AlDubayan,[1,2] Isaac Wakiro,3 Mei-Ju Su,4 Kristen Felt,[5] Evisa Gjini,[5] Lauren K. Brais,[1] Asaf Rotem,[3] Michael H. Rosenthal,[6] Orit Rozenblatt-Rosen,[2] Scott Rodig,[5,7] Kimmie Ng,[1] Eliezer M. Van Allen,[1,2,3] Steven M. Corsello,[1,2] Shuji Ogino,[2,8,9,10] Aviv Regev,[2] Jonathan A. Nowak,[8,9] and Marios Giannakis[1,2]

1. Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
2. Broad Institute of MIT and Harvard, Cambridge, MA, USA
3. The Center for Cancer Precision Medicine, Dana-Farber Cancer Institute, Boston, MA, USA
4. Biotherapeutic and Medicinal Sciences, Biogen, Cambridge MA, USA
5. Center for Immuno-Oncology, Dana-Farber Cancer Institute, Boston, MA, USA
6. Department of Radiology, Dana-Farber Cancer Institute, Brigham and Women's Hospital, and Harvard Medical School, Boston, MA, USA
7. Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA

8. Department of Oncologic Pathology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA, USA
9. Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
10. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

3.2 **Abstract**

Immunotherapy with checkpoint inhibitors, such as the programmed death-1 (PD-1) antibodies pembrolizumab and nivolumab, are effective in a variety of tumors, yet not all patients respond. Tumor microsatellite instability-high (MSI-H) has emerged as a biomarker of response to checkpoint blockade, leading to the tissue-agnostic approval of pembrolizumab in MSI-H cancers. Here we describe a patient with MSI-H colorectal cancer that was treated with this immune checkpoint inhibitor and exhibited progression of disease. We examined this intrinsic resistance through genomic, transcriptional, and immunohistochemical characterization of the patient's tumor and the associated immune microenvironment. The tumor had typical MSI-H molecular features, including a high neoantigen load. We also identified biallelic loss of the gene for β_2-microglobulin (*B2M*), whose product is critical for antigen presentation. Immune-infiltration deconvolution analysis of bulk transcriptome data from this anti–PD-1–resistant tumor and hundreds of other colorectal cancer specimens revealed a high natural killer (NK) cell and M2 macrophage infiltration in the patient's cancer. This was confirmed by single-cell transcriptome analysis and multiplex immunofluorescence. Our

study provides insight into resistance in MSI-H tumors and suggests immunotherapeutic strategies in additional genomic contexts of colorectal cancer.

3.3 Introduction

Immune checkpoint inhibitors, such as programmed cell death 1 (PD-1, *PDCD1*) antibodies, have revolutionized cancer treatment by demonstrating long-lasting responses in patients with several types of malignancies[122]. However, only a subset of patients experience benefit from these agents and complete response remains uncommon. In this context, tumor DNA mismatch repair deficiency (dMMR) and a high-level of microsatellite instability (MSI-H) have emerged as powerful genomic markers of response to immune checkpoint inhibitors across malignancies[123,124], leading to the tissue-agnostic FDA-approval of the PD-1 antibody pembrolizumab in refractory dMMR/MSI-H solid malignancies and to the approval of PD-1 antibody nivolumab with or without the CTLA-4 antibody ipilimumab in dMMR/MSI-H colorectal cancer (CRC) after fluoropyrimidine, oxaliplatin and irinotecan-based chemotherapy. The leading proposed reason for the immunogenicity of dMMR tumors is their high mutational and neoantigen burden[25]; however, only 30–55% of patients with such cancers respond to immune checkpoint blockade with another 10–28% of patients remaining primarily refractory to immunotherapy[123–126]. To date, the molecular and microenvironmental features of dMMR/MSI-H tumors that are intrinsically resistant to immune-checkpoint blockade remain unknown. Their characterization could provide insights for novel combination

immunotherapies in this subset of tumors and also inform resistance, and strategies to overcome it, in additional genomic contexts.

Here, we describe a patient with metastatic dMMR CRC who was treated with pembrolizumab after combination chemotherapy. Despite having confirmed dMMR/MSI-H status and a high neoantigen load, her disease progressed on pembrolizumab. To analyze the basis of this intrinsic immune checkpoint inhibitor resistance, we performed bulk and single-cell characterization of her tumor and the associated immune microenvironment.

3.4 Materials and Methods

3.4.1 Patient study

The patient provided written consent to participate in research protocols for additional core biopsies and research testing. All biopsies and molecular testing were performed in accordance with protocols approved by the IRB at the Dana-Farber Cancer Institute.

3.4.2 Statistical analyses

We used R-3.4.4 to perform the statistical analyses. For two-group comparisons, significance was evaluated by the Mann–Whitney *U* test for

non-normal distributions, and with a two-tailed student *t* test otherwise. *P* values of < 0.05 were considered statistically significant.

3.4.3 Bulk sequencing

DNA and RNA extractions from Formalin Fixed Paraffin Embedded (FFPE) sections and peripheral blood were carried out using standard methods[127]. Whole-exome sequencing (WES) was performed as detailed previously[128] on the pre-immunotherapy tumor and peripheral blood, with mean depth of coverage of 270× and 101×, respectively. For bulk whole-transcriptome sequencing (RNA-seq), we used the TCap (Transcriptome Capture) protocol (genomics.broadinstitute.org/products/whole-transcriptome-sequencing), which is optimal for low-input and degraded samples such as FFPE samples. Using this method, RNAseq was performed on the pre-treatment tumor with >22.000 genes and 99.4% exons detected.

3.4.4 Single-cell sequencing

The core biopsy was received in additive free M199 media (Thermofisher Scientific; #11150059). To generate a cell suspension for single-cell RNA-seq (scRNA-seq), the core was minced into smaller ~1-mm pieces, which were then dissociated by a combination of mechanical and enzymatic digestion with Accumax (Innovative Cell Technologies; #AM105) at room temperature for 10 minutes. Following dissociation, cells were strained

through a 100 µm strainer, washed with ice cold PBS (Ca/Mg free) with 2% FCS and resuspended in 0.04% BSA (Thermofisher Scientific #AM2616) with PBS. From this suspension, two channels were loaded on 10x; one with 4000 cells and the other with 6000 cells. Libraries were prepared using established protocols. Droplet-based massively parallel scRNA-seq was performed using Chromium Single cell 3' Reagents Kits (v.1) according to the manufacturers protocols (10x Genomics). The generated scRNA-seq libraries were sequenced using 100 cycle Illumina HiSeq. After quality control, 595 resulting cells were used for further analyses.

3.4.5 Variant calling

Tumor somatic mutations were called from WES using standardized pipelines including MuTect for somatic SNV inference and Strelka for small insertion/deletions. We corrected for FFPE and oxoguanine artifacts, and used a panel of normal filter as previously described[87].

Tumor purity and ploidy were inferred using ABSOLUTE, and cancer cell fraction (CCF) of mutations (i.e. the proportion of tumor cells with the mutation) estimated. Allelic copy number alterations were inferred using an adaptation of a circular binary segmentation[129] and corrected for tumor purity and ploidy. The mutations discussed were orthogonally validated by a next-generation CLIA-certified sequencing panel[130]. In order to study the mutational signatures in the tumor of the patient, we used DeconstructSig based on linear combination analysis of preexisting signatures. POLYSOLVER was used to

detect the HLA type of the patient, which enabled neoantigen prediction using NetMHCpan as previously described[87].

3.4.6 MLH1 methylation testing

DNA methylation patterns in the CpG island of the MLH1 promoter gene were determined by chemical (bisulfite) modification of unmethylated cytosines to uracil and subsequent PCR using primers specific for either methylated or the modified unmethylated DNA[131]. The PCR products were analyzed by capillary gel electrophoresis.

3.4.7 Gene expression analysis

For bulk RNA-seq analysis, STAR and RSEM was used for alignment and gene expression quantification, respectively. Immune cell subset deconvolution was performed using CIBERSORT to assess the relative and total abundance of 22 immune cell types. For single-cell analysis, gene expression counts were obtained by aligning reads to the GRCh38 genome using Cell Ranger analysis pipeline (https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/2–1). The consensus molecular subtypes (CMS)[132] were called using “CMScaller” R package.

3.4.8 Immunohistochemistry and Multiplex Immunofluorescence

Tumor sections were deparaffinized and stained for Beta-2-Microglobulin (Polyclonal rabbit anti-human β_2-microglobulin, Dako A007202–2) and MMR proteins (as described in[133]) with standard immunohistochemistry protocols. Staining for multispectral imaging analysis was performed on a BOND RX automated stainer (Leica Biosystems) utilizing 5-µm thick section of FFPE tissue. After deparaffinization, rehydration and antigen retrieval, slides were serially stained with primary antibodies to Cytokeratin (clone AE1/AE3, DAKO), CD3 (Polyclonal, DAKO A0452), CD56 (clone 123C3; DAKO), followed by incubation with an anti-rabbit polymeric horseradish peroxidase secondary IgG (Poly-HRP, BOND Polymer Refine Detection Kit, Leica Biosystems). Signal for antibody complexes was labeled and visualized by Opal Fluorophore Reagents (PerkinElmer). Image acquisition was performed using the Mantra multispectral imaging platform (Vectra 3.0, PerkinElmer, Hopkinton, MA). Representative intratumoral regions of interest were chosen by a gastrointestinal pathologist (J.N.), and 3–5 fields of view (FOVs) were acquired at 20x resolution as multispectral images. Cell identification was performed as previously described[134]. In short, after image capture, the FOVs were spectrally unmixed and then analyzed using supervised machine learning algorithms within Inform 2.3 (PerkinElmer). Immune cell densities were then calculated based upon phenotyped cell counts and tissue areas.

3.5 Results

3.5.1 Case history

A 78-year-old woman with metastatic colon adenocarcinoma was admitted to the hospital with abdominal pain. CT imaging revealed a large heterogeneously enhancing paracolic mass (**Figure 3.1A**). The patient had a history of three metachronous early-stage colon adenocarcinomas: a stage II (pT3N0) descending colon primary, a stage I (pT1N0) ascending colon primary, and a stage III (pT3N1b) transverse colon primary. She previously underwent sequential left hemicolectomy, right hemicolectomy, and completion colectomy over a six-year period. Molecular testing of her stage III tumor showed a *BRAF* c.1799T>A (p.V600E) mutation and loss of MMR proteins MLH1 and PMS2 by immunohistochemistry. Following her completion colectomy, she received 12 cycles of adjuvant 5-fluorouracil, leucovorin, and oxaliplatin (FOLFOX). However, one year later she had disease recurrence in the upper abdominal mesentery and went on to receive 5-fluorouracil, leucovorin, and irinotecan (FOLFIRI) with bevacizumab for metastatic CRC. She tolerated this poorly and was changed to FOLFOX and bevacizumab with subsequent progression of disease after several months of treatment.

Biopsy of the recurrent tumor was recommended to confirm dMMR and MSI-H status. Ultrasound-guided abdominal mass biopsy was performed. Pathology revealed poorly differentiated adenocarcinoma with loss of nuclear MLH1 and PMS2 staining by immunohistochemistry, microsatellite instability in five of the five genomic markers tested and methylation of the MLH1 promoter. She was started on pembrolizumab (200 mg every 3 weeks) with restaging scans after 2

months interpreted as progression of disease. Given the possibility of pseudoprogression with immunotherapy, the patient was maintained on therapy, but another set of scans after 5 months of treatment showed clear disease progression (**Figure 3.1A**).

3.5.2 Mutation, copy number, and neoantigen analyses

To investigate for mechanisms of intrinsic resistance to immune checkpoint blockade in this dMMR tumor, we performed preimmunotherapy tumor and matched normal WES, immunohistochemistry, and multiplex immunofluorescence pathologic analyses as well as bulk and single-cell RNA-seq (**Figure 3.1B**). WES revealed a high mutation load with a total of 1857 somatic single nucleotide variants (SNVs) and small indels (**Supplementary Table S1**), a high neoantigen load and a quiet copy number landscape (**Supplementary Figure 3.S1A and 3.S1B**), as typical for dMMR/MSI-H CRC[135]. Mutational signature analysis demonstrated that the majority of mutations were stemming from a mutational signature associated with dMMR[43](**Supplementary Fig. 3.S1C**).

WES also confirmed the presence of a *BRAF* c.1799T>A (p.V600E) mutation as well as mutations in *RNF43*, a gene that is mutated in approximately 50% of MSI-H CRC[128]. To evaluate the possibility of an inherited cancer risk allele in this patient with a history of multiple tumors, germline coding variants in 14 established CRC risk genes as well as 40 cancer risk genes that are part of the DNA repair machinery (**Supplementary Table S2**) were called and evaluated for pathogenicity as previously described[136]. Our assessment showed no known

pathogenic or likely pathogenic germline mutations in neither the CRC risk gene set nor the DNA repair set. There was also no germline MLH1 promoter methylation. Thus, to the best of our current knowledge, the tumor appears to be sporadic.

The patient's tumor harbored a *B2M* frameshift deletion p.V47Afs*6 (CCF = 0.76) and had loss of heterozygosity (CCF = 0.97). The inactivation of *B2M* in this tumor is consistent with a clonal event in its evolution (**Figure 3.2A**). To validate B2M loss, we performed immunohistochemistry on the pre-immunotherapy tumor sample for B2M and confirmed complete loss of expression in the tumor cells (**Figure 3.2B and Figure 3.2C**). There were no biallelic inactivation events in other genes of the antigen presentation machinery/interferon-gamma pathway (*JAK1, JAK2, STAT1, STAT2, STAT3, CD274, PDCD1, PDCD1LG2, HLA-A, HLA-B, HLA-C, TAP1, TAP2, IFNGR1, IFNGR2*).

3.5.3 Gene expression and infiltrating immune cell deconvolution

To characterize the tumor's transcriptional state as well as the tumor immune microenvironment in this intrinsically resistant dMMR CRC, we performed bulk RNA-seq (**Supplementary Table S1**) and compared the results to transcriptional profiles of 594 TCGA CRCs that are publicly available on cBioPortal (www.cbioportal.org/, version: coadread_tcga_pan_can_atlas_2018). Subtyping efforts in CRC have showed the existence of four distinct groups, based on gene expression data[132]. As expected, RNA-seq-based CMS classification of this patient's tumor showed a

CMS1 gene expression pattern that is typical of MSI-H tumors[132]. We also employed RNA-seq immune cell subset deconvolution and found that the patient's tumor had a high inferred immune infiltrate abundance compared to other CRCs from TCGA (within top 2%), as expected given MSI-H status, but also had a significantly higher infiltration with activated natural killer (NK) cells and M2 macrophages (**Figure 3.3A**). These results held true when we restricted the comparison to MSI-H (n = 73, **Supplementary Figure 3.S2A**) or advanced (stage III and IV) tumors (n = 240, **Supplementary Figure 3.S2B**). Among the 12 CRCs that were MSI-H with an advanced stage, the tumor of the reported patient had the highest infiltration of both activated NK cells and M2 macrophages. We did not further stratify tumors by *B2M* status, as biallelic inactivation of *B2M* was not validated by immunohistochemistry in TCGA.

3.5.4 Tumor–immune microenvironment at a single-cell resolution

To orthogonally validate the above findings, we performed multiplex immunofluorescence to quantify T cells [$CD3^+CD56(NCAM1)^-$] and NK cells [$CD3^-CD56(NCAM1)^+$] in the tumor microenvironment (**Figure 3.3B**) and compared the results from this intrinsically resistant tumor to an dMMR/MSI-H tumor from another patient with metastatic CRC that responded to pembrolizumab. This was a 70-year-old woman who had been previously treated with surgery and FOLFOX/bevacizumab chemotherapy prior to receiving a PD-1 inhibitor (nivolumab) and showing response by imaging and tumor marker. We found that there was a higher number of intraepithelial NK cells within the center of the PD-1–resistant cancer (17.4 cells/mm^2 (± 5.6) in the resistant tumor and 1.4 cells/mm^2 (± 1.4) in the responding tumor, *P* value =

0.032, two-tailed student *t* test). There were no significant differences in T-cell infiltration. To further assess the presence of NK cells in this intrinsically resistant tumor, and in order to specifically interrogate the transcriptional activation state of these immune cells, we performed single-cell analysis of the pre-checkpoint inhibition specimen using scRNA-seq. Our results confirm the presence of activated NK cells in this tumor and high expression of activated NK cell markers including *NKG7*, *GZMB*, *GZMA*, and *GNLY* (**Figure 3.3C**).

3.6 Discussion

This patient with dMMR/MSI-H CRC was treated with pembrolizumab but exhibited primary resistance to immune checkpoint blockade with progression of her disease on restaging scans. Her tumor had a high neoantigen load, a mutational signature consistent with dMMR, *BRAF* and *RNF43* mutations and other molecular features typical of dMMR tumors. On WES prior to initiation of pembrolizumab, we identified somatically acquired biallelic loss of *B2M*, a critical component of the antigen presentation machinery and MHC class I expression. The second hit in this locus was due to loss of heterozygosity, despite an expected overall low copy number alteration burden. This was consistent with work describing enrichment of inactivating antigen presentation machinery mutations in MSI-H primary CRCs[137]. We confirmed complete loss of B2M protein expression through immunohistochemistry. Loss of *B2M* has been previously implicated in acquired resistance in melanoma, lung cancer, and MSI-H CRC[124,138,139] and intrinsic resistance in melanoma[140], but it has not been previously described as source

of intrinsic resistance in dMMR tumors. This case suggests that MHC class I and B2M expression may need to be considered prior to initiation of PD-1 inhibition to identify patients with dMMR/MSI-H cancers who may not respond to therapy.

To further elucidate the tumor–immune microenvironment, we performed bulk and single-cell transcriptomic analysis. We found that this tumor had the highest inferred activated NK cell and M2 macrophage infiltration when compared to hundreds of CRC specimens from the TCGA with available bulk transcriptional data. We validated these findings through multiplex immunofluorescence against T and NK cell markers and demonstrated enrichment of NK cells in the intrinsically-resistant MSI-H tumor compared to one that responded to PD-1 inhibition. We also identified transcriptionally activated NK cells in this immune checkpoint–resistant tumor through single-cell RNA-seq analysis of the tumor and tumor-associated immune cells from the pre-immunotherapy biopsy specimen.

NK cells can recognize and eliminate cells lacking MHC class I expression[141,142], but are also continuously tuned by classical and non-classical MHC class I molecules, in addition to MHC I–independent mechanisms that instruct NK cells to acquire appropriate missing self-recognition capacity, a process termed NK cell “education”[143,144]. This could at least partially explain the lack of NK cell–mediated tumor control/elimination in a completely B2M-deficient tumor microenvironment, despite an otherwise activated NK cell phenotype. In addition, M2 macrophages have been shown to exert an immunosuppressive role, in particular by impairing NK cells degranulation during cancer progression[145].

These findings suggest that NK cell–based immunotherapies, such as the transfer of “educated” NK cells to patients, could offer an attractive option for

MSI-H tumors that are resistant to immune checkpoint inhibition due to lack of antigen presentation. Our findings also further support the development of immunotherapeutic strategies that aim to shift the balance between M2 and M1 macrophages[146,147]. More broadly, our results have implications for primary resistance to immune checkpoint blockade and novel immunotherapeutic approaches through modulation of the innate immune response in cancer patients.

3.7 Acknowledgements

M.G. and this research was supported by a Conquer Cancer Foundation of ASCO Career Development Award, the Project P-Fund, the Cancer Research UK C10674/A27140 Grand Challenge Award, and a Stand Up to Cancer Colorectal Cancer Dream Team Translational Research Grant (Grant Number: SU2C-AACR-DT22-17). Stand Up to Cancer is a division of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, a scientific partner of SU2C. This work was also supported by National Institute of Health (NIH) grants RO1 CA205406 (K.N.) and P50 CA127003 (M.G.).

3.8 Figures

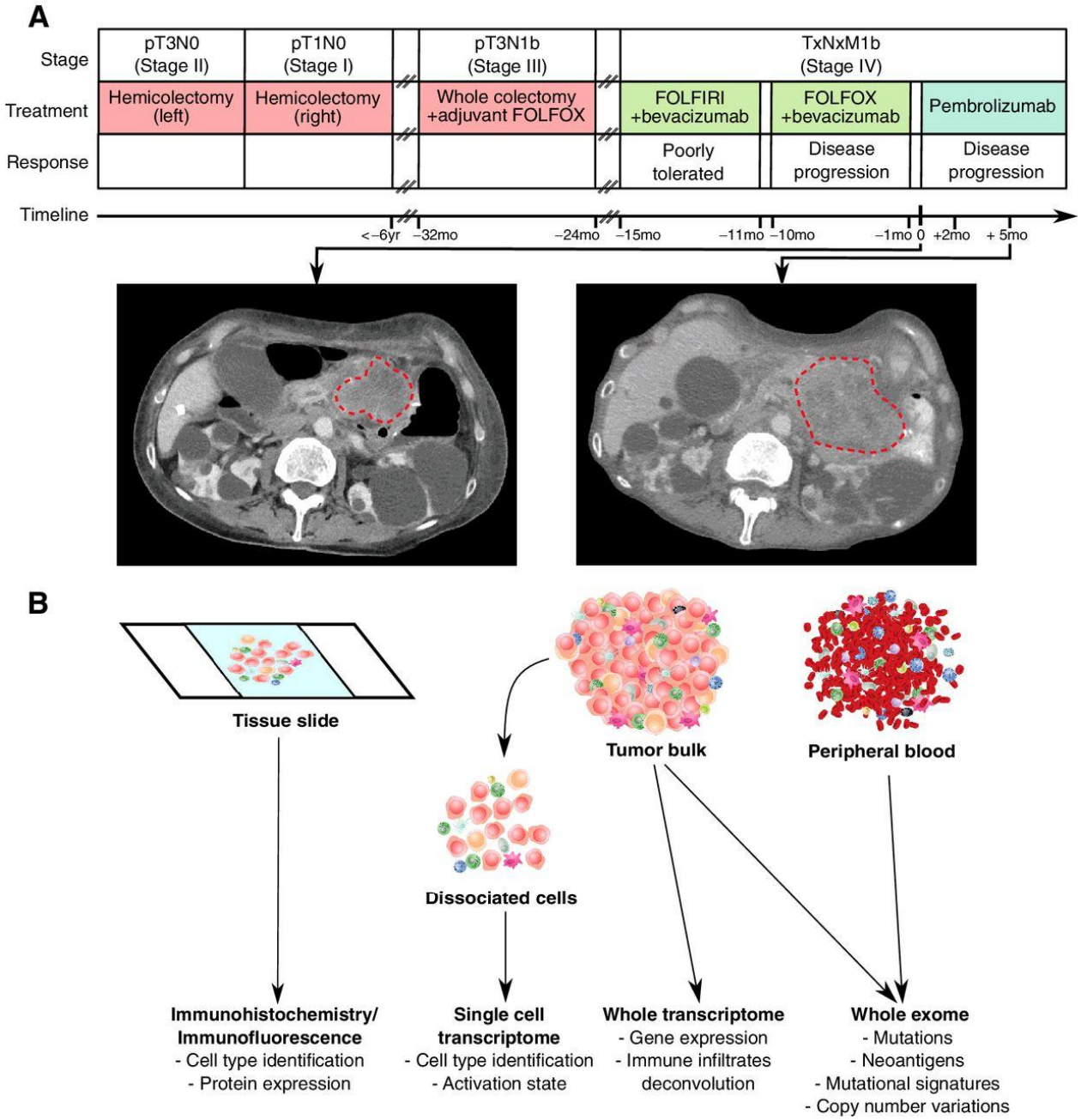


Figure 3.1: Patient disease and treatment course.

A, The stage, treatments, and response of the patient's cancer are shown. Timepoint zero in the event timeline indicates the start of immunotherapy. CT

scans at the indicated timepoints are shown at the bottom. Dotted lines delineate the tumor. B, Flowchart of specimens used, data generated, and analyses performed.

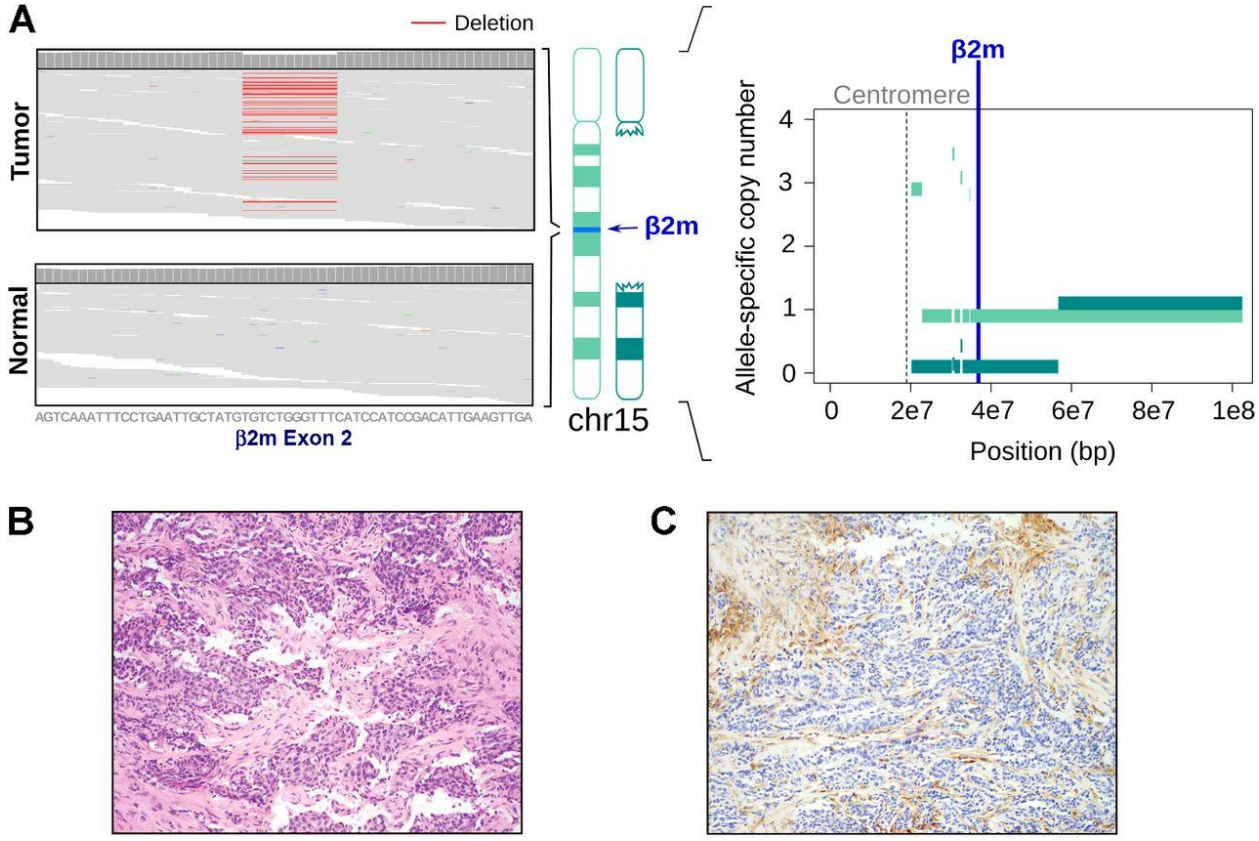


Figure 3.2: Impairment of the antigen presentation machinery through biallelic loss of β2-microglobulin.

A, Mutation analysis results. The left schematic shows the Integrated Genomics Viewer (IGV) panels of both tumor and normal tissue. Deleted regions are shown in red. The figure on the right shows the allele-specific copy-number profile: the y axis represents the copy number of the allele, and the x axis shows the genomic localization on chromosome 15. The left figure shows the frameshift deletion in exon 2 of B2M, whereas the right figure shows the LOH of a segment of 15q surrounding B2M. B, Hematoxylin and eosin staining of

tumor. C, B2M expression. There is complete loss of expression of membranous B2M in tumor cells, with retained expression in surrounding nonneoplastic stromal cells.

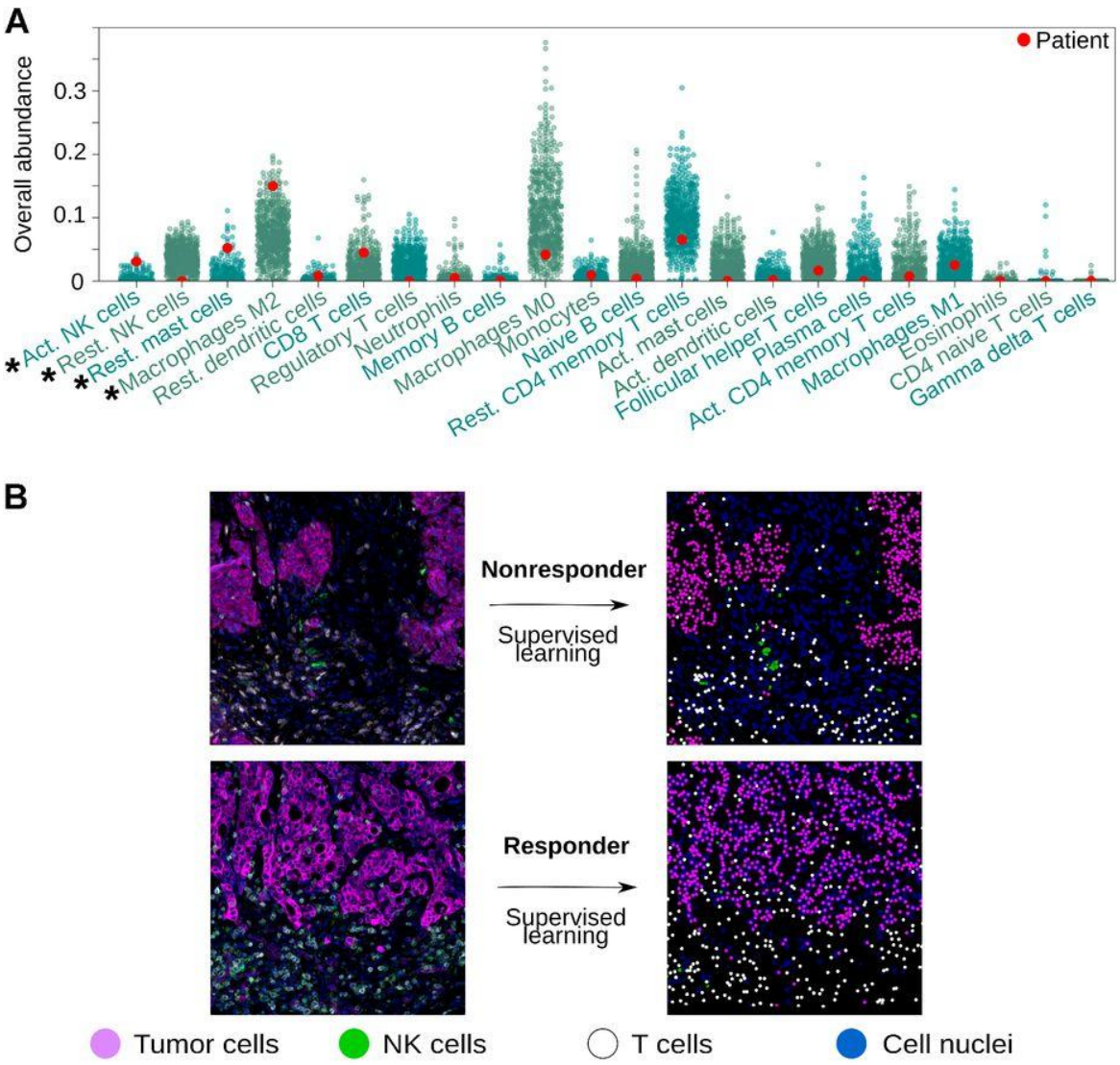

A
Overall abundance
0.3
0.2
0.1
0
Patient
*Act. NK cells
*Rest. NK cells
*Rest. mast cells
*Macrophages M2
Rest. dendritic cells
CD8 T cells
Regulatory T cells
Neutrophils
Memory B cells
Macrophages M0
Monocytes
Naive B cells
Rest. CD4 memory T cells
Act. mast cells
Act. dendritic cells
Follicular helper T cells
Plasma cells
Act. CD4 memory T cells
Macrophages M1
Eosinophils
CD4 naive T cells
Gamma delta T cells
B
Nonresponder
Supervised learning
Responder
Supervised learning
Tumor cells
NK cells
T cells
Cell nuclei
C
Cluster 4
Cluster 6
Cluster 5
Cluster 3
Cluster 2
Cluster 1
tsne 2
tsne 1
Clusters
6 5 4 3 2 1
10 0 −10
Log2 fold change
LTB
CXCL13
TNFRSF4
TRBC2
CD3E
TRBC1
CD3D
CD2
TRAC
CD7
TNFRSF18
CTSW
HOPX
GNLY
GZMB
GZMA
CCL5
NKG7
IL32
Activated T cells
Activated NK cells

Figure 3.3: NK cell infiltration in the tumor–immune microenvironment.

A, Immune infiltrates deconvolution from RNA-seq data. The immune infiltrates abundances, as a total cell fraction, of the immune checkpoint–resistant tumor (red point) are compared with 594 other colorectal cancer tumors from TCGA. The immune infiltrates (x axis) were sorted by P values from left to right. The immune infiltrates for which the patient was in the top or bottom 5% are denoted with an asterisk. B, Multiplexed immunofluorescence imaging analysis of the tumor–immune microenvironment. Left panels, representative immunofluorescent expression of CD3 (white), CD56 (green), cytokeratin (purple), and DAPI (blue, marking nuclei). Right panels, results from image analysis, driven by machine learning, that identifies $CD3^{+}CD56^{-}$ T cells (white dots) and $CD3^{-}CD56^{+}$ NK cells (green dots) within tumor regions. C, Single-cell transcriptional analysis of the tumor. t Distributed Stochastic Neighbor Embedding (t-SNE) visualization (left) of the scRNA-seq data from 595 cells. The heatmap (right) shows significantly differentially expressed genes of interest between the nonepithelial cell clusters.

3.9 Supplementary Figures

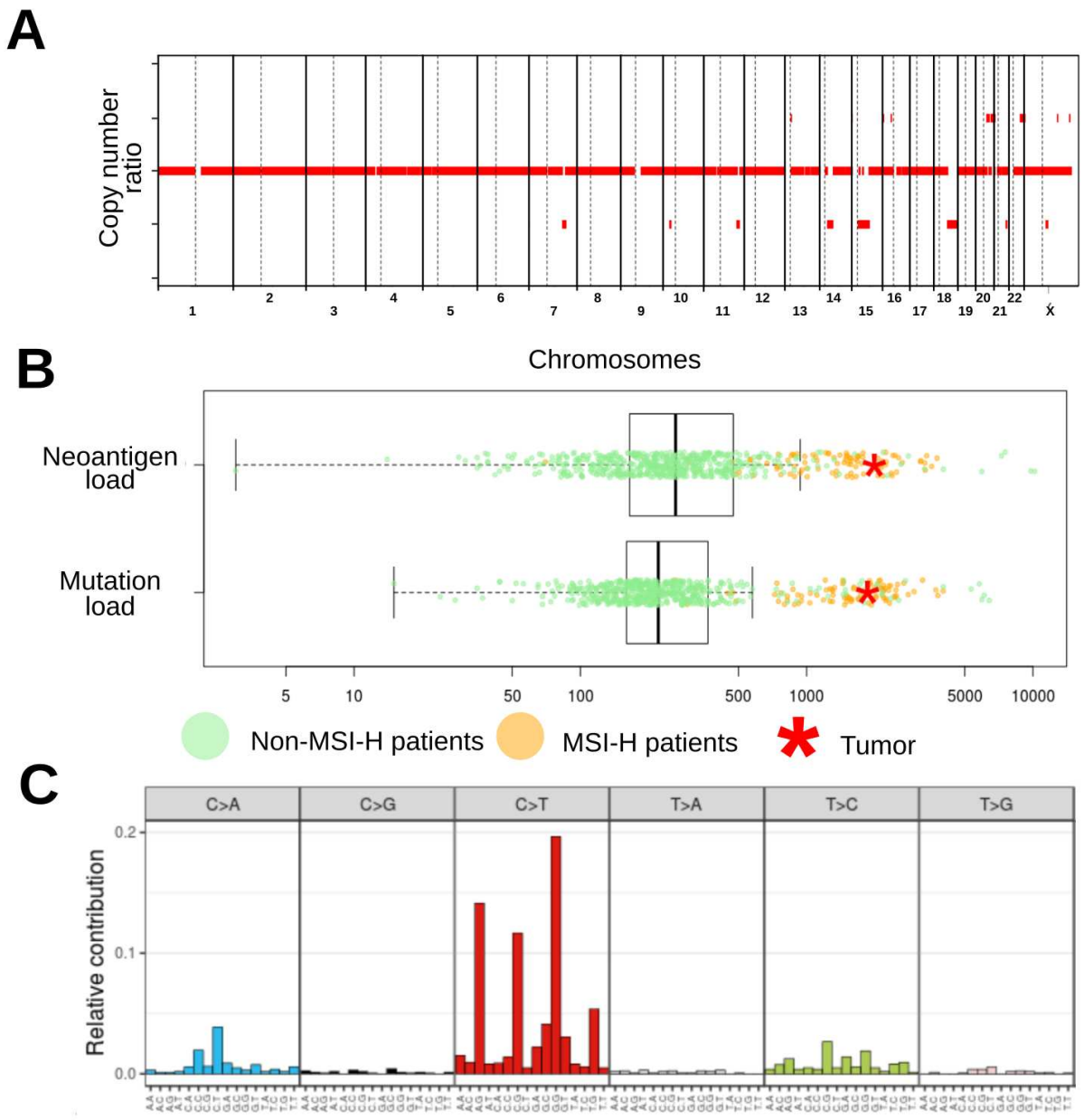


Figure 3.S1: Genomic characterization of the PD1-resistant MSI-H tumor

(A) Copy number ratio profile. The x axis indicates the chromosomal position and the y axis indicates the copy number ratio. (B) Neoantigen and mutation load for the tumor (red asterisk) compared to 619 CRCs. The x-axis shows the number of mutations and neoantigens on a logarithmic scale. The superimposed 1D scatter plots show the neoantigen load and mutation load distributions of MSI-H (orange dots) and non-MSI-H (green dots) CRC patients. (C) Mutational signature analysis. The y-axis indicates the percentage of mutations attributed to each substitution type. The latter were defined by the sequence context immediately 3' and 5' to the mutated base. Mutational signature analysis showed a contribution of 55% of Signature 6, which is associated with defective DNA mismatch repair and found in MSI-H cancers. Signature 1A, which is found in most cancer samples and correlates with age of cancer diagnosis, showed a contribution of 43%

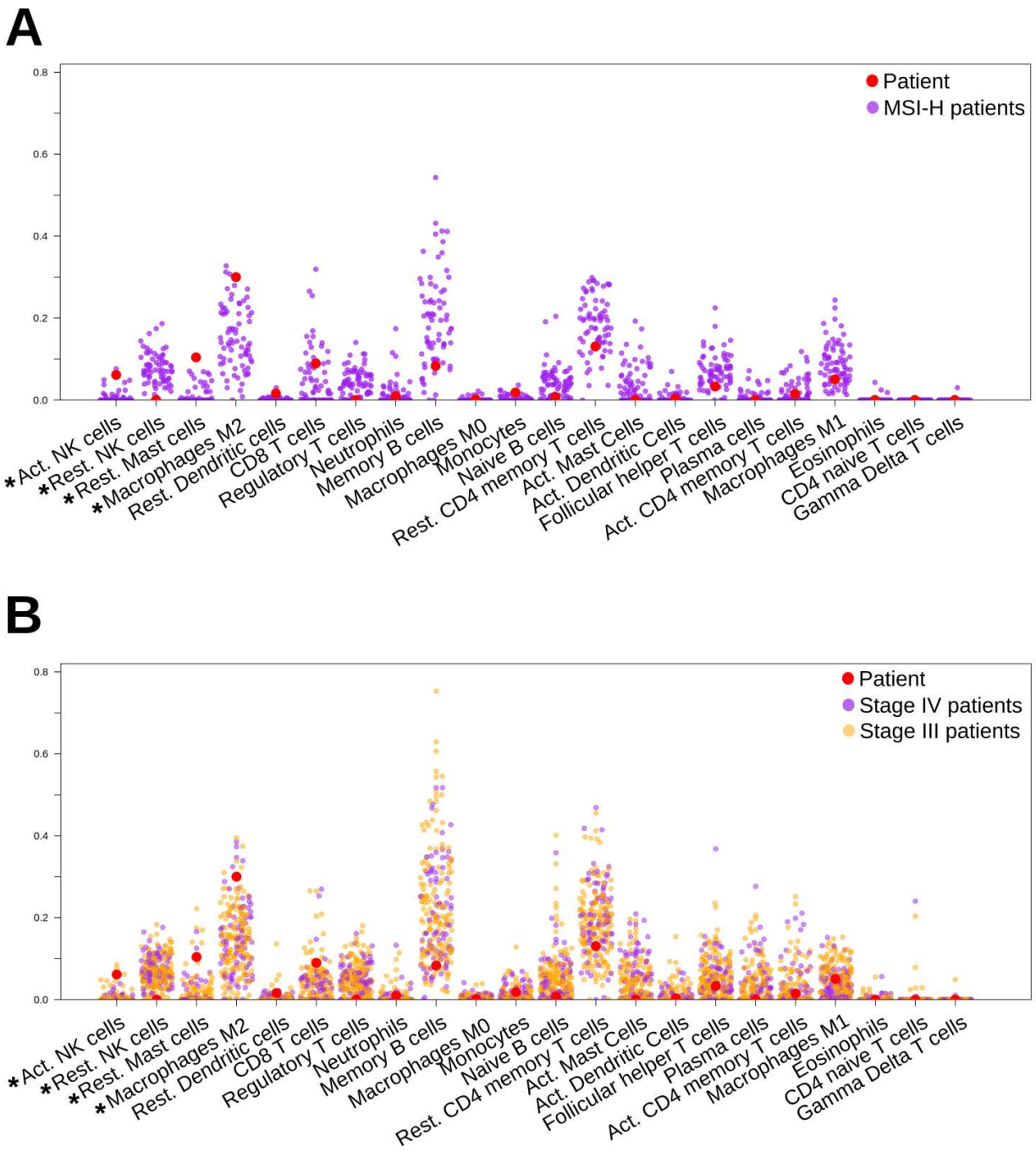


Figure 3.S2: MSI status and tumor staging stratifications for Immune infiltrates deconvolution

(A) Stratification by MSI status. Comparison of the immune infiltrates abundances, as a total cell fraction, among 73 MSI-H TCGA patients. MSI-H status was evaluated for 557 TCGA tumors as previously described (19). The immune infiltrates for which the patient was in the top or bottom 5% are denoted with an asterisk. (B) Stratification by tumor stage Comparison of the immune infiltrates abundances, as a total cell fraction, among 240 advanced stage tumors (stage III and stage IV). The immune infiltrates for which the patient was in the top or bottom 5% are denoted with an asterisk.

4. Discovery and features of an alkylating signature in colorectal cancer

Published in Cancer Discovery on June 17, 2021.

4.1 Author information

Carino Gurjao [#1,2] , Rong Zhong [#3,4] , Koichiro Haruki [#3], Yvonne Y Li [1,2], Liam F Spurr [1,2,5], Henry Lee-Six [6], Brendan Reardon [1,2] , Tomotaka Ugai [3,7], Xuehong Zhang [8,9], Andrew D Cherniack [1,2], Mingyang Song [7,8,9,10,11], Eliezer M Van Allen [1,2], Jeffrey A Meyerhardt [1,2], Jonathan A Nowak [12] , Edward L Giovannucci [7,8,9], Charles S Fuchs [13], Kana Wu [#9], Shuji Ogino [#2,3,7,12], Marios Giannakis [#14,2]

1. Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts.
2. Broad Institute of MIT and Harvard, Cambridge, Massachusetts.
3. Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts.
4. Department of Epidemiology and Biostatistics and Ministry of Education Key Lab of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China.
5. Pritzker School of Medicine, Biological Sciences Division, University of Chicago, Chicago, Illinois.

6. Wellcome Sanger Institute, Hinxton, United Kingdom.
7. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts.
8. Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts.
9. Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, Massachusetts.
10. Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts.
11. Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts.
12. Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts.
13. Yale Cancer Center, Yale School of Medicine, Smilow Cancer Hospital, New Haven, Connecticut.
14. Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts. Marios_Giannakis@dfci.harvard.edu.

[#] Contributed equally.

4.2 Abstract

Several risk factors have been established for colorectal cancer, yet their direct mutagenic effects in patients' tumors remain to be elucidated. Here, we leveraged whole-exome sequencing data from 900 colorectal cancer cases that had occurred in three U.S.-wide prospective studies with extensive dietary and lifestyle information. We found an alkylating signature that was previously undescribed in colorectal cancer and then showed the existence of a similar mutational process in normal colonic crypts. This alkylating signature is associated with high intakes of processed and unprocessed red meat prior to diagnosis. In addition, this signature was more abundant in the distal colorectum, predicted to target cancer driver mutations *KRAS* p.G12D, *KRAS* p.G13D, and *PIK3CA* p.E545K, and associated with poor survival. Together, these results link for the first time a colorectal mutational signature to a component of diet and further implicate the role of red meat in colorectal cancer initiation and progression.

SIGNIFICANCE: Colorectal cancer has several lifestyle risk factors, but the underlying mutations for most have not been observed directly in tumors. Analysis of 900 colorectal cancers with whole-exome sequencing and epidemiologic annotations revealed an alkylating mutational signature that was associated with red meat consumption and distal tumor location, as well as predicted to target *KRAS* p.G12D/p.G13D.

4.3 Introduction

Most tumor mutations are passengers that have little to no functional role in cancer. However, their positional context in the genome may reveal information about the underlying mutational processes[43]. Snapshots of these processes, called mutational signatures, were originally deconvoluted using a nonnegative matrix factorization (NMF) approach[148] on a large collection of whole-genome sequencing and whole-exome sequencing (WES) data [46]. Mutational signatures may elucidate the roles of mutagens in cancer and inform prevention and treatment efforts. Several studies have been conducted to associate mutational signatures with cellular processes or exposures. These include rare cancer predisposition syndromes[48], environmental agents[49], and microbiota[50]. Such association studies have relied on either DNA-sequencing data sets or preclinical models, such as organoids. However, although many lifestyle-related factors have been linked to colorectal cancer[149], larger and more comprehensive data sets are needed to enable the discovery of the associated signatures. Consequently, past efforts have not been able to capture the cumulative effect of putative mutagens, such as dietary components, over decades. In particular, red meat consumption has been consistently linked to the incidence of colorectal cancer[150–152]. The suggested mechanism is mutagenesis through alkylating damage induced by N-nitroso-compounds (NOC), which are metabolic products of blood heme iron or meat nitrites/ nitrates[153]. Nevertheless, this mutational damage is yet to be observed directly in patients' tumors.

4.4 Results

4.4.1 Active Mutational Signatures in Colorectal Tumors and normal Colonic Crypts

To address this gap, we leveraged a database of incident colorectal cancer cases that had occurred in three U.S.-wide prospective cohort studies, namely the Nurses' Health Studies (NHS) I and II and the Health Professionals Follow-up Study (HPFS)[154]. Study participants (more than 230,000 women and 50,000 men) repeatedly provided data on diet, lifestyle, and other factors without knowing their future colorectal cancer diagnosis, if any. We performed WES on matched primary untreated tumor–normal pairs in 900 patients with colorectal cancer with adequate tissue materials (**Figure 4.1A; Supplementary Table 4.S1**). NMF signal separation revealed the existence of seven mutational processes (see **Methods** and **Figure 4.1B** and **4.1C**; **Supplementary Figure 4.S1**). We confirmed the robustness of the deconvolution by using another signature assignment program (SigProfiler[46]); we again found seven mutational processes (**Supplementary Figure 4.S2**, left) that are highly similar to the ones obtained using the standard NMF approach (**Supplementary Figure 4.S2**, right). To uncover the etiology of these colorectal signatures (that we name c-signatures), we first used a cosine similarity metric (cossim) to compare the deconvoluted signatures to reference COSMIC Single Base Substitution (SBS) signatures[46]. The seven de novo signatures displayed the highest similarity with four known mutational processes (**Supplementary Figure 4.S3**), namely POLE deficiency (c-POLEa/SBS10a, cossim = 0.95 and c-POLEb/SBS10b, cossim = 0.86), aging (c-Age/SBS1, cossim = 0.95), deficient mismatch repair (dMMR; c-dMMRa/ SBS15, cossim = 0.90 and c-dMMRb/SBS26, cossim = 0.90), and exposure to alkylating agents

(c-Alkylation/SBS11, cossim = 0.94). c-SBS40 matched the closest to SBS40 (cossim =0.84), which is a featureless signature with unknown etiology and found in most cancers[46].

We substantiated the etiology of the four mutational processes by integrating clinical, pathology, and methylation data (**Figure 4.2A**). Tumors harboring a POLE exonuclease domain mutation were significantly enriched in signatures c-POLEa and c-POLEb (P = 2.3 × 10–5 and P = 1.8 × 10–6, respectively, Mann–Whitney U test). Similarly, patients with orthogonally assessed microsatellite instability (MSI)–high status were significantly enriched in signatures c-dMMRa and c-dMMRb (P < 2 × 10–16 for both, Mann–Whitney U test). Signature c-Age also displayed a significant association with patients' age at diagnosis (P = 1.7 × 10–5, Mann–Whitney U test). Last, we support the etiology of the alkylating-like signature, not previously described in colorectal cancer, by assessing the MGMT (O-6-methylguanine-DNA methyltransferase) promoter methylation status in tumors from the NHS/HPFS cohorts. *MGMT* is a central gene in the repair of alkylating lesions. Among the sequenced specimens with available MGMT promoter methylation data, we observed that tumors with methylated MGMT promoters were enriched in the signature c-Alkylation (P = 6.6 × 10–3, Mann–Whitney U test; **Figure 4.2A**), further supporting that this signature represents the biological consequence of increased alkylating damage. Of note, SBS18, which is associated with MUTYH-associated polyposis[46], is absent in the tumor samples we sequenced. We believe this is the case because of the low occurrence of MUTYH deficiency generally in colorectal cancer (less than 1%[155]), as well as further undersampling of patients with germline predisposition mutations as only healthy individuals were enrolled prospectively in NHS/HPFS. NMF signal separation in The Cancer Genome Atlas (TCGA) colorectal tumors (n = 540)

revealed the existence of seven signatures (**Supplementary Figure 4.S4 and 4.S5** for SigProfiler results) similar to the ones found in NHS/HPFS (**Supplementary Figure 4.S6**), thus suggesting the existence of the same underlying mutational processes in all colorectal cancer cohorts. Analysis of the TCGA colorectal tumors (**Figure 4.2B**) substantiated the same etiologies for the POLE signatures c-POLEa and c-POLEb ($P = 6.2 \times 10–7$ and $P = 7.3 \times 10–7$, respectively, Mann–Whitney U test), as well as dMMR signatures c-dMMRa and c-dMMRb ($P < 2 \times 10–16$ for both, Mann–Whitney U test). We also observed that TCGA tumors with MGMT promoter methylation were enriched in signature c-Alkylation ($P = 9.7 \times 10–5$, Mann–Whitney U test). Of note, in TCGA, signature c-Alkylation displayed the highest similarity with SBS30 (cossim = 0.81), followed by SBS11. Conversely, SBS30 was the second most similar signature to the c-Alkylation one in the NHS/HPFS cohorts (**Figure 4.2C; Supplementary Figure 4.S3**). SBS30 resembles SBS11 (cossim of 0.76, **Figure 4.2C**) and is attributed to base excision repair (BER) deficiency[46], which is also a central pathway in repairing damage from alkylated bases. We nevertheless found no association between germline polymorphisms in NHTL1 and other genes of the BER pathway and the alkylating signature in the TCGA specimens (see **Methods** and **Supplementary Figure 4.S7**). The presence of SBS30 ahead of SBS11 in the TCGA colorectal cancer data set could instead be attributed to a smaller sample size of colorectal cancers in TCGA compared with NHS/HPFS (see “Undersampling Simulations” in Methods and **Supplementary Figure 4.S8**). The Fanconi anemia (FA) and translesion synthesis (TLS) DNA damage repair pathways also do not show an association with the alkylating signature (see **Methods** and **Supplementary Figure 4.S9A and 4.S9B**).

We also estimated the effect size for the Mann–Whitney U tests by calculating the rank-biserial correlation r_{rb} for each mutational signature and the respective molecular or clinical phenotype shown in **Figure 4.2B**. We observed that the effect sizes were similar for the alkylating signatures and the aging signature (r_{rb} = 0.14 and r_{rb} = 0.16, respectively) and smaller than the hypermutator dMMR and POLE signatures (r_{rb} > 0.8 for dMMR and POLE signatures in both TCGA and NHS/HPFS).

Interestingly, a previously published survey of mutational signatures in normal colorectal crypts [156] from the European Genome–phenome Archive (EGA) showed the existence of a signature (named SBSC) that we found to be similar to the alkylating one that we observed in NHS/HPFS colorectal cancers (cossim = 0.85). Of note, SBSC matched closely to SBS23, which, similar to SBS30, also resembles SBS11 (cossim of 0.77; **Figure 4.2C**). The hierarchical clustering of SBSC with the seven signatures deconvoluted from NHS/HPFS and TCGA confirmed the similarity of EGA SBSC with the alkylating imprints (**Figure 4.2C**).

4.4.2 Dietary Patterns of Alkylation Damage

To test whether dietary components contributed to the alkylating signature in colorectal cancer, we leveraged prospectively collected repeated measurements of meat, poultry, and fish consumption in grams per day in the NHS and HPFS cohorts. All available red meat variables showed significant positive associations between prediagnosis intakes and alkylating damage in colorectal cancers (**Figure 4.3A**; overall red meat, P = 0.017/r_{rb} = 0.14; unprocessed red meat, P = 7.8 × 10–3/r_{rb} = 0.16; and processed red meat, P =

7.3 × 10–3/r_{rb} = 0.16, Mann–Whitney U test). Other dietary variables (fish and chicken intake, **Figure 4.3B**) and lifestyle factors (body mass index, alcohol consumption, smoking, and physical activity in **Supplementary Figure 4.S10**) did not show any significant association with the alkylating signature. In addition, no other colorectal cancer mutational process showed a significant association with red meat intake (**Supplementary Figure 4.S11**). Of note, MGMT promoter methylation did not differ by red meat consumption (two-sided Mann–Whitney U test, P = 0.51; **Supplementary Figure 4.S12**). When adjusted for red meat intake, there was no difference in alkylating damage between male and female patients with colorectal cancer (two-sided Mann–Whitney U test, P = 0.27 for patients with high overall red meat consumption).

Previous studies[150,151] showed a positive association between processed red meat and colorectal cancer incidence in the distal colon. Thus, we also investigated how the alkylating damage might differ by tumor location. We found that, compared with the proximal colon, the distal colorectal specimens exhibited higher alkylating damage in tumors (P = 1.4 × 10–4 in NHS/HPFS and P = 1.9 × 10–8 in TCGA, Mann–Whitney U test) and normal crypts (P = 0.022, Mann–Whitney U test; **Figure 4.3B**).

4.4.3 Carcinogenicity of Alkylation Damage

Mutational processes increase the likelihood of specific driver mutations in certain trinucleotide contexts. To find such driver mutations that associate with the alkylating signature, we devised a simple model (**Figure 4.4A**; see **Methods**) that predicts the relative likelihood of mutational processes to target

colorectal cancer recurrent drivers in non–MSI-high, non–POLE-mutated tumors. In particular, the alkylating signature appeared to be the dominant one that targets KRAS p.G13D (relative likelihood = 1) and KRAS p.G12D (relative likelihood = 0.91; **Figure 4.4A**). This is due to p.G12D and p.G13D being in trinucleotide contexts (ACC>ATC and GCC>GTC, respectively) mainly targeted by the alkylating signature. PIK3CA p.E545K (TCA>TTA) is also predicted to be predominantly targeted by the alkylating signature (relative likelihood = 0.87). Supporting this, we showed that colorectal cancers having KRAS p.G12D, KRAS p.G13D, or PIK3CA p.E545K-mutant colorectal cancers were enriched with the alkylating signature compared with all other tumors (**Figure 4.4B**, P = 0.013, Mann–Whitney U test). Last, we examined patient survival across ordinal alkylating mutational signature quartiles and found that patients whose tumors have high alkylation damage (top quartile) had a worse colorectal cancer–specific survival (log-rank test Ptrend = 0.036; **Figure 4.4C**; **Supplementary Tables 4.S2 and 4.S3**). Furthermore, higher alkylating signature contribution was associated with worse colorectal cancer–specific survival in both univariable and multivariable Cox proportional hazards regression analyses (Ptrend = 0.015 and Ptrend = 0.036, respectively, **Figure 4.4D** and **Supplementary Table 4.S3**).

4.5 Discussion

Our work demonstrated the presence of a novel alkylating mutational signature, which we deconvoluted directly from WES of colorectal tumors. Interestingly, this signature is highly similar to SBS11, which was originally discovered in patients with prior exposure to temozolomide[43]. Temozolomide is an alkylating agent used as a treatment of brain gliomas with MGMT promoter methylation[43] and induces the same lesions as dietary NOCs and in the same

proportions (80% of N7-methylguanine and N3-methylguanine, as well as 10% of O6-methylguanine [157,158]). Previous attempts have shown the existence of alkylating lesions in normal colorectal mucosa, notably caused by NOCs[159]. The latter can be formed endogenously after nitrosylation of heme iron from blood[159,160] but have also been associated with red meat intake in a small cohort of participants[161]. However, these previous studies were based on limited data sets (small sample sizes and/or use of laboratory methylating agents) and lack comprehensive sequencing that would enable the discovery of the full mutational spectrum induced by red meat. Crucially, past efforts have focused on normal colorectal tissues and not examined colorectal cancer.

Our analysis reveals the existence of an alkylating signature in colorectal cancer, which is associated with high prediagnosis intake of processed and unprocessed red meat. Earlier work also hypothesized that the distal colon has increased DNA damage from exposure to dietary carcinogens, as a result of feces storage and water resorption in this portion of the large intestine[162]. This is believed to explain the association observed between distal cancer incidence and red meat consumption[151,152,162]. Consistently, we found an enrichment in tumors and normal crypts in the distal colon and rectum. In support of the International Agency for Research on Cancer (IARC) Monograph Working Group, which classified processed meat as carcinogenic[150], our results provide molecular evidence of this dietary factor's mutagenic impact. In addition, our analyses further implicate unprocessed meat intake and suggest MGMT as a factor of susceptibility to red meat–induced damage. The existence of a similar alkylating signature in normal colorectal crypts also suggests that mutational changes due to such damage may start to occur early in the path of colorectal carcinogenesis. Our analysis predicted KRAS p.G12D, KRAS p.G13D, and PIK3CA p.E545K to be mainly targeted by the alkylating signature in non

hypermutated colorectal cancers. We showed that there was indeed higher alkylating damage in tumors harboring these driver mutations. Independent epidemiologic analyses have also shown a positive association between high consumption of red meat products and KRAS p.G12D and KRAS p.G13D[163,164]. Although the number of mutations due to alkylation damage was lower than other mutational processes, we showed that alkylation might have considerable carcinogenic potential by targeting driver mutations in KRAS and PIK3CA. We also demonstrated a significantly worse survival for patients with high levels of the alkylation signature contribution. Our study has leveraged a comprehensive data set with repeated dietary measures over years, without patients knowing their upcoming colorectal cancer diagnosis, and WES on a large collection of colorectal tumors. It provides unique evidence supporting the direct impact of dietary behaviors on colorectal carcinogenesis. Moreover, the presence of a similar alkylating signature in normal mucosa advocates for the utility of early dietary interventions and suggests potential precision prevention approaches in MGMT-methylated premalignant tissue. Similarly, the association of the signature with cancer driver mutations, such as KRAS and PIK3CA ones, may offer future potential therapeutic opportunities. More generally, our study exemplifies the potential role of large-scale molecular epidemiologic studies in elucidating cancer pathogenesis[165] and guiding prevention efforts through lifestyle modifications, such as dietary interventions.

4.6 Methods

4.6.1 Study Population, Specimens, and Sequencing

We used data from three prospective cohort studies in the United States: the Nurses' Health Study I (NHS1, including 121,701 women ages 30 to 55 years at enrollment who had been followed since 1976), the Nurses' Health Study II (NHS2, including 116,429 women ages 25 to 42 years who had been followed since 1989), and the HPFS (including 51,529 men ages 40 to 75 years followed since 1986[154]). The study participants had been sent questionnaires biennially to update information on lifestyle factors and newly diagnosed diseases, including colorectal cancer. The follow-up rate had been more than 90% for each follow-up questionnaire cycle in the three cohort studies. The patients were followed until death or end of follow-up (January 1, 2016, for HPFS; June 1, 2016, for NHS1; and June 1, 2015, for NHS2), whichever came first. Study physicians, who were blinded to exposure data, reviewed medical records of 4,855 incident colorectal cancer cases to confirm the disease diagnosis and to collect data on tumor size, tumor anatomic location, and disease stage. Archival formalin-fixed, paraffin-embedded (FFPE) tissue blocks of tumor and normal colon were collected in a subset of colorectal cancer. We previously showed that in our cohorts, demographic features of cases did not differ appreciably by tissue availability[166]. The study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health (Boston, MA) and those of participating registries as required. Written informed consent was obtained from all patients with colorectal cancer. We prioritized relatively more recent

colorectal cancer cases for sequencing to mitigate the potential impact of FFPE artifacts. Given the number of NHS versus HPFS participants (2:1 female/male ratio), we also sequenced relatively more specimens from male patients to obtain more balanced sequencing data. Supplementary Table S4 shows the clinical and pathologic characteristics of the 4,855 patients with colorectal cancer. WES was carried as previously described[25]. Briefly, using guide hematoxylin and eosin–stained slides, tumor areas were selected to extract tumor-enriched DNA from tissue sections of tumor FFPE blocks. Normal DNA was extracted from resection margins or other areas free from tumors. DNA specimens underwent hybrid capture with SureSelect v.2 Exome bait (Agilent Technologies), followed by sequencing on Illumina HiSeq 2000 instruments. The obtained average coverage was 85× in tumors and matched adjacent normal colon tissue (see Supplementary Table S5).

4.6.2 Dietary Variables

Ascertainment of diet was carried out as previously described[151]. To assess dietary intake in each cohort, food frequency questionnaires (FFQ) were initially collected in 1980 for NHS and in 1986 for HPFS. For the NHS, a 61-item semi quantitative FFQ was used at baseline[167], which was expanded to approximately 130 food and beverage items in 1984, 1986, and every 4 years thereafter. For the HPFS cohorts, baseline dietary intake was assessed using a 131-item FFQ that was also used for updates generally every 4 years subsequently[168]. In particular, unprocessed red meat consumption was evaluated based on forms on the intake of “beef or lamb as main dish,” “pork as main dish,” “hamburger,” and “beef, pork, or lamb as a sandwich or mixed

dish." Processed meat diets included "bacon"; "beef or pork hot dogs"; "salami, bologna, or other processed meat sandwiches"; and "other processed red meats such as sausage, kielbasa, etc." Consumption of red meat, chicken, poultry, and fish was evaluated in grams per day. For the remainder of our analysis, we considered the top decile of each variable to determine the "high-intake" patients and considered the rest as "low-intake" patients, because only the top-decile patients show a substantial difference in overall red meat intake (**Supplementary Figure 4.S13A** and **4.S13B**). Data were based on the most recent prediagnosis reported intake for each patient.

4.6.3 MGMT Promoter Methylation, MSI, and POLE Deficiency Status

MGMT promoter methylation analysis in the NHS/HPFS cohorts was carried out using bisulfite conversion and real-time PCR as previously described[169]. MSI status was evaluated using 10 microsatellite markers (D2S123, D5S346, D17S250, BAT25, BAT26, BAT40, D18S55, D18S56, D18S67, and D18S487) as formerly detailed[154]. POLE deficiency was assessed by sequencing and manual Integrated Genome Viewer curation of POLE exonuclease domain mutations in hypermutated non–MSI-high tumors (>400 mutations).

4.6.4 Somatic Variant Calling

We have used the Cancer Genome Analysis (CGA) WES characterization pipeline (https://github.com/broadinstitute/CGA_Production_Analysis_Pipeline) developed at the Broad Institute of MIT and Harvard to call, filter, and annotate somatic mutations. All analyses were carried out on the human genome build hg19. The pipeline employs the following tools: MuTect[19], ContEst[170], Strelka[171], DeTiN[172], AllelicCapSeg [173], MAFPoNFilter[174], RealignmentFilter, GATK[175], and PicardTools. FFPE-specific artifacts are filtered similarly to previous publications[25,127]. Briefly, FFPE artifacts arise from formaldehyde deamination of cytosines resulting in C-to-T transition mutations, which presents itself as an "Orientation bias" (excess of C>T sites in F1R2 read pairs and an excess of G>A in F2R1 read pairs). In the pipeline we used, the "Orientation Bias Filter" tool[20] filters out FFPE-specific artifacts. To further filter spurious single-nucleotide variant calls, we used Burrows–Wheeler Aligner BWA-MEM (http://bio-bwa.sourceforge.net/) to realign sequenced reads associated with the mutations to a set of sequences derived from the human reference assembly. The Panel of Normal was created using normal samples with less than 1% of crosssample contamination (as evaluated by Contest[170]) and less than 1% of tumor in normal (as outputted by DeTIN[172]). We illustrate the variant calling pipeline in **Supplementary Figure 4.S14**.

4.6.5 TCGA Data Analysis

Clinical, methylation, and somatic mutation data from TCGA were downloaded from the Data Coordination Center (DCC) data portal at https://dcc.icgc.org/releases/current/Projects/COAD-US and

https://dcc.icgc.org/releases/current/Projects/READ-US (as of March 2020). For consistency, only WES data sets were used. Altogether, we pooled 540 TCGA patients with somatic mutation data, among whom 523 patients also had methylation data. We evaluated MGMT promoter methylation status using the MGMT-STP27 prediction model[176]. In short, two probes (cg12434587 and cg12981137) were used to predict MGMT promoter methylation. An M value cutoff of 0.358, which empirically maximized the sum of sensitivity and specificity, was then used to discriminate MGMT promoter methylation status (**Supplementary Figure 4.S15**).

4.6.6 Nonnegative Matrix Factorization

Mutations were deconvoluted into separate signatures based on the number of mutations in each of 96 possible trinucleotide contexts. Deconvolution was carried out with a standard NMF method based on Kullback–Leibler divergence using the "NMF" R package[177]. This method is particularly adapted for mutational signature analysis as recent studies demonstrated[178]. A critical parameter in NMF is the estimation of the rank (i.e., the number of expected mutational signatures). To determine this, we performed quality measures on a range of ranks (n = 2 to 10) for the 900 colorectal cancer exomes in the NHS/HPFS cohorts. This showed a sharp increase in the cophenetic (i.e., the stability of the NMF classes) and dispersion (i.e., the reproducibility of the class assignments) metrics after rank = 7. For this rank, we also observed that the residual sum of squares (RSS) reached a lower plateau (**Supplementary Figure 4.S1**). A similar rank survey on an independent cohort of 540 colorectal cancer exomes from the TCGA

(**Supplementary Figure 4.S4**) revealed the same dispersion and cophenetic peaks at rank = 7 and a lower plateau RSS. For the rest of the analysis, we consequently used rank = 7. We confirmed the robustness of these seven signatures by running NMF with different variant allele frequency (VAF) cutoffs (**Supplementary Figure 4.S16**). This demonstrates that the signature discovery is not affected by low VAF mutations, which are more likely to represent sequencing artifacts, such as those due to FFPE preservation. SigProfiler was run on NHS/HPFS and TCGA colorectal cancer exomes as previously described[46].

4.6.7 Undersampling Simulations

To show that the difference in sample size between TCGA (n = 540) and NHS/HPFS (n = 900) can explain the presence of SBS30 instead of SBS11 in the former cohort, we (i) randomly sampled 540 patients of the 900 from NHS/HPFS; (ii) extracted seven signatures from the 540 patients and found their closest fit among SBS1 (aging signature), SBS10a and SBS10b (POLE signatures), SBS15 and SBS26 (dMMR signatures), and SBS11 and SBS30; and (iii) repeated steps (i) and (ii) a hundred times.

4.6.8 Crypt Mutational Signature Analysis

Mutational signatures from normal colonic crypts[156] were used in our analysis. These signatures were extracted from WGS data from 571 crypts

from 42 individuals from the EGA[156]. Deconvolution was performed using a hierarchical Dirichlet process, which produces results similar to NMF[156].

4.6.9 Analysis of Recurrent Hotspot Mutations

To compute the relative likelihood of mutational processes to target a specific hotspot, we (i) localized the trinucleotide context of the hotspot, (ii) extracted the signatures contribution for the specific trinucleotide context, and (iii) normalized the contribution of each signature, such that the sum became 1. Recurrent hotspots were defined as specific point mutations occurring in at least 25 patients.

4.6.10 TCGA Germline Polymorphisms Analysis

TCGA genotyping data (Affymetrix SNP 6.0 array platform) were used to select germline variants from genes in the BER, FA, and TLS pathways extracted from the GSEA database (https://www.gsea-msigdb.org/gsea/msigdb/[179,180]). We imputed autosomal variants for TCGA samples using IMPUTE2[181], with haplotypes of 1000 Genomes Phase 3[182] as the reference panel. We used the following criteria to select SNPs with the plink software[183]: (i) average imputation confidence score, also called INFO score, ≥0.4; (ii) minor allele frequency ≥5%; (iii) SNP missing rate <5% for best-guessed genotypes at posterior probability ≥0.9; and (iv) Hardy–Weinberg equilibrium P value > 1 × 10−6. After imputation, 2,041

variants were included in our subsequent analysis. We tested for an additive effect (genotype 0,1,2 as a continuous variable) for each SNP and found no association with the alkylating signature [**Supplementary Figure 4.S7** and **Supplementary Figure 4.S9**, FDR-adjusted P value (q value) less than 0.1 for all SNPs tested].

4.6.11 Statistical Analysis

We used R version 3.6.2 to perform statistical analyses. Significance for two-group comparisons was evaluated by a one-sided Mann–Whitney U test unless otherwise indicated. $P < 0.05$ was considered statistically significant. For the comparisons of the alkylating signature by age in the NHS/HPFS cohorts and TCGA colorectal cancer database, the patients' median age (70 and 67 years, respectively) was used as the cutoff. Eight hundred eighty-two patients with available colorectal cancer survival data were subsequently used for survival analyses. Univariable and multivariable-adjusted Cox proportional hazards regression analysis as used to calculate the HR of colorectal cancer–specific survival and overall survival according to ordinal alkylating mutational signature quartiles (Q1–Q4). The multivariable Cox regression model initially included sex (female vs. male), age at diagnosis (<60, 60–64, 65–69, and ≥70 years), year of diagnosis (1995 or before, 1996–2000, 2001–2005, and 2006–2014), family history of colorectal cancer (present vs. absent), current smoking status (never smoking, past smoking, 1–14 pack-years, 15–24 pack-years, ≥25 pack-years), alcohol consumption (women: 0–<0.15, 0.15–<2.0, 2.0–<7.5, and ≥7.5 g/day; men: 0 to <1, 1–<6, 6–<15, and ≥15 g/day), tumor location (proximal colon vs. distal colon vs. rec- tum), CpG

island methylator phenotype (high vs. low/negative[184]), KRAS mutation (mutant vs. wild-type[185]), BRAF mutation (mutant vs. wild-type[185]), tumor differentiation (well to moderate vs. poor), disease stage (I/II vs. III/IV), microsatellite instability status (MSI-high vs. non-MSI-high[184]), and long-interspersed nucleotide element 1 (LINE-1) methylation level (continuous[186]). A backward elimination with a threshold P of 0.05 was used to select variables for the final models. Cases with missing data were assigned to the majority category of a given categorical covariate to limit the degrees of freedom, except for cases with missing LINE-1 methylation, for which we assigned a separate indicator variable. We confirmed that excluding the cases with missing information in any of the covariates did not substantially alter results.

4.6.12 Data Availability

WES data have been deposited in dbGAP (accession number phs000722). WES quality metrics and a subset of clinical annotations are included in this article. Additional clinical and epidemiology data from the NHS1, NHS2, and HPFS can be requested through the NHS/HPFS consortia.

4.6.13 Code Availability Statement

All analysis scripts are available upon request.

4.6.14 Authors' Disclosures

Y.Y. Li reports other support from g.Root Biomedical Services outside the submitted work. A.D. Cherniack reports other support from Bayer outside the submitted work. E.M. Van Allen reports grants from Novartis and BMS; personal fees from Tango Therapeutics, Genome Medical, Invitae, Monte Rosa Therapeutics, Manifold Bio, Illumina, Enara Bio, and personal fees from Janssen outside the submitted work; in addition, E.M. Van Allen has a patent for institutional patents filed on chromatin mutations and immunotherapy response, and methods for clinical interpretation pending. J.A. Meyerhardt reports personal fees from COTA Healthcare and personal fees from Taiho Pharmaceutical outside the submitted work. C.S. Fuchs reports personal fees from Amylin Pharma, AstraZeneca, Bain Capital, CytomX Therapeutics, Daiichi-Sankyo, Eli Lilly, Entrinsic Health, Evolveimmune Therapeutics, Genentech, Merck, Taiho, and personal fees from Unum Therapeutics outside the submitted work; in addition, C.S. Fuchs serves as a director for CytomX Therapeutics and owns unexercised stock options for CytomX and Entrinsic Health; is a cofounder of Evolveimmune Therapeutics and has equity in this private company; has provided expert testimony for Amylin Pharmaceuticals and Eli Lilly. C.S. Fuchs is now an employee of Genentech and Roche. S. Ogino reports grants from NIH during the conduct of the study. M. Giannakis reports grants from CRUK, SU2C, and grants from NIH/NCI during the conduct of the study; grants from Bristol-Myers Squibb, Merck, Servier, Janssen, and grants from ASCO Conquer Cancer Foundation outside the submitted work. No disclosures were reported by the other authors.

4.6.15 Authors' Contributions

C. Gurjao: Data curation, formal analysis, investigation, visualization, methodology, writing–original draft, writing–review and editing. R. Zhong: Data curation, formal analysis, writing–review and editing. K. Haruki: Data curation, formal analysis, writing–review and editing. Y.Y. Li: Writing–review and editing. L.F. Spurr: Writing–review and editing. H. Lee-Six: Resources, writing–review and editing. B. Reardon: Writing–review and editing. T. Ugai: Writing–review and editing. X. Zhang: Writing–review and editing. A.D. Cherniack: Writing–review and editing. M. Song: Writing–review and editing. E.M. Van Allen: Writing–review and editing. J.A. Meyerhardt: Resources, writing–review and editing. J.A. Nowak: Resources, writing–review and editing. E.L. Giovannucci: Resources, data curation, writing–review and editing. C.S. Fuchs: Resources, funding acquisition, writing–review and editing. K. Wu: Data curation, funding acquisition, writing–review and editing. S. Ogino: Resources, data curation, funding acquisition, writing–review and editing. M. Giannakis: Conceptualization, resources, supervision, funding acquisition, investigation, methodology, writing–original draft, writing–review and editing.

4.6.16 Acknowledgments

We thank N. Abdennur and S. Abraham for technical feedback, as well as W.L. Chiu, J. Elhai, and L. Fossecave for useful comments. This work was supported by the NIH grants P01 CA87969 (to R.M. Tamini), UM1 CA186107 (to M.J. Stampfer), P01 CA55075 (to W.C. Willett), UM1 CA167552 (to W.C.

Willett), U01 CA167552 (to W.C. Willett and L.A. Mucci), U01 CA176726 (to W.C. Willett), U54 HG003067 (to E.S. Lander and S.B. Gabriel), P50 CA127003 (to B. Wolpin), P30CA016359 (to C.S. Fuchs), R01 CA118553 (to C.S. Fuchs), R01 CA169141 (to C.S. Fuchs), R35 CA197735 (to S. Ogino), R01 CA151993 (to S. Ogino), K07 CA190673 (to R. Nishihara), K07 CA188126 (to X. Zhang), R21 CA238651 (to X. Zhang), R03 CA197879 (to K. Wu), R21 CA222940 (to K. Wu and M. Giannakis), and R21 CA230873 (to K. Wu and S. Ogino); by Cancer Research UK.

Grand Challenge Award (UK C10674/A27140, to M. Giannakis and S. Ogino); by Nodal Award (2016-02) from the Dana-Farber Harvard Cancer Center (to S. Ogino); by the Stand Up To Cancer Colorectal Cancer Dream Team Translational Research Grant (SU2C-AACR-DT22-17, to C.S. Fuchs and M. Giannakis), administered by the American Association for Cancer Research, a scientific partner of SU2C; and by grants from the Project P Fund, The Friends of the Dana-Farber Cancer Institute, Bennett Family Fund, and the Entertainment Industry Foundation through National Colorectal Cancer Research Alliance. Stand Up To Cancer is a division of the Entertainment Industry Foundation. K. Haruki was supported by fellowship grants from the Uehara Memorial Foundation and the Mitsukoshi Health and Welfare Foundation. X. Zhang was supported by American Cancer Society Research Scholar Grant (RSG NEC-130476). X.Zhang was supported by the Dana-Farber Harvard Cancer Center (DF/HCC) GI SPORE Developmental Research Project Award (P50CA127003), DF/HCC Nodal Award (Cancer Center Support Grant, P30CA006516-55), the Karin Grunebaum Cancer Research Foundation, and the Zhu Family PEER Award. J.A. Meyerhardt is supported by the Douglas Gray Woodruff Chair fund, the Guo Shu Shi Fund,

Anonymous Family Fund for Innovations in Colorectal Cancer, Project P fund, and the George Stone Family Foundation. M. Giannakis was supported by a Conquer Cancer Foundation of ASCO Career Development Award. T. Ugai was supported by a grant from Overseas Research Fellowship (201960541) from Japan Society for the Promotion of Science. R. Zhong was supported by a fellowship grant from Huazhong University of Science and Technology, Wuhan, Hubei, China. We thank the participants and staff of the Nurses' Health Studies and the Health Professionals Follow-up Study for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, and WY. The authors assume full responsibility for analyses and interpretation of these data.

4.7 Figures

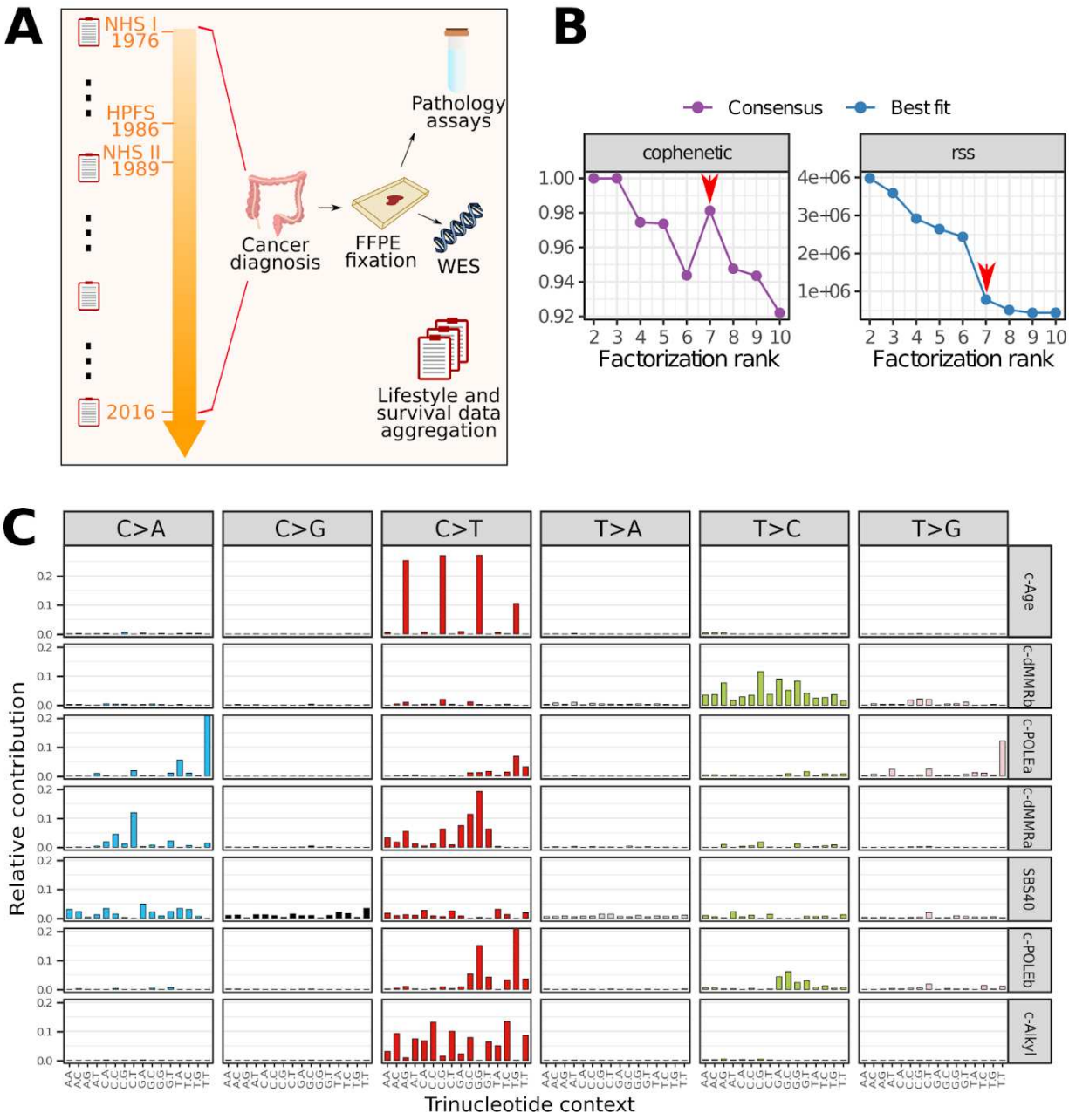


Figure 4.1: De novo signature deconvolution in NHS/HPFS CRCs

(A) Cohort and data overview. NHS: Nurses' Health Studies (NHS I and NHS II). HPFS: Health Professionals Follow-up Study. (B) Quality measures for NMF in NHS/ HPFS, Arrows indicate the estimated rank of mutational signatures. rss: residual sum of squares (C) The consensus seven signatures found by NMF in NHS/HPFS.

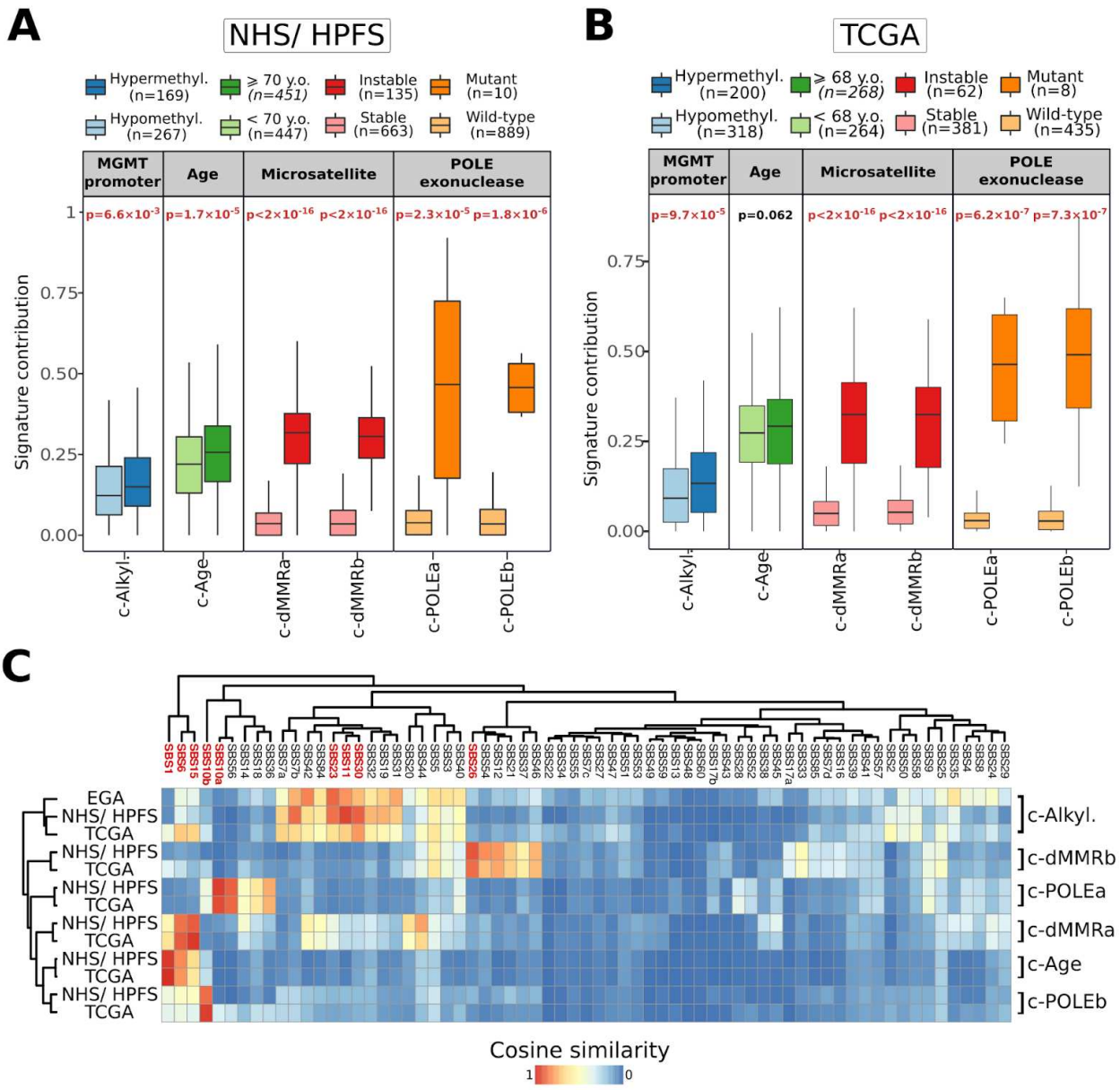


Figure 4.2: Active mutational signatures in colonic cells

Proportion of mutations assigned to de novo extracted signatures in CRCs from NHS/ HPFS (A) and TCGA (B), segregated by MGMT promoter methylation status, POLE exonuclease mutations, microsatellite instability and age at diagnosis. Boxplot outliers not shown. (C) Heatmap of the similarity scores between colorectal tumour (from TCGA and NHS/ HPFS) signatures - clustered

on the y axis- and reference COSMIC signatures, clustered on the x axis. COSMIC signatures found in either NHS/ HPFS or TCGA are bolded. The alkylating normal colon (from EGA) signature is also shown. Clustering has been performed according to cosine similarity. EGA: European Genome-phenome Archive.

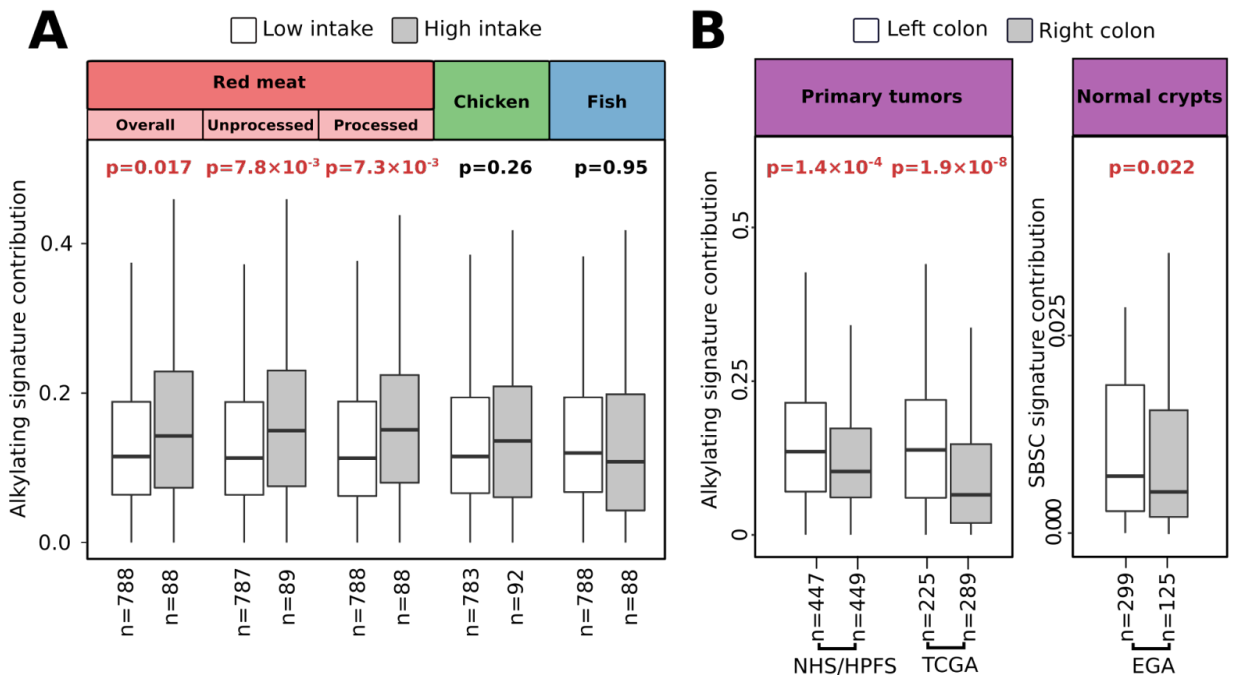


Figure 4.3: Epidemiology and distribution of alkylating damage

(A) Proportion of mutations assigned to alkylating damage in NHS/ HPFS, segregated by intake (top decile in grams per day versus the rest) of overall, processed and unprocessed red meat, as well as chicken and fish. (B) Proportion of mutations assigned to alkylating damage in CRC and normal colon, segregated by tumour location. Boxplot outliers not shown.

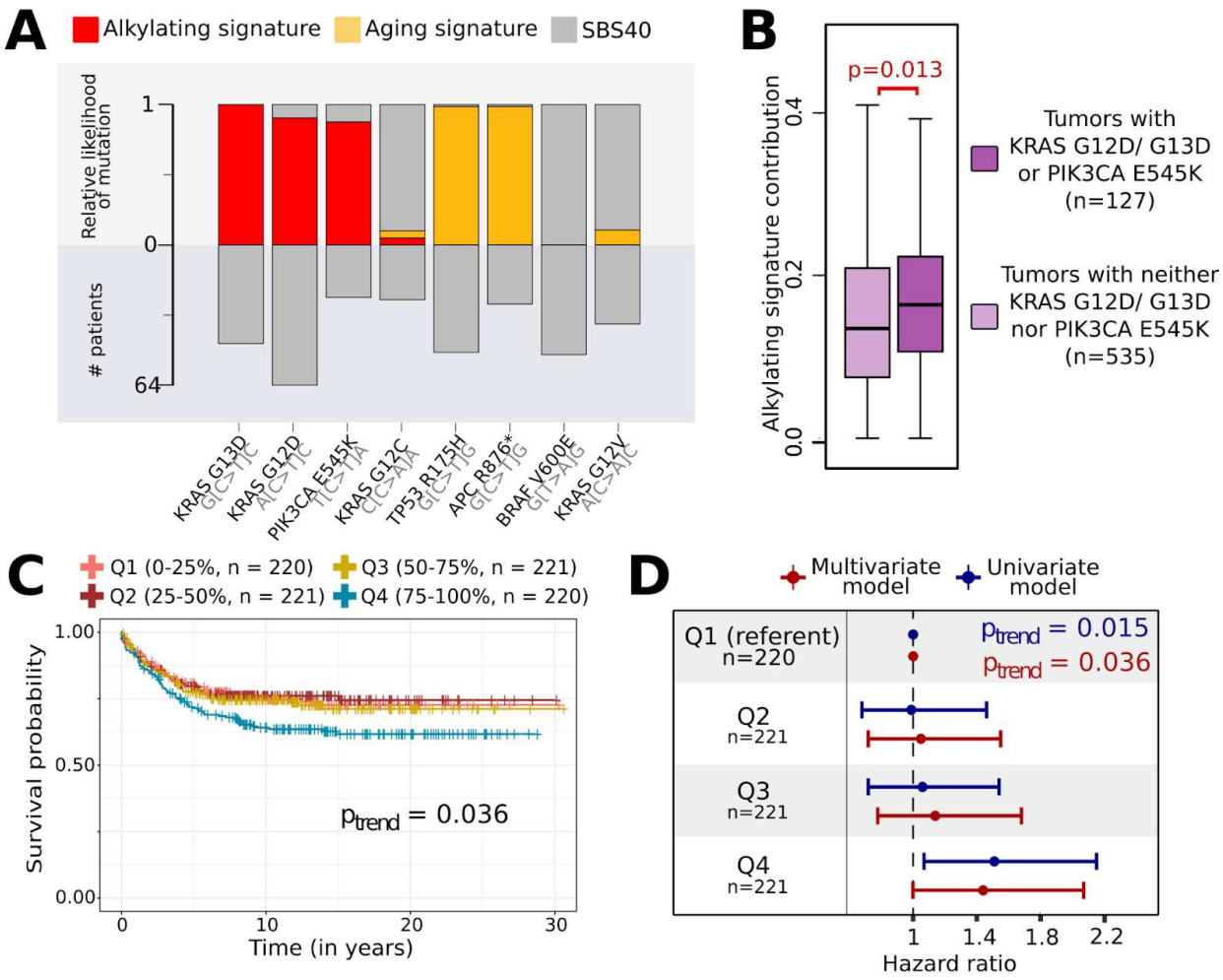
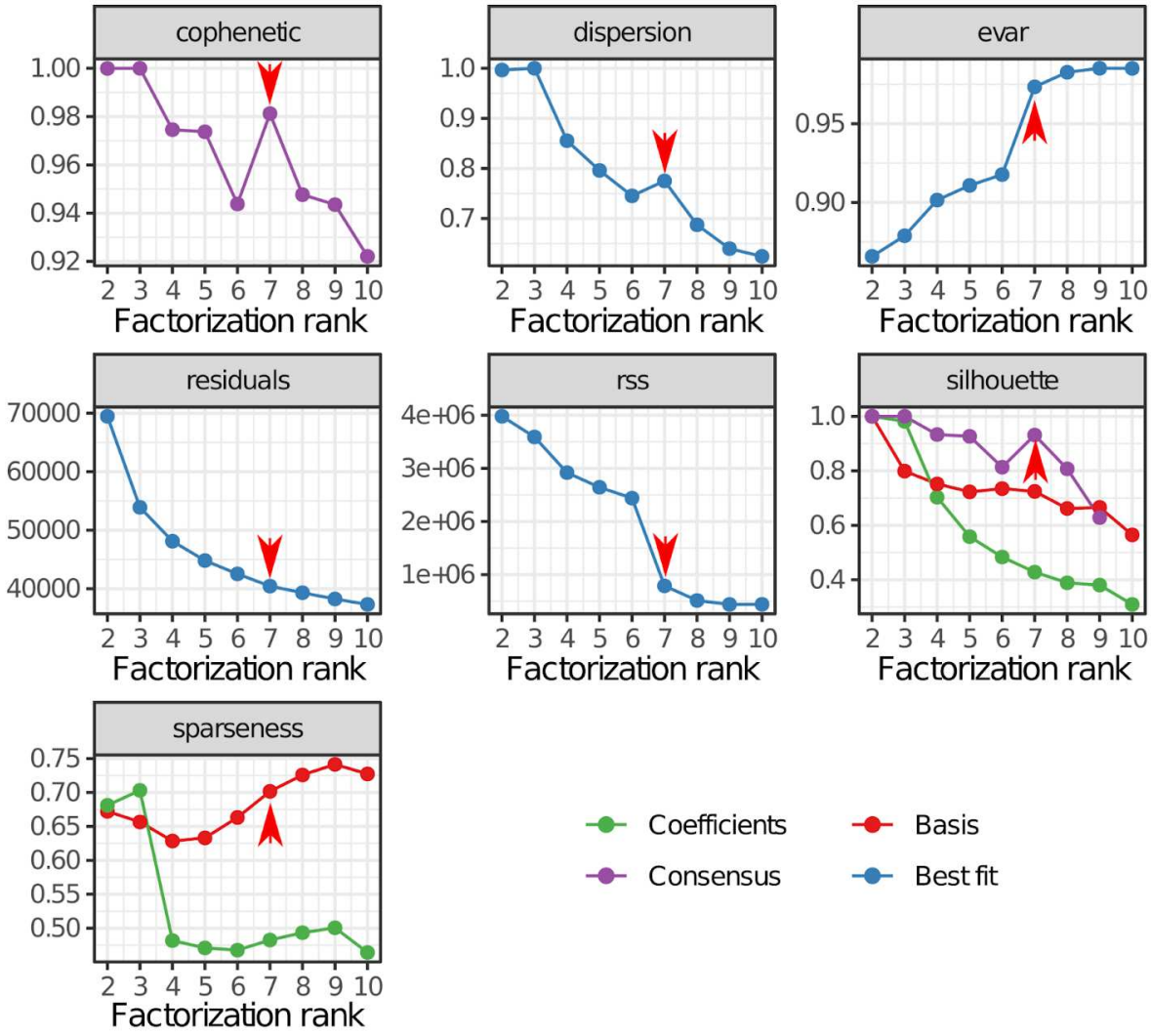


Figure 4.4: Carcinogenic potency of alkylating damage

(A) Relative likelihood of mutational processes to target recurrent hotspots in non-hypermutated CRC. As hotspots we considered all point mutations that were present in at least 25 patients with non-hypermutated (non-MSI-high, non-POLE mutated) CRC. Each stacked bar represents the relative likelihood of a given signature to target a given hotspot. (B) Proportion of mutations assigned to alkylating damage in NHS/ HPFS, TCGA CRCs, segregated by KRAS G12D/ KRAS G13D/ PIK3CA E545K mutation status. Boxplot outliers not shown. (C) Kaplan–Meier plot illustrating colorectal cancer-specific survival

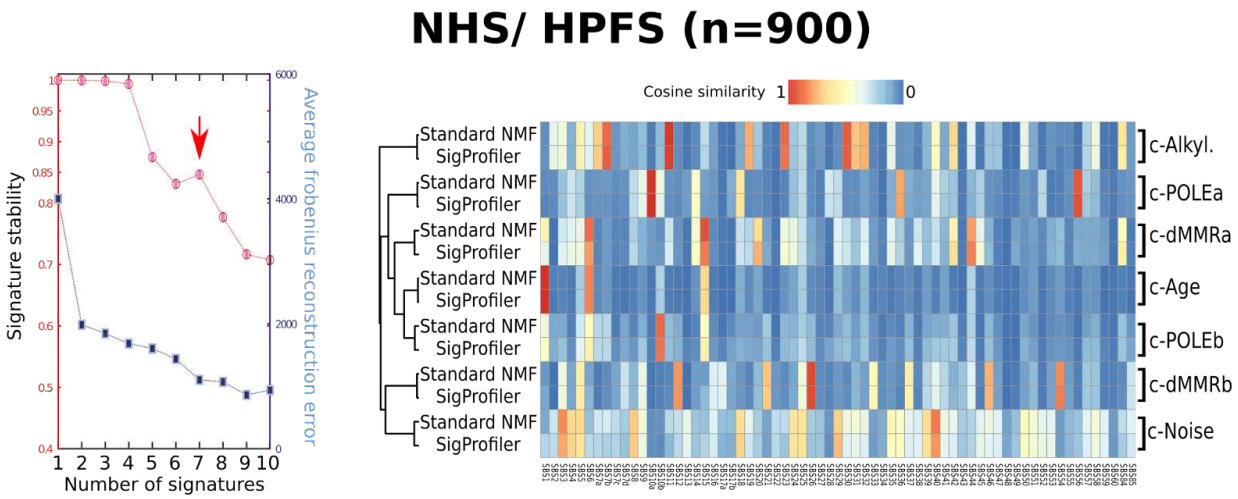
of the patients stratified into quartiles of alkylating signature contribution. (D) Forest plot of the association between the colorectal specific survival and quartiles of alkylating signature contribution in univariable and multivariable Cox regression models.

4.8 Supplementary Figures



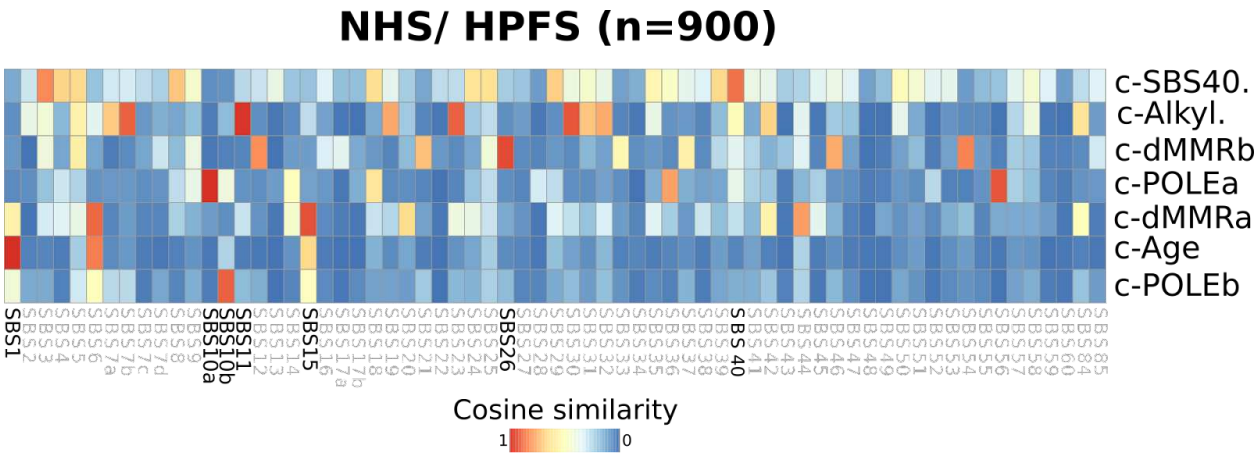
Supplemental Figure 4.S1: Non-negative matrix factorization rank survey in NHS/HPFS

Quality measures for Non-negative matrix factorization in NHS/ HPFS, as described in the R package "NMF". Arrows indicate the estimated rank of mutational signatures.



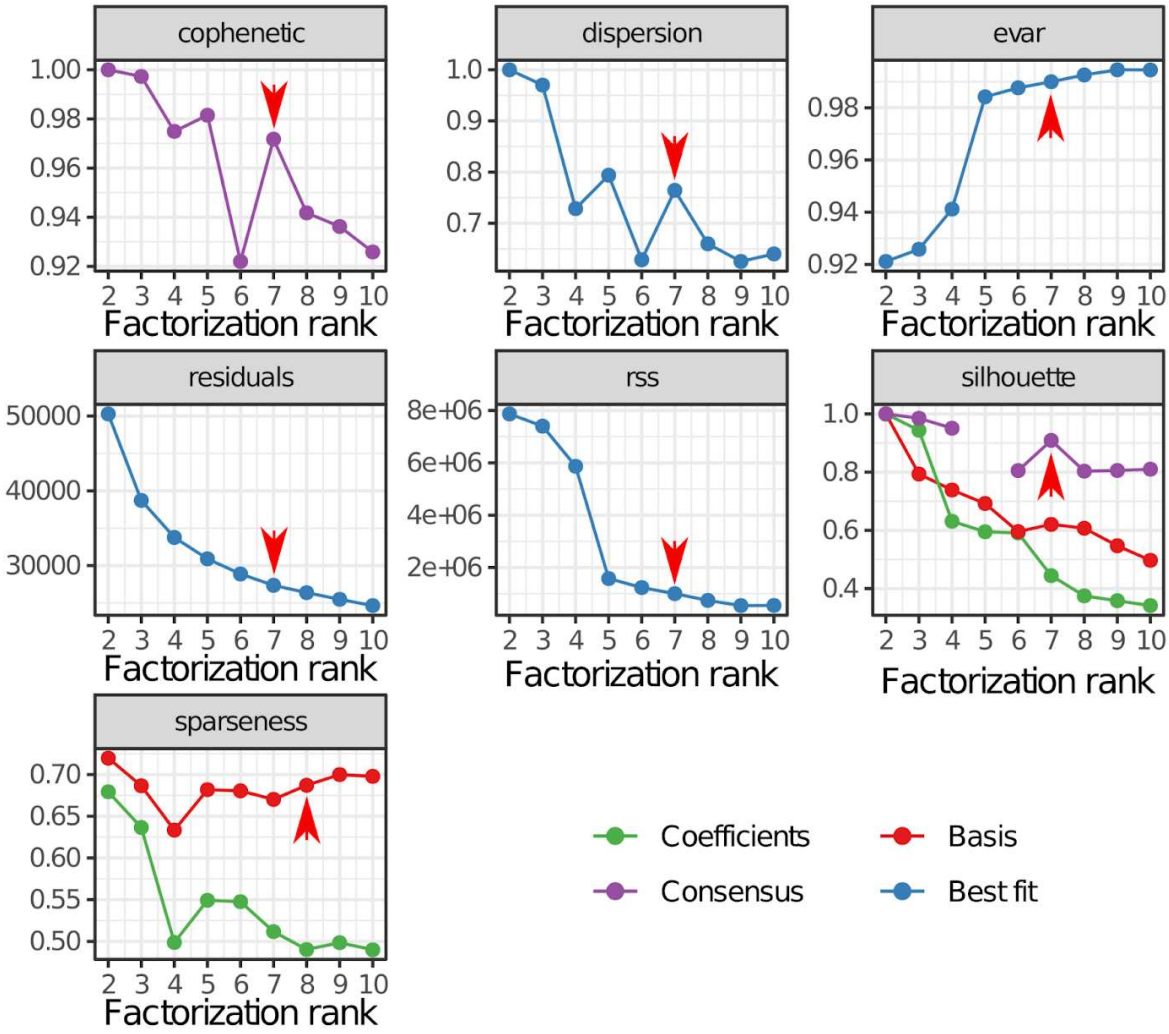
Supplemental Figure 4.S2: Comparison between SigProfiler and standard NMF approach for NHS/ HPFS

Rank survey (left panels) and cosine similarity heatmaps between SigProfiler and standard NMF signatures (right panels) in NHS/ HPFS.



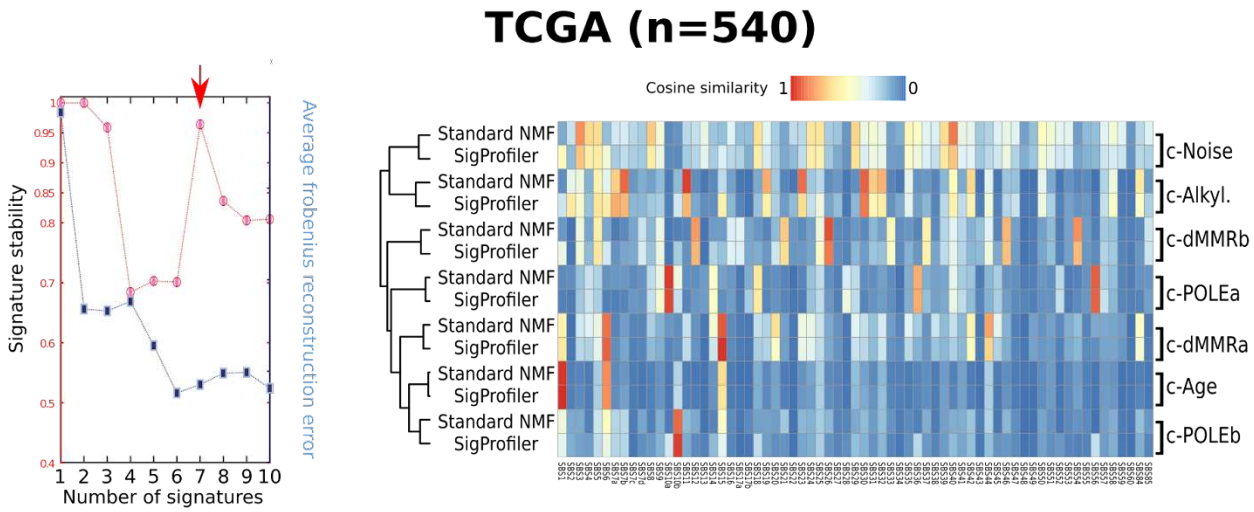
Supplemental Figure 4.S3: Comparison between NHS/ HPFS CRC and COSMIC signatures

Cosine similarity heat map between de novo signatures found in NHS/ HPFS (y axis) and COSMIC signatures (x axis). NHS/ HPFS signatures were named after their closest COSMIC mutational process match (in bold)



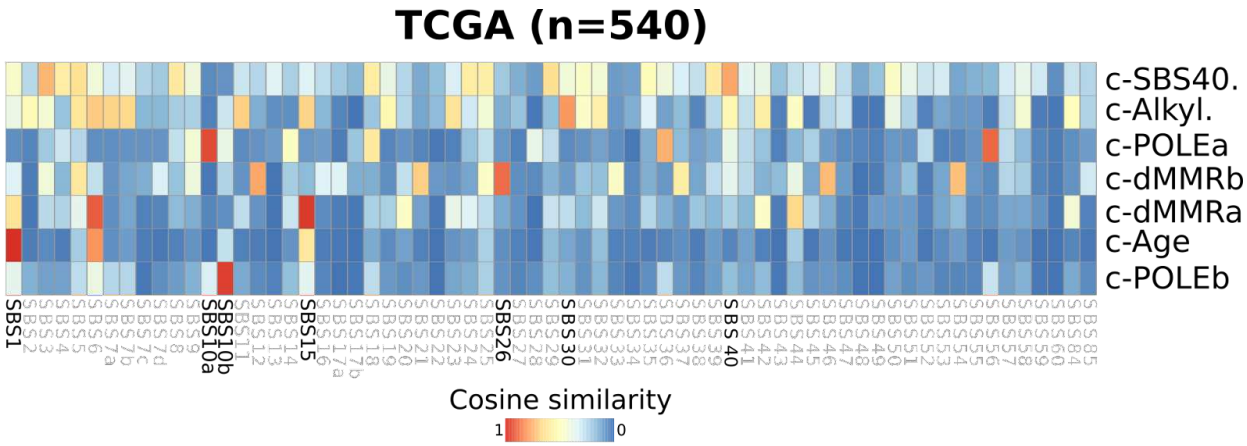
Supplemental Figure 4.S4: Non-negative matrix factorization rank survey in TCGA COAD/READ

Quality measures for Non-negative matrix factorization in TCGA COAD/READ, as described in the R package "NMF". Arrows indicate the estimated rank of mutational signatures.



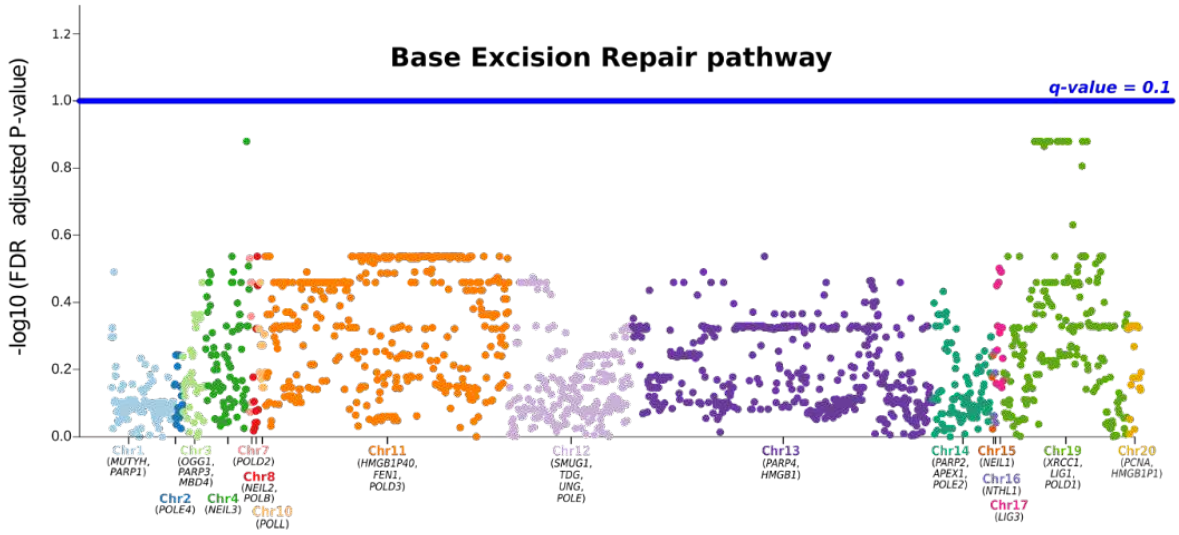
Supplemental Figure 4.S5: Comparison between SigProfiler and standard NMF approach for TCGA

Rank survey (left panels) and cosine similarity heatmaps between SigProfiler and standard NMF signatures (right panels) in TCGA.



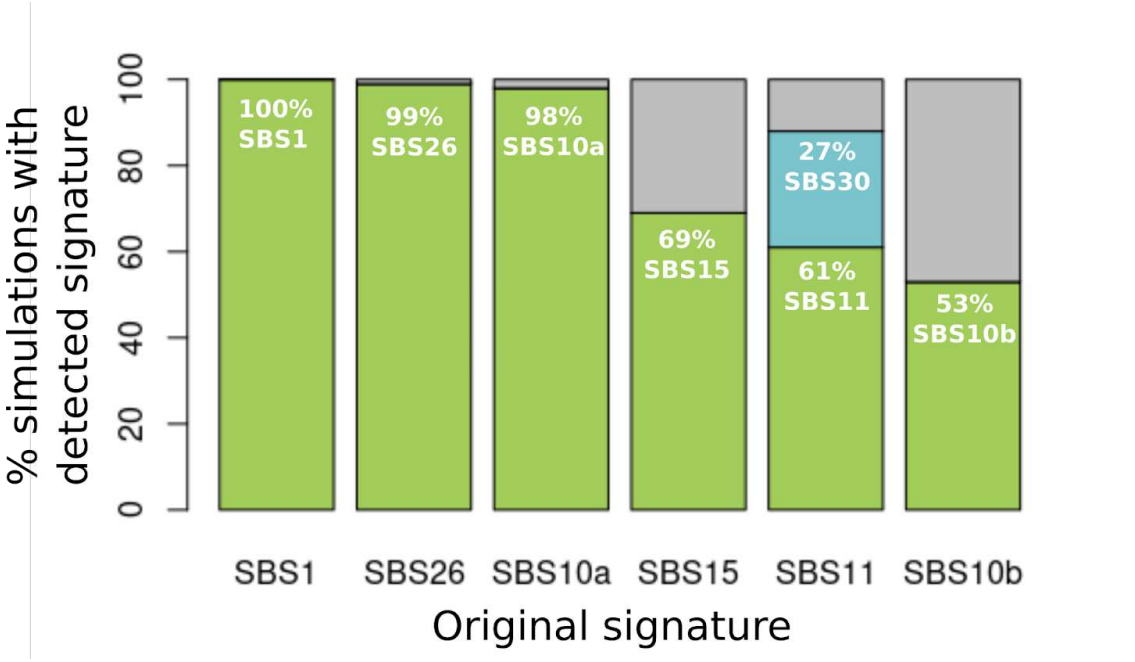
Supplemental Figure 4.S6: Comparison between TCGA CRC and COSMIC signatures

Cosine similarity heat map between de novo signatures found in TCGA (y axis) and COSMIC signatures (x axis). TCGA signatures were named after their closest COSMIC mutational process match (in bold)



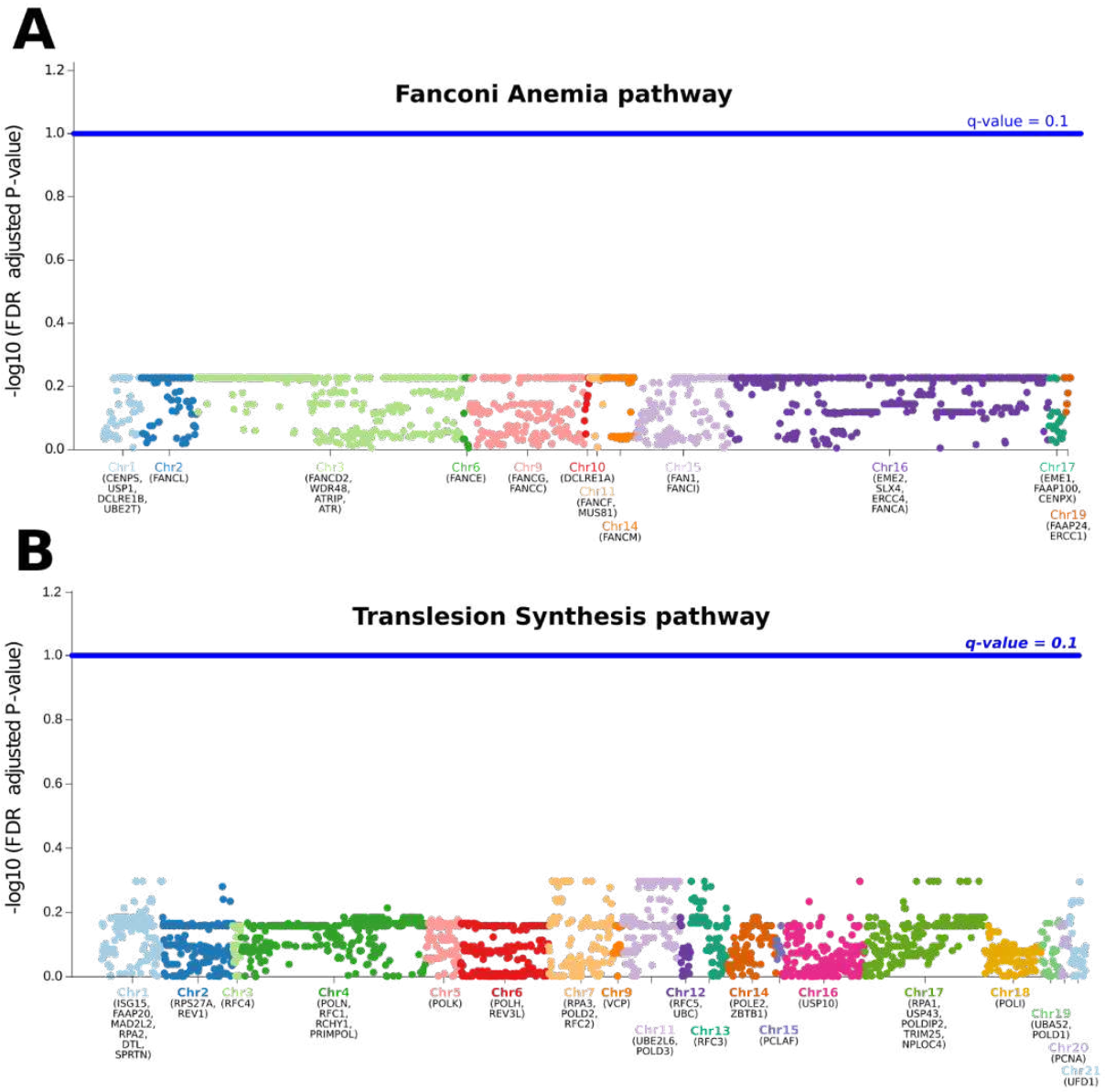
Supplemental Figure 4.S7: Lack of association of the alkylating signature with germline SNPs in BER genes

The False Discovery Rate (FDR) adjusted p-value (y axis) for SNPs in each BER (x axis) colored by chromosome location. The blue horizontal line indicates the significance threshold of <0.1. BER: Base Excision Repair



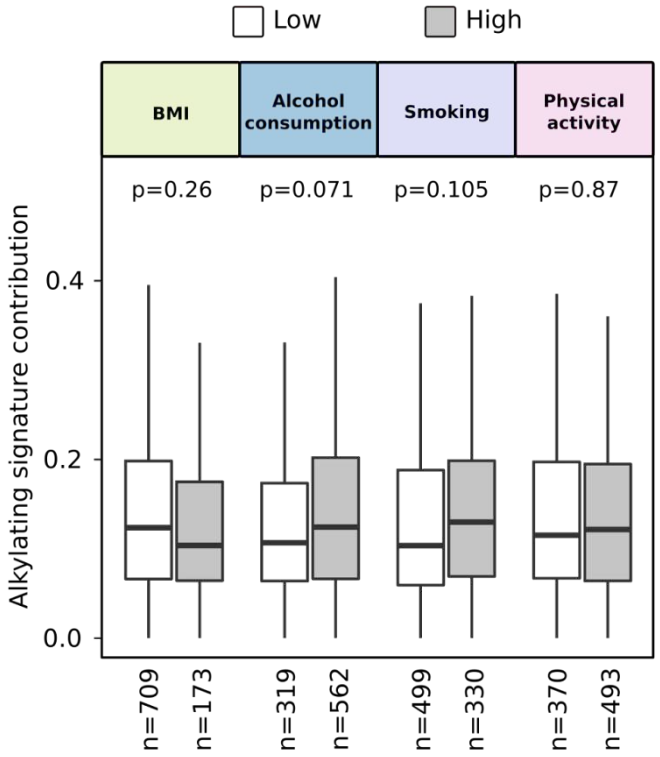
Supplemental Figure 4.S8: Signature assignment after under sampling

Under sampling simulations from 900 (NHS/HPFS sample size) to 540 (TCGA CRC sample size). For each signature (x axis) we count each type of closest signatures (y axis) deconvoluted after under sampling.



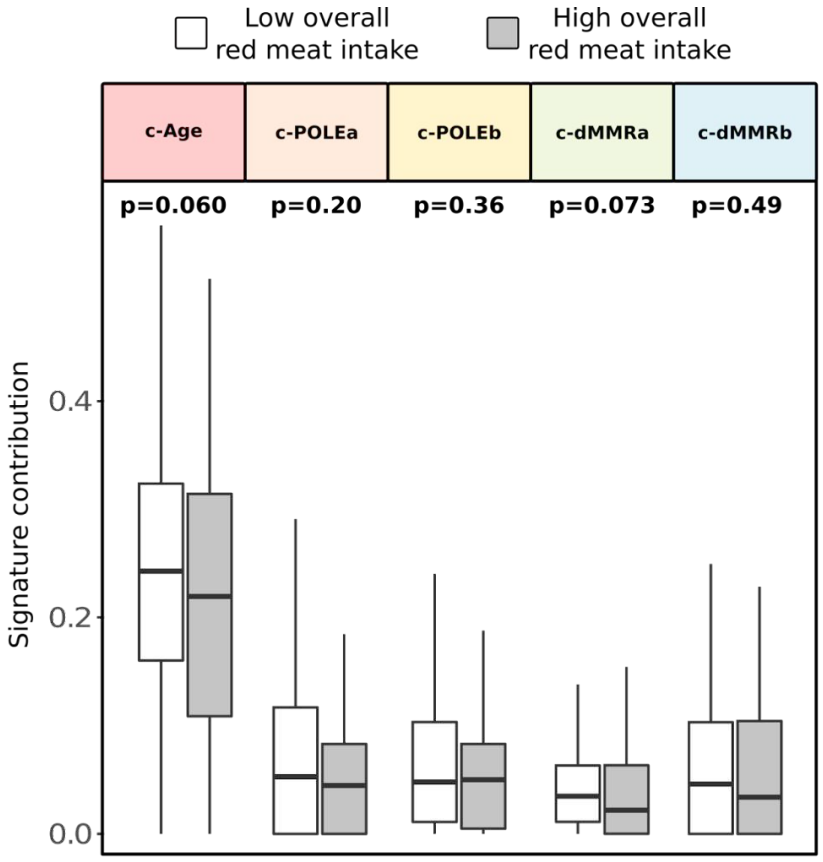
Supplemental Figure 4.S9: Lack of association of the alkylating signature with germline SNPs in FA and TLS genes

The False Discovery Rate (FDR) adjusted p-value (y axis) for SNPs in each FA (A) and TLS (B) gene (x axis) colored by chromosome location. The blue horizontal line indicates the significance threshold of <0.1. FA: Fanconi Anemia, TLS: Translesion synthesis.



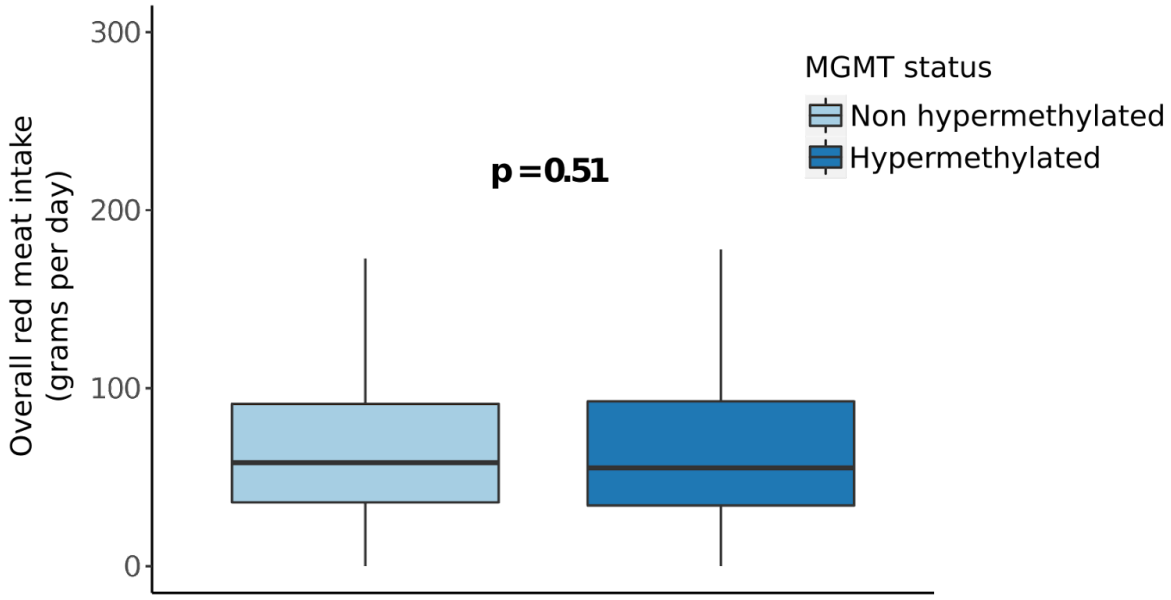
Supplemental Figure 4.S10: Alkylating signature and lifestyle factors

Proportion of mutations assigned to alkylating damage in NHS/ HPFS, segregated by alcohol consumption (alcohol consumption versus none), smoking (current or former smoker vs never smoker, obesity (BMI > 30 versus BMI < 30) and physical activity [Metabolic equivalent of task (METS)-hours/week > 10 versus < 10. The two-sided Mann-Whitney test p-value is indicated. BMI: Body Mass Index.



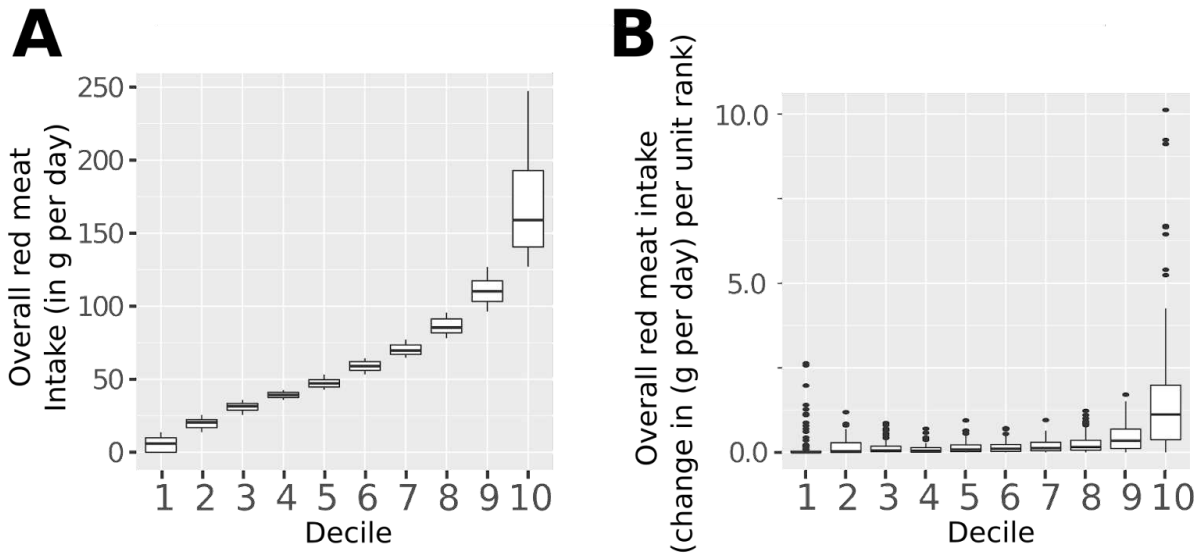
Supplemental Figure 4.S11: Association of red meat intake with other CRC mutational processes

Proportion of mutations assigned to alkylating damage segregated by overall red meat intake (top decile in grams per day versus the rest, as defined in the manuscript) for the other CRC mutational processes. The two-sided Mann-Whitney test p-value is indicated.



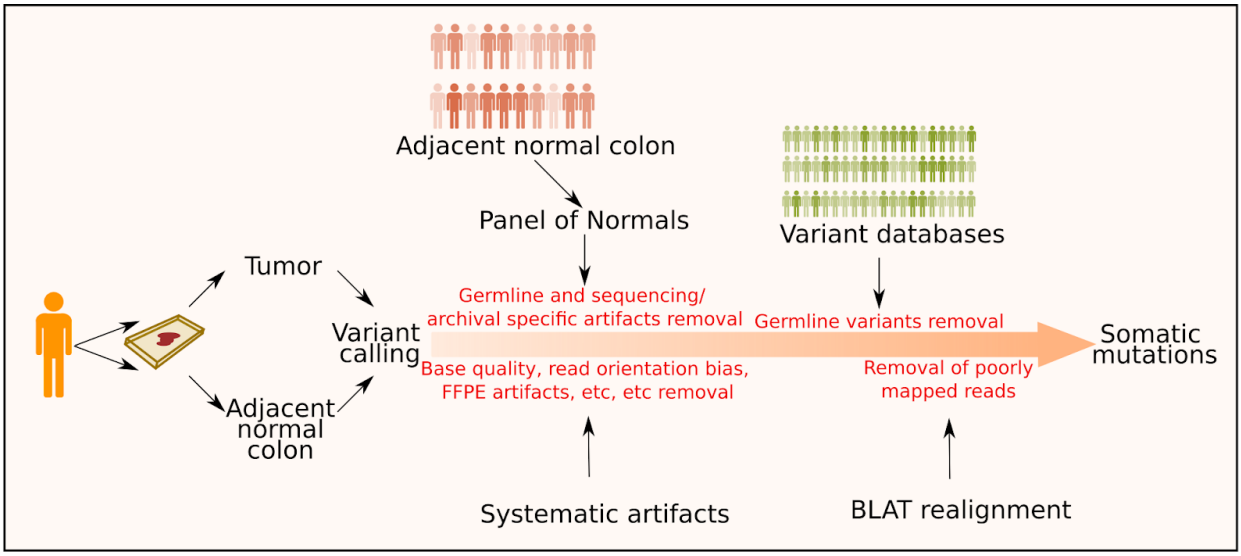
Supplemental Figure 4.S12: MGMT promoter methylation status and red meat intake

Red meat intake for NHS/ HPFS patients, stratified by MGMT promoter methylation status. The two-sided Mann Whitney U test p-value is indicated.



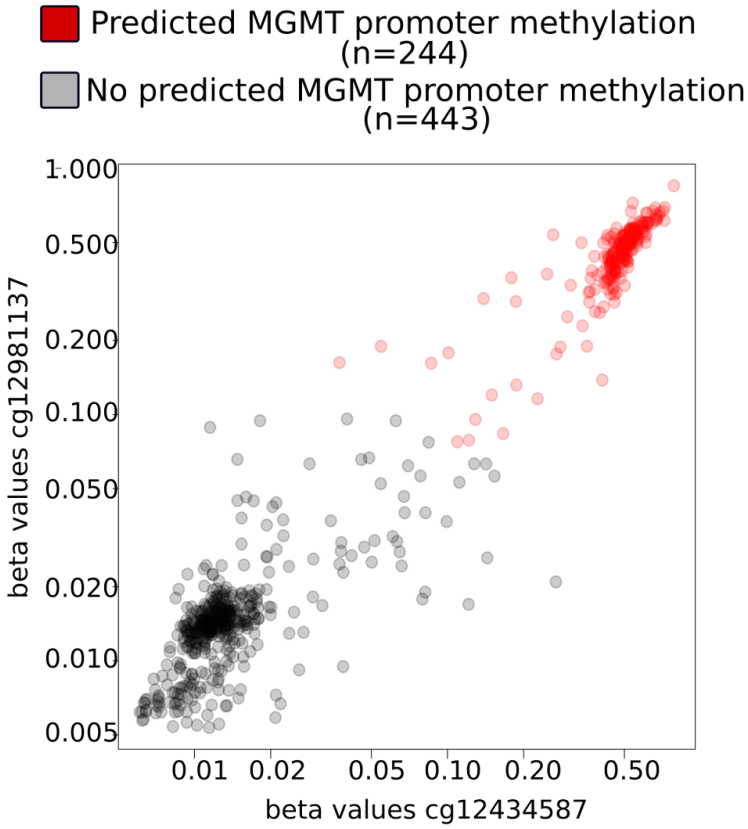
Supplemental Figure 4.S13: Distribution of overall red meat intake

(A) Overall red meat intake (in grams per day) decile distribution (B) First derivative (per unit rank) of overall red meat intake (in grams per day)



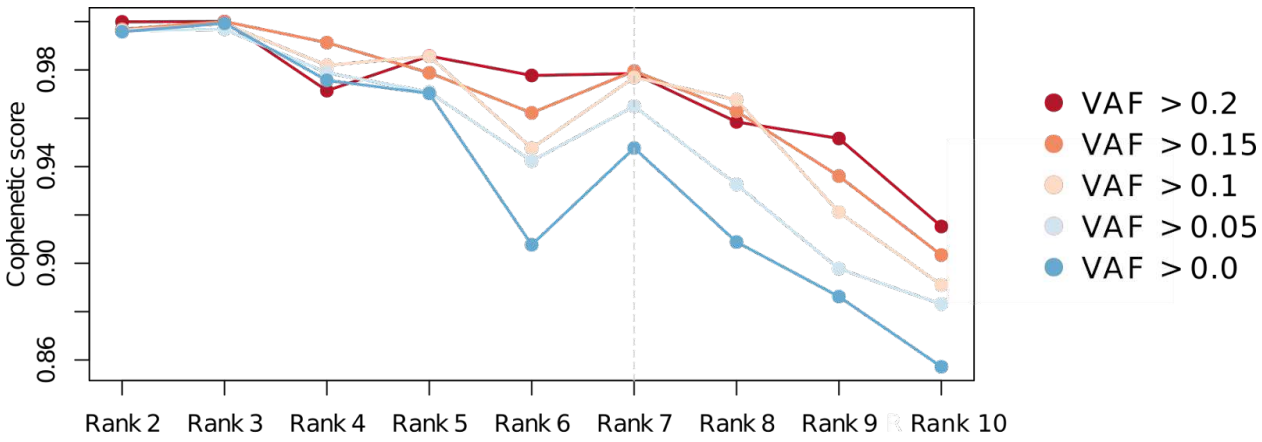
Supplemental Figure 4.S14: Variant calling pipeline

Overview of the variant calling pipeline as described in CGA WES Characterization pipeline (https://github.com/broadinstitute/CGA_Production_Analysis_Pipeline) developed at the Broad Institute of MIT and Harvard.



Supplemental Figure 4.S15: MGMT promoter methylation status in TCGA COAD/ READ

MGMT promoter methylation status in TCGA COAD/READ as predicted by the MGMT-STP27 model. The beta values of two probes (cg12434587 and cg12981137, resp. on the x and y axis) are plotted, as are the predicted methylated status.



Supplemental Figure 4.S16: Impact of VAF on NMF rank selection

The cophenetic score after NMF in NHS/HPFS was computed for different ranks (2 to 10, on the x axis) and different variant allele fraction thresholds (0 to 20).

5. Discussions

Recent technological advances have facilitated the fast sequencing of tumors at a low cost. Because driver mutations are at the root of carcinogenesis, their analysis has been the primary goal of sequencing studies. However, the cancer genome can harbour more than a million somatic mutations, a majority of which are passengers with no role in cancer progression. Recent advances have highlighted the use of passenger mutations as molecular fingerprints of past exposure to mutagens, as well as potential targets for immunotherapy treatments.

This thesis aimed at further expanding the use of passenger mutations in informing cancer prevention and treatment. This chapter provides a discussion on the importance of the papers presented in this work, as well as a critical reflection on the results.

5.1 Significance of the work

5.1.1 Refining the use of mutational load in the clinic

Tumour mutational burden is the second FDA-approved tissue-agnostic biomarker of response in solid tumors; it represents an important advancement in the field of oncology by further supporting the use of genomic testing to drive personalized medicine for cancer patients. However, more analyses are needed to refine the use of TMB for therapy selection.

The first chapter of this dissertation investigated the predictiveness of TMB across cancer types. We found that evidence of an association between TMB and response to immunotherapy relies largely on data for two cancer types - melanoma and non-small cell lung cancer - which are likely confounded by cancer subtypes.

After multiple hypothesis testing, lacking in the original studies, we did not find any TMB cutoff that could distinguish a group with significantly increased survival. TMB is overall a very poor biomarker: even in the best-case scenario, 25% of responders fall below the treatment prioritization threshold and are thus discouraged from receiving a potentially efficacious and life-extending treatment.

Last, we put forward a mathematical model that expands the neoantigen theory and reconciles it with the clinical data. First, our model presents a simple mathematical underpinning of the neoantigen theory and is consistent with the lack of association observed between TMB and response/ survival after

immunotherapy. Second, the model reproduces the effect of immunodominance hierarchies of T-cell responses observed *in vivo* [111]. Finally, it explains why immunoediting (i.e. negative selection against immunogenic mutations) is inefficient and allows tumours to accumulate a high TMB.

Overall, the results of the first chapter caution against using TMB in a clinical setting, as it could skew access to immunotherapy for patients who might benefit from it. Since the publication of these results, other cancer-specific studies have supported the lack of predictiveness of TMB [187–189].

To further understand the genomic and immune underpinnings of response to immunotherapy in high TMB patients, the second part of this work examined the case of a tumour with extremely high TMB due to microsatellite instability (MSI-H).

This case report is the first comprehensive molecular description of intrinsic resistance to checkpoint inhibitors in MSI-H tumours and can further help therapy selection for cancer patients. We found that the patient's tumor had a loss of Beta-2-Microglobulin (B2M) which is part of the Major Histocompatibility Complex class I (MHC I). Subsequently, the loss of B2M halted the presentation of tumor antigens at the cell surface. Thus, although the tumor presented a high TMB, the resulting neoantigens could not bind to T-cells and trigger an immune response (**Figure 1.6**).

However, cells with missing B2M are also expected to be selectively eliminated by Natural Killer (NK) cells [190]. Consistently, RNAseq and multiplex immunofluorescence experiments confirmed that the patient's tumor presented

a high infiltration of NK cells. Yet, the presence of an immunosuppressive environment might explain the lack of active tumor killing.

Altogether, the second chapter refined our understanding of TMB in a clinical setting. In particular, the detection of mutations in the antigen-presenting pathway can further help identify patients that are the most likely to derive benefit from immunotherapy.

In conclusion, the first two chapters demonstrate the limits of using mutational load as-is for therapy selection. Nevertheless, our analysis suggests that the use of mutational load in the clinic can be informed by specific driver mutations (e.g. *B2M* loss, as presented in Chapter 2) to refine patients' therapy selection.

5.1.2 Mechanistic link between red meat and colorectal cancer

The landscape of passenger mutations is not shaped by selection. Therefore, it constitutes a track record of the mutagenic processes the tumour underwent before their sequencing. The study of passenger mutations can help identify the molecular underpinnings of cancer development associated with lifestyle habits, which is the first step towards earlier detection, better therapies and improvement of patients' survival.

However, prior studies lacked comprehensive epidemiological annotations that would give insights into how lifestyle exposures impact tumour molecular features. Although several behaviours such as dietary habits have

been labelled as carcinogenic, proof of their mutagenic effect has not been observed directly in patients' tumours.

The third chapter provided for the first time evidence of a novel mutational signature in CRC. Our study demonstrated that the signature is the biological consequence of DNA alkylation. In addition, we found this signature in normal colonic cells, which suggests that the alkylating signature is present early during cancer progression.

Furthermore, we showed that this signature is associated with high intakes of processed and unprocessed red meat, which have been hypothesized to yield alkylating damage. This is the first time an alkylating signature is linked to a component of the diet in CRC, and this further supports the carcinogenicity of red meat consumption.

Consistently, we observed that many of the known cancer features related to red meat intake are also faithfully reproduced by this alkylating signature, mainly a higher incidence in distal tumours and enrichment in *KRAS p.G12D* mutations.

Overall, the last chapter further supported the use of passenger mutations to confirm the carcinogenicity of putative mutagens and guide cancer prevention guidelines. Prior to this study, the presupposed DNA damage induced by red meat intake had never been observed directly in human tumors. In addition, our mutational signature analysis exemplified how leveraging passengers can predict how mutagenic processes affect driver mutations and initiate cancer.

5.2 Limits and future directions

Here, we discuss the limits of focusing only on the passenger mutational landscape and potential future directions for the research projects covered in this dissertation .

5.1.1 Leveraging passengers for precision medicine

5.1.1.1 Refinement of TMB-derived predictors of response to immunotherapy

The poor predictiveness of TMB has stimulated the search for other determinants of immunotherapy response. This includes advanced fitness cost models [115] that predict patient survival, and TMB-derived measures based on mutation clonality. The latter is based on the observations that clonal mutations are more likely recognized by T-cells [191]. Recent advances also include machine learning models [189] incorporating TMB and predicting clinical response with higher sensitivity and specificity.

5.1.1.2 Alternative treatments for tumors with high TMB

Although passenger mutations have individually little to no role in cancer progression, they can collectively steer its trajectory [192]. To examine passenger mutations' role in light of cancer treatment, we solely focused on leveraging

them to predict response to immunotherapy. However, other treatments can also leverage high mutational burden.

Indeed, the accumulation of mutations is a hallmark of cancer progression. Consequently, cancer cells often manifest stress-related phenotypes caused by DNA damage, which can be exploited in the clinic [193]. In particular, an excess of mutations can lead to proteotoxic stress caused by the accumulation of misfolded proteins. The unfolded protein response (UPR) is a signalling pathway that allows cells to cope with the accumulation of unfolded proteins through adaptive programs to reduce misfolded protein accumulation [194]. Hence, inhibitors of UPR could potentially unleash the deleterious effect of accumulated passengers and be a potential alternative to immunotherapy in tumours with high TMB [195,196].

5.1.2 Leveraging passengers for precision prevention

5.1.2.1 All carcinogens are not direct mutagens

Although our study exemplifies the identification of mutagenic lifestyle factors (here, red meat consumption) to potential carcinogenic effects, it is important to note that carcinogens are not all direct mutagens. An alternative model [197,198] is that carcinogens affect selective constraint, which lead to the clonal expansion of pre-existing mutations. This is supported by recent [198] mice experiments of exposures to known human carcinogens which showed no associated mutational signature.

Non-mutagenic carcinogens can also act by weakening immunosurveillance. HIV (Human Immunodeficiency Virus) positive patients can present low T-cell counts and subsequently develop lymphomas associated with EBV (Epstein–Barr Virus) or KSV (Kaposi sarcoma-associated herpesvirus) infection[199].

Moreover, some mutational signatures are non-specific to biological processes. This is often the case when the biological process is an indirect mutagen. For example, chronic inflammation leads to indirect mutagenic damage by generating nitric oxide in affected tissues [200]. HBV (Hepatitis B Virus) and HCV (Hepatitis C Virus) are well-documented examples of inflammatory agents which can lead to hepatocellular carcinoma [201].

Thus, while mutational signature analysis is an invaluable tool to understand carcinogenesis and potential treatments, it might not provide the whole picture.

5.1.2.2 Identification of inherited predispositions

Although there are general guidelines for red meat consumption, there are no tailored dietary recommendations based on inter-individual variability. In that sense, the identification of a novel alkylating signature associated with red meat consumption offers an exciting opportunity to identify patients that are most susceptible to dietary-induced alkylating damage.

In particular, a specific polymorphism [202] (*rs16906252*) present in 12% of the global population [203], has been previously associated with an impairment of *MGMT*. MGMT is a central gene to alkylating damage repair. As our study

showed an increase of alkylating damage in tumors with loss of *MGMT,* patients harboring this particular polymorphism might be particularly sensitized to dietary-induced alkylating damage.

Bibliography

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
2. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* (2015) doi:10.1126/science.1260825.
3. Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43–47 (2015).
4. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
5. Anand, P. *et al.* Cancer is a preventable disease that requires major lifestyle changes. *Pharm. Res.* **25**, 2097–2116 (2008).
6. Lynch, H. T. *et al.* Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin. Genet.* **76**, 1–18 (2009).
7. Wu, S., Zhu, W., Thompson, P. & Hannun, Y. A. Evaluating intrinsic and non-intrinsic cancer risk factors. *Nat. Commun.* **9**, 3490 (2018).
8. Lortet-Tieulent, J. *et al.* State-Level Cancer Mortality Attributable to Cigarette Smoking in the United States. *JAMA Intern. Med.* **176**, 1792–1798 (2016).
9. Popkin, B. M. Understanding global nutrition dynamics as a step towards

controlling cancer incidence. *Nat. Rev. Cancer* **7**, 61–67 (2007).

10. on Cancer, I. A. F. R. & Others. IARC monographs on the evaluation of carcinogenic risk of chemicals to man. *IARC Monogr. Eval. Carcinog. Risk Chem. Man* **1**, (1972).
11. Kemp, C. J. Animal Models of Chemical Carcinogenesis: Driving Breakthroughs in Cancer Research for 100 Years. *Cold Spring Harbor Protocols* vol. 2015 db.top069906 (2015).
12. Mortelmans, K. & Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* **455**, 29–60 (2000).
13. Mardis, E. R. & Wilson, R. K. Cancer genome sequencing: a review. *Hum. Mol. Genet.* **18**, R163–8 (2009).
14. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
15. Koboldt, D. C. Best practices for variant calling in clinical sequencing. *Genome Med.* **12**, 91 (2020).
16. Xiao, W. *et al.* Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. *Nat. Biotechnol.* **39**, 1141–1150 (2021).
17. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
18. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* vol. 45 1113–1120 (2013).
19. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

20. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
21. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
22. Kumar, S. *et al.* Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell* **180**, 915–927.e16 (2020).
23. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **89**, 311–336 (2007).
24. Tilk, S., Curtis, C., Petrov, D. A. & McFarland, C. D. Most cancers carry a substantial deleterious load due to Hill-Robertson interference. doi:10.1101/764340.
25. Giannakis, M. *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep.* **17**, 1206 (2016).
26. Chow, A. Y. Cell cycle control by oncogenes and tumor suppressors: driving the transformation of normal cells into cancerous cells. *Nature Education* **3**, 7 (2010).
27. Casey, S. C., Baylot, V. & Felsher, D. W. The MYC oncogene is a global regulator of the immune response. *Blood* **131**, 2007–2015 (2018).
28. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034–1035 (2018).
29. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).

30. Gundem, G. *et al.* IntOGen: integration and data mining of multidimensional oncogenomic data. *Nat. Methods* **7**, 92–93 (2010).
31. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
32. Dietlein, F. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
33. Vogelstein, B. & Kinzler, K. W. The path to cancer—three strikes and you're out. *N. Engl. J. Med.* **373**, 1895–1898 (2015).
34. Hubbard, T., Wooster, R., Rahman, N. & Stratton, M. R. A census of human cancer genes. *Nat. Rev.* (2004).
35. Thomas, R. K. *et al.* High-throughput oncogene mutation profiling in human cancer. *Nat. Genet.* **39**, 347–351 (2007).
36. Eroglu, Z. & Ribas, A. Combination therapy with BRAF and MEK inhibitors for melanoma: latest evidence and place in therapy. *Ther. Adv. Med. Oncol.* **8**, 48–56 (2016).
37. Le, T. & Gerber, D. E. Newer-Generation EGFR Inhibitors in Lung Cancer: How Are They Best Used? *Cancers* **11**, (2019).
38. Hartmann, J. T., Haap, M., Kopp, H.-G. & Lipp, H.-P. Tyrosine kinase inhibitors - a review on pharmacology, metabolism and side effects. *Curr. Drug Metab.* **10**, 470–481 (2009).
39. Yuan, M., Huang, L.-L., Chen, J.-H., Wu, J. & Xu, Q. The emerging treatment landscape of targeted therapy in non-small-cell lung cancer. *Signal Transduct Target Ther* **4**, 61 (2019).

40. Alexandrov, L. B. Signatures of mutational processes in human cancer. (2014) doi:10.13140/2.1.4037.2164.
41. De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169–185 (2004).
42. Islami, F. *et al.* Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. *CA Cancer J. Clin.* **68**, 31–54 (2018).
43. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
44. Vöhringer, H., Van Hoeck, A., Cuppen, E. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nat. Commun.* **12**, 3628 (2021).
45. Bian, X. *et al.* Comparing the performance of selected variant callers using synthetic data and genome segmentation. *BMC Bioinformatics* **19**, 429 (2018).
46. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
47. Tomkova, M., Tomek, J., Kriaucionis, S. & Schuster-Böckler, B. Mutational signature distribution varies with DNA replication timing and strand asymmetry. *Genome Biol.* **19**, 129 (2018).
48. Grolleman, J. E. *et al.* Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-tumor Phenotype. *Cancer Cell* **35**, 256–266.e5 (2019).
49. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821–836.e16 (2019).

50. Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks+ E. coli. *Nature* (2020) doi:10.1038/s41586-020-2080-8.

51. Kocakavuk, E. *et al.* Radiotherapy is associated with a deletion signature that contributes to poor outcomes in patients with cancer. *Nat. Genet.* **53**, 1088–1096 (2021).

52. Temko, D., Tomlinson, I. P. M., Severini, S., Schuster-Böckler, B. & Graham, T. A. The effects of mutational processes and selection on driver mutations across cancer types. *Nat. Commun.* **9**, 1857 (2018).

53. Cannataro, V. L., Mandell, J. D. & Townsend, J. P. Attribution of Cancer Origins to Endogenous, Exogenous, and Actionable Mutational Processes. *bioRxiv* 2020.10.24.352989 (2021) doi:10.1101/2020.10.24.352989.

54. Rosendahl Huber, A., Van Hoeck, A. & Van Boxtel, R. The Mutagenic Impact of Environmental Exposures in Human Cells and Cancer: Imprints Through Time. *Front. Genet.* **12**, 760039 (2021).

55. Jiang, T. *et al.* Tumor neoantigens: from basic research to clinical applications. *J. Hematol. Oncol.* **12**, 93 (2019).

56. Wei, S. C., Duffy, C. R. & Allison, J. P. Fundamental Mechanisms of Immune Checkpoint Blockade Therapy. *Cancer Discov.* **8**, 1069–1086 (2018).

57. Hodi, F. S. *et al.* Combined nivolumab and ipilimumab versus ipilimumab alone in patients with advanced melanoma: 2-year overall survival outcomes in a multicentre, randomised, controlled, phase 2 trial. *The Lancet Oncology* vol. 17 1558–1568 (2016).

58. Postow, M. A., Callahan, M. K. & Wolchok, J. D. Immune Checkpoint Blockade in

Cancer Therapy. *J. Clin. Oncol.* **33**, 1974–1982 (2015).

59. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **363**, 711–723 (2010).
60. Dang, T. O., Ogunniyi, A., Barbee, M. S. & Drilon, A. Pembrolizumab for the treatment of PD-L1 positive advanced or metastatic non-small cell lung cancer. *Expert Rev. Anticancer Ther.* **16**, 13–20 (2016).
61. Ribas, A. & Wolchok, J. D. Cancer immunotherapy using checkpoint blockade. *Science* **359**, 1350–1355 (2018).
62. Osipov, A. *et al.* Tumor Mutational Burden, Toxicity, and Response of Immune Checkpoint Inhibitors Targeting PD(L)1, CTLA-4, and Combination: A Meta-regression Analysis. *Clin. Cancer Res.* (2020) doi:10.1158/1078-0432.CCR-20-0458.
63. Riaz, N. *et al.* The role of neoantigens in response to immune checkpoint blockade. *Int. Immunol.* **28**, 411–419 (2016).
64. Matsushita, H. *et al.* Cancer exome analysis reveals a T-cell-dependent mechanism of cancer immunoediting. *Nature* **482**, 400–404 (2012).
65. Heemskerk, B., Kvistborg, P. & Schumacher, T. N. M. The cancer antigenome. *EMBO J.* **32**, 194–203 (2013).
66. Hugo, W. *et al.* Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* **168**, 542 (2017).
67. Snyder, A. *et al.* Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).
68. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in

metastatic melanoma. *Science* **350**, 207–211 (2015).

69. Rizvi, N. A. *et al.* Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
70. Hellmann, M. D. *et al.* Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. *Cancer Cell* **33**, 843–852.e4 (2018).
71. Food and Drug Administration. Highlights of prescribing information: KEYTRUDA. *https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/125514s068lbl.pdf*.
72. Wu, H.-X., Wang, Z.-X., Zhao, Q., Wang, F. & Xu, R.-H. Designing gene panels for tumor mutational burden estimation: the need to shift from ‘correlation’ to ‘accuracy’. *J. Immunother. Cancer* **7**, 206 (2019).
73. Fancello, L., Gandini, S., Pelicci, P. G. & Mazzarella, L. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. *J Immunother Cancer* **7**, 183 (2019).
74. Gurjao, C., Tsukrov, D., Imakaev, M., Luquette, L. J. & Mirny, L. A. Limited evidence of tumour mutational burden as a biomarker of response to immunotherapy. *BioRxiv* (2020).
75. Gurjao, C. *et al.* Intrinsic Resistance to Immune Checkpoint Blockade in a Mismatch Repair–Deficient Colorectal Cancer. *Cancer Immunol Res* **7**, 1230–1236 (2019).
76. Gurjao, C. *et al.* Discovery and Features of an Alkylating Signature in Colorectal Cancer. *Cancer Discov.* (2021) doi:10.1158/2159-8290.CD-20-1656.
77. Rahib, L., Wehner, M. R., Matrisian, L. M. & Nead, K. T. Estimated Projection of

US Cancer Incidence and Death to 2040. *JAMA Netw Open* **4**, e214708 (2021).

78. Chan, T. A. *et al.* Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* **30**, 44–56 (2019).
79. Cristescu, R. *et al.* Pan-tumor genomic biomarkers for PD-1 checkpoint blockade–based immunotherapy. *Science* vol. 362 eaar3593 (2018).
80. Samstein, R. M. *et al.* Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).
81. Yarchoan, M., Hopkins, A. & Jaffee, E. M. Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *N. Engl. J. Med.* **377**, 2500–2501 (2017).
82. Sharma, P., Hu-Lieskovan, S., Wargo, J. A. & Ribas, A. Primary, Adaptive, and Acquired Resistance to Cancer Immunotherapy. *Cell* **168**, 707–723 (2017).
83. Heeke, S. & Hofman, P. Tumor mutational burden assessment as a predictive biomarker for immunotherapy in lung cancer patients: getting ready for prime-time or not? *Translational Lung Cancer Research* **7**, 631 (2018).
84. Sicklick, J. K. *et al.* Molecular profiling of cancer patients enables personalized combination therapy: the I-PREDICT study. *Nat. Med.* **25**, 744–750 (2019).
85. McGranahan, N. & Swanton, C. Neoantigen quality, not quantity. *Sci. Transl. Med.* **11**, (2019).
86. Miao, D. *et al.* Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors. *Nat. Genet.* **50**, 1271–1281 (2018).
87. Miao, D. *et al.* Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* **359**, 801–806 (2018).
88. Braun, D. A. *et al.* Interplay of somatic alterations and immune infiltration

modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. *Nat. Med.* **26**, 909–918 (2020).

89. Snyder, A. *et al.* Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: An exploratory multi-omic analysis. *PLoS Med.* **14**, e1002309 (2017).
90. Liu, D. *et al.* Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nat. Med.* **25**, 1916–1927 (2019).
91. Riaz, N. *et al.* Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934–949.e16 (2017).
92. Litchfield, K. *et al.* Meta-analysis of tumor and T cell intrinsic mechanisms of sensitization to checkpoint inhibition. (2020).
93. Wood, M. A., Weeder, B. R., David, J. K., Nellore, A. & Thompson, R. F. Burden of tumor mutations, neoepitopes, and other variants are weak predictors of cancer immunotherapy response and overall survival. *Genome Med.* **12**, 33 (2020).
94. Patel, N. A., Vokes, N. I., Elmarakeby, H., Hanna, G. J. & Van Allen, E. M. Abstract 5859: Genomic correlates of response to immune checkpoint inhibitors in advanced head and neck squamous cell carcinoma. *Molecular and Cellular Biology / Genetics* (2020) doi:10.1158/1538-7445.am2020-5859.
95. Motzer, R. J. *et al.* Avelumab plus axitinib versus sunitinib in advanced renal cell carcinoma: biomarker analysis of the phase 3 JAVELIN Renal 101 trial. *Nat. Med.* (2020) doi:10.1038/s41591-020-1044-8.
96. Pepe, M. S., Feng, Z., Janes, H., Bossuyt, P. M. & Potter, J. D. Pivotal evaluation

of the accuracy of a biomarker used for classification or prediction: standards for study design. *J. Natl. Cancer Inst.* **100**, 1432–1438 (2008).

97. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* **545**, 175–180 (2017).
98. Mark, N. M. *et al.* Chronic Obstructive Pulmonary Disease Alters Immune Cell Composition and Immune Checkpoint Inhibitor Efficacy in Non-Small Cell Lung Cancer. *Am. J. Respir. Crit. Care Med.* **197**, 325–336 (2018).
99. Wang, W. *et al.* Impact of COPD on prognosis of lung cancer: from a perspective on disease heterogeneity. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 3767–3776 (2018).
100. Salehi-Rad, R. & Dubinett, S. M. Understanding the Hurdles in Lung Cancer Immunotherapy in the Context of Chronic Obstructive Pulmonary Disease. *American journal of respiratory and critical care medicine* vol. 198 835–837 (2018).
101. Lim, J. U. *et al.* Chronic Obstructive Pulmonary Disease-Related Non-Small-Cell Lung Cancer Exhibits a Low Prevalence of EGFR and ALK Driver Mutations. *PLoS One* **10**, e0142306 (2015).
102. Suzuki, M. *et al.* Molecular Characterization of Chronic Obstructive Pulmonary Disease-Related Non-Small Cell Lung Cancer Through Aberrant Methylation and Alterations of EGFR Signaling. *Annals of Surgical Oncology* vol. 17 878–888 (2010).
103. Garassino, M. C. *et al.* Evaluation of blood TMB (bTMB) in KEYNOTE-189: Pembrolizumab (pembro) plus chemotherapy (chemo) with pemetrexed and

platinum versus placebo plus chemo as first-line therapy for metastatic nonsquamous NSCLC. *J. Clin. Orthod.* **38**, 9521–9521 (2020).

104. Bortolomeazzi, M. *et al.* Immunogenomics of colorectal cancer response to checkpoint blockade: analysis of the KEYNOTE 177 trial and validation cohorts. *Gastroenterology* (2021) doi:10.1053/j.gastro.2021.06.064.

105. Westfall, P. H. & Stanley Young, S. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. (John Wiley & Sons, 1993).

106. Dudoit, S., Shaffer, J. P. & Boldrick, J. C. Multiple Hypothesis Testing in Microarray Experiments. *Stat. Sci.* **18**, 71–103 (2003).

107. Shlyakhter, A., Mirny, L., Vlasov, A. & Wilson, R. Monte Carlo modeling of epidemiological studies. *Human and Ecological Risk Assessment: An International Journal* **2**, 920–938 (1996).

108. Xiao, Y. & Freeman, G. J. The microsatellite instable subset of colorectal cancer is a particularly good candidate for checkpoint blockade immunotherapy. *Cancer discovery* vol. 5 16–18 (2015).

109. Van den Eynden, J., Jiménez-Sánchez, A., Miller, M. L. & Larsson, E. Lack of detectable neoantigen depletion signals in the untreated cancer genome. *Nat. Genet.* **51**, 1741–1748 (2019).

110. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* **18**, 1009–1021 (2017).

111. Zamora, A. E. *et al.* Pediatric patients with acute lymphoblastic leukemia generate abundant and functional neoantigen-specific CD8+ T cell responses. *Sci. Transl.*

Med. **11**, (2019).

112. Marabelle, A. *et al.* Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol.* **21**, 1353–1365 (2020).
113. Spranger, S. *et al.* Density of immunogenic antigens does not explain the presence or absence of the T-cell–inflamed tumor microenvironment in melanoma. *Proceedings of the National Academy of Sciences* vol. 113 E7759–E7768 (2016).
114. Sarkizova, S. & Hacohen, N. How T cells spot tumour cells. *Nature* vol. 551 444–446 (2017).
115. Łuksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).
116. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, l1 (2013).
117. Feng, X., Li, L., Wagner, E. J. & Li, W. TC3A: The Cancer 3' UTR Atlas. *Nucleic Acids Res.* **46**, D1027–D1030 (2018).
118. Efremova, M. *et al.* Targeting immune checkpoints potentiates immunoediting and changes the dynamics of tumor evolution. *Nat. Commun.* **9**, 32 (2018).
119. McFarland, C. D., Mirny, L. A. & Korolev, K. S. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15138–15143 (2014).
120. Persi, E., Wolf, Y. I., Leiserson, M. D. M., Koonin, E. V. & Ruppin, E. Criticality in

tumor evolution and clinical outcome. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11101–E11110 (2018).

121. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* **49**, 1785–1788 (2017).

122. Topalian, S. L. *et al.* Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. *New England Journal of Medicine* vol. 366 2443–2454 (2012).

123. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

124. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* (2017) doi:10.1126/science.aan6733.

125. Overman, M. J. *et al.* Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *The Lancet Oncology* vol. 18 1182–1191 (2017).

126. Overman, M. J. *et al.* Durable Clinical Benefit With Nivolumab Plus Ipilimumab in DNA Mismatch Repair-Deficient/Microsatellite Instability-High Metastatic Colorectal Cancer. *J. Clin. Oncol.* **36**, 773–779 (2018).

127. Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* **20**, 682–688 (2014).

128. Giannakis, M. *et al.* RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat. Genet.* **46**, 1264–1266 (2014).

129. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary

segmentation for the analysis of array-based DNA copy number data. *Biostatistics* vol. 5 557–572 (2004).

130. Garcia, E. P. *et al.* Validation of OncoPanel: A Targeted Next-Generation Sequencing Assay for the Detection of Somatic Variants in Cancer. *Arch. Pathol. Lab. Med.* **141**, 751–758 (2017).
131. Grady, W. M., Rajput, A., Lutterbaugh, J. D. & Markowitz, S. D. Detection of aberrantly methylated hMLH1 promoter DNA in the serum of patients with microsatellite unstable colon cancer. *Cancer Res.* **61**, 900–902 (2001).
132. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
133. Nowak, J. A. *et al.* Detection of Mismatch Repair Deficiency and Microsatellite Instability in Colorectal Adenocarcinoma by Targeted Next-Generation Sequencing. *J. Mol. Diagn.* **19**, 84–91 (2017).
134. Carey, C. D. *et al.* Topological analysis reveals a PD-L1-associated microenvironmental niche for Reed-Sternberg cells in Hodgkin lymphoma. *Blood* **130**, 2420–2430 (2017).
135. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
136. AlDubayan, S. H. *et al.* Inherited DNA-Repair Defects in Colorectal Cancer. *Am. J. Hum. Genet.* **102**, 401–414 (2018).
137. Grasso, C. S. *et al.* Genetic Mechanisms of Immune Evasion in Colorectal Cancer. *Cancer Discov.* **8**, 730–749 (2018).
138. Zaretsky, J. M. *et al.* Mutations Associated with Acquired Resistance to PD-1

Blockade in Melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).

139. Gettinger, S. *et al.* Impaired HLA Class I Antigen Processing and Presentation as a Mechanism of Acquired Resistance to Immune Checkpoint Inhibitors in Lung Cancer. *Cancer Discovery* vol. 7 1420–1435 (2017).

140. Sade-Feldman, M. *et al.* Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8**, 1136 (2017).

141. Porgador, A., Mandelboim, O., Restifo, N. P. & Strominger, J. L. Natural killer cell lines kill autologous β2-microglobulin-deficient melanoma cells: Implications for cancer immunotherapy. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 13140–13145 (1997).

142. Wagner, A. K. *et al.* Retuning of Mouse NK Cells after Interference with MHC Class I Sensing Adjusts Self-Tolerance but Preserves Anticancer Response. *Cancer Immunology Research* vol. 4 113–123 (2016).

143. Anfossi, N. *et al.* Human NK cell education by inhibitory receptors for MHC class I. *Immunity* **25**, 331–342 (2006).

144. Shifrin, N., Raulet, D. H. & Ardolino, M. NK cell self tolerance, responsiveness and missing self recognition. *Semin. Immunol.* **26**, 138–144 (2014).

145. Nuñez, S. Y. *et al.* Human M2 Macrophages Limit NK Cell Effector Functions through Secretion of TGF-β and Engagement of CD85j. *J. Immunol.* **200**, 1008–1015 (2018).

146. Noy, R. & Pollard, J. W. Tumor-associated macrophages: from mechanisms to therapy. *Immunity* **41**, 49–61 (2014).

147. Kaneda, M. M. *et al.* Corrigendum: PI3Kγ is a molecular switch that controls immune suppression. *Nature* **542**, 124 (2017).

148. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

149. Platz, E. A. *et al.* Proportion of colon cancer risk that might be preventable in a cohort of middle-aged US men. *Cancer Causes Control* **11**, 579–588 (2000).

150. Bouvard, V. *et al.* Carcinogenicity of consumption of red and processed meat. *Lancet Oncol.* **16**, 1599–1600 (2015).

151. Bernstein, A. M. *et al.* Processed and Unprocessed Red Meat and Risk of Colorectal Cancer: Analysis by Tumor Location and Modification by Time. *PLoS One* **10**, e0135959 (2015).

152. Larsson, S. C., Rafter, J., Holmberg, L., Bergkvist, L. & Wolk, A. Red meat consumption and risk of cancers of the proximal colon, distal colon and rectum: the Swedish Mammography Cohort. *Int. J. Cancer* **113**, 829–834 (2005).

153. Bastide, N. M., Pierre, F. H. F. & Corpet, D. E. Heme iron from meat and risk of colorectal cancer: a meta-analysis and a review of the mechanisms involved. *Cancer Prev. Res.* **4**, 177–184 (2011).

154. Liao, X. *et al.* Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. *N. Engl. J. Med.* **367**, 1596–1606 (2012).

155. Lubbe, S. J., Di Bernardo, M. C., Chandler, I. P. & Houlston, R. S. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J. Clin. Oncol.* **27**, 3975–3980 (2009).

156. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).

157. Zhang, J., Stevens, M. F. G. & Bradshaw, T. D. Temozolomide: mechanisms of

action, repair and resistance. *Curr. Mol. Pharmacol.* **5**, 102–114 (2012).

158. Fahrer, J. *et al.* Dose–response of alkylation-induced colorectal carcinogenesis in MGMT-proficient and -deficient mice. *Toxicology Letters* vol. 221 S71 (2013).

159. Povey, A. C. *et al.* DNA alkylation and repair in the large bowel: animal and human studies. *J. Nutr.* **132**, 3518S–3521S (2002).

160. Bingham, S. A. *et al.* Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study. *Lancet* **361**, 1496–1501 (2003).

161. Billson, H. A. *et al.* Dietary variables associated with DNA N7-methylguanine levels and O6-alkylguanine DNA-alkyltransferase activity in human colorectal mucosa. *Carcinogenesis* **30**, 615–620 (2009).

162. Chao, A. *et al.* Meat consumption and risk of colorectal cancer. *JAMA* **293**, 172–182 (2005).

163. Brink, M. *et al.* Meat consumption and K-ras mutations in sporadic colon and rectal cancer in The Netherlands Cohort Study. *Br. J. Cancer* **92**, 1310–1320 (2005).

164. Gilsing, A. M. J. *et al.* Dietary heme iron and the risk of colorectal cancer with specific mutations in KRAS and APC. *Carcinogenesis* **34**, 2757–2766 (2013).

165. Song, M., Vogelstein, B., Giovannucci, E. L., Willett, W. C. & Tomasetti, C. Cancer prevention: Molecular and epidemiologic consensus. *Science* vol. 361 1317–1318 (2018).

166. Nishihara, R. *et al.* Aspirin use and risk of colorectal cancer according to BRAF mutation status. *JAMA* **309**, 2563–2571 (2013).

167. Willett, W. C. *et al.* Reproducibility and validity of a semiquantitative food frequency questionnaire. *Am. J. Epidemiol.* **122**, 51–65 (1985).

168. Rimm, E. B. *et al.* Reproducibility and Validity of an Expanded Self-Administered Semiquantitative Food Frequency Questionnaire among Male Health Professionals. *American Journal of Epidemiology* vol. 135 1114–1126 (1992).

169. Ogino, S., Kawasaki, T., Brahmandam, M. & Cantor, M. Precision and performance characteristics of bisulfite conversion and real-time PCR (MethyLight) for quantitative DNA methylation analysis. *of molecular diagnostics* (2006).

170. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).

171. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).

172. Taylor-Weiner, A. *et al.* DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).

173. Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).

174. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).

175. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

176. Bady, P. *et al.* MGMT methylation analysis of glioblastoma on the Infinium

methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. *Acta Neuropathologica* vol. 124 547–560 (2012).

177. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
178. Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nature Cancer* vol. 1 249–263 (2020).
179. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
180. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
181. van Leeuwen, E. M. *et al.* Population-specific genotype imputations using minimac or IMPUTE2. *Nature Protocols* vol. 10 1285–1296 (2015).
182. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
183. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
184. Ogino, S. *et al.* Evaluation of markers for CpG island methylator phenotype (CIMP) in colorectal cancer by a large population-based sample. *J. Mol. Diagn.* **9**,

305–314 (2007).

185. Nosho, K. *et al.* PIK3CA mutation in colorectal cancer: relationship with genetic and epigenetic alterations. *Neoplasia* **10**, 534–541 (2008).

186. Ogino, S. *et al.* LINE-1 hypomethylation is inversely associated with microsatellite instability and CpG island methylator phenotype in colorectal cancer. *Int. J. Cancer* **122**, 2767–2773 (2008).

187. McGrail, D. J. *et al.* High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Ann. Oncol.* **32**, 661–672 (2021).

188. Rousseau, B. *et al.* The Spectrum of Benefit from Checkpoint Blockade in Hypermutated Tumors. *N. Engl. J. Med.* **384**, 1168–1170 (2021).

189. Chowell, D. *et al.* Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-01070-8.

190. Bix, M. *et al.* Rejection of class I MHC-deficient haemopoietic cells by irradiated MHC-matched mice. *Nature* **349**, 329–331 (1991).

191. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).

192. McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R. & Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2910–2915 (2013).

193. De Raedt, T. *et al.* Exploiting cancer cell vulnerabilities to develop a combination therapy for ras-driven tumors. *Cancer Cell* **20**, 400–413 (2011).

194. Clarke, R. *The Unfolded Protein Response in Cancer*. (Springer, 2019).

195. Jordan, D. M., Ramensky, V. E. & Sunyaev, S. R. Human allelic variation: perspective from protein function, structure, and evolution. *Curr. Opin. Struct. Biol.* **20**, 342–350 (2010).

196. Sheltzer, J. M. & Amon, A. The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends Genet.* **27**, 446–453 (2011).

197. Lopez-Bigas, N. & Gonzalez-Perez, A. Are carcinogens direct mutagens? *Nature genetics* vol. 52 1137–1138 (2020).

198. Riva, L. *et al.* The mutational signature profile of known and suspected human carcinogens in mice. *Nat. Genet.* **52**, 1189–1197 (2020).

199. Morales-Sánchez, A. & Fuentes-Pananá, E. M. Human viruses and cancer. *Viruses* **6**, 4047–4079 (2014).

200. Jaiswal, M., LaRusso, N. F., Burgart, L. J. & Gores, G. J. Inflammatory cytokines induce DNA damage and inhibit DNA repair in cholangiocarcinoma cells by a nitric oxide-dependent mechanism. *Cancer Res.* **60**, 184–190 (2000).

201. Petruzziello, A. Epidemiology of Hepatitis B Virus (HBV) and Hepatitis C Virus (HCV) Related Hepatocellular Carcinoma. *The Open Virology Journal* vol. 12 26–32 (2018).

202. Ogino, S. *et al.* MGMT germline polymorphism is associated with somatic MGMT promoter methylation and gene silencing in colorectal cancer. *Carcinogenesis* **28**, 1985–1990 (2007).

203. Phan, L. *et al.* ALFA: Allele Frequency Aggregator. *National Center for Biotechnology Information, US National Library of Medicine* **10**, (2020).

Extended summary (in French)

(A) Introduction

Les avancées technologiques récentes ont permis le séquençage rapide du génome à très bas coût. Le cancer étant une maladie du génome, ces nouvelles technologies ont conduit à la création de larges bases de données ADN de tumeurs afin de mieux caractériser le cancer.

En particulier, la recherche en oncologie s'est longtemps focalisée sur l'analyse des mutations dites *drivers*, qui sont des altérations du génome participant à la croissance tumorale. Néanmoins, les tumeurs peuvent contenir jusqu'à plusieurs millions d'autres mutations sans aucun impact sur la progression du cancer. Cette majorité de mutations sans rôle, appelée mutations *passengers*, sont perçues comme étant des mutations collatérales et ont été peu étudiées en oncologie.

Ma thèse s'intéresse à ces mutations *passengers*, et leur potentielle utilité pour la prévention et le traitement du cancer. En particulier, ma thèse explore deux attributs des mutations: leurs nombre total (appelé charge mutationnelle) et leurs localisations dans l'ADN (par l'analyse de 'signature mutationnelle'). Les projets présentés dans cette thèse comprennent (i) une méta-analyse de la charge mutationelle comme biomarqueur de réponse à l'immunothérapie (ii) le cas d'un patient dont le cancer colorectal sans reponse therapeutique a l'immunothérapie malgré une forte charge mutationelle (iii) L'analyse de signatures mutationelle d'une cohorte de patients atteint du cancer colorectal.

(B) La charge mutationelle comme biomarqueur de réponse à l'immunothérapie

Le premier chapitre de cette thèse remet en question un paradigme central en onco-immunologie: la notion qu'un nombre élevé de mutations ainsi que les néo-antigènes qui en résultent (reconnus comme « corps étrangers » par les lymphocytes T) mènent à une meilleure réponse immunitaire antitumorale. Plusieurs articles ont ainsi suggéré que les tumeurs à forte charge mutationelle (c.-à-d. le nombre total de mutations) réagissent mieux à l'immunothérapie. Ces études ont été à la base de l'approbation par la FDA (Agence fédérale américaine des produits alimentaires et médicamenteux) de la charge mutationelle comme facteur pronostique de la réponse à l'immunothérapie. La charge mutationnelle tumorale est ainsi devenu le deuxième marqueur de réponse “tissu-agnostique” (c.-à-d. Ne dépendant pas du tissu cancéreux approuvé par la FDA pour les tumeurs solides. Ceci représente une avancée importante dans le domaine de l'oncologie, notamment pour la médecine dite de précision où la sélection de traitement thérapeutique est optimisée selon des caractéristiques propres au patient. Néanmoins, des analyses supplémentaires sont nécessaires pour affiner l'utilisation de la charge mutationelle pour les patients atteints du cancer.

Nous re-examinons le paradigme de la charge mutationnelle en menant une méta-analyse de plus de 2500 patients cancéreux ayant reçu un traitement immunotherapeutique. Nous constatons que la preuve d'une association entre charge mutationelle et réponse à l'immunothérapie repose en grande partie sur les données de deux types de cancer - le mélanome et le cancer bronchique non à petites cellules. Ces deux types de cancer contiennent des sous-groupes pouvant confondre l’association entre charge mutationelle et réponse à l'immunothérapie.

Pour ré-analyser l’association entre charge mutationelle et réponse à l'immunothérapie, nous utilisons en premier lieu des métriques standard de performances diagnostiques (par exemple, l'analyse de courbes ROC i.e. fonction d’efficacité du récepteur). De manière générale, nous observons que la charge mutationelle est un très mauvais biomarqueur: dans le meilleur des cas, 25 % des patients bénéficiant de l'immunothérapie sont en dessous du seuil de priorisation du traitement de la FDA. En d’autre termes, près de 25% des patients peuvent donc être potentiellement privés d’un traitement vital.

De plus, notre analyse propose une correction statistique pour hypothèses multiples lorsque les tests ne sont pas indépendants. Après correction pour hypothèses multiples, absente des autres études en faveur de la charge mutationnelle comme facteur pronostique, nous ne trouvons aucun seuil de charge mutationelle pouvant être utile en clinique.

Enfin, nous construisons un modèle mathématique de la théorie des néo-antigènes reflétant nos observations de l'association entre charge mutationelle et réponse au traitement immunotherapeutique. Notre modèle présente un fondement mathématique cohérent avec l'absence d'association observée entre la charge mutationelle et la survie après immunothérapie. De plus, le modèle est cohérent avec l'effet d'immunodominance des cellules T observées in vivo. Enfin, notre modèle explique pourquoi l'immunoédition (c'est-à-dire la sélection négative de mutations immunogènes) est inefficace et permet aux tumeurs d’avoir une charge mutationelle élevée.

En résumé, ce chapitre démontre une utilité limitée de la charge mutationelle comme facteur pronostique de réponse au traitement immunothérapeutique.

(C) Analyse d'un patient traité par immunothérapie avec une charge mutationnelle tumorale élevée

Les tumeurs à charge mutationelle élevée, en particulier les tumeurs présentant des défauts de réparation de l'ADN tels que l'instabilité des microsatellites (MSI), sont un type de tumeur pour lequel le traitement immunotherapeutique est particulièrement recommandé par la FDA. Le deuxième chapitre de cette thèse vise à affiner notre compréhension du lien entre charge mutationelle et réponse à l'immunothérapie. Pour cela, nous menons l'analyse exploratoire d'un patient dont la tumeur MSI n'a pas répondu à l'immunothérapie. Ce chapitre a ainsi pour objectif de mettre en évidence les mécanismes de résistance intrinsèque à l'immunothérapie en mettant en lien les mutations *passengers* et *drivers* dans le but de prédire la réponse au traitement.

Pour mieux comprendre les fondements génétiques et immunitaires de la réponse à l'immunothérapie chez les patients à charge mutationelle élevée, nous effectuons d'abord le séquençage de l'exome pour caractériser les mutations *passengers* et *drivers* du patient. Nous constatons que la tumeur du patient présente une perte de beta-2-microglobulin (B2M) une des composantes du complexe majeur d'histocompatibilité de classe I (MHC I). Cette perte de B2M prohibe la présentation des néo-antigènes à la surface des cellules tumorales. Ainsi, bien que la tumeur ait une charge mutationelle élevée, les néo antigènes résultants ne peuvent pas être reconnus par les lymphocytes T: ces néo-antigènes ne peuvent ainsi pas déclencher une réponse immunitaire.

Cependant, les cellules sans marqueur du “soi” (en particulier, les marqueurs MHC I) devraient être éliminées par les Lymphocyte NK. Par analyse du transcriptome et immunofluorescence, nous observons effectivement une forte infiltration de lymphocyte NK dans la tumeur. La présence d’un environnement immunosuppresseur (i.e. infiltration de macrophages M2) pourrait expliquer l'absence d’élimination des cellules tumorales par les lymphocytes NK.

En résumé, cette étude de cas est la première description moléculaire de la résistance intrinsèque à l'immunothérapie pour un patient atteint d’un cancer colorectal MSI. La détection de mutations pouvant empêcher la présentation d'antigènes peut en outre aider une meilleure sélection des patients pour l'immunothérapie.

En conclusion, les deux premiers chapitres démontrent les limites de l'utilisation de la charge mutationnelle pour la sélection de patients pour l'immunothérapie. Néanmoins, notre analyse suggère que l'utilisation de la charge mutationnelle en clinique peut être améliorée par l'incorporation d'éléments génomiques additionnels. En particulier, certaines mutations drivers (par exemple, la perte de B2M, comme présenté au chapitre 2) peuvent aider à affiner la sélection des patients pour l'immunothérapie.

(D) Empreintes moléculaires dans le cancer colorectal

Le cancer colorectal est le troisième cancer le plus fréquent chez l'homme. Sa fréquence augmente de manière inquiétante chez les jeunes adultes: d'ici 2040, le cancer colorectal sera la principale cause de décès par cancer chez les personnes âgées de 20 à 40 ans. Bien que les raisons de cette

tendance soient encore mal comprises, l'alimentation -en particulier la consommation de viande rouge- est suspectée d'y contribuer. Néanmoins, à ce jour, aucun rôle mutagène de la viande rouge n'a été confirmé dans l'initiation et la progression du cancer colorectal.

Les mutations *passengers* constituent un historique des processus mutagènes subis par la tumeur lors de son évolution. L'étude des *passengers* peut ainsi aider à identifier les fondements moléculaires du développement du cancer associés à certains modes de vie. Le troisième chapitre de cette thèse présente l'analyse des empreintes mutationnelles d'une large cohorte de tumeurs colorectales: 900 tumeurs ont été séquencées à partir d'échantillons prélevés depuis les années 70. Cette cohorte prospective dispose ainsi d'informations complètes sur le mode de vie d'individus avant d'être diagnostiquée du cancer colorectal.

L'analyse de cette cohorte nous a permis de découvrir une nouvelle signature mutationnelle dans les tumeurs colorectales.Nous montrons que cette signature est la conséquence biologique de l'alkylation de l'ADN. Cette signature existe aussi dans les cellules coliques d'individus sains, ce qui suggère que cette signature alkylante est présente lors de l'initiation du cancer. Nous montrons ensuite que cette signature est associée à une consommation élevée de viande rouge transformée et non transformée, qui ont longtemps été suspecté d'alkyler l'ADN.

Ce troisième chapitre démontre ainsi pour la première fois qu'une empreinte mutation est liée à une composante de l'alimentation, ce qui confirme le rôle carcinogène de la consommation de viande rouge.

De plus, nous observons que de nombreuses caractéristiques cancéreuses connues liées à la consommation de viande rouge sont

fidèlement reproduites par cette signature alkylante. En particulier, nous observons une plus forte abondance de la signature alkylante dans (i) les tumeurs colorectales distales (proches du rectum) (ii) les tumeurs présentant une mutation *KRAS* p.G12D.