Genomic signatures of conditional selection in different copy number states in human cancers

Elizaveta Besedina

TESI DOCTORAL UPF / ANY 2022

DIRECTOR DE TESI
Fran Supek

Department de Medicina i Ciències de la Vida

upf. Universitat Pompeu Fabra Barcelona

Nothing in biology makes sense
except in the light of evolution

Dobzhansky, 1973

To my mother and my cariño
...and everyone who has supported me along the way

Acknowledgments

This thesis has not only been a scientific journey but also a life journey for me. I want to take a moment to thank all people that made this journey more enjoyable and easy for me.

First of all, I want to thank my supervisor Fran Supek for allowing me to do research in the lab, for his feedback on my thesis, and for regular challenging. I also would like to express my thanks to lab members, past and current, for providing me feedback on my thesis, teaching me, and helping me during these long four years. Special thanks to David and Marina, who made my work in the lab comfortable since the first day and who would help me every time I needed it.

I would like to take this opportunity to thank my thesis committee, Ben Lehner, Tomas Marques-Bonet, and Patrick Aloy, for their regular feedback on my project and advice on how to improve it. Besides, I also want to thank members of the jury for taking the time to evaluate my work: Solip Park, Maria del Mar Albà Soler, Martin Shaefer, David Alejandro Juan Sopeña, and Aleix Bayona.

I cannot express enough how thankful I am to Olivier. Thank you so much for being there for me during all these years – this thesis would not happen without your daily support and love. The value you added to this thesis cannot be overrated, and I want to take a moment to tell you this once more. Thank you for helping me.

I want to also thank all my friends whom I met along the way, both in Barcelona and from Russia.

Аня, хочу поблагодарить тебя за поддержку, которую ты оказывала мне все долгие годы, что мы знаем друг друга. Несмотря на расстояние, твоя дружба и участие были и остаются очень ценными для меня. Осознание, что где-то есть родной человек очень помогает в трудные минуты. Для меня таким человеком являешься ты. Шлю тебе теплый привет и желаю удачи с твоим проектом!

I would like to express my deepest thanks to Loïc, for his help and support. The warmth that I feel when you are around has been motivating and helping me in various ways. I feel lucky that you are in my life. Wishing

you luck with your project, I am sure it will be outstanding!

Federica and Anamaria, the moments we spent together laughing and talking, are stored in a special place in my heart. I am happy I have met you. Thank you for the fun times, there will be more to come!

Мама, без тебя эта диссертация не была бы написана. Бесконечная благодарность за все то, что ты мне дала, за условия которые позволили мне учиться и за поддержание интереса к науке. Спасибо за твою поддержку, твою энергию, любовь и мотивацию которые я чувствую. Люблю тебя.

Lastly, I would like to express my deep and sincere acknowledgment to all the justice fighters around the globe – but especially in my long-suffering home country. The example you set for others has been motivating me in many situations. Thank you for fighting and giving hope for a better future.

Resumen

Determinar los genes que están bajo selección positiva o negativa durante la evolución del cáncer puede explicar cómo se forman y progresan los tumores y además puede sugerir nuevos objetivos terapéuticos. Sin embargo, la detección de selección utilizando datos de mutaciones somáticas es un desafío debido a la heterogeneidad de la tasa de mutación en todo el genoma y las limitaciones de los algoritmos existentes. En particular, se necesitan métodos estadísticos que puedan identificar y cuantificar la selección somática condicional, es decir, cambios en la selección debido a la influencia de otros factores. En este estudio, nos centramos en el efecto de las alteraciones del número de copias somáticas en la selección de mutaciones.

Presentamos MutMatch, una metodología estadística que puede estimar tanto la fuerza de selección en diferentes condiciones como las posibles interacciones. Nuestro método también puede cuantificar el cambio en la selección de mutaciones somáticas condicionado a otros factores y su significación estadística. El método MutMatch incluye la heterogeneidad de la tasa de mutación en el genoma, la ambigüedad generada por alteraciones en las dosis de genes y los patrones de mutación de trinucleótidos. Lo logra estimando la tasa base de mutación a partir de genes vecinos no seleccionados o regiones no restringidas dentro del mismo gen. Aplicamos MutMatch para estudiar los efectos de selección específicos en diferentes dosis de genes impulsores en cánceres humanos utilizando conjuntos de datos de secuenciación del genoma, exoma y de paneles de genesl a gran escala. A continuación, buscamos patrones genómicos de selección negativa en regiones poco mutadas de oncogenes conocidos y en genes identificados como esenciales en experimentos *in vitro* con líneas celulares. Finalmente, caracterizamos el panorama del efecto de selección dependiente a la dosis de los genes en cada tipo de tumor a través de un análisis de reducción de dimensionalidad. Este análisis identificó al menos cuatro tendencias que varían independientemente en los efectos de selección en los genes impulsores, lo que puede proporcionar una clasificación de los mecanismos de activación de oncogenes e inactivación de genes supresores mediante combinaciones de alteraciones genéticas en tumores humanos.

Abstract

Determining which genes are under positive or negative selection during cancer evolution can explain how tumors form and progress, and suggest new therapeutic targets. However, the detection of selection using somatic mutation data is challenging because of severe confounding by mutation rate heterogeneity across the genome and limitations of existing algorithms. In particular, there is a need for statistical methods that can identify and quantify conditional somatic selection, meaning changes in selection due to the influence of other factors. In this study, we focus on the effect of somatic copy number alterations on the selection of mutations.

We present MutMatch, a statistical methodology that can estimate both the selection strength across different conditions and possible interactions. Our method can also quantify the condition-specific change of somatic mutation selection and its statistical significance. The MutMatch method controls for mutation rate heterogeneity across the genome, gene dosage confounding, and trinucleotide mutation signatures. It achieves this by deriving a mutation rate baseline estimated from neighboring non-selected genes or non-constrained regions within the same gene. We applied MutMatch to study selection effects specific to different gene copy number states on driver genes in human cancers using large-scale genome, exome, and panel sequencing data sets. Next, we searched for genomic signatures of negative selection in lowly mutated regions of known oncogenes, and in genes identified as essential in *in vitro* experiments with cell lines. Finally, we characterized the landscape of gene copy number-dependent selection effect in cancer genes across each cancer type via a dimensionality reduction analysis. This identified at least four independently varying trends in selection effects in driver genes, which can provide a classification of mechanisms of oncogene activation and tumor suppressor inactivation by combinations of genetic alterations in human tumors.

Contents

1 Introduction **1**

1.1 Cancer statistics: incidence and mortality 1
1.2 Cancer as an evolutionary model 2
1.3 Hallmarks of cancer 3
1.4 Discovery of cancer-driver events 7
1.4.1 Somatic alterations 7
1.4.2 Mutation processes 15
1.4.3 Mutation rate heterogeneity 15
1.4.4 Computational approaches for identifying selection . 22
1.5 Somatic selection 25
1.5.1 Negative selection in cancer 25
1.5.2 Epistatic interactions in cancer 27
1.6 Aim and objectives 34

2 Development and evaluation of the MutMatch method **37**

2.1 Overview 37
2.2 Design of the MutMatch method 38
2.2.1 Mutation rate modeling 38
2.2.2 Input data 41
2.2.3 Output interpretation 43
2.2.4 Mutation spectra control 43
2.2.5 Baseline mutation rate models 45
2.2.6 Filtering of genomic sites 48
2.3 Correction of regression estimates 49

2.3.1 Data sparsity leads to estimation biases 49
2.3.2 Randomization approach to debias selection estimates 53
2.4 Benchmarking and evaluation 55
2.4.1 Size of the neighborhood 55
2.4.2 Mutation rate outliers in neighboring genes 57

3 Oncogenes and essential genes are under purifying selection in human tumors 61
3.1 Overview 61
3.2 Results 62
3.2.1 Negative and positive oncogene selection shape tumor evolution 62
3.2.2 Opposing selection forces lead to underestimation of positive and negative selection 65
3.2.3 Essential genes are negatively selected in soma . . . 74
3.3 Chapter methods 78
3.3.1 Gene classification 78
3.3.2 Gene selection models 79
3.3.3 Post-processing of regression estimates of selection . 80

4 Epistatic interaction between mutations and copy number alterations in the same gene 81
4.1 Overview 81
4.2 Results 82
4.2.1 Conditional selection upon hemizygous gene loss . . 82
4.2.2 Conditional selection upon copy number gain 87
4.2.3 Validation in an independent dataset 92
4.3 Chapter methods 93
4.3.1 Gene selection models 93
4.3.2 Post-processing of regression estimates of selection . 94
4.3.3 VAF testing 94

5 Mechanistic classification of cancer genes based on selection effects 95
5.1 Overview 95

5.2 Results . . . 96
5.2.1 Dosage and stoichiometry of the wild-type and mutant alleles affect tumor fitness . . . 96
5.2.2 Inferring gene clusters from the patterns of selection 100
5.2.3 Selection signatures are differentially active between cancer types . . . 105
5.3 Chapter methods . . . 108
5.3.1 PCA on selection estimates . . . 108

6 Discussion 111
6.1 Negative selection in cancer genomes . . . 111
6.1.1 Negative selection on cell-essential genes . . . 113
6.1.2 Near-neutral selection acting in somatic cells on population-constrained genes . . . 115
6.1.3 Signatures of negative selection in oncogenes . . . 117
6.2 Mutant allele imbalance of cancer genes . . . 120
6.2.1 Allele imbalance of OGs in cognate cancer types . . . 121
6.2.2 Allele imbalance of TSGs in cognate cancer types . . . 123
6.3 Classification of genes based on the selection patterns across copy number states . . . 124

7 Methods 127
7.1 Mutation and copy number data collection and processing . 127
7.2 Mutation frequencies . . . 129
7.3 Annotation of cognate cancer types . . . 129
7.4 Simulation of bias in regression estimates . . . 130
7.5 CADD scores . . . 131
7.6 Genomic filters . . . 131
7.6.1 Hotspot detection . . . 131
7.6.2 NMD-detected and NMD-evading regions . . . 133

8 Supplementary Figures 135

9 Supplementary Tables 147

Acronyms

AUROC Area Under the Receiver Operating Characteristics. 38, 56, 59, 139

BER Base Excision Repair. 15

CADD Combined Annotation-Dependent Depletion. 47, 48, 131

CERES Computational correction of copy-number effect in CRISPR-Cas9 essentiality screens. 75, 76, 77, 78, 114, 115

CGC Cancer Gene Census. 38, 72, 78, 79, 105, 108, 146

CNA Copy Number Alteration. 10, 11, 12, 24, 30, 42, 59, 79, 80, 82, 92, 93, 98, 99, 101, 120, 122, 123, 125, 129, 130

CPTAC The Clinical Proteomic Tumor Analysis Consortium. 128

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats. 25, 27, 35, 75, 76, 78, 114, 115

DFCI Dana Farber Cancer Institute. 128

DSB Double-Stranded Break. 12, 13

GENIE Genomics Evidence Neoplasia Information Exchange. 92, 93, 128

GLM Generalized Linear Model. 38, 40, 41

GoF Gain-of-Function. 9, 32, 61, 64, 68, 84, 99, 117, 118, 121, 122

HGNC HUGO Gene Nomenclature Committee. 42

HMF Hartwig Medical Foundation. 127

LOEUF Loss-of-function Observed over Expected Upper bound Fraction. 75, 76, 77, 78, 108

LoF Loss-of-Function. 7, 9, 75, 105, 115, 116, 123, 143

LOH Loss Of Heterozygosity. 10, 12, 33, 120

MC3 The Multi-Center Mutation Calling in Multiple Cancers. 127, 129

MMEJ Microhomology-Mediated End Joining. 12

MMR Mismatch Repair. 13, 15, 17, 19

MS96 Trinucleotide Mutation Spectra. 23, 44, 45, 72, 79, 140, 141, 142

NER Nucleotide Excision Repair. 15, 18, 19

NHEJ Non-Homologous End Joining. 12

NMD Nonsense-Mediated mRNA Decay. 36, 64, 73, 112, 117, 133

OG Oncogene. 9, 25, 27, 31, 32, 33, 61, 64, 65, 66, 68, 69, 77, 78, 81, 82, 83, 84, 86, 87, 91, 92, 93, 103, 117, 118, 120, 121, 122, 123

PC Principal Component. 96, 101, 102, 103, 104, 105, 106, 107, 108, 146

PCA Principal Component Analysis. 95, 101, 103, 109, 124

PCAWG Pan-Cancer Analysis Of Whole Genomes. 127

POG570 Personal Oncogenomics project. 128

PTC Premature Termination Codon. 133

REVEL Rare Exome Variant Ensemble Learner. 48

SNV Single Nucleotide Variant. 8, 41, 81, 127, 128

TAD Topologically Associating Domain. 46

TCGA The Cancer Genome Atlas Program. 59, 85, 89, 94, 127

TF Transcription Factor. 19, 22

TSG Tumor Suppressor Gene. 4, 9, 11, 12, 14, 25, 30, 31, 32, 33, 61, 63, 66, 77, 78, 79, 81, 82, 83, 84, 85, 86, 87, 89, 91, 92, 93, 103, 105, 117, 118, 120, 121, 122, 123, 144

VAF Variant Allele Frequency. 84, 85, 87, 89, 90, 94, 129

WES Whole Exome Sequencing. 93, 127, 128

WGD Whole Genome Duplication. 12, 26, 32, 33, 85, 120

WGS Whole Genome Sequencing. 45, 93, 127, 128

Chapter 1

Introduction

1.1 Cancer statistics: incidence and mortality

Cancer is one of the leading causes of death worldwide, killing one out of eight people [1]. In 134 countries, cancer is the leading cause (first or second rank) of premature deaths in the cohort between 30 and 69 years old. According to the yearly studies of the American Cancer Society, the expected number of new cancer diagnoses in 2022 is around 1.9 million people; over 600 000 people die from cancer every year in only USA [2].

Although cancer can develop in anyone, age is a a major risk factor: over 80 % of cancer patients in the US are older than 55 years old. Regardless, each year, about 270 thousand cases are diagnosed in children. Having 18 million cases in 2018 and a growing and aging population, it is predicted that the total burden of cancer in the world will reach 29 million cases by 2040 [3].

Understanding the mechanisms that underlie tumor formation and its early development, transformation into more malignant subtypes, tumor drug response, and different trajectories that the evolution of each tumor can take is an important task in battling this disease and constructing personalized or patient-specific therapy.

1.2 Cancer as an evolutionary model

Cancer is an evolutionary process, given that somatic cells within an organism are subject to selection. The nature of this evolutionary process is similar to that of unicellular microorganisms lacking sexual reproduction [4, 5, 6].

Accumulation of mutations in cancer genomes provides a substrate for subsequent action of selection pressures. Positive selection increases the proportion of variants that give a fitness advantage for a tumor clone, while negative selection, also known as purifying, removes mutations that decrease cellular fitness compared to other clones [7, 8].

The lack of genetic recombination in somatic cells implies an inability of selection to remove or fix individual variants. The unit of selection, therefore, is the entire cancer genome, which will be removed from a tumor cellular pool or not depending on its relative aggregate fitness compared to other clones.

Mutations in a small proportion of the genome may lead to phenotypic changes important for cancer cells, which correspond to 5 %–10 % of all genes, according to different estimates [9]. Impactful mutations in these genes can be passed to the daughter cells and fixed in the population with time, or removed from the pool of mutations in cancer, depending on whether they increase or decrease cellular fitness. Cancer-driving events, including somatic mutations in cancer driver genes, have a positive impact on cellular fitness and are selected positively. On the contrary, harmful mutations in genes whose function is essential for tumor survival will be negatively selected.

On the other hand, the vast majority of mutations do not change cellular fitness or change it lightly. As a result, they are not selected or weakly selected. The loss or fixation of such mutations in the cellular population entirely depends on the genetic background of the cell. Because of that, these mutations are called passenger mutations. Clones with a fit background have a certain amount of passenger mutations; the exact number and the characteristics of passenger mutations are highly specific to cancer types, which can be used to classify tumors [10].

The inability to uncouple any two mutations (also known as Hill-Robertson interference) not only leads to the fixing of passenger or lightly deleterious mutations in very fit backgrounds but also is the reason for the ineffective selection of driver mutations if they appear in particularly unfit backgrounds.

There are two Hill-Robertson interference processes: hitchhiking and Muller's ratchet [11, 12, 13]. Hitchhiking is a way for a mutation to stay in a cell population and increase its allele frequency through a genomic linkage to a positively selected driver mutation.

Muller's ratchet mechanism is a process of continuous accumulation of deleterious mutations in an asexual genome [12, 13]. The chance of reverse mutations is very low, which makes the ratchet turn almost only one way. When a critical number of deleterious mutations is acquired, cells undergo a mutational meltdown [14, 11]. At this stage, first, the size of the cell population decreases because of the negative selection acting on those unfit genomes. In small-sized populations, genetic drift (changing the allelic frequencies because of a random chance, or sampling error) plays a significant role. The combination of Muller's ratchet mechanism with genetic drift becomes the main factor in the clone's survival; with time, a cell population might be lost as more fit clones outcompete them.

The rate of the two described Hill-Robertson interference processes depends on the genomic mutation rate [11]. A high mutation rate leads to an accumulation of higher numbers of passenger and deleterious mutations (and higher proportions of these mutations): therefore, in tumors with a high mutation rate and high mutation burden, the efficacy of the selection is decreased [11].

1.3 Hallmarks of cancer

All the diversity of neoplastic diseases has certain common alterations in cell physiology that allow the grouping of hundreds of known cancer types and subtypes together. These alterations are essential for the pathological development of the cell on its way to becoming malignant. Some of these features are associated with changes that are independent of the signals

that cells receive from the environment, while others are coupled to these signals [15].

The changes happening in cancerous cells were analyzed in the work of Hanahan & Weinberg in 2000, where they proposed that all the complexity of human tumor diseases are summarized in six hallmarks of cancer [15]. These traits include self-sufficiency in growth signals, insensitivity to growth-inhibitory signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis.

The final objective of all these phenotypical changes is to increase the cell division rate, decrease the probability of cell death and help to compete with other subclones. The listed traits are acquired gradually during the development of human tumors and reflect biological capabilities needed at different stages of tumor initiation and progression [15].

- **Sustaining Proliferative Signaling.** Normally, cell proliferation is carefully regulated by the growth-promoting signals that are received by the cell. Cancer cells have multiple ways to deregulate this system to increase their division rate. In particular, they might produce growth factors themselves, or send signals to their microenvironment that stimulate the release of such growth factors by noncancerous cells. Overexpressing receptors of such growth factors has the same downstream effect: oversensitivity to the levels of the growth factors and increased proliferation [16].

- **Evading Growth Suppressors.** Besides increasing proliferation, cells need to resist negative signals downregulating the rate of cell division rate. Tumors can suppress such programs of negative regulation by inactivation of Tumor Suppressor Genes (TSGs), inhibition, or counterbalancing [16].

- **Resisting Cell Death.** Programmed cell death, or apoptosis, is a way an organism gets rid of heavily damaged, unneeded or unhealthy cells. This cell process is started after various physiologic stresses, which are exploited by anticancer therapies. However, tumors have several strategies to limit the efficacy of apoptosis, including inactivation or loss of *TP53* tumor suppressor function, downregulation of proapoptotic factors, and upregulation of antiapoptotic factors [16].

- **Enabling Replicative Immortality.** For most of the cells in the human organism, only a limited number of cell divisions are available (Hayflick's limit), which is determined by the activity of telomerase. If this protein is not expressed or not active, each consequent division inevitably leads to a shortening of the chromosome ends, loss of the essential genes, and, ultimately, to cell death. Tumors demonstrate a capability to divide endlessly, which is promoted by upregulation of telomerase activity or, less commonly, through a recombination-like mechanism or telomeres maintenance [16].

- **Inducing Angiogenesis.** Constant growth of tumors leads to the increased need for oxygen and nutrient supply, as well as the efficient evacuation of metabolites. Expectedly, pre-tumor vasculature does not provide a good infrastructure to satisfy the newly formed and growing needs in transportation. An "angiogenic switch'' upregulating angiogenic signals is an essential step in forming many tumor types, which causes new vessels to grow through the tumor to more effectively deliver components needed for the tumor cells. The neovascularization mechanisms vary between tumor types and may be activated by the same mechanisms that are involved in proliferative signaling [16].

- **Activating Invasion and Metastasis.** Later development stages of various tumors are characterized by the local invasion in the surrounding tissues, and even deattaching, spreading, and colonizing other places in the body, including different organs. These changes are facilitated and accompanied by alterations in cell shapes and loss of the cell connections to the other cells or extracellular matrix. In particular, upregulation and downregulation of different cadherins were shown to be associated with metastasis in carcinomas [16].

Since 2000, new observations have been added to the original hallmarks of cancer that helped clarify them. Currently, two new **emerging hallmarks** have been added to the six original hallmarks [16].

- **Deregulating cellular energetics.** Tumor cells reorient cellular energy production pathways such that glucose catabolism does not

use oxygen. The breakdown of glucose into pyruvate by glycolysis produces energy and reduces electron carriers. In healthy cells, electrons are then transferred to the electron transport chain to yield additional energy and are ultimately transferred to oxygen. However, even in the presence of oxygen, tumors preferentially ferment pyruvate into lactate to restore the redox balance. This phenomenon is known as the Warburg effect. Although this way of obtaining energy is much less efficient than oxidative phosphorylation that takes place in mitochondria, it is believed that glycolytic intermediates are used to synthesize new amino acids and nucleotides, essential for the macromolecules of daughter cells in rapidly doubling tumor cells [16].

- **Avoiding immune destruction.** The majority of cancer cells are recognized by the immune system at early stages and eliminated from the body by the immune system. How tumors escape immune surveillance is still largely unknown. Some studies have recently shown that cancer cells may inhibit the activity of Cytotoxic T lymphocytes (CTLs) and Natural Killer (NK) cells by secreting immunosuppressive factors (including TGF-β). Others argue that cancer cells may attract immune cells, for example, regulatory T-cells, using their immunosuppressive activity to escape immune response [16].

Additionally, two **enabling characteristics** that help to acquire these phenotypic changes are also commonly accepted in the field [16]:

- **Genome instability and mutation.** To acquire the changes listed above, a cell should undergo certain alterations at DNA level. An important requirement for such changes should be that they are inheritable (thus certain epigenetic changes might also be included) and give a selective advantage to a subclone. In this paradigm, tumor evolution is a chain of clonal expansions, where each clone overcomes the previous one by obtaining an additional cancer hallmark. To facilitate the acquisition of these traits and compensate for the activity of DNA-repair systems, mutation rates are often higher in the tumor genome compared to normal cells. This is achieved

by the alterations that lead to a Loss-of-Function (LoF) in DNA repair pathways, mutations that increase the probability of mistakes inserted by DNA polymerases. Another feature of cancer genomes is widespread copy number alterations and aneuploidy [16].

- **Tumor-promoting inflammation.** Paradoxically, inflammation accompanying tumor tissues has a tumor-promoting effect. A variety of bioactive molecules taking part in the inflammation process have a tumor-promoting effect, such as growth factor molecules, antiapoptotic factors, pro-angiogenic factors, factors facilitating angiogenesis, and many others. Reactive oxygen species released by inflammatory cells additionally act as mutagenic factors, increasing the speed of accumulation of mutations [16].

1.4 Discovery of cancer-driver events

In the context of the above-listed cancer hallmarks, it is of primary importance to understand how cancer cells acquire traits that give a selective advantage and drive the course of tumor evolution, and which processes facilitate them. One such enabling characteristic is genomic instability. Here we will give a brief review of the main sources of genomic instability in cancer cells, their types and prevalences, and how they can change cell fitness.

1.4.1 Somatic alterations

Increasing genetic variability in a living system provides a substrate for natural selection, that will select for the fittest genetic background and remove the unfit ones from the population. Mutations and recombination are the changes that increase genetic variability. The recombination process happens mainly during meiosis, although some examples of somatic recombination are known, including V(D)J recombination, which takes part in the immune response formation. In the absence of recombination in normal somatic cells, accumulation of mutations and also genomic instability that leads to a loss, gain, or rearrangement of genomic loci

are the main sources of variability. A special type of genomic instability is a catastrophic cellular event, chromothripsis, or chromosomal shattering. Chromothripsis is characterized by large-scale (tens to thousands) chromosomal rearrangements occurring simultaneously early in tumor development [17, 18].

1.4.1.1 Mutations

The most common type of genetic change in cancer is small mutations – changes in the DNA sequence that occur locally. Mutations can be classified according to various characteristics that can take into account the scale of the change, the cause, and their consequence.

Most of the small mutations are point mutations, meaning that one mutation leads to a change of a single nucleotide. However, sometimes a single mutation event leads to a set of changes in the DNA, changing the identity of several neighboring nucleotides at once. While they are relevant for cancer development, they are not as well studied as point mutations, partially because of their relative rareness. Here and later, I will mainly refer to mutations having in mind point mutations, specifically Single Nucleotide Variants (SNVs).

From the mechanistic point of view, small mutations can be split into **substitutions** (substitutions of one nucleotide are called SNVs) and **indels** (insertions or deletions).

Indels of lengths that are not multiples of three change the codons downstream are called a **frameshift** mutations. Frameshift mutations are likely to generate a premature stop codon in this new frame, which results in the synthesis of a truncated protein. Otherwise, in-frame indels only lead to the insertion of a few additional amino acids in the protein or a loss of the amino acids.

SNVs can be classified according to the consequence at the phenotypic level. First, a nucleotide change within a protein-coding region may lead or may not lead to the change of the amino acid. Approximately two-thirds of mutations are **nonsynonymous** and generate changes in the DNA sequence. In contrast, about one-third of all mutations are **synonymous**

at the level of amino acid because of the redundancy in the genetic code. Nonetheless, somatic synonymous mutations might still be not silent and affect the cellular fitness level. This can happen because of a bias in codon usage in a cell: not every triplet encoding the same amino acid is used in translation at the same frequency. Some trinucleotides are avoided, while others are preferentially used to encode an amino acid. As a result, not optimal usage of the codons can affect the rate and efficacy of translation. Another effect of synonymous changes is that they may alter splicing sites and therefore be under purifying selection, which was reported for the human genome [19]. In cancer, synonymous changes sometimes affect splicing in Oncogenes (OGs) and TSGs and are under selection [20].

Nonsynonymous changes, in its turn, can be further split into **nonsense**, which generate premature stop-codon and truncated protein, and **missense** mutations that change the protein sequence. Rarely, stop-codons can be lost, which results in the translation of some additional amino acids before encountering a stop-codon in the 3' UTR of the mRNA.

At the bigger phenotypical scale, mutations may be **beneficial**, **neutral**, or **deleterious** for the cell. The frequency of beneficial mutations is very low. However, it is these mutations that drive the course of tumor evolution and are positively selected. Most of the somatic mutations are neutral or deleterious, having no or weak functional impact on cellular fitness.

Both deleterious and beneficial mutations are mainly generated by nonsynonymous changes. Both LoF and Gain-of-Function (GoF) mutations can be beneficial, depending on the cancer gene function. Genes whose activation is beneficial for the tumor are called oncogenes; genes that have to be switched off in cancer are called TSGs. Expectedly, GoF mutations are beneficial when they appear in TSGs and detrimental to oncogenes or essential genes. Same way, GoF mutations in TSGs are not beneficial for a cell and are positively selected for in oncogenes.

1.4.1.2 Copy number changes

While there might not be many, somatic Copy Number Alterations (CNAs) affect the biggest fraction of the genome compared to other changes that happen in cancer [21, 1]. Highly sensitive identification of somatic CNAs achieved by analyzing the allele imbalance at germline heterozygous loci reports that 94 % of the tumors have megabase-scale genomic copy number alterations [22].

Apart from short deletions and insertions that are part of small mutations, bigger genomic loci can be lost or gained during genome evolution. Loss or gain of the whole chromosome leads to aneuploidy; more local copy number events (focal changes or segments of a chromosome) affect only a limited part of the chromosome, from a few kilobases to the entire chromosome arm [23].

Deletions, or genomic losses, are opposed to genomic gains or amplifications. Depending on the number of the events, CNA of genomic regions can be a **homozygous deletion** (two copies are lost), a **loss** or Loss Of Heterozygosity (LOH) when only one copy is lost, a **gain** (an additional copy is gained) or **amplification**, or high-level gain, defined as two or more copies or eight or more copies, depending on the source [23]. A specific case of LOH is a copy-neutral LOH (CN-LOH), a lesion that results in the presence of two identical alleles [24, 25].

Some of the CNAs contribute to the tumorigenesis, while others appear randomly and accumulate under a neutral selection. CNAs contributing to the tumorigenesis will have an elevated rate, which, however, is complicated to estimate without an adequate model for background rates of CNAs under neutral selection [21]. Mechanisms that lead to CN-LOH can be different, including uniparental disomy (both copies are received from the same parent), the identity of both parental alleles due to the close relatedness of parents, or duplication of the allele after a genomic loss [26]. Temporal ordering of CNAs events, spatial distribution, and selection pressures can elucidate the functional role of CNAs and therefore are of great interest in cancer genomics.

Contribution of copy number changes to tumorigenesis

The expected straightforward effect of copy number change is a change in gene expression on the mRNA and protein level. This gene dosage effect, in reality, however, is often compensated by another mechanism. For example, dosage compensation through X-chromosome inactivation in females and transcriptional upregulation helps to maintain homeostasis. On the other hand, copy number effects in autosomal genes are less clear. Generally, there is a correlation between a ploidity of a gene and the transcription level, and the correlation becomes weaker for the protein levels as many additional mechanisms and feedback loops participate in maintaining a perfect proportion of the proteins [27, 28]. Nevertheless, in some genes dosage effect of amplified or lost genes is not compensated, which has a big role in cancer evolution [29, 30].

Contribution of CNA events to cancer evolution was discovered in 1971 when Knudson proposed the "two-hit hypothesis'' [31]. In this model, inactivation of a TSG requires two mutations that happen in two gene copies normally present in the genome. This indicated a recessive nature of the disease because a single gene copy can rescue a phenotype. Inactivation of the gene can be achieved by a point mutation, a gene deletion, epigenetic inactivation, or a combination of them. Homozygous deletions of two-hit TSGs or a hemizygous deletion accompanied by inactivation of the second allele by one of the alternative mechanisms are common in tumors. Examples of well-known two-hit TSGs include *RB1*, *PTEN*, *NF1* and others. Interestingly, some TSGs may switch from one-hit to two-hit drivers between cancer types [32].

Other TSGs demonstrate a haploinsufficient nature: inactivation of only one gene copy is enough to provoke a cancer phenotype. Observing a recurrent hemizygous deletion of the gene without an additional mutation in the second gene copy can be an indicator of this type of inactivation mechanism. Similarly, dominant-acting inactivation mutations will have the same pattern of deletions in genes. A study of tumor samples with hemizygous deletions published in 2012 suggested that gene islands enriched for the genes that negatively regulate proliferation and depleted in genes that promote tumorigenesis are often hemizygously deleted in tumors because of the cumulative haploinsufficiency of such gene groups

[30].

Alternatively, homozygous deletions also can shape the evolution of cancer genomes through a negative selection: essential genes located close to the TSGs are limiting the size of the homozygous deletion while tolerating a hemizygous deletion [33].

The interplay between copy number changes and genome ploidy can bias the results of analysis when seeking selection signals in CNAs distribution. Tumors that underwent a genome duplication event have a significantly lower proportion of the genome affected by a LOH. Moreover, a homozygous loss of essential genes is not negatively selected in tumors with Whole Genome Duplication (WGD) opposing to a negative selection in non-WGD tumors. WGD not only masks a deleterious effect of deleterious mutations in essential genes but also decreases the fitness cost of passenger mutations (that still may be slightly deleterious) in non-essential genes. The bigger the negative fitness cost of a passenger alteration, the more frequently WGD will be selected in tumors, and the bigger fraction of WGD samples will have a cancer type [34].

Mechanisms of copy number changes

Variability in lengths of genome loci affected by CNA suggests that there are multiple mechanisms of origin. Homologous recombination and non-homologous repair together play a key role in the generation of CNAs. Both mechanisms are used to repair DNA damage: a sequence that is used to repair DNA is homologous in one case and is not homologous or only has microhomology in another case. Various DNA damages, breaks, or gaps, including Double-Stranded Break (DSB), stalled replication forks induce **homologous recombination**. If a sequence that was chosen to repair a DNA break or damage was incorrect, the consequence of it is a deletion or a gain.

Another source of copy number variation is **non-homologous repair** pathways. Non-homologous repair **outside of replication** process consists of Non-Homologous End Joining (NHEJ) pathway and non-canonical Microhomology-Mediated End Joining (MMEJ) pathway. NHEJ repairs DNA breaks by rejoining DSB ends without requiring a homological sequence, and MMEJ needs a sequence that has a microhomology. Often

the end-joining process is not done correctly and small sequences might be lost. Alternatively, free mitochondrial DNA or parts of retrotransposons can be inserted in the place of DSB [35].

DSB can also lead to the loss of a telomere, in which case two sister chromatids without telomeres are likely to fuse, forming a dicentric chromosome. In the telophase, those two centrosomes will be pulled in two opposite directions, until inevitably another break does not happen at a random position. Until the telomere is acquired, a **breakage–fusion–bridge** cycle can repeat, causing multiple inverted duplications [35].

Replication slippage and **template switching** along the exposed DNA template during replication can be the reason for short deletions or duplications (with the upper length limit restricted by the length of Okazaki fragments) [35, 36, 37, 38]. It happens when the lagging strand has short inverted repeats, which causes the formation of secondary structures (for example, DNA hairpins). Replication machinery fails at copying these secondary structures and resumes the synthesis after it, or, optionally, can switch templates. The frequency of replication slippage is higher in regions with a high density of repeats, or in mutants with compromised Mismatch Repair (MMR) system or enzymes involved in DNA replication.

1.4.1.3 Other types of somatic alterations

Apart from copy number changes and mutations, cancer genomes obtain a variety of other somatic alterations.

Structural variants in cancer include **inversions**, **translocations** within and between chromosomes, and **complex events** combining them with duplications and deletions. An interesting example of a catastrophic event that happens in a quarter of bone cancers is **chromothripsis** and has implications in many other cancer types, characterized by many chromosomal rearrangements in a single cell [39]. The special term **chromoplexy** has been proposed for simultaneous combinatorial structural rearrangements in multiple chromosomes at once [40].

Interestingly, exogenous virus DNA may be acquired in cancer genomes with the most known examples of human papillomavirus, Epstein Barr

virus, hepatitis B virus and human herpesvirus [1].

Disruption of epigenetic regulation also has implications in tumorigenesis. The role of **DNA methylation** in the regulation of transcription levels of genes has been actively studied in normal cells. DNA methylation mainly occurs at the 5' position of the cytosine ring in CpG dinucleotides and has an opposite effect when it is in the gene promoter or a gene body. DNA methylation is a negative transcription regulator in transcription factor binding sites (in promoters and enhancers) and correlated positively with the transcription levels when DNA methylation is in the gene body where it is a marker of intronic sequences, repetitive sequences, or retrotransposons [41, 42, 43]. Moreover, DNA methylation influences the spatial organization of the genome and the formation of hetero- and euchromatin by attracting histone modifiers.

Overall, cancer cells show genome-wide hypomethylation, which results in genomic instability [44]. Notably, promoters in cancer cells often are hypermethylated in CpG islands compared to normal tissue of the same origin. In particular, promoter methylation is used to repress transcription and decrease protein levels of tumor suppressor genes. Silencing of TSGs via promoter methylation was shown for *CDKN2A*, *MLH1*, *BRCA1* and *VHL*. On the other hand, hypomethylation in *IGF2* oncogene promoter leads to the transcription upregulation in cancer [43].

Additionally, the change of the landscape of **histone modifications** that regulate DNA packing, and gene expression is a hallmark of cancer that is associated with increased potential of malignant transformations [43]. Enzymes that are involved in recognizing, adding, and removing histone marks are altered in many cancer types [43].

Epigenetic alterations, or epimutations, are a natural way of gene inactivation that can often act as a second hit in two-hit genes following the two-hit Knudson hypothesis [31, 45].

1.4.2 Mutation processes

Mutations in a cell in normal conditions appear spontaneously, or as a consequence of errors made during replication or recombination. Normally, these mistakes happen in any organism with a certain frequency that depends on the fidelity of a particular polymerase. The majority of the mistakes in replication do not pass to the daughter cells because of the DNA repair enzymes that recognize and fix them. However, when DNA repair systems are failing, errors in DNA become fixed mutations that will be passed to the next generations. Interestingly, most mutagenic lesions in DNA are not resolved within one cell cycle and can last a few cell generations before being repaired or turning into a mutation [46].

Some factors can increase the probability of errors and mutations in somatic cells. Endogenous factors start from changes that happen inside a cell, and exogenous factors comprise different chemical or physical, or biological mutagenic agents.

Many changes increasing a mutation rate in cells are cancer type specific. For example, expression of APOBEC enzymes that catalyze cytosine deamination leads to an increase in the frequency of specific mutation patterns in breast cancers [47, 48, 49]. Mutations in the exonuclease domain of DNA polymerase ε cause hypermutation phenotype characterized by the abundance of DNA lesions and high mutation burden of tumors in some cohorts of patients with colorectal, brain, and uterus cancer [50, 51]. Defects and dysfunction of DNA repair systems, such as MMR, Base Excision Repair (BER), Nucleotide Excision Repair (NER) are also associated with an elevated level of tumor mutation burden [52].

A variety of different factors, both endogenous and exogenous, affect the mutation rate at different scales. A more detailed overview of them is presented further.

1.4.3 Mutation rate heterogeneity

As mentioned above, the vast majority of nonsynonymous mutations do not change the fitness of the somatic cell. Such mutations are called passenger mutations as opposed to driver mutations that happen in cancer

genes and directly influence the course of the disease. As passenger mutations do not change the fitness of tumors, the speed of their accumulation entirely depends on the activity of selectively neutral mutational processes in a cell, both exogenous and endogenous.

Interestingly, the number of somatic passenger mutations correlates with the age of a patient, which together with an observation that a part of somatic mutations is shared between healthy and cancer tissues of the same origin suggests that at least half of somatic mutations in cancer appear before tumor initiation [53]. A later study by Tang et al. (2020) conducted on melanoma confirmed that mutation burdens of cancer cells were similar to mutation burdens of the neighboring normal cells [54]. This pattern, however, is not universal. For example, Reorink et al. (2018) reported that most mutations in colorectal cancer cells were acquired during the latest clonal expansion and were absent in the normal colorectal cells [55]. The differences might be driven by the activation of an additional mutational process after an acquisition of a driver alteration.

Tumor mutation rate can be estimated by accessing the regional mutation density that not only depends on the cancer type and stage of the tumor but also varies across the genome. Heterogeneity of mutation rate is present in cancer cells at many scales, from the level of a single nucleotide to a megabase level [52].

1.4.3.1 Domain-scale variability

Replication time

Large-scale variability of somatic mutation rate is correlated with the state of the chromatin and the DNA replication time. Openness/closeness of the chromatin, its enrichment with repressive or active chromatin marks, and replication time are largely correlated factors: all of them are associated with the gene expression level and clusters of genes that are highly or lowly expressed. However, careful analysis of the effects of each of the mentioned factors with controlling for confounders has shown that replication time can explain a big proportion of mutation rate variability across the genome and cell types [52, 56, 57]. The association at a megabase scale between replication time and DNA accessibility to the DNAse enzymes,

estimated via density of DNAse hypersensitive sites, explains a correlation between the lower density of DNase I hypersensitive (DHS) sites and higher mutation rates, which does not hold at smaller scale changes in DNA accessibility. Similarly, active and repressive chromatin marks alone have a subtle effect on changes in mutation rate, confirming that the major factor shaping the regional mutation rate is replication time.

However, a strong association between replication time and mutation rate does not reflect a causal relationship. There is growing evidence that DNA replication time likely affects the mutation rate through the process of differential DNA repair efficiency in late and early replicating regions. MMR-deficient tumors have more flat distribution of mutation rates between early and late replicating regions, supporting that MMR preferentially targets early replicating regions [58]. This was further validated with direct experiments that have shown that artificially induced MMR failure in human cell lines results in losing some variability between mutation rates in regions with different replication times. Additionally, DNA damage may preferentially target less active domains of heterochromatin [59, 60]. One of the proposed mechanisms is that the peripheral location in the nucleus of heterochromatin makes it more accessible for UV light and damage induced by UV exposure [59].

1.4.3.2 Between-genes and within-gene variability

H3K36me3 histone mark

Smaller scale variability at a level from a kilobase to hundreds of kilobases concerns differential mutation rates between genes and across different gene body parts. Mechanisms of such differences are associated with differential gene transcription levels. Highly expressed genes have a lower local mutation rate in both transcribed and untranscribed strands. Enrichment of highly expressed genes with a specific histone mark H3K36me3 was shown to be causal for this association. In particular, H3K36me3 attracts mismatch recognition protein to DNA in G1 and early S phases, which ensures faster correction of DNA damage in genes that have modification on histones of this type [61]. The abundance of H3K36me3 is generally increased in the 3'-ends of the genes compared to the 5' ends,

which determines the gradient of mutation rate changes along the gene body.

Coding and non-coding regions

Within-gene mutation rate has been observed to differ between exonic and intronic gene parts. Partially, this can be explained by higher levels of H3K36me3 histone marks in exons. Stringent analysis controlling for confounding factors such as repeat frequencies, GC-content, alignability of short reads, and excluding low-quality mutation calls demonstrated no statistically significant difference between mutation rates in exonic and intronic parts. On the other hand, selection forces acting on the coding gene parts may change the observed local mutation rate in exons. For example, a study published in 2012 has found that exons of cancer genes are enriched with mutations compared to the intronic regions and other untranslated DNA segments [62]. Another study has shown that spicing-associated sequences have a depleted mutation rate [63]. Oxidative damage in genic parts was also observed to be depleted compared to the intergenic regions [52].

Strand assymetry

The local mutation rate within a gene also depends on the strand. At least two mechanisms cause an asymmetry in mutation rates between strands. One of them is a transcription-coupled NER that preferentially removes DNA damage on the transcribed strand. Another one, less understood, is transcription-coupled damage that preferentially targets non-transcribed DNA strand [64].

Strand asymmetry can be created by a single mutational process combined with the subsequent selection of variants [46]. This is the case if mutations are created by a mutational process that does not have a flat mutational profile (that is, specifically mutating certain nucleotides) and if its intensity is high enough to generate multiple lesions within the same cell cycle. Here, each strand is damaged similarly, but the changes are asymmetrical: a damaged nucleotide in one strand will correspond to a normal nucleotide in another and vice versa. Because DNA lesions can remain in a cell for some time, they can pass the first round of division without being repaired or resolved into mutations [46]. This way, lesions

in sister chromatids become separated in daughter cells, where they are then resolved. Some of those mutations can be driver mutations and this will have a substantial fitness advantage for a cell carrying it. Statistically, driver mutations are rare, so only one of the daughter cells will likely have it. This will lead to the clonal expansion of only one of the daughter cells and with that, an enrichment of mutations of a certain type if called on the forward strand of the reference genome [46].

Interestingly, strand asymmetry spans regions that are smaller than a whole chromosome, as expected under the assumption that there is no recombination in somatic cells. However, the direction of the strand bias changes every several megabases for autosomes and is absent in the X chromosome in males. This suggests that homologous recombination plays a key role in resolving such clustered lesions, and is an additional mechanism of increasing the genetic diversity of somatic cells by combinatorial assembly of genetic variants [46].

1.4.3.3 Local variability at the sub-gene level

Transcription factors binding sites

Variability at the scale of tens of base pairs to kilobases is often associated with the enrichment of binding sites of Transcription Factors (TFs). For example, binding sites of the CTCF/cohesin complex that plays a key role in chromatin architecture and transcription are hypermutable. In many cancer types, such as stomach, colorectal, skin, liver, and melanoma, there was shown more than a 3-fold increase of local mutation rates in CTCF/cohesin binding sites [65, 66, 67]. One explanation of this is the physical exclusion of the repair systems from binding sites, both MMR and NER [65, 66].

Interference of proteins bound to the DNA with the NER is also observed for other TFs, especially in cancers where exposure to mutagens is more common: lung cancer and skin cancer. Disentangling which TFs have such interesting effects is challenging because many of them form complexes. Moreover, TF binding sites often overlap with CTCF binding sites. Another curious example is proteins of the ETS family, whose binding sites demonstrated an elevated mutation rate due to the differ-

ential DNA damage in those sites, as opposed to differential DNA repair. DNA bound to ETS proteins decreases the resistance to the UV-induced damage [68, 69].

Nucleosomes

Mutation rates have been shown to follow the periodicity of nucleosome occupancy in the genome, both at the population and somatic levels [70, 71, 72, 73]. Regular nucleosomal architecture influences the local mutation rate at the two periodicity lengths. Strong periodic pattern 10 basepairs length repeats the changes in DNA minor groove orientation, in particular, whether it faces towards the nucleosome or away from it. While a minor groove facing out the histones is theoretically more accessible for the DNA repair enzymes, mutagens might preferentially target the same nucleotides. Which factor plays a bigger role depends on the mutagen: UV-induced DNA damage is accumulating faster than the DNA repair systems acts upon them [72].

Another regularity in local mutation rates is observed at the level of 200 nucleotides, which corresponds to the internucleosomal distances. Internucleosomal, or linker DNA, is more exposed to DNA damage and DNA repair machinery. Similarly, type of the DNA damage influences the final distribution of mutation rates: UV-induced damage is enriched in nucleosomes, while cytosine deamination is nearly absent in nucleosomal DNA [74, 71].

1.4.3.4 Nucleotide-scale variability

Mutational signatures

The smallest scale of mutation rate heterogeneity is associated with the probability of having a mutation in a particular nucleotide context. Analyzing the patterns of relative frequencies of mutation types in thousands of tumor samples made it possible to reconstruct mutational signatures defined as relative frequencies of mutation types in a particular trinucleotide context [75, 76]. A linear combination of mutational signatures determines the final picture of mutation frequencies in a tumor sample.

Although the mutational signature is a mathematical construct [75], it is

expected (and shown in some cases) that a unique mutational process generates a mutational signature. Mutational processes can be of endogenous origin (deficiencies in repair pathways, mutations in polymerases, and so on) or can be as well a result of the activity of some exogenic mutagen. For example, signatures 2 and 13 that have an excess of C>T and C>G mutations are associated with the deaminase activity of proteins from the APOBEC family cytidine deaminases [75]. Signature 7 is generated by UV-induced DNA damage with C>T changes, and signature 4 is correlated with smoking (C>A mutations) [75]. Unfortunately, out of more than 50 known signatures, only for a small fraction there is a clear understanding of a process that lies behind it, and even for a smaller fraction of the proposed etymologies was proved experimentally.

One limitation of the method is that certain mutational processes might generate not only single nucleotide substitutions but as well result in short and long indels, copy number changes, or can be better differentiated by using penta-, heptanucleotide or even longer mutational context [77, 78]. Including these features, as well as the strand identity, can increase the number of signatures and characterize them better. Certain progress has been made in this direction, however, the limited number of mutations available makes this more complex and elaborated analysis challenging.

Mutational hotspots

Recurrently mutated positions in cancer, or hotspots, are thought to be under selection and to have a functional impact [79]. For example, one of the most common and well-known cancer mutations are V600E mutation in *BRAF* gene, G12/13 position in *KRAS*, Q61 in *NRAS* that activate oncogenic pathways [80, 81]. Selected hotspots are common in cancers, comprising more than 1 % of all the mutations in cancer [82]

However, many genomic positions frequently mutated in cancers are not selected; here and later we will refer to them as passenger or mutational hotspots. Small mutations or bigger-scale structural variants in such mutational hotspots are caused by specific mutational processes that target local sequences, many of which we have covered in the current section. For example, DNA hairpins that are formed by palindromic sequences are targeted by ABOPEC activity [83]. Multiple secondary structures and non-canonical forms of DNA such as G-quadruplexes, Z- and H-forms of

DNA are associated with an elevated local mutation rate [84]. Mutational hotspots are also associated with the TF binding sites [52, 65, 66, 67].

1.4.4 Computational approaches for identifying selection

One of the great challenges of cancer genomics is identifying genes that are under selection in cancer and promoting tumor evolution. This knowledge is crucial for an understanding of the principles of cancer development, selection pressures constraining and driving tumorigenesis, and identifying cancer vulnerabilities that can be targeted by cancer therapies. Recent advances in sequencing large-scale genomic datasets provided a large number of cancer genomes for studying somatic selection [85, 86, 87, 88]. Some of the proposed methods to study selection in cancer genomes are inspired by the methods used in population and comparative genetics, for example, the dN/dS ratio that compares the rate of nonsynonymous changes with the rate of synonymous changes that is presumably reflecting the rate of neutral evolution [9, 89, 90, 91].

Despite obvious similarities between the evolution of organisms and somatic cells within a body, cancer evolution has particular features that make a straightforward application of methods developed for population genetics inadequate. The major difference relies on the factor of mutational heterogeneity, which leads to many false-positives hits, both for positive and negative selection [92, 93, 94].

A variety of statistical methods have been developed for discovering the genes under selection in tumor genomes. A big proportion of them exploits the principle that the rate of driver mutations should be higher than the rate of passenger mutations.

Direct comparison of mutation rates between genes is not considered to be the best practice because of the big variability of mutation rates in the genome and other confounding factors. While the top-mutated genes in tumors are usually cancer-driving genes, the prediction of cancer genes based on the analysis of recurrent mutations in the tumor without control for confounders gives a lot of false-positive hits. They appear in the analysis because these genes are enriched with mutational hotspots due to certain local features of DNA [92].

Accurate estimation of the rate of passenger mutations for each gene is quite challenging. State-of-the-art methods (*dNdSCV*, *MutSigCV*) are modeling background mutation rates using mutation rate covariates, such as DNA replication timing or gene transcription activity while controlling for Trinucleotide Mutation Spectra (MS96) [92, 9]. This is extremely important to control for the activity of different mutational processes while modeling background mutation rate; it is often done through mutation rate prediction separately for each mutation type in trinucleotide context.

While the inclusion of mutation rate covariates is a working approach, one can not be entirely sure that cell-line data of DNA replication time and gene expression does not change when a cell undergoes a malignant transformation. On the contrary, one of the cancer hallmarks is epigenetic changes that modify histone marks and the pattern of DNA methylation that affects how DNA is packed, and, therefore, the local mutation rate. Probabilistic modeling of mutation counts via the Bayesian framework does not have the limitation of having wrong point estimates and relies fully on the tumor data. The *CBaSe* method estimates a distribution of per-gene mutation probabilities by fitting synonymous mutation counts across the entire set of genes. Implicitly, this form of equation considers all known and unknown covariates of mutation rate, also controlling gene-specific parameters that describe local heterogeneity of mutation rate and synonymous target size [91].

MutPanning method uses the assumption that functionally important mutations in driver genes (although likely to be generated by the same mutational processes that are active in passenger genes) will deviate from the characteristic contexts around passenger mutations. Following this logic, genes that have mutations in unusual contexts are likely to be driver genes [95]. Standard MS96 or extended nucleotide contexts can be used depending on the data. A limitation of the approach is that it cannot be applied to find selected genes in cancer types with low background mutation rates. MutPanning implements a combination of this feature with the signals normally used for the detection of cancer genes [95].

A heuristic approach for discovering and classifying cancer genes was proposed by Vogelstein et al. in 2013 [96]. Genes that have at least 20 % of mutations that are recurrent missense mutations in the same amino

acids were classified as oncogenes. To be classified as a tumor suppressor, at least 20 % of mutations in a gene are required to be inactivation (nonsense or frameshift) mutations. *"20/20 rule''* was shown to work well for known cancer genes, accurately separating tumor suppressor genes from oncogenes [96]. However, the sensitivity of this method to detect novel cancer genes might not be very high, because a large number of mutations is required [96, 97].

A similar idea based on the fact that oncogenes are activated mainly by mutations clustered in specific amino acid residues is exploited in other methods. Mutation clustering was interpreted as a signal of positive selection in several publications [98, 99, 100]. Relying not only on the information of clustering mutations in primary DNA sequence but including data about 3D clustering of mutations in protein further increases the sensitivity of this approach [100, 101]. Another way to enhance the signal is to focus on certain domains of the protein, functional sites (phosphorylation sites, protein-protein interaction surfaces) [102, 103, 104] or search for the bias in clustering of mutations with high functional impact [105].

Analysis of distributions and frequencies of CNA events also give an insight into which genes are essential for tumor development (such genes will have an increased level of amplifications and a decreased rate of deletions) and which genes need to be inactivated (opposite trend) [106, 30, 107]. A major problem is in constructing an accurate model that can predict a baseline level of copy number events under the assumption of neutral evolution. Another difficulty comes from the fact that many CNAs affect several genes at once, which makes it harder to interpret the results.

Interestingly, combining several types of data in one analysis allows one to find a selection by looking at the coordination between copy number events, for example, the length of hemizygous and homozygous deletions in the locus [33]. Interaction between mutation rate or copy number event can also provide an insight into the function of a gene [31, 89, 32, 108].

1.5 Somatic selection

1.5.1 Negative selection in cancer

Various studies showed that negative germline selection on nonsynonymous mutations in different populations prevails over a positive selection, with a very low average dN/dS metric ranging between 0.01 and 0.1 [109, 110, 9]. Analysis of common variation in germline mutations of the human population estimated the dN/dS ratio for nonsense mutations to be 0.08, which indicates that 92 % of nonsense mutations are removed by negative selection [111].

In contrast to germline selection, purifying selection in cancer cells is not playing a global role with a normalized dN/dS ratio being slightly above one [9]. Moreover, similar values of dN/dS are observed in healthy somatic tissues from multiple organs (skin, small intestine, liver, colon, blood).

In frequently mutated cancer genes, dN/dS can reach very high values. Depending on the type of selected mutations, cancer genes are separated into two main groups: TSGs and OGs. In TSGs inactivating mutations (nonsense, frameshift, or missense mutations) are under a strong positive selection. OGs contribute to cancer by acquiring gain-of-function missense mutations that cluster in functionally important gene parts.

Intriguingly, negative selection does not seem to play a big role in cancer cells. Only a tiny fraction of genes were reported to have dN/dS values significantly lower than one [9]. Martincorena et al. (2017) argue that it was possible to miss signals of negative selection if numbers of positively and negatively selected mutations in a given gene are exactly balanced – which is unlikely to happen for a large number of studied genes. The negative selection was extensively searched for in multiple studies that suggested a few gene groups that are essential for the survival of the tumor. In particular, signals of negative selection were enriched in a group of cell-essential genes identified with *in vitro* by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) screens [91]. Importantly, among genes with the strongest signal of negative selection, authors have found a group of genes with a reported oncogenic role in cancer [91]. Consistent

with this finding, truncating mutations in oncogenes were reported to be under negative selection [112, 97]. These genes have to be activated to drive tumorigenesis; inactivation of such genes, on the contrary, inhibits proliferation and is not tolerated in cancer. The apparent lack of negative selection can be explained by various reasons. Detection of depletion of mutation rates is a particularly difficult task because of the data sparsity and requires much larger datasets compared to the number of mutations needed to find a mutational excess [91]. Indeed, cancers that had higher mutational densities exhibited stronger signals of negative selection [91, 113].

Purifying selection can be less effective in the presence of a wild-type allele in the case of heterozygous recessive mutations. In hemizygous genomic regions, weak signals of negative selection were reported for nonsense and high-impact missense mutations and not for low-impact missense and synonymous mutations [89, 9]. Consistent with this, tumors that have undergone WGD show no purifying selection on essential genes, while tumors before genome duplication have a negative selection on essential genes in hemizygous loci [34]. Redundancy of cellular pathways is another reason why deleterious mutations inactivating any gene can, to a certain extent, be tolerated.

Another reason for having weak signals of negative selection in cancer is a lack of recombination, which facilitates the accumulation of weakly deleterious mutations in fit genetic backgrounds with strong cancer driver mutations through hitchhiking and Muller's ratchet mechanisms [11, 9]. Consistent with this, tumors with a lower mutation load (that is, a low number of mutations in a tumor sample) have been shown to have stronger signals of negative selection. The efficacy of purifying selection is decreased in tumors that have a higher mutation rate because of genetic linkage between driver and passenger mutations [11].

Last, but not least, multiple bottlenecks throughout tumor evolution make the factor of genetic drift very substantial. Indeed, random sampling of a small-sized population increases the chance of fixing in populations weakly deleterious mutations that are more numerous compared to the cancer-driver mutations because they are more likely to arise during the random mutational process.

In this study (Section 3.2.3), we will address the question to which extent negative selection can play a role in determining tumor fitness using essential genes from CRISPR screens and essential genes identified with germline variants. Furthermore, we will test the hypothesis that for some genes (specifically, for OGs), both postive and negative selection affect the distribution of mutation rates, thus making signals of selection weaker and harder to detect (Section 3.2.2).

1.5.2 Epistatic interactions in cancer

Epistatic interactions were first reported in 1909 by Bateson as interactions among individual genes [114]. In that context, the main visual mark of epistasis was a distortion of Mendelian ratios of genotypes. Later, the term reappeared in statistical genetics and was used to describe a non-linear additive effect of contributions of single genes on a phenotype [115].

Epistasis is an important feature of complex biological systems. Multiple examples of epistasis can all be summarised as deviations from the additive expected effect of two or more genetic features on the response variable, which is a phenotype of interest [116]. Epistasis in general terms is a property of a fitness landscape – a mapping function from a set of genotypes to fitness [117, 118]. A basic example of non-additive effects of two factors on phenotype is dominant and recessive gene alleles: combining unique effects of each of the alleles does not predict a phenotype produced by a heterozygous combination of them.

Hundreds of genomic and epigenomic alterations arise in cancer genomes with complex combinatorial effects on phenotype. Despite long and extensive research in the field, the combinatorial effects of most genetic alterations are still poorly defined. Understanding how selection changes depending on the genetic background of the cell is one of the open questions. The interplay of genomic alterations with exogenous environmental factors, such as the type of treatment that is prescribed to a patient, and the selection of evolutionary trajectories depending on the activities of mutational processes, can navigate cancer therapies to choose the most effective. Existing cancer-specific treatments reflect the complex interactions of genetic networks that exist within each cancer type.

Given two genetic alterations – for example, mutations in genes A and B that decrease cellular fitness by 30 % and 50 %, respectively, – and knowing their unique and combinatorial effects, we can expect several modes of interaction. If mutations do not interact, the expected fitness of double mutant can be calculated by multiplying the fitnesses of single mutants: $0.7 \times 0.5 = 0.35$. Any other estimate for cellular fitness in a double mutant implies a genetic interaction (Figure 1.1).

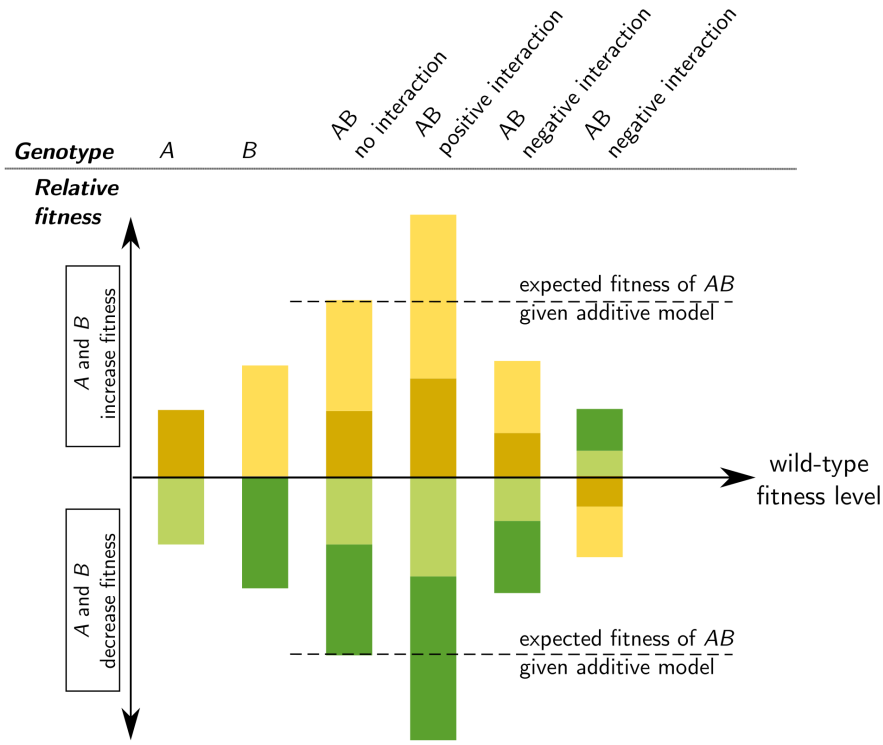

Figure 1.1: Interaction between mutations in genes *A* and *B*. Illustration to the epistatic interaction given the individual fitness effects of mutations in genes A and B.

Genetic interaction can be positive (enhancement, or synergistic effect of two mutations) or negative (suppressive effect). Furthermore, positive interaction for this case can cause a **synthetic lethality** – 100 % decrease in fitness, or cell death. The less strong epistatic effect is **synthetic**

sick – decrease in fitness bigger than 65 % and less than a 100 % (for the particular case of 30 % and 50 % of fitness decrease of genes A and B). **Synthetic viable**, on the contrary, is a type of interaction when two mutations are antagonistic and mask deleterious effects of each other. Double mutant regains maximal fitness. In the case of incomplete **genetic suppression**, the deleterious effect is less profound in the double mutant than expected and can take any value between 0 % and 65 %.

These complex interactions are very important as mechanisms to stabilize the cellular microenvironment and ensure homeostasis. The emergence of new alleles in a cell, most likely non-optimal, will probably not have a strong deleterious effect on fitness and results in functional buffering. However, the next genetic alteration may disrupt the network: the more complex genetic aberrations, the more likely that they are deleterious.

1.5.2.1 Epistatic interactions between mutations

All types of genetic interaction have been reported in cancer cells. Tumors often have mutations in pairs of cancer genes. For example, mutations in *RAS* are often accompanied by the *MYC* activation (genetic enhancement). The toxic effect of mutations in *BRCA1/BRCA2* (arrest in G2/M checkpoint) is suppressed by the inactivation of *TP53*. This way, DNA damage in a cell does not cause apoptosis and can lead to the accumulation of the next driver mutations. Similarly, inactivation mutation in tumor suppressor gene *RB1* helps the cell to suppress a negative effect of mutations in another tumor suppressor gene *VHL* [119].

However, some types of observed relationships can have a second layer and may be not what they seem. For example, the interaction between oncogenic driver mutations *ATM* and *RAS* genes was previously viewed as a genetic enhancement effect. Surprisingly, later studies have shown that oncogenic mutations in *RAS* gene cause deleterious effects of oncogene-induced senescence, which includes *ATM* activation. This deleterious effect is, however, outweighed by the strong selective advantage that receives a cell with a mutated *RAS* gene. The deleterious effect can be mitigated by a second mutation that inactivates the *ATM* gene. Therefore, this is an example of synthetic viability rather than enhancement effect [119].

The synthetic lethal relationship is usually explained by the functional redundancy of two genes. The loss of function of one gene can be compensated by the activity of a second; however, in the absence of the second gene cell cannot perform the needed function and dies. Gene pairs exhibiting such relationships are of great interest, as they suggest targets for targeted therapies in patients with compromised function of one gene. Not only this makes such therapies more effective but also it decreases the harm made to normal cells due to drug toxicity. Only cells with mutations in one of the synthetic lethal partners are susceptible to such therapies.

The first therapy exploiting this mechanism was inhibitors of PARP in BRCA1- or BRCA2-deficient tumors [119]. This example, however, illustrates not the redundancy of cellular functions but the induced essentiality of PARP-polymerase in the perturbed organization of biological networks within cells where *BRCA1* or *BRCA2* gene is inactivated.

Given the toxic result of mutations in some cancer genes, it is reasonable to assume that not only a combination of two mutations needed to rescue cancer phenotype but also the order of the mutations is important. For example, a mutation in *TP53* should precede a mutation in *BRCA1/BRCA2*, and inactivation of *RB1* should happen before *VHL* inactivation.

Not only the presence of two mutations, but also the order of their acquisition can play a role in tumor evolution. Furthermore, the first mutation can decrease the number of evolutionary paths that the cell can take further, as some mutations in this genetic background might be deleterious.

1.5.2.2 Epistatic interactions between mutations and CNAs

A classical example of epistasis in cancer is Knudson's two-hit hypothesis that we have described in the previous section. In short, two alleles of the TSGs need to be inactivated to acquire cancer phenotypes [31]. The exact way how alleles are inactivated in a cell can vary. Two independent somatic point mutations switching off the TSG, although not forbidden, happen rarely. A fraction of the population has a genetic predisposition to cancer as they carry a pre-existing germline mutation in TSG. Only one additional somatic mutation is needed to completely inactivate a TSG in

such people (which is more likely to happen during a person's lifetime), in agreement with the two-hit model.

Successful inactivation of both allele copies can be done through a combination of different somatic alterations: point mutation + deletion of a second allele, point mutation + epimutation that downregulates the transcription of the second copy, or epimutation in the background of deletion of another allele.

Not all TSGs have this type of relationship between selection on the two gene copies. The exact estimations of what proportion of genes operate under the two-hit model are not available and the scale of it might be quite limited. Among the known genes that act as two-hit genes at least in some cancers, are *RB1*, *TP53*, *CDKN2A* genes [120].

In contrast to two-hit genes, hemizygous inactivation mutation in one-hit TSGs is sufficient to acquire a malignant phenotype. One-hit TSGs may be caused by at least two scenarios: dominant-active mechanism of action of inactivating (negative) mutations or **haploinsufficiency** of the gene. Haploinsufficiency refers to the phenomenon when one copy of the wild-type allele present in a cell is not enough to perform its function. Haploinsufficient TSGs might also be sensitive to the increased dosage of the gene (**triploisensitivity**) [120].

Cumulative haploinsufficiency and triplosensitivity were proposed to drive aneuploidy patterns in the cancer genome. It was reported that hemizygously deleted regions are enriched in for "STOP" genes (negative regulators of cell proliferation, largely TSGs) and depleted of "GO" genes (positive regulators of proliferation, OGs and essential genes) [30]. Furthermore, haploinsufficiency and triplosensitivity of gene groups together with their tumorigenic potency can explain the pattern of copy number alterations in the genome at the scale of whole chromosomes and chromosome arms [120].

Reversing the expectations of the researchers, the same study reported that X-chromosome encodes 86 % more TSGs than expected by chance. Oncogenes, on the other hand, were not overrepresented in X-chromosome [120]. X-chromosome is functionally haploid in all sexes because of dosage compensation, which implies that mutations in TSGs can not be masked

by a second allele, thus contributing to cancer. The reasons for the observed pattern are not yet known; however, natural selection did not act to remove TSGs from X-chromosome, as cancer usually develops in post-reproductive age.

Essential genes in hemizygous regions show signs of weak selection [89, 9]. Having more than one wild-type allele of a gene destroys the negative selection pressure on the essential gene, which corresponds to the samples without a gene loss or samples with duplication in a genome [34]. Genetic buffering can be also provided by gene paralogs as demonstrated by De Kegel et al (2019) [121]. Differential activity of the paralog pairs in different cell lines provides a possible explanation why the essentiality of a gene sometimes depends on the genetic context (conditionally essential genes) [121]. Interestingly, paralogs that arose as a consequence of WGD events during the course of human evolution are less likely to be essential compared to the paralogs from small-scale genetic duplications [121].

Focal homozygous deletions are limited in length, and the length depends on the local density of the "GO" genes in the neighborhood. An essential gene located close to a TSG limits the region of homozygous deletion, which gives a selective advantage to a cell and is not affecting to the same extent the length of hemizygous deletions [33].

1.5.2.3 Mutant allele imbalance

The interplay between mutations and copy number alterations, namely deletions, has been extensively studied for a two-hit class of TSGs. GoF mutations in oncogenes were long thought to be dominant-acting and therefore not dependent on the copy number status [122]. Indeed, many oncogenic mutations in proto-oncogenes are heterozygous. Most commonly focally amplified oncogenes are rarely mutated, and frequently mutated oncogenes are rarely focally amplified, suggesting functional redundancy of the two types of GoF alterations.

Few exceptions to this rule exist: the amplification of the mutated version of a gene can additionally increase cellular fitness [123, 124, 125]. A class of two-hit gain genes identified by Park et al. (2021) describes a group of cancer genes (6 OGs and 2 TSGs) for which co-occurrence be-

tween mutations and gains in tumor samples was higher than expected by chance [32]. For these genes, by analyzing allele frequencies in samples, it was also shown that the amplified allelic variant is the mutated one, independently confirming the results.

Selection for dosage increase of oncogenic mutations was carefully studied by Bielski et al. (2018) [108]. The authors reported an allelic imbalance for 45 % of all oncogenic mutations across 69 OGs. Consistent with the current knowledge that focal amplifications are rare for mutated oncogenes, only 12 % of allelic imbalance was caused by focal amplifications. A bigger proportion of allele imbalance (33 %) was due to the loss of the wild-type allele through LOH or CN-LOH [108]. Similarly, single-copy gains accounted for 33 % of allelic imbalance in OGs. In contrast with OGs, allele imbalance in TSGs was mainly explained by loss of the wild-type allele, accounting for the 84 % of allelic imbalance.

Comparing the number of the wild-type and the mutated alleles, it is possible to infer which allele is selected through allele imbalance and whether the allelic imbalance is driven by positive or negative selection on the mutated allele. Few oncogenes in tumors with hemizygous hotspot driver mutations which had experienced WGD event and subsequent single-copy loss were analyzed: *EGFR*, *BRAF*, *KRAS*, *NRAS* and *MAP2K1*. For these genes, in a deletion event, the allele that was kept in a tumor with the mutated one. This result indicated a positive selection pressure for mutant allele imbalance [108].

A broader analysis of selection acting on mutant allele imbalance across all mechanisms of allelic imbalance confirmed this finding. Positive selection for the mutant alleles across tumors varied for different genes, being less common for some genes (15 % for *IDH1*) and more common for other genes. Despite this, in tumors with mutant allele imbalance, positive selection pressure was strong – 74 % mutant allele-specific selection for *IDH1* or almost always (>93 % of cases) for *MTOR* and *MET* genes.

There were, however, few exceptions – genes *U2AF1*, *SF3B1* and *SRSF2* encoding members of spliceosome – showing an opposite trend: retaining a wild-type copy of the oncogene in tumors with hemizygous deletions. It indicates a negative selection pressure that ensures the presence of at least one wild-type copy of haplo-essential genes in the cancer cell. Low-level

genomic gains also preferentially targeted wild-type alleles rather than the mutated allele.

Allelic imbalance in oncogenes favoring mutant alleles was proposed to have two mechanistic explanations [29] based on their functional consequence. The first scenario suggests that the selection favors an increase in the dosage of the mutant allele, and the number of wild-type alleles present in a cell does not affect cellular fitness. Oncogenic mutations that are selected by this mechanism may be not sufficient to initiate a tumor on their own, and additional overexpression is needed to acquire a malignant phenotype. Alternatively, additional drivers in the same gene or other oncogenes can also lead to the increase of cumulative oncogenic potency, as shown for *PIK3CA*, *HER2* and *EGFR* among others [29]. This mechanism should be manifested via mutant allele imbalance genomic gains and amplifications.

The second mechanism proposes that there is an antagonistic relationship between the wild-type allele and the mutant allele, which gives a selective advantage to the subclone that loses the wild-type allele. Stoichiometry or the ratio between the number of wild-type alleles and mutant alleles in a cell is a factor that determines the fitness landscape rather than the number of mutant alleles alone. Change in stoichiometry can be achieved through a loss of the wild-type allele (predominantly) or gains of the mutant allele [29].

Mechanisms of antagonistic and tumor suppressive role of wild-type alleles remain to be unclear. Some experiments suggest that inhibition of the mutant allele can be achieved through the formation of a heterodimer with a protein product of the wild-type allele, which leads to the blocking of the oncogenic potential of the mutant allele. Indeed, dimerization of *KRAS* was shown to be a crucial step in cellular transformation and activation of *MAPK* [126, 127, 128].

1.6 Aim and objectives

Understanding how the interplay between copy number and mutation rate in genes shapes somatic evolution lacks a systematic approach that can

detect and quantify conditional somatic selection across multiple groups of genes. A variety of methods have been proposed and developed to detect and quantify selection in somatic cancer cells [9, 92, 91, 90, 95, 129]. The majority of them, however, are not suitable for detection and quantifying epistatic effects in the somatic selection and estimating to which extent selection might depend on other factors.

Thus, the aim of this thesis was to **develop a statistical methodology** that aims to detect and quantify condition-specific selection and apply it to **analyze epistasis between selection on point mutations and gene copy number status in tumor**.

In particular, we are interested in studying how the selection landscape in known cancer genes changes between tissues, between types of mutation, and between copy number states of the gene. This would provide a theoretical ground to group similar cancer genes by the molecular mechanism of (in)activation, or identify genes under negative selection, which may be therapeutically relevant.

Along with this, we want to classify cancer genes in an unsupervised manner using their spectra of selection across tissues, and epistatic interaction between selected mutations and gene deletion or amplification.

Therefore, the objectives of this thesis are:

- To develop and test a statistical methodology for measuring conditional selection in somatic cells, while rigorously controlling for confounding by mutation rate heterogeneity.
- Apply the method to systematically characterize copy number state-specific selection across >17 000 human cancer genomes, further validating the results in an independent data set(s).
- Test the hypothesis that negative selection is overlooked in tumors:
  - Search for signals of negative selection in genes that were essential in CRISPR screening experiments in cell lines [130] and genes constrained in human populations [131].
  - Test the hypothesis that many mutations in oncogenes may sometimes be under negative selection in cancer and that this

is obscured by simultaneously acting positive selection.

- Derive a new global, systematic classification of cancer genes by mechanisms of (in)activation, inferred from the patterns of selection on missense vs nonsense mutations in different parts of genes (in/out of hotspots for missense mutations and Nonsense-Mediated mRNA Decay (NMD)-detected versus NMD-evading regions for nonsense mutations), and in different copy number states.

Chapter 2

Development and evaluation of the MutMatch method

2.1 Overview

In this chapter, we introduce the developed method MutMatch which is able to identify selection and condition-specific change of selection while controlling for the mutation rate heterogeneity across the genome and mutation signatures. To do so, it estimates the baseline mutation rate from neighboring genes or non-constrained regions within the same genes. Additionally, it controls for mutation rate changes that are driven by differences in trinucleotide composition and activities of different mutation processes (Section 2.2.1).

Additionally, we aim to reduce the effect of all factors that change the rate of mutation accumulation locally. This is achieved by removing from the neighborhood genes that we consider to be mutation rate outliers (Section 2.4.2).

We have also studied the limitations of the method, which requires a high number of mutations to accurately estimate selection effects. Biases

of estimates appear because of the size differences between the group used to model the baseline mutation rate, and the tested genomic region. To overcome this problem, we propose to use randomized selection distributions and correct selection estimates by subtracting the median coefficient of the randomized selection effects (Section 2.3).

The MutMatch method was shown to accurately predict known mutated cancer genes from Cancer Gene Census (CGC) [132], with Area Under the Receiver Operating Characteristics (AUROC) score = 0.77 for non-synonymous mutations. Overall, we believe that the MutMatch method can be applied for addressing many open questions regarding epistatic interactions between mutations and other factors in cancers, for example, conditional selection of genes in tumors showing hypermutated phenotype, or in metastatic tumors.

2.2 Design of the MutMatch method

2.2.1 Mutation rate modeling

To detect the selection signal in a gene, we compare the mutation rate in the gene of interest with the baseline mutation rate. We adopted several approaches where the baseline mutation rate can be modeled with different sets of mutations that are presumably under neutral selection (details in Section 2.2.5).

To describe the variability in mutation counts in a genomic locus, we model raw mutation counts Y using the following Generalized Linear Model (GLM):

$$\log E[Y] = \lambda = \omega t + \sum_i m_i \mu_i + \sum_j z_j \beta_j + \alpha + \log r \qquad (2.1)$$

where t is a dummy variable (target variable) used to distinguish mutations accumulated in a genomic area that is currently being tested for a selection signal and the control group. The effect of selection is quantified by ω.

The coefficient ω (or selection effect) reflects the log-fold change of mutation rates in the tested genomic area (where $t = 1$) compared to the region used to model a baseline mutation rate (where $t = 0$). Positive ω estimate indicates enrichment of mutations and positive selection, while negative ω estimate corresponds to the mutation depletion in the area of interest and negative selection.

In somatic evolution, it is crucial to control for different activities of mutational processes on different oligonucleotides [93, 133, 9]. This is controlled by mutation type variables m and corresponding effects μ.

Other variables can be used to account for differences in mutation rate because of differential usage of mutation processes between cancer types, thus including cancer type as a variable for pan-cancer analysis. Other types of optional variables can be included to control for confounders or batch effects, such as the source of the data. We denote these variables as z and write their corresponding effects as β.

We include the base mutation rate α as the intercept of the model. Finally, to adjust the mutation counts to the maximal number of mutations that can happen, we include the number of nucleotides at risk r as an exposure variable. This allows us to model rates using mutation counts. Practically, it accounts for different DNA length in genomic loci.

The general form of the method allows us to easily extend a model to search for differential or condition-specific selection. Estimation of the effect of exogenous and endogenous factors on selection strength in the cancer genome may lead to the discovery of cancer vulnerabilities, or help with understanding which mechanisms are involved in the activation or inactivation of cancer genes in different patients.

In this work, we focus on the effect of copy number changes on the selection of genes, with minor changes, the same approach can identify selection forces that are specific to the tumors with hypermutation phenotype, cancer-specific selection changes, changes in selection between the stages of the tumor.

When looking for conditional selection signals in genes we use an extended version of the regression that includes a condition variable c encoding the state of the genomic region with respect to the condition (for

example, copy number state of a gene) and the interaction term of the target variable t with the condition variable c.

$$\log E[Y] = \lambda = \omega t + \sum_i m_i \mu_i + \sum_j z_j \beta_j + c\gamma + \delta\, tc + \alpha + \log r \qquad (2.2)$$

It should be noted that here, we present a general formula that does not incorporate all variations. The modular organization of the workflow allows the inclusion of many conditional variables and testing for the interaction of any pair of variables, as well as estimation of higher-order interactions between multiple variables.

2.2.1.1 Different regression models

At this stage, we try to fit different GLMs to describe the observed distribution of mutation counts in a studied region while accounting for user-defined variables.

In the case of an abundance of mutation counts, a negative binomial with the mean λ and the overdispersion parameter δ can be used:

$$Y \sim \mathrm{NB}(\lambda, \delta) \qquad (2.3)$$

When the number of mutations is limited, Poisson distribution with the mean λ is better suited.

$$Y \sim \mathrm{Poisson}(\lambda) \qquad (2.4)$$

However, detecting statistical interactions requires a large number of mutations for regression to be able to fit the parameters. Cancer mutation datasets that are available for analysis often do not have enough mutations because of controlling for many parameters, which leads to data sparsity. Low mutation counts and uneven distribution of DNA lengths result in errors in the estimation of selection.

Moreover, the bias is enhanced due to the problem of perfect separation when a linear combination of the predictors is perfectly predictive of the outcome. In this case, absolute values of regression estimates are pushed very far which makes them unlikely and untrustable [134, 135]. To reduce this problem, we use an extension of the classical GLM by adding a weakly informative prior distribution of regression coefficients (e.g. selection estimates) [136]. This assumes that the majority of effects are centered at zero and that there is a low probability that there will be a change greater than 5 on a logarithmic scale. The parameters for independent prior distributions (for the regression intercept and predictors) can be changed in the implementation of this approach in the *bayesglm* function from *arm* package in R [136].

2.2.2 Input data

The method is implemented in a way that allows to easily substitute variables that a user wants to control for, input mutations, and annotation of samples. The MutMatch program requires a set of mandatory files as input:

1. **Mutation data file**. Mutations should be formatted in 1-base, in VCF-like format with columns: CHROM, POS, ALT, REF, SAMPLE. Only Single Nucleotide Variants (SNVs), in particular substitutions, are considered by the MutMatch method.

2. **Genome-wide sample annotation**. This file should include the information about each variable that a user wants to either include in the regression model or to be using it to run fit regression separately for each value of the annotation. This annotation should not change depending on the position in a genome. An example would be a cancer type.

3. **Gene-specific annotation**. Some variables (for example, a condition variable c that distinguishes genes with a different copy number status in a cell) have gene-specific values and are not consistent within one sample. A gene-specific annotation should be provided in a separate file. The file should contain columns SAMPLE and

HUGO Gene Nomenclature Committee (HGNC) (gene name). A dummy variable CNA in this example should denote the state of the gene in each tumor sample.

4. **Variables to include in the regression.** File with a list of variables defining the regression model. This should always include:

   - the target variable t that distinguishes mutations between tested and control loci (named $isTarget$ in the code implementation).
   - the mutation type variable m (named $MutationType$ in a file).

   The remaining variables are user-defined. If sample annotation contains a variable that is not used in this file, this is interpreted as a signal that regression should be fitted separately for each level of this variable (or a combination of variables in case more than one variable is missing). The inclusion of condition-specific selection is made by adding a term $variable_1 : variable_2$. For example, if sample annotation provides Copy Number Alteration (CNA) status of a gene, conditional selection associated with a change of copy number status is denoted with $isTarget : CNA$.

5. **HGNCsymbol.** A name of a gene for which selection estimates should be calculated. It is possible to use an alias for a gene name.

6. **Clustering level for mutation spectra.** It ranges from 0 to 96 where 1 means no separating mutation types and 96 stratifying mutations into 96 mutation types with controlling of 1 nucleotide upstream and downstream. 0 stands for relative pentanucleotide frequency matching procedure, described in Section 2.2.4.2.

   Briefly, we account for the different pentanucleotide compositions of tested and control regions by subsampling regions and proportionally decreasing the number of mutations in them until the composition between genomic regions is the same. Therefore, the differences in mutation rate between the tested and control regions cannot be due to the differential contribution of mutational processes rather than selection.

2.2.3 Output interpretation

Coefficient estimates ω,μ_i, β_i, γ, and δ are obtained after fitting a regression model searching for conditional selection. Each estimate tells by how much the mutation rate changes when a corresponding variable changes its value from the basal to the tested. Some coefficients, such as μ_i and β_i, are disregarded in further analyses. Coefficients ω and δ, on the other hand, are the key estimates that reflect the selection forces.

For example, if the coefficient ω associated with the t variable is equal to 0.5, this means that the mutation rate is increased by $e^{0.5} = 1.649$ in the tested genomic area compared to the area used to model a baseline mutation rate, given that all the other variables are in the base level. In other words, $\sim$65 % of mutations in the tested genomic locus are driver mutations and are positively selected. On the contrary, $\omega = -0.5$ means that $\sim$40 % ($e^{-0.5} = 0.607$) of mutations were lost due to the negative selection (Figure S8.1).

A coefficient estimate δ that is associated with an interaction between a selection variable t and a condition variable c shows the effect of the condition change on the selection. This is a deviation from the expected behavior in the assumption that the selection variable t and the condition variable c have an independent effect on the mutation rate.

2.2.4 Mutation spectra control

The nucleotide context of a DNA sequence plays a big role in the process of acquiring DNA damage that leads to mutations. It is widely accepted to define mutational processes by relative proportions of mutated bases in trinucleotide context [75], although some works are using an expanded sequence context and control for penta- or heptanucleotide mutation context [78, 50].

2.2.4.1 Control for trinucleotide mutation spectra

To control for the context-dependent mutagenesis, the MutMatch method stratifies all the mutations into different classes according to the trinu-

cleotide context of the point mutation (Trinucleotide Mutation Spectra (MS96)). A combination of 6 main substitution classes (C>A, C>G, C>T, T>A, T>C, T>G) collapsed strand-symmetrically with upstream and downstream adjacent nucleotides results in $6 \times 16 = 96$ mutational contexts. These classes are included in a regression model as dummy variables.

2.2.4.2 Matching of relative mutation frequencies

However, when accounting for the trinucleotide context, not all mutation types may have a sufficient number of mutations to properly estimate selection. To overcome this challenge, specific algorithms have to be utilized.

One approach is to match the relative pentanucleotide composition between the central gene and its neighbors. To do that we first calculate the proportions of different pentanucleotides within each region, leaving for the downstream analysis only those pentanucleotides that are present in both groups. Next, the target frequencies of each pentanucleotide are determined depending on how much they differ between the central and neighboring genes. If a particular pentanucleotide is more frequent in the neighboring area (but no more than 3 times more frequent), then the target frequency will be taken as in the central genes, and otherwise if not.

After that, we sample sites from the genome to comply with the target proportions of each pentanucleotide. Simultaneously, the number of mutations in these pentanucleotide contexts is proportionally decreased and then sampled using a multinomial distribution (Hansen-Hurwitz method, *UPMultinomial* function from the *sampling* package in R) [137]. Finally, all mutation counts are summarized in one number by taking a sum, as well as all nucleotides at risk.

This approach produces a set of summarized mutation counts and summarized length of each region that are representing the expected number of mutations in case neighboring genes and central genes had the same pentanucleotide composition. This procedure is repeated fifty times for excluding the influence of random factors in the multinomial sampling of sites. Later, summary tables are used for estimation of mutation enrichment in a gene of interest.

2.2.4.3 Clustering of mutation types

An alternative way to decrease the number of estimated parameters in regression is clustering similar mutation types. The similarity of mutation types of evaluated based on their relative frequencies across samples in the non-coding parts of the genome using Whole Genome Sequencing (WGS) data. We clustered 96 mutation types (separating MS96) using hierarchical clustering.

This makes it possible to choose the number of groups of mutation types that will be controlled in a regression, ranging from 1 to 96. For example, cutting the hierarchical clustering at level 1 will not differentiate between mutations, cutting the clustering at level 2 will separate NCG>T mutations from all other types (NDH>N, where D is A, G or T). Clustering level of 96 means that MS96 is treated as a variable (described in 2.2.4.1).

Regardless of the chosen parameter, the algorithm separates mutations from six main mutation classes. Additional mutation types will be extracted from the data based on the chosen level of cutting the clustering. In the example of cutting the hierarchical clustering at level 1, seven mutation types will be used to stratify the data: NCG>T, NCH>T (H is A, C or T), C>A, C>G, T>A, T>C, T>G. Similarly, 7 substitution types (A>T, A>C, A>G, C>A, C>T, C>G, or CpG>N) were used in the paper of Zapata et al. (2018) [133].

2.2.5 Baseline mutation rate models

Estimation of a selection strength for a gene is performed using gene annotation in the reference human genome assembly GRCh37 (hg19). Only mutations located in exonic parts of the most expressed transcript variant are used [138]. To account for the selection in splice sites at the 5' and 3' ends of introns, the coordinate of each exon are extended by 5 nucleotides upstream and downstream inside the intronic region.

2.2.5.1 Neighboring genes baseline

The mutation rate baseline is mainly shaped by the DNA replication time domains and/or other features varying on the same scale of 100 kb – 1 Mb such as Topologically Associating Domains (TADs) [52, 56, 57]. Having this in mind, it is reasonable to assume that the background mutation activity in a close gene neighborhood is very similar, with minor differences driven by changes in oligonucleotide composition and selection forces. On the other hand, the majority of the genome is evolving under neutral selection, and a randomly chosen gene will most likely be not selected [9, 139]. Therefore, exonic parts of genes in the close neighborhood (below 1 Mb) can be used to model a background mutation rate [20].

The neighborhood size can, in principle, be used as a parameter with the default value set up to be 500 kb (0.5 Mb) both upstream and downstream [20]. The choice of the neighborhood length is substantiated in Section 2.4.1. In gene-poor areas, if no gene is located in such a neighborhood, the neighborhood is extended until at least one gene on each side is found (excluding the case when one side is limited by a chromosome end).

Similarly, as for the central gene, only mutations located in the exonic parts with 5-nucleotide extension for splice sites of the most expressed transcripts for neighboring genes are used [138]. If a gene-specific annotation is used in the analysis, a neighboring gene is excluded from the analysis if the annotation is not matching between the central gene and the neighboring gene. For instance, if one copy of a central gene is deleted in a sample but deletion is focal and only a part of the genes in the neighborhood are deleted, only those that are deleted will be further used to model a background mutation rate. Given the trivial effect of the copy number changes on the DNA amounts and, consequentially, on the number of mutations, this way we ensure that the modeled mutation rate derived from gene neighborhoods is not confounded by the copy number state changes.

2.2.5.2 Effectively synonymous exonic sites using CADD genomic score

Some individual mutations have a very strong functional impact on the transcript or the protein level (driver mutations or deleterious mutations), while other mutations may have no phenotypic effect. This is often the case, but not always [20], when a mutation is synonymous or near-synonymous, leading to a retaining the same amino acid or to a change to an amino acid that is similar in its physicochemical properties. A drastic change in amino acid sequence can have no effect or a very small effect when it is located in a non-conserved part of a gene, supposedly having no functional involvement in gene activity. Since the phenotypic effect of such mutations is weak or absent, such mutations are not subject to a selection in somatic evolution and can be used to model a background mutation rate [140, 105, 141, 133].

We use Combined Annotation-Dependent Depletion (CADD) genomic score to distinguish between genomic regions where changes are likely to have a functional impact and those where mutations have no deleterious effect [142]. Details are in Section 7.5.

2.2.5.3 Other mutation rate baselines

In addition to using neighboring genes, the MutMatch framework allows using other baselines for modeling background mutation rates. These include:

- **Invex-like method**. Introns and flanking non-coding parts of a gene (including intronic parts of neighboring genes if a gene has no introns) are used to estimate an expected number of mutations in exonic parts, similar to Hodis et al. (2012) [62].
- **Trans-neighbors**. Genes are grouped into clusters based on the mutation rate level in their non-coding parts. For every gene from a cluster, the selection is estimated by comparing the mutation rate in the exonic parts of a tested gene with intronic parts of all genes in the same mutation rate cluster, similar to Supek et al. (2014) [20].

- **dN/dS**. Mutations are classified by their phenotypic effect: the rate of nonsynonymous changes is compared with the rate of synonymous mutations [9, 91, 90].
- **Other functional impact scores (e.g. REVEL)**. Mutations are separated into groups based on additional ranking scores, similar to CADD baseline. For example, a Rare Exome Variant Ensemble Learner (REVEL) score that predicts the pathogenicity of mutations can be used to compare the rate of pathogenic mutations with the rate of proxy-neutral mutations [143].

2.2.6 Filtering of genomic sites

2.2.6.1 Mutation rate outlier genes

Mutation rate heterogeneity at a smaller level (generated by mutational hotspots on the gene and subgene scale) can confound our framework and lead to the detection of differences in mutation rates between control and test sequences that can be interpreted as selection [92, 93]. To avoid such situations and decrease the false-positive rate we exclude "outlier" genes from the neighborhood that might be enriched or depleted with mutational hotspots compared to the level in the gene of interest. Details are described in Section 2.4.2.

2.2.6.2 Regions with uncertainty in mapping and conversion-unstable positions

We also filtered regions to avoid errors in mutation-calling or mapping or reads. For each genomic region, only nucleotides that are mappable according to the CRG75 Alignability track were considered. Additionally, we removed positions that were unstable when converting between GRCh37 and GRCh38 (conversion-unstable positions) [144, 145].

2.2.6.3 User-defined genomic regions

In specific cases, one can be interested in measuring selection in particular parts of a gene. For instance, selection acting on nonsense mutations in NMD-detected regions, on all nonsynonymous mutations inside, inside or outside of hotspots (details on the used filters see in Section 7.6) [112, 82]. It is possible to specify a set of genomic coordinates that should be used to collect mutations exclusively or, on the contrary, excluded from the analysis.

2.3 Correction of regression estimates

2.3.1 Data sparsity leads to estimation biases

Controlling for confounding factors is a mandatory task in searching somatic evolution due to the high heterogeneity of mutation rates across the genome, copy number states, and other factors [9, 92, 89]. However, stratifying by a large number of such variables simultaneously, using our MutMatch regression framework leads to a small number of mutations in each category.

In this case, an important role in estimating regression parameters play additional factors such as relative lengths of tested and control genomic regions (i.e. the central gene, and the neighboring genes). The combination of low mutation counts and unequal sizes of these groups leads to the inaccurate estimation of the regression parameters. The estimates may be biased to the negative side or positive side, depending on which group (control or test) has more chances of acquiring a mutation solely due to the DNA length.

Let us illustrate it using an example when a selection in gene G is estimated using the average mutation rate modeled with N neighboring genes V_i, $i \in \{1, \cdots, N\}$.

If a gene G is selected with the selection strength equal to ω, it means that given a sufficient number of mutations they will be observed e^{ω} times more frequently in the gene G than in genes V_i. Assuming there is no selec-

tion ($\omega = 0$), the mutation counts will be distributed randomly generating the same mutation rate in the control and the tested groups. However, if the number of mutations is low, a mutation is more likely to be observed in the group of neighboring genes V_i. The reason is that the total amount of DNA in these genes is bigger than in a gene G which results in an underestimation of the selection. An estimate of the selection strength $\hat{\omega}$ shows a false-negative selection.

Controlling for variables that have "typical" values and values that are less common in the population further leads to disproportionate distributions of DNA sequences between cohorts. For example, if one wants to separate samples by the copy number state of a gene, it will result that only a small fraction of samples will have a non-diploid state. This leads to the subsequent misestimation of the conditional selection (Figure 2.1).

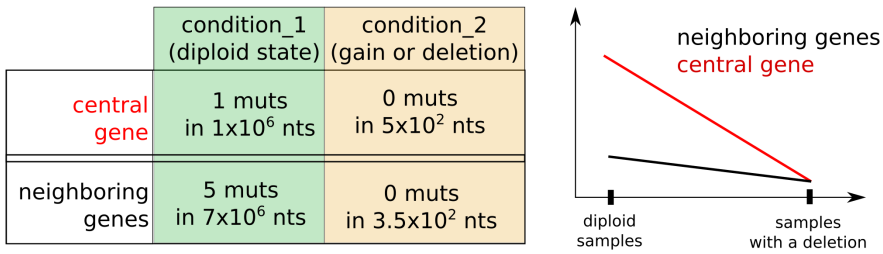

Figure 2.1: Bias in estimating regression coefficients. Low number of mutations leads to high noise-to-signal ratio, inaccuracy of estimation of mutation rate, and bias in regression estimates.

Increasing the number of mutations in regression leads to removing the bias. Interestingly, there is a non-monotonous relationship between the number of mutations and the estimation error using the particular implementation of the Bayesian regression method that we applied (Figure 2.2) [136].

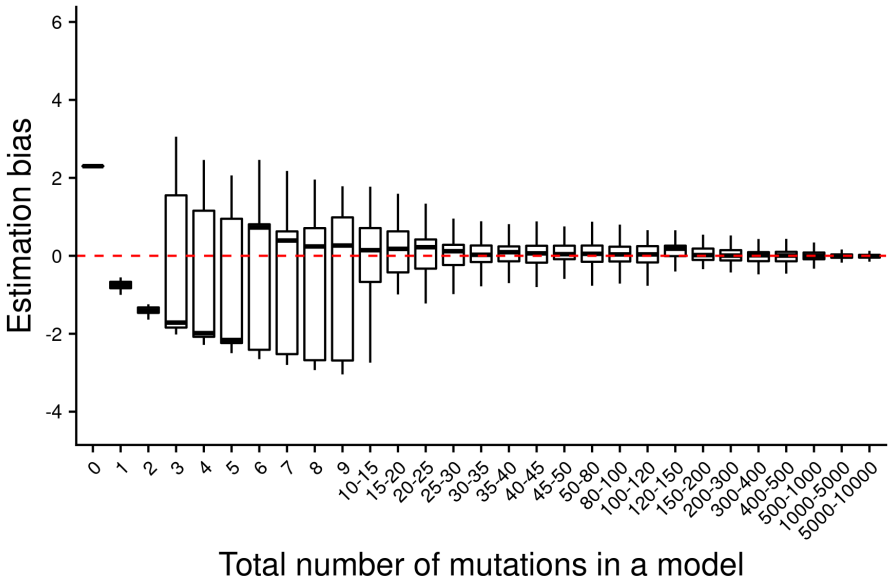

Figure 2.2: Estimation error of regression. The data was simulated using a neutral selection model ($\omega = 0$, red dotted line) with $CGF = 10$. Estimated coefficients $\hat{\omega}$ are shown as a function of the total mutation number in a model (the sum of the mutation counts in the tested and control groups).

When zero or only a few mutations are present, each of them changes the mutation rate in the tested and control groups in a different proportion. The group with a longer DNA sequence (usually, this is the neighboring genes group in the case of the neighbors' method) will have a proportionally lower mutation rate, thus regression will return a positive coefficient. Next mutations appearing in the region will more likely target the neighboring group of genes with the probability proportional to the control group factor (CGF), that is, how much more DNA sequence is in the control group (in this case, in the neighboring genes V_i) compared to the tested group (central gene G). This will create a negative estimation error.

$$CGF = \frac{L_c}{L_t}, \tag{2.5}$$

where L_c is the total length of the DNA sequence in the control group (neighboring genes V_i in this example), and L_t is the length of the DNA sequence in the tested group (central gene G in this example).

Therefore, although these estimates have large standard errors (Figure S8.2), they introduce a bias in selection estimates because the length of the control group is not equal to the length of the tested group.

The described relation depends on the relative DNA lengths of the tested and control groups; the bias is negative when the control group is bigger than the tested ($CGF > 1$) and positive in the opposite scenario ($CGF < 1$). For example, using neighboring genes to model the baseline mutation rate leads to underestimation of selection effects, while coefficients from the CADD baseline, which has longer sequences in the tested group tend to be overestimated.

The extent of misestimation of regression coefficients strongly depends on the total number of mutations (which is influenced by background mutation rate) and the extent of CGF. High absolute CGF values cause a larger bias and require a bigger number of mutations in regression for accurate ω estimation (Figure 2.3).

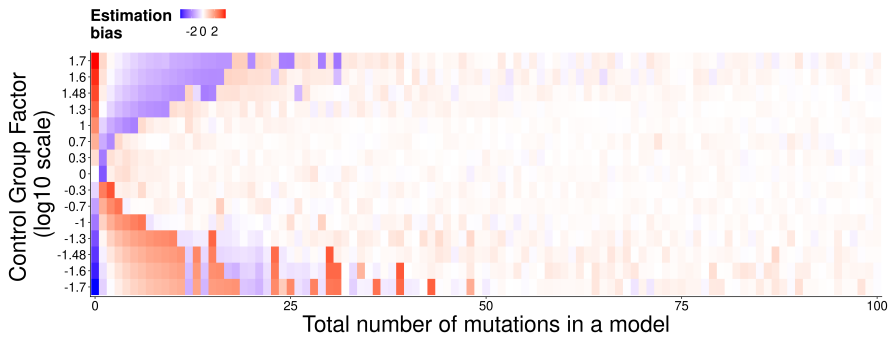

Figure 2.3: Estimation error depends on the CGF. Bigger difference between the sizes of the tested and control groups requires more mutations to accurately estimate regression coefficients.

Taking this into account, one cannot consider the raw coefficients that regression produces as estimates of selection. Therefore, a procedure to correct these biases due to low mutation counts was needed.

2.3.2 Randomization approach to debias selection estimates

To remove the bias in selection estimates, we generate a null distribution of selection estimates for ω and δ coefficients using a randomization procedure. The parameters of the null distribution are used to check if the estimate from the actual data is effectively different from the median of the distribution of random data (Figure 2.4).

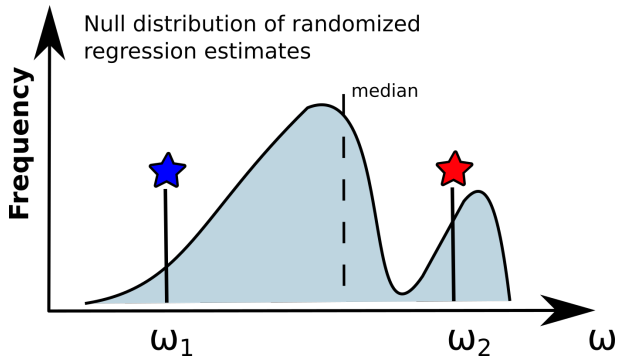

Figure 2.4: Estimation of the real selection strength. The deviation between the real estimates and the median of the null distribution of randomized estimates is interpreted as the corrected selection effect. The ω_1 is effectively negative (is lower than the median of the null distribution) and the ω_2 is effectively positive (is greater than the median of the null distribution). The absolute values of a null distribution are not affecting the procedure.

To generate a null distribution of regression estimates we first calculate the total number of mutations M_{total} in the genomic region (including the tested locus and in the control genomic region) while accounting for the mutation type and other variables (for instance, cancer type). Then, mutations are shuffled between the tested group and the baseline group randomly, with the probability of receiving a mutation proportional to the

DNA length of each region. Each generated table is then used to estimate selection effects under randomness. This procedure is repeated 50 times. The parameters of the distribution of the randomized selection effects are then used as described above (Figure 2.5).

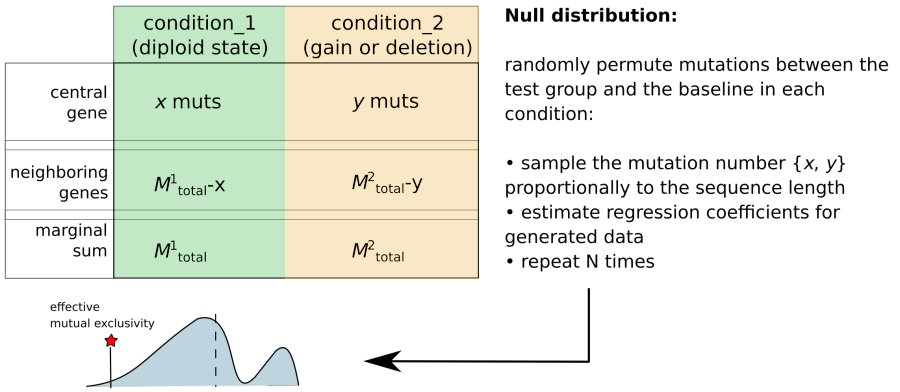

Figure 2.5: Randomization procedure to correct biases in the data. Repeated $n = 50$ times to generate a null distribution of regression estimates of selection (ω and δ).

We compared uncorrected selection estimates with the corrected ones (Figure 2.6) across different copy number states for random genes (neutral baseline), cancer genes, and essential genes. For random genes, the distribution of all selection estimates after correction for the bias was centered at 0, as expected. On the contrary, there was a negative bias of the random genes for uncorrected selection estimates.

Selection estimates corrected for the bias can be used for further analysis.

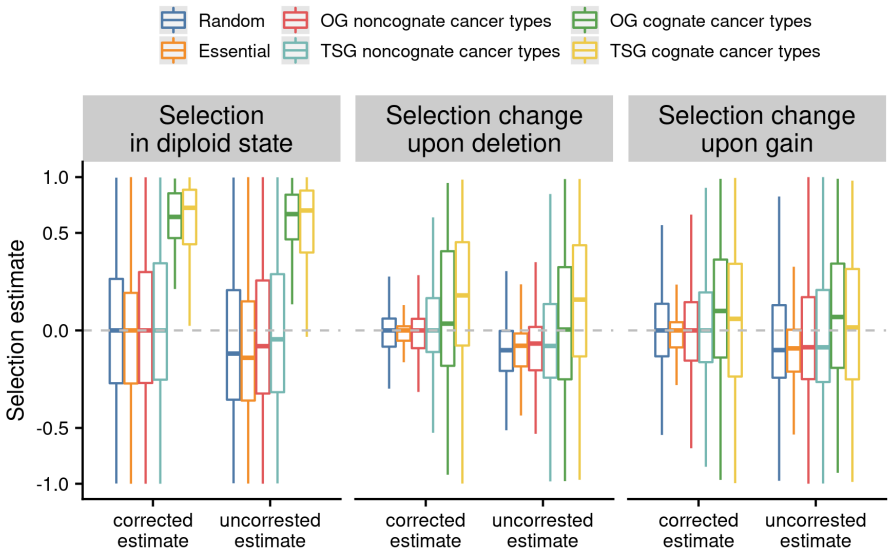

Figure 2.6: Selection estimates before and after correction. Estimates for selection on nonsynonymous mutations in a diploid state and a conditional selection in a hemizygous state or cells with copy number gain. Uncorrected estimates tend to be more negative than expected, based on the assumption that random genes should be under neutral selection. Bias-corrected estimates for random genes are centered at zero for all copy number states, as expected.

2.4 Benchmarking and evaluation

2.4.1 Size of the neighborhood

We wish to determine the appropriate size of the neighborhood to model the expected mutation rate in a genomic locus. While the neighborhood should not be too big to be able to reflect local changes in a mutation rate, it should have enough genes and acquired mutations to precisely model mutation rates.

Given that the average human gene length is about 2.7×10^4 nucleotides (Figure S8.3) and the genome length is 3.2×10^9, one gene can be found

on average every 1.2×10^5 nucleotides [146]. This distance gives us an idea of the minimal neighborhood length that is reasonable to have: it cannot be smaller than 100 kb.

To determine the optimal neighborhood length, we benchmarked different neighborhood sizes against the test's ability to distinguish between the positive selection of known cancer genes [132] in cognate cancer types and neutral selection in random genes (Figure 2.7).

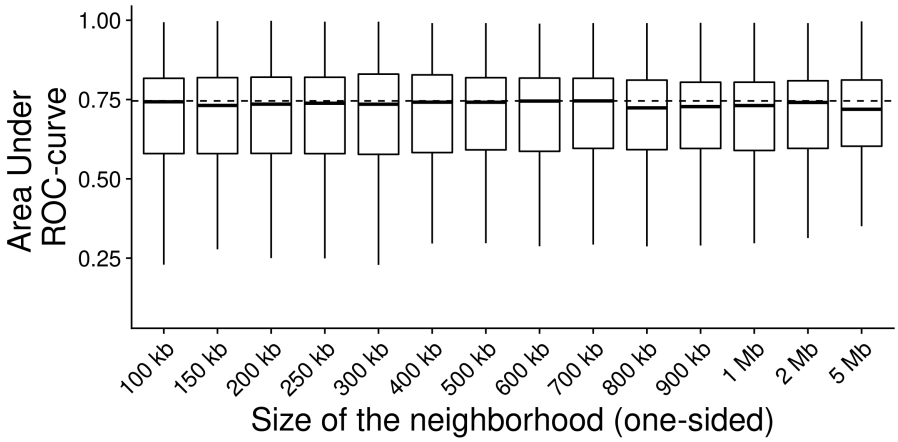

Figure 2.7: AUROC scores for each neighborhood size. The separation of debiased selection effects ω was used to predict mutated Census cancer genes in cognate cancers versus random genes. One dot corresponds to the AUROC score in one cancer type. The dotted line is the maximal median AUROC score per group in tested sizes (corresponds to 700 kb).

An additional factor in determining the size of the neighborhood was the average length of deletions and amplification in tumor samples (Figure 2.8). Although the median length of heterozygous deletions and low-level gains was above 1 Mb, up to 25 % of low-level copy number changes were shorter than 0.5 Mb.

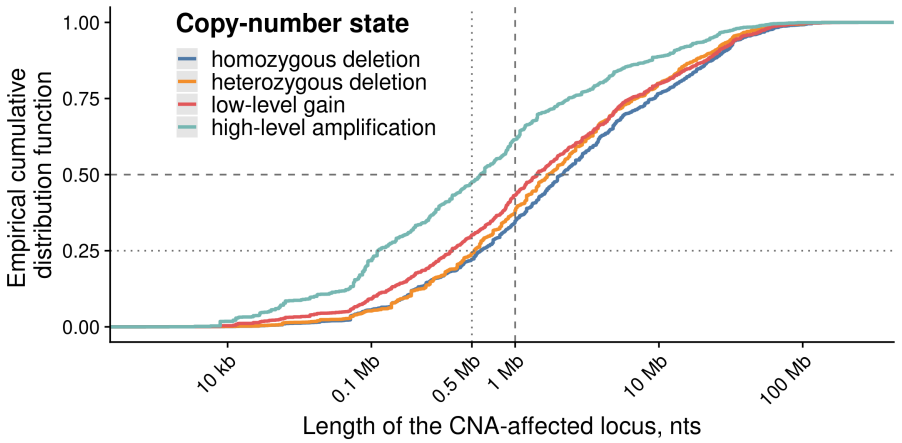

Figure 2.8: Cumulative distribution of lengths of different copy number states in a genome. Based on the TCGA copy number annotation of samples derived from SNP-array data using GISTIC [106].

2.4.2 Mutation rate outliers in neighboring genes

Sometimes neighboring genes have a different mutation profile than the central gene, implying that neighboring genes do not provide an adequate baseline to model a mutation rate. This phenomenon can occur when neighboring genes are selected, have different transcription levels, have different transcription factor sites abundance, or have different sets of chromatin marks [147, 148, 149, 9, 92].

To overcome this challenge, we exclude from the neighborhoods those genes that we consider outliers in mutation rate. For this, we compare the intronic mutation rate between the central gene and every gene in the neighborhood while accounting for trinucleotide mutation spectra with a matching procedure. Preprocessing of the data and calculation of the outlier scores S_{outlier} was performed by Marina Salvadores [150].

In this procedure, a 20 kb sequence around the central position of the gene is taken excluding exons, CRG75 unalignable regions, CTCF binding sites, ETS TF, and APOBEC hairpins since they were shown to influence mutation rates) [52]. If the sequence remaining after this filtering step was

shorter than 5 kb, the transcript was discarded. In the same way, transcripts that had no mutations were discarded. In total 18 227 transcripts passed this filter.

For each gene, its sequence was processed so that the difference between window trinucleotide composition and the reference trinucleotide composition (whole genome trinucleotide frequencies) was minimized. This was achieved by removing the trinucleotide positions in an iterative manner (10 000 iterations) for a gene-trinucleotide combination with the biggest deviation in proportion from the reference proportion. At the end of the procedure, the difference in relative trinucleotide composition of genes and the reference trinucleotide frequencies was not higher than 0.035 – which was considered to be negligible.

Next, the total number of mutations acquired in each gene was compared to generate a matrix of differences in mutation rate between genes x and y. S_{outlier} measures how well the gene's mutation rate is representative of the genomic neighborhood it resides in and was calculated using the formula 2.6.

$$S_{\text{outlier}} = \log \frac{M_x/L_x}{M_y/L_y}, \tag{2.6}$$

where M is the number of mutations observed in a gene, and L is the gene length.

Additionally, 11 transcripts without mutations were excluded after a matching procedure to avoid infinite values of S_{outlier}. In the end, an S_{outlier} value was calculated for 18 214 transcripts.

The distribution of this score depending on the distance between same-chromosome transcripts is shown in Figure 2.9.

We tested different thresholds using the S_{outlier} to exclude mutation rate outliers from the neighborhood: after applying the strictest threshold (0.2), on average, 44% of the genes are excluded from the neighborhood (Figure 2.10). The positive selection benchmark was used to find the optimal value of S_{outlier} to separate neighboring genes with similar mutation rates from those that were not. We found that the value of 0.2), which corresponds to 20 % of variability in the mutation rate is the optimal

threshold.

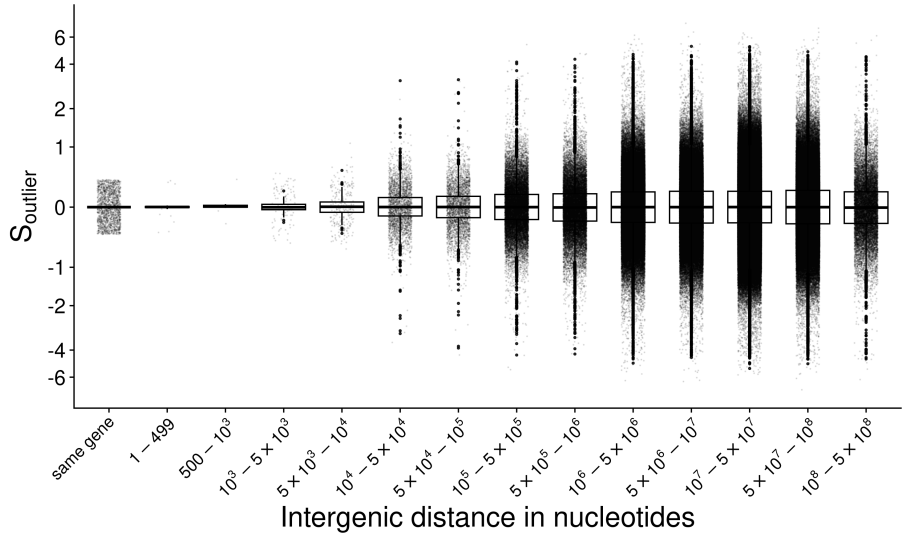

Figure 2.9: Cumulative distribution of lengths of different copy number states in a genome. The length of CNAs was calculates using the The Cancer Genome Atlas Program (TCGA) annotation of tumor samples.

Finally, with the chosen neighborhood size of 500 kb and $S_{\text{outlier}} = 0.2$, we estimated that CGF for a set of cancer genes was about 4.8, meaning the size of the control group to draw the expected mutation rate was 4.8 times bigger than the size of a tested gene. According to our simulations, at least 10 mutations should be in the model to avoid biases when $CGF = 5$ (Figure S8.4).

Last, we have tested the ability of the MutMatch method to distinguish between neutrally evolving genes (random genes) and positively selected genes (known cancer genes in cognate cancers [132]) using different sets of mutations. The difference in selection strength between groups was the strongest when considering the selection signal derived from all nonsynonymous mutations with AUROC score = 0.77 (Figure S8.5).

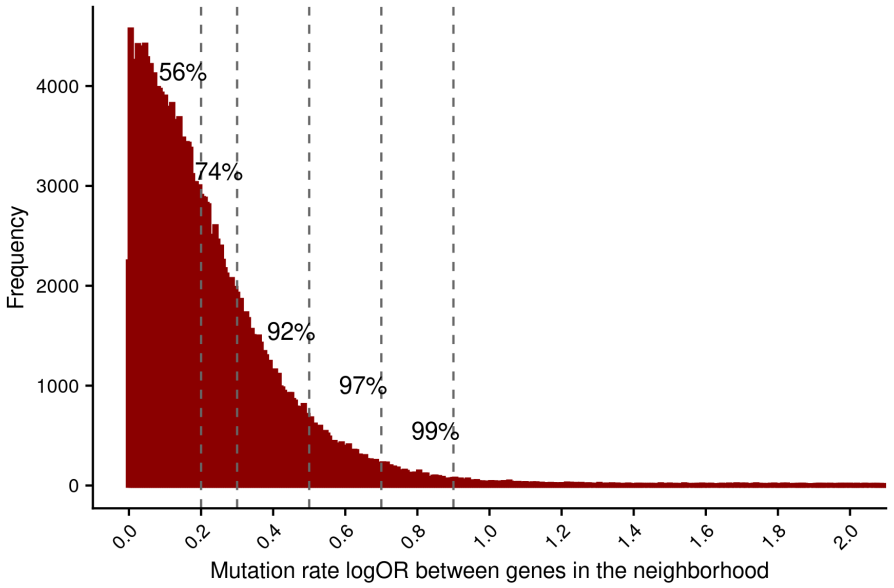

Figure 2.10: The distribution of S_{outlier} between genes not further than 500 kb apart. The dotted lines show the tested thresholds (0.2, 0.3, 0.5, 0.7, 0.9); the proportion of the genes in the neighborhood after filtering out outliers defined by each threshold is shown.

Overall, our results suggest that the MutMatch method is a flexible approach that can detect selection in human cancers. Although it has certain limitations, their effect can be reduced by using a randomization approach, filtering out mutation rate outlier genes, and via other techniques.

Based on the results of this chapter, we conclude that the MutMatch method can be used to test for changes in selection strength that appear under certain conditions. In the following two chapters, we will focus on the signals of selection in tumors that can be used to find cancer vulnerabilities, as well as the epistatic relationship between mutations and copy number changes in the human genome.

Chapter 3

Oncogenes and essential genes are under purifying selection in human tumors

3.1 Overview

In this chapter, we analyze selection estimates obtained by MutMatch to genomic mutation and copy number variation data. We estimated selection acting on a set of cell-essential genes [130], Tumor Suppressor Genes (TSGs) and Oncogenes (OGs) from the Cancer Gene Census list, and a set of random genes used as a control group. More specifically, these selection estimates for each gene were obtained for each cancer type, separately in the diploid state and across different copy number states. Finally, we estimated pan-cancer selection by adjusting for cancer identity.

The main results reported in this chapter include:

- Opposing selection forces shape mutational profile in oncogenes in tumors: positive selection in hotspots bearing Gain-of-Function (GoF) mutations, and negative selection in nonsense mutations and non-synonymous mutations outside of hotspots.

- Mixture of positive and negative selection leads to underestimation of genes positively selected in cancers, as well as obscuring signals of negative selection.
- Focusing on regions that are putatively under positive selection increases the sensitivity of cancer gene discovery. Negative selection on oncogenes is prevalent and may be exploited for therapeutic purposes.
- Cell-essential genes are under weak but nonetheless measurable negative selection in tumors; the signal becomes more evident with larger sample sizes which increases the number of mutations.
- Genes essential at the population level do not exhibit signatures of negative selection.

3.2 Results

3.2.1 Negative and positive oncogene selection shape tumor evolution

We first looked for signs of selection across different genes (random, core essential genes, cancer genes) in a diploid state (Figure 3.1). To accomplish this, we included copy number data as a covariate in the model and fitted a model with the conditional selection term (as detailed in Methods of this chapter, equation 3.2). Coefficients that were associated with the gene in the diploid state (the reference state in the analysis) were then analyzed.

For each cancer gene, we divided cancer types into two categories: those where a gene was positively selected (cognate cancer types) and those where it was not selected (noncognate cancer types). Positive selection to separate cognate and non-cognate cancer types was estimated per gene across all copy number states (as detailed in Section 7.3).

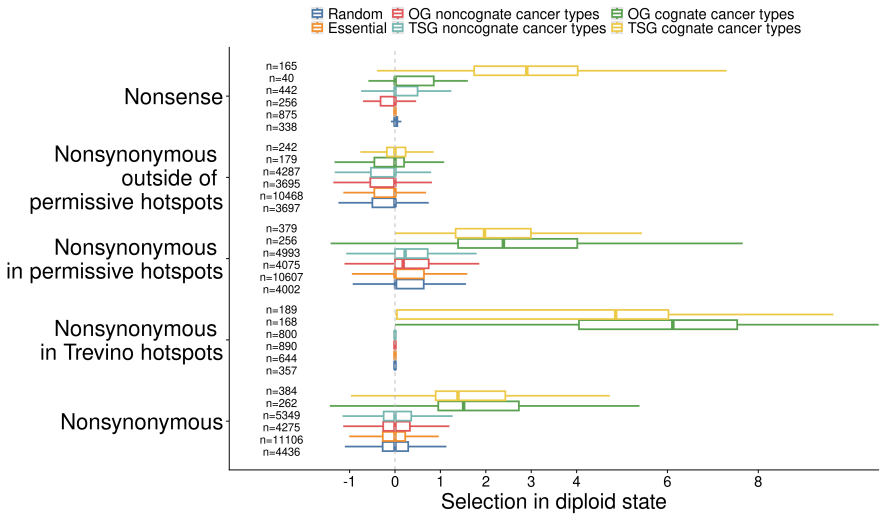

Figure 3.1: Selection in diploid state. Debiased selection estimates for the neutral copy number status of a gene obtained on the discovery cohort. The number of gene-tumor pairs used to produce each of the box plots is written in the left part of the plot.

Expectedly, we found a strong positive selection of nonsynonymous mutations in cancer genes in cognate cancer types but not in other gene groups. An even stronger positive selection was observed in hotspot sites, with positive selection acting in strict hotspots defined by Trevino et al. (2020) [82] and weaker selection in permissive hotspots detected in this work (Section 7.6.1). Selection in hotspots of noncognate cancer types was significantly lower than in cognate cancer types. However, selection estimates were higher than in random genes and essential genes. Likely, some gene-tissue combinations from this group are selected positively, although selection pressure is weaker than in the group of cognate cancer types.

Nonsense mutations in cancer genes in cognate cancer types were selected positively. TSGs in cognate cancer types had a strong positive selection on nonsense mutations, while the weaker (but still positive) selection was also shown for TSGs in noncognate cancer types. Similarly, there was an enrichment of nonsense mutations in some cognate cancer

types for OGs. While unexpected, this enrichment could be explained by rare GoF truncating mutations [151, 152, 153].

Interestingly, signs of negative selection were found for OGs in noncognate cancer types. The lower quartile of the distribution of selection in oncogenes for nonsense mutations was more negative compared to random genes. Purifying selection for Nonsense-Mediated mRNA Decay (NMD)-inducing nonsense somatic mutations in a group analysis of oncogenes was previously shown in the work of Lindeboom et al. (2016) [112]. While the power afforded by the data analyzed here did not allow us to show significant individual hits under negative selection (best observed FDR was 0.78), we highlight the potential candidates where the absolute value of negative selection was the greatest: *IL6ST* in BRCA-Lum cancer, *KMT2A* in COREAD-POLE cancer, *BIRC6* in LUSC and others (Figure 3.2).

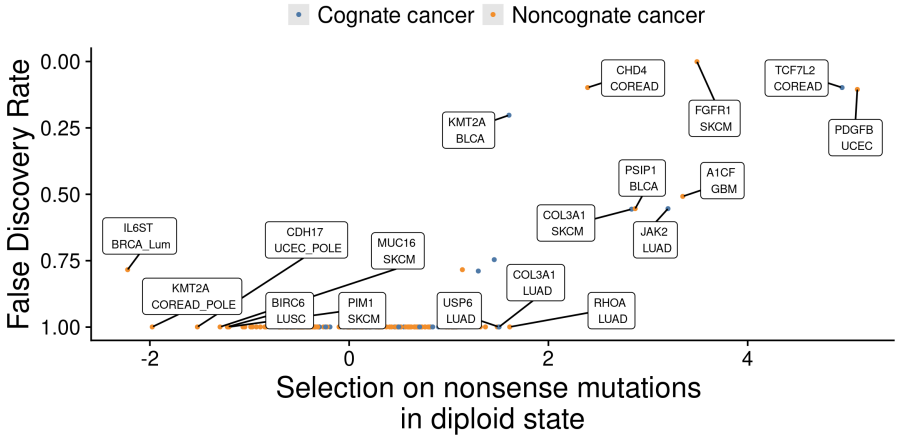

Figure 3.2: Selection on nonsense mutations in the diploid state.

We next asked what fraction of cancer gene mutations are in hotspots. We estimated what is the proportion of mutations in hotspots for each mutation type (Figure 3.3). The fraction of mutations in hotspots for oncogenes was higher than in tumor suppressor genes. The exact proportion of hotspot mutations per gene depends on the hotspot definition: it is, expectedly, much lower when using a Trevino set of hotspot sites and larger using a self-defined and more permissive set of hotspots (Section

7.6.1). The proportion of hotspot mutations in tumor suppressor genes was much larger than in random genes; we estimated that at least 85.7 % of all mutations for 3/4 of tumor suppressor genes in cognate cancer types are in hotspots.

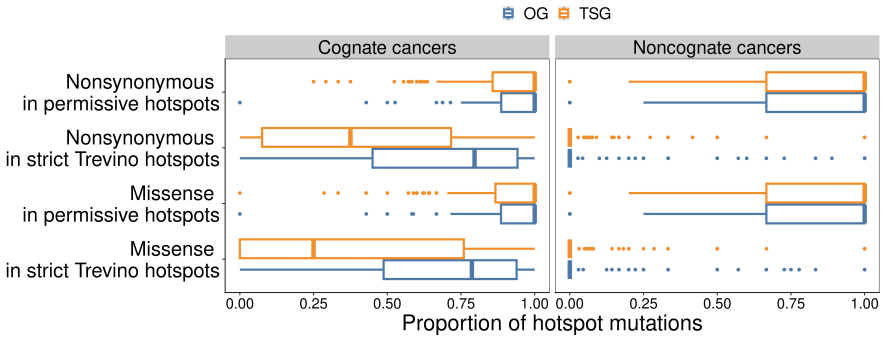

Figure 3.3: Fraction of mutations located in hotspots for different gene and mutation types. One data point represents the value from one gene-tumor combination.

3.2.2 Opposing selection forces lead to underestimation of positive and negative selection

To obtain more accurate estimations of selection in cancer types, we estimated selection across different copy number states (Figure 3.4). While results were overall similar to the selection estimated for the diploid state, certain patterns became more evident. In particular, we found negative selection on nonsynonymous mutations in essential genes, which was not observed in the previous analysis, suggesting that not controlling for copy number state of genes may increase the power to detect positive or negative selection acting on point mutations.

We hypothesized that the pattern of mutations on OGs might be explained not only by positive selection but also by a negative selection that simultaneously acts to remove deleterious mutations in regions located outside of hotspots. To examine this, we estimated selection in hotspots-free gene regions using a permissive set of hotspots for exclusion (7.6.1). Indeed, nonsynonymous mutations in OGs in noncognate cancer types

were selected more negatively in comparison to the set of random genes (in random genes, the effect sizes were distributed with the first quartile $Q1 = -0.4$ and median $Q_2 = 0$, in OGs $Q_1 = -0.6$ and $Q_2 = -0.01$). On the other hand, in cognate cancer types, the difference was not statistically significant ($Q_1 = -0.5$ and $Q_2 = 0$).

Similarly, for TSGs in noncognate cancer types, we found negative selection outside of permissive hotspots ($Q_1 = -0.4$ and $Q_2 = 0$), and for cognate cancer types the difference was not significant ($Q_1 = -0.2$ and $Q_2 = 0$).

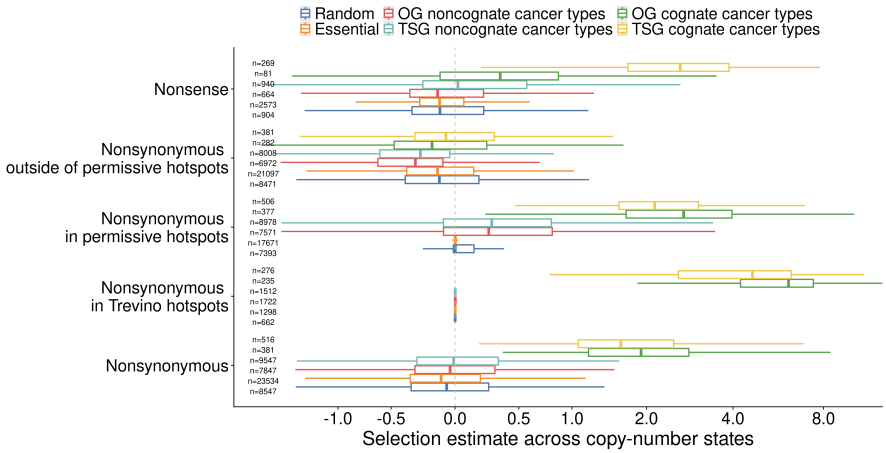

Figure 3.4: Selection across all copy number states. Debiased selection estimates ω across all copy number states of a gene obtained on the discovery cohort. The number of gene-cancer-type pairs used to produce each of the boxplots is written in the left part of the plot. One data point corresponds to one gene-tumor combination.

Although the effect is small, it can reflect negative selection on regions of genes that are essential for the activity of a gene. Stronger sign of signals of negative selection in noncognate cancer types together with weak signals of positive selection in hotspots leads us to think that some of these noncognate cancer types might be wrongly annotated. In other words, the negative selection can offset some signals of positive selection on the same gene, reducing the power to detect positive selection and wrongly

annotating the gene as noncognate.

To test this, we compared selection estimates for hotspots with selection estimates for hotspot-free areas (Figure 3.5).

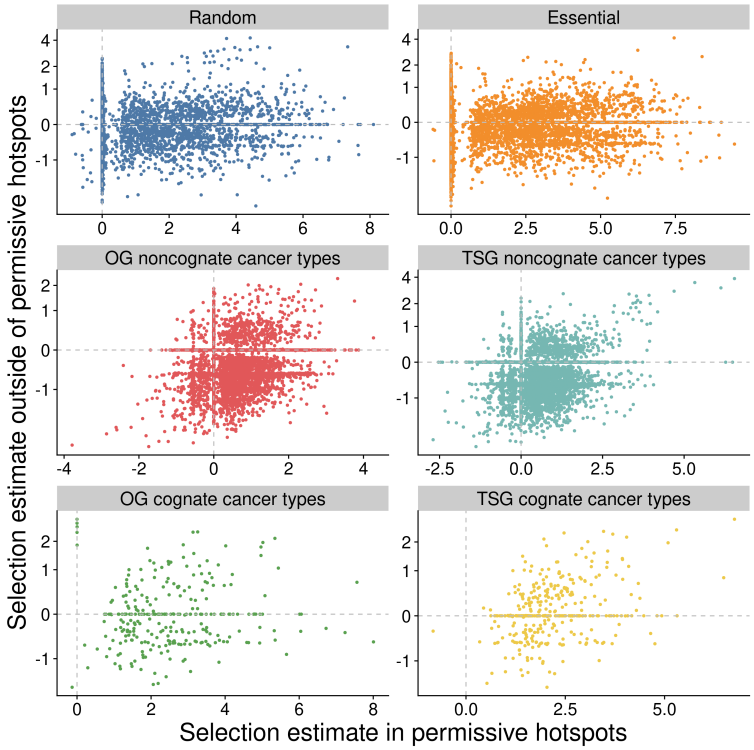

Figure 3.5: Selection acting on different areas in a gene. Debiased selection estimates ωacross all copy number states of a gene obtained on the discovery cohort. One data point corresponds to one gene-tumor combination.

A substantial proportion of genes in noncognate cancer types with negative selection outside of hotspots were positively selected in hotspots. For these genes, the total signal of selection will be the result of the balancing forces and will be neutral or weakly positive or negative, depending on the proportion of the hotspot sites in a gene. Our definition of cognate and noncognate cancer types, although adjusted to account for weak positive

signals of selection (for cognate cancer types FDR $\leq 10\,\%$ and noncognate FDR $\geq 75\,\%$, as detailed in Section 7.3), was missing such cases.

3.2.2.1 Noncanonical oncogene addiction

We examined if negative selection acting simultaneously at gene level and reducing the number of positively selected mutations could also be detected in the analysis while adjusting for copy number states. Results obtained using selection estimates across copy number states were similar to results from selection estimates for the diploid state. We compared signals of selection inside strict Trevino hotspots against hotspot-free areas outside of the permissive set of hotspots (Figure 3.6) focusing on oncogenes that had such strict hotspots.

As seen in Figure 3.6, OGs formed two clusters. First group, where hotspots mutations were not selected, predominately contained genes in noncognate cancer types. The second group consisted mainly, but not only, of cognate cancer types and had a strong selection in hotspots.

Negatively selected genes were present in both groups, which proposes that there are two types of negative selection removing deleterious mutations from oncogenes. The first type of negative selection acts on driver oncogenes and illustrates a phenomenon known as oncogene addiction: dependence of a cell on a single oncogene, which is otherwise usually activated by GoF mutations. The second type of negative selection, however, demonstrates that even tumors depend on the oncogene function even in some cases where that specific oncogene is not a driver of tumorigenesis in that particular tumor. Herein we term this phenomenon "non-canonical oncogene addiction".

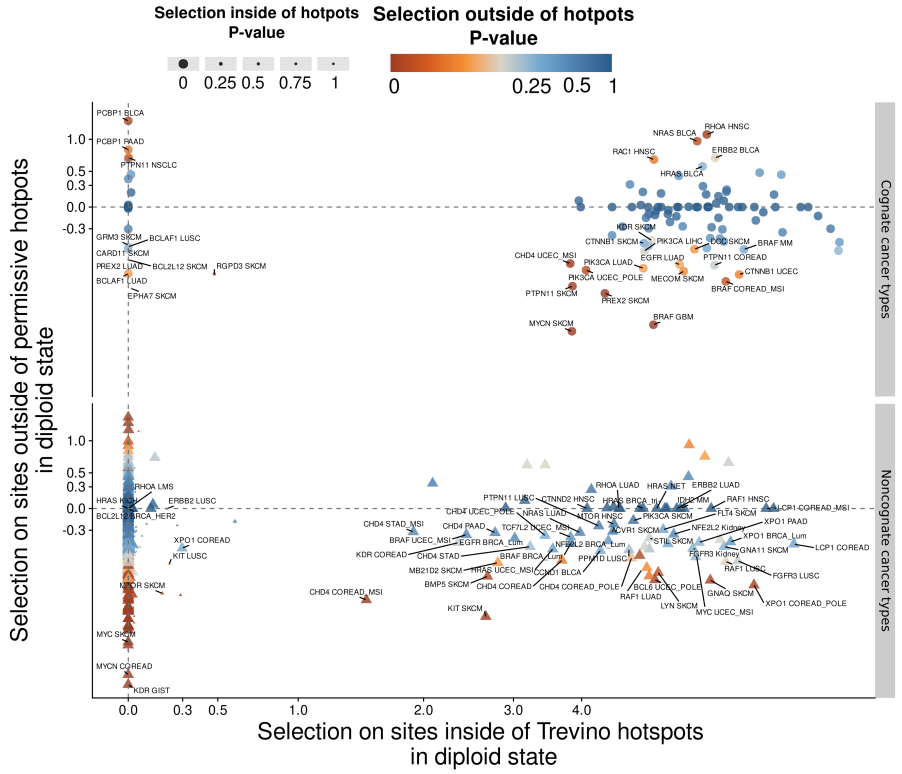

Figure 3.6: Two mechanisms of oncogene negative selection cancer types. Non-canonical oncogene addiction (no positive selection in hotspots) and misannotated cancer types (hotspot sites are positively selected). One data point corresponds to one gene-tumor combination.

Non-canonical oncogene addiction would be a new mechanism that makes cell dependent on the function of an oncogene even without its activation. Possibly, we speculate that this dependency on the healthy function of OG may exist in healthy tissues as well, which would mean that some OGs are essential genes. *MYCN*, *MYC*, *MTOR*, *BRAF*, *KIT*, *KDR* and *XPO1* genes are examples of genes from this group showing negative selection signals in noncognate cancer types.

3.2.2.2 *EGFR* is selected in BRCA, HNSC, and ESAD cancers

As shown above, interfering signals of positive and negative selection in different genic areas can lead to an underestimation of positive selection strength, thus increasing the number of false negative identifications of driver genes. This problem is important to address for genes that have inhibitors passing the stage of clinical trials or are already used in targeted therapies for cancer types.

An important example that fits this description is *EGFR* gene. Known cognate cancer types where *EGFR* is positively selected include brain and LUAD [95, 85, 9, 154]. Collective results of our analyses suggest that *EGFR* is selected in BRCA-Lum (a luminal subtype of breast cancer), ESAD, ESCA, GBM, HNSC, LGG, LUAD. In particular, the gene level positive selection was not captured in BRCA-Lum because of negative selection outside of hotspots.

To additionally illustrate this, we plotted raw mutation rates in the *EGFR* gene and its neighbors in a set of tissues including those known from the literature *EGFR* cognate cancer types (lung and brain) (Figure 3.7A) and presumably noncognate cancer types (Figure 3.7B). We compared the mutation rates inside of hotspots, outside, and overall in the whole gene.

Taken together, our results demonstrate that mutations in *EGFR* gene in HNSC and BRCA-Lum cancer are positively selected inside of hotspots (the mutation rate is higher in the *EGFR* gene than in the neighboring genes). Furthermore, selection in BRCA-Lum cancer *EGFR* appears to be subtype-specific. For this subtype *EGFR* is negatively outside of hotspots.

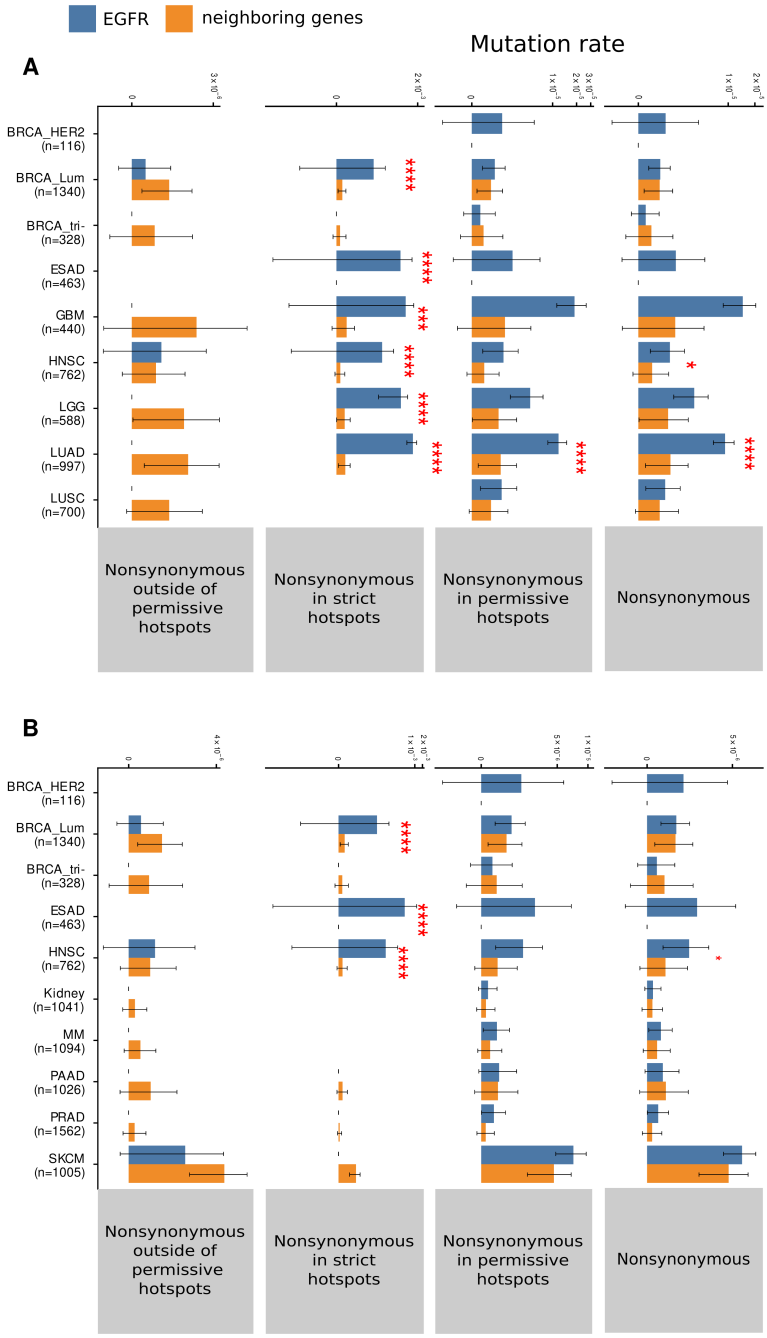
EGFR
neighboring genes
Mutation rate
A
BRCA_HER2 (n=116)
BRCA_Lum (n=1340)
BRCA_tri- (n=328)
ESAD (n=463)
GBM (n=440)
HNSC (n=762)
LGG (n=588)
LUAD (n=997)
LUSC (n=700)
Nonsynonymous outside of permissive hotspots
Nonsynonymous in strict hotspots
Nonsynonymous in permissive hotspots
Nonsynonymous
B
BRCA_HER2 (n=116)
BRCA_Lum (n=1340)
BRCA_tri- (n=328)
ESAD (n=463)
HNSC (n=762)
Kidney (n=1041)
MM (n=1094)
PAAD (n=1026)
PRAD (n=1562)
SKCM (n=1005)
Nonsynonymous outside of permissive hotspots
Nonsynonymous in strict hotspots
Nonsynonymous in permissive hotspots
Nonsynonymous

Figure 3.7: Mutation rate of *EGFR* gene and its neighbors across copy number states. Mutation rate is calculated mutation counts divided by the length of the gene without controlling for Trinucleotide Mutation Spectra (MS96) or copy number state. The estimated mutation rate is plotted with 95 % confidence intervals. Asteriscs indicate cancer types where the gene was positively selected. Y-axis is log-transformed. **A.** Potentially new cognate cancer types together with previously known: LUAD, GBM, LGG. **B.** Potentially new cognate cancer types together with noncognate cancer types: Kidney, MM, PAAD, PRAD, SKCM.

EGFR might also be positively selected in ESAD cancer. Although this gene-tissue combination did not pass our mutation number filter in the copy number analysis (at least 2 mutations required in the model per copy number state), it passes the filter considering the mutations only in the diploid state. Importantly, EGFR was not identified as a significantly mutated driver gene in ESAD, HNSC, and BRCA in the two recent comprehensive studies of similar-sized datasets [95, 85, 154].

Apart from *EGFR*, we additionally compared raw mutation rates between genes *MTOR*, *XPO1*, *GNAQ*, and their respective neighboring genes. Similarly, elevated mutation rates were present in hotspots for cancer types that are not considered cognate cancer types according to the Cancer Gene Census (CGC) or MutPanning annotations (Figures S8.6, S8.7, S8.8). This provides additional examples of genes that may have been missed in previous efforts to catalog driver genes because of simultaneous positive and negative selection acting on them.

3.2.2.3 Re-annotation of cognate cancer types for known cancer genes

We addressed the question of how many driver cancer types-gene pairs may have been lost due to the opposing forces that remove deleterious mutations and keep beneficial ones. We estimated selection strength across different copy number states from different mutation types and regions. We found that more than 1800 unique pairs had positive selection estimates (FDR $< 25\,\%$) when considering the cancer types with the largest number of mutations. We excluded cancer subtypes with microsatellite instability and hypermutation phenotype to avoid spurious correlations because of the extended mutational signatures (i.e. wider than the trinucleotide,

which we rigorously control for in our method).

Out of over 1800 driver gene-tissue pairs, less than a third could be detected using a standard approach (selection signal using nonsynonymous mutations from the whole gene), and almost all the pairs (over 1500 combinations) were found using a permissive definition of hotspots (Figure 3.8).

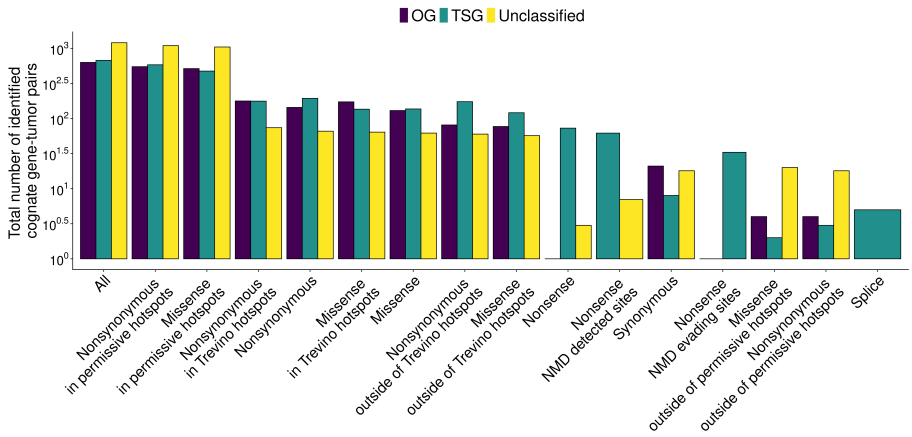

Figure 3.8: Total number of selected gene-tumor pairs discovered using different mutation classes. Cognate gene-tumor pairs identified without controlling for copy number state in 13 cancer types with the largest number of mutaitons.

The largest overlap, as expected, was between combinations found while estimating selection signals in permissive hotspots for all nonsynonymous mutations and missense mutations (Figure 3.9). Additionally, 183 cognate pairs were found only using nonsynonymous mutations in permissive hotspots and 46 with missense mutations in permissive hotspots. Furthermore, 25 cognate pairs were discovered with mutations in strict hotspots with nonsynonymous mutations, two pairs – with missense mutations in strict hotspots, and two using nonsense mutations in NMD-detected sites were missed in all other setups.

Overall, restricting the analysis to only a specific gene parts results in the larger number of cognate cancer types than focusing on one of the mu-

tation types. For example, even for oncogenes where nonsense mutations are not expected to be positively selected, a model that was including all nonsynonymous mutations (missense and nonsense together) predicted more cognate cancer types.

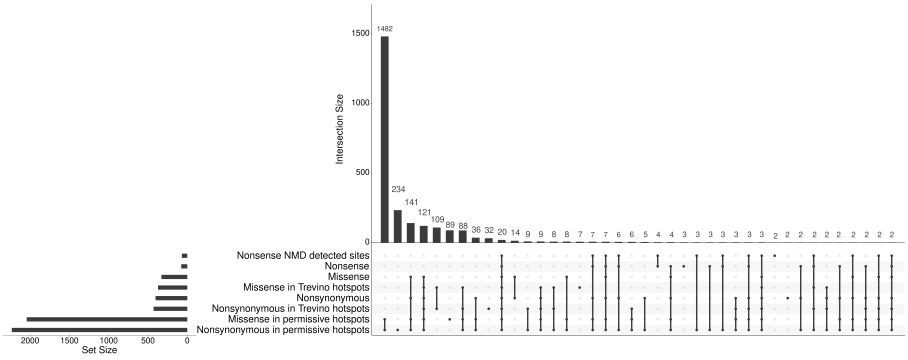

Figure 3.9: Number of overlapping and unique combinations of cognate cancer type-gene pairs found by different approaches.

3.2.3 Essential genes are negatively selected in soma

We analyzed selection on nonsynonymous mutations in a pan-cancer analysis across 13 major cancers with the largest numbers of mutations (BLCA, BRCA-Lum, COREAD, ESAD, HNSC, Kidney, LGG, LUAD, LUSC, MM, PAAD, PRAD, SKCM).

Our results show that cell-line essential genes (CEG2) are negatively selected in the pan-cancer analysis. Regardless, there was not enough power in the dataset to detect the negative selection of essential genes in each cancer type separately.

To exclude the influence of any cancer-type specific mutation signature that might create a bias in the estimation of selection, we repeated the analysis while removing each cancer type one by one (Figure 3.10). Although removing melanoma had a cancer type created the biggest impact on the selection estimates, in all these analyses selection estimates of essential genes were lower than selection estimates of random genes (FDR $< 1\,\%$).

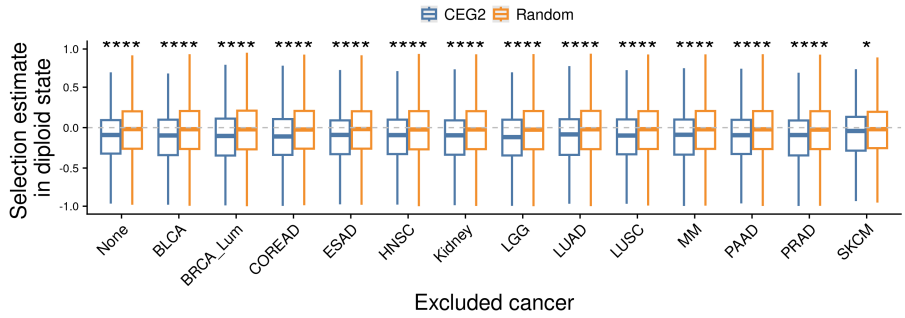

Figure 3.10: Selection in the diploid state in a pan-cancer analysis. Selection in the two gene groups – random and core essential genes (CEG2) – estimated in a pan-cancer analysis [130]. Only the cancer types with the largest number of mutations were considered. One data point corresponds to the selection estimate in one gene across all cancer types excluding one (X-axis). Asterisks indicate the level of significance of the difference between gene groups while controlling for the multiple testing: '*' for FDR ≤ 0.05, '****' for FDR $\leq 1 \cdot 10^{-4}$.

We sought to understand whether known essentiality metrics correlate with selection estimates across the genome, that is, whether genes are under stronger negative selection in tumor genomes if they are ranked as more essential by previous methods.

We compared the selection estimated with the mean cell-essentiality scores across different cell lines (Computational correction of copy-number effect in CRISPR-Cas9 essentiality screens (CERES) score by Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas9 approach, [155]) and essentiality score derived from a population data (selection of Loss-of-Function (LoF) germline point mutations summarized in Loss-of-function Observed over Expected Upper bound Fraction (LOEUF) score, [156]).

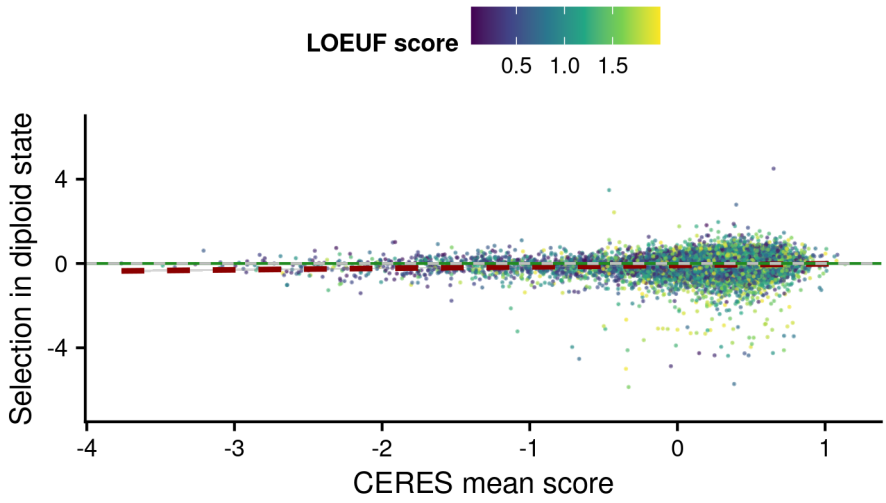

Figure 3.11: Somatic cell lines derived and population essentiality scores do not correlate with somatic selection estimated in tumors. Lower values of CERES and LOEUF scores correlate with a higher chance of being essential in cell-line knockout experiments or with haploinsufficiency at the population level. We observe no correlation between these essentiality scores and estimates of selection derived with the MutMatch method (in the diploid state, pan-cancer analysis) across all genes.

No significant correlation was observed between essentiality defined by CRISPR-Cas9 experiments with cancer cell lines and tumors (Pearson's correlation = 0.07). Similarly, our results showed no correlation between pan-cancer selection estimates for genes and their negative selection of heterozygous, loss-of-function variants at the population level that is measured by LOEUF score (Figure 3.11).

Moreover, we tested if the most essential genes according to the LOEUF metric had a shifted distribution of selection effects quantified with the MutMatch method (top-10 %, top-500, top-200, top-100, top-50 of the most genes were tested). Surprisingly, our data showed no variability in the coefficient estimates depending on the gene ranking (Figure 3.12).

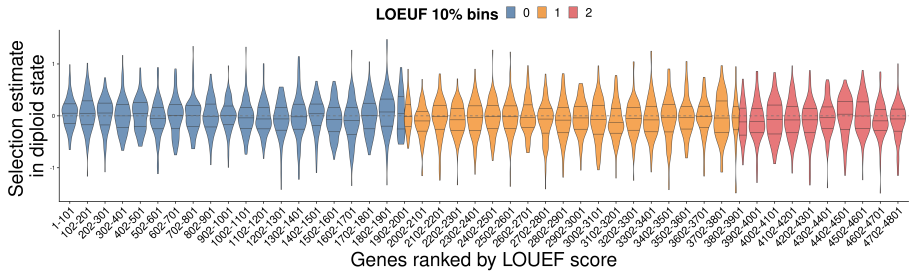

Figure 3.12: The most essential genes according to the LOEUF score are not more negatively selected than less essential genes. Only top-25 % of genes with the most essential LOEUF score are presented here.

We also compared the distributions of essentiality scores in different gene groups: essential genes, cancer genes, and random genes (Figure S8.9).

Interestingly, there was an enrichment of genes under purifying selection (lower values or bins of examined scores) in cancer genes and essential genes compared to the random set of genes (Figure S8.9A). The enrichment of low CERES scores in the group of core cell essential genes compared to the random genes is expected by the design of the essential gene set ($Q_1 = -2.64$ and $Q_2 = -2.16$ for essential genes and $Q_1 = 0.11$ and $Q_2 = 0.32$ for random genes). In contrast, lower than in random genes parameters of the distribution were not expected for cancer genes ($Q_1 = -0.17$ and $Q_2 = 0.21$ for TSGs and $Q_1 = -4 \cdot 10^{-3}$ $Q_2 = 0.27$ for OGs).

Similarly, cancer genes and essential genes had a bigger proportion of the genes with the low LOEUF score compared to the expected (Figure S8.9B). Almost half (49.9 %) of all essential genes were having the lowest 30 % LOEUF scores versus 28 % for random genes. In cancer genes, 63 % and 65 % of OGs and TSGs, respectively, had the LOEUF scores from the lowest 3 deciles.

This proves that functions of cell essential and cancer genes are vital not only at the level of the cell or a tumor but also at the organismal level. However, constraints on the mutation profile of those genes found with essentiality screens only partially overlap with somatic negative selection. This may be because somatic selection estimates are noisy, which is likely to improve with larger data sets.

Altogether, we find core cell-essential genes that were previously identified via CRISPR screening experiments in cell lines [130] are more related to the somatically negatively selected genes than those genes whose evolution is constrained in the population (by LOEUF score). Two independent scores are weakly correlated (Pearson's correlation = 0.15, Figure 3.13), and so they provide orthogonal metrics to gene essentiality.

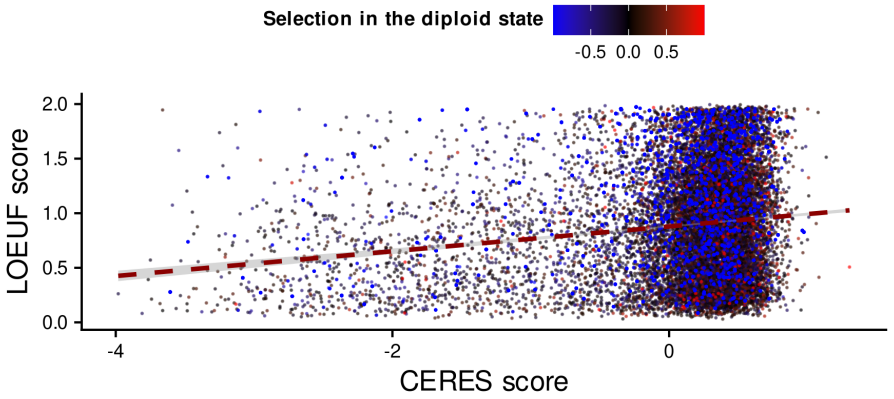

Figure 3.13: LOEUF and CERES scores are weakly correlated.

3.3 Chapter methods

3.3.1 Gene classification

Cancer genes from CGC list [132] were classified by their role in cancer in three groups: OGs, TSGs, and unclassified genes. The latter group was excluded from the analysis.

First, we used the classification provided by CGC to separate OGs from TSGs [132]. For genes with mixed annotation, we used TUSON q-values [120] with the following thresholds:

1. $q\text{-}value_{OG} < 0.4$ and $q\text{-}value_{TSG} > 0.6$ to be classified as an OG
2. $q\text{-}value_{TSG} < 0.4$ and $q\text{-}value_{OG} > 0.6$ for TSGs

Next, for the genes left in the unclassified category, we calculated the proportion of nonsense mutations from all the mutations in a gene. Following the "20/20 rule" [96], genes with nonsense mutations comprising more than 20 % of all mutations were classified as TSGs.

After this, our cancer genes dataset contained 249 oncogenes, 305 tumor suppressor genes, and 123 unclassified genes with no identified or mixed role in cancer.

To have a positive control set where a negative selection can be expected, we used a list of CEG2 core essential genes from Hart et al. (2017) [130]. We also included a set of 300 random genes (excluding known cancer genes from CGC and MutPanning [95]) in our analysis to have a neutral baseline.

3.3.2 Gene selection models

For cancer-specific estimates of selection across the copy number state we used the following regression formula to model raw mutation counts Y:

$$\log E[Y] = \omega t + \sum_i m_i \mu_i + \alpha + \log r \tag{3.1}$$

where trinucleotide-specific differences in mutation rate m are controlled using MS96 stratification of mutations, and the t variable separates mutations from a baseline (neighboring genes) and the gene of interest and is the key variable in estimating a selection strength. For the details on the formula see Section 2.2.1.

For cancer-specific estimates of selection controlling for Copy Number Alteration (CNA) we used the following regression formula:

$$\begin{aligned} \log E[Y] = \omega t + \sum_i m_i \mu_i + c\gamma \\ + \delta\, tc + \alpha + \log r \end{aligned} \tag{3.2}$$

Here, condition variable c separates diploid versus deleted state of a gene in samples for estimation of selection change upon deletion δ, or diploid

versus a gain state of a gene in samples for estimation of selection change upon gene gain δ.

For pan-cancer analysis, we included cancer variable z to account for cancer-driven variations in mutation rate and CNA:

$$\begin{aligned}\log E[Y] = \omega t + \sum_i m_i \mu_i + \sum_j z_j \beta_j + c\gamma \\ + \delta\, tc + \alpha + \log r\end{aligned} \tag{3.3}$$

3.3.3 Post-processing of regression estimates of selection

For all selection estimates (ω – selection in the diploid state, and δ – selection change upon a CNA event), we subtracted the median of the selection estimates null distribution generated using a randomization procedure to correct the estimates for the bias (Section 2.3.2).

We filtered gene-tissue pairs based on the number of mutations in each state (diploid state, gain, or deletion state). We required at least two mutations to be observed in an analyzed genomic region (with a theoretical possibility of having one mutation in a central gene and neighboring genes) for regression in cancer-specific analysis and ten mutations for pan-cancer analysis. We required at least six mutations to be observed in an analyzed genomic region in the analysis across all copy number states.

Chapter 4

Epistatic interaction between mutations and copy number alterations in the same gene

4.1 Overview

In this chapter, we highlight the most interesting results we found studying how selection changes depending on whether a copy of a gene was duplicated or lost. We estimated selection in two independent datasets using neighboring passenger genes or low-impact mutations to model the baseline mutation rate, while stringently controlling for the confounding effect of gene dosage on mutation burden. The result of this and additional analyses demonstrate:

- For most Tumor Suppressor Genes (TSGs) and Oncogenes (OGs), tumors that either lost or gained a gene copy had a stronger selection for driver Single Nucleotide Variant (SNV) mutations (missense and nonsense for TSGs and missense for OGs).

- In samples with a copy gain, normally, the mutant allele gets amplified rather than the wild-type allele for both OGs and TSGs in cognate cancer types. Some individual genes are the exceptions from this pattern, showing strong negative selection for amplifying a mutant allele.
- Mutant allele imbalance in samples with gene copy-gains shown for TSGs suggests that a dominant negative mechanism of inactivating mutations is common in TSGs.
- Mutant allele imbalance achieved by deletions or copy-gains reflects different selection forces that favor either removal of the wild-type allele or increasing the dosage of the alleles.

4.2 Results

4.2.1 Conditional selection upon hemizygous gene loss

4.2.1.1 Selection estimates

To evaluate how somatic Copy Number Alterations (CNAs) affect the selection of somatic point mutations in genes, we performed two analyses. In the first analysis, we estimated conditional selection on point mutations associated with a gene loss (i.e. a change in selection strength that is observed in tumor genomes where a gene copy is deleted). In the second analysis, we estimated selection change upon a gene copy-gain.

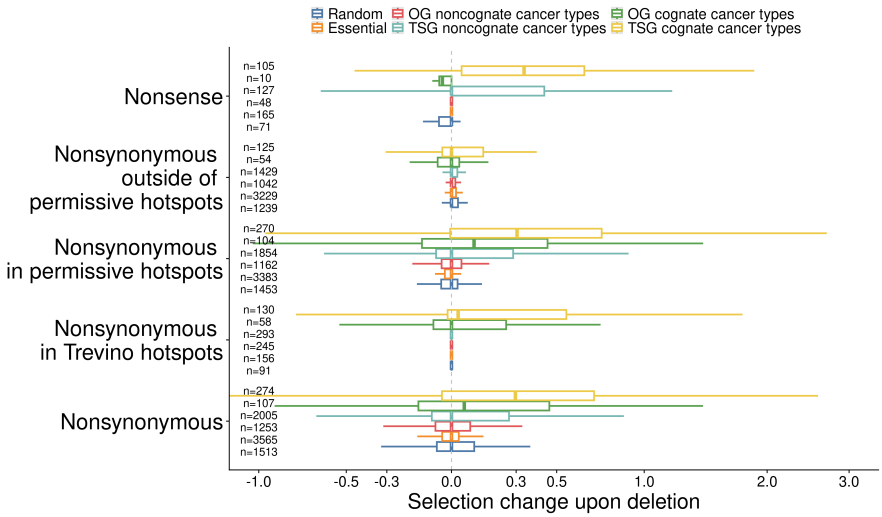

Figure 4.1: Regression coefficient $\delta_{deletion}$ on the interaction term between the selection variable t and copy number variable c. Debiased conditional selection estimates obtained for each gene using the different mutation classes of the discovery cohort. The change of selection strength $\delta_{deletion}$ between samples where genes are in the diploid state and where a gene copy was lost was estimated using neighboring genes as a mutational baseline. One data point corresponds to one gene-tumor combination. The number of gene-tumor pairs used to produce each of the boxplots is written in the left part of the plot.

In agreement with the two-hit model of gene inactivation (Figure 4.1) [31], nonsense mutation and deletions were cooccurring more than expected by chance in TSGs. Interestingly, although the effect was stronger for cognate cancer types, a long tail of positive interaction was also shown for noncognate cancer types. This confirms the previous statement that some of the non-cognate cancer types (identified using selection estimates across copy number states) are false positives and are selected.

We observed a depletion in the rate of nonsense mutations in OGs upon deletion compared to the baseline, which, however, was not significant – likely due to the small sample size (Figure 4.1). Additionally, 25th percentiles derived from distributions of conditional selection estimates outside of permissive hotspots in OGs and TSGs were significantly lower

than from random genes. We believe this represents signs of negative selection in hemizygous regions of some genes (at least a quarter of them, judging by the 25th percentile), especially for OGs.

Both in OGs and TSGs, selection on all nonsynonymous mutations was stronger in samples with a gene loss (Figure 4.1). The estimates of conditional selection were higher for genes under selection: for random genes the upper quartile $Q_3 = 0.1$ and median $Q_2 = 0$, while for OGs $Q_3 = 0.46$ and median $Q_2 = 0.05$ and for TSGs $Q_3 = 0.7$ and median $Q_2 = 0.3$. This effect was even more pronounced in the case of mutations in permissive hotspots: $Q_3 = 0.45$ for OGs, $Q_3 = 0.74$ for TSGs, $Q_3 = 0.02$ for random genes.

It was expected for two-hit TSGs, which need to inactivate both alleles to produce a cancer phenotype. However, for OGs, enrichment of mutations with gene loss in a cell was not anticipated.

There can be two explanations for the observed increased mutation rate in oncogenes. If a gene loss happens after the Gain-of-Function (GoF) activating mutation, losing the wild-type allele is beneficial because (1) losing an allele with a GoF mutation may lead to the tumor regression and therefore is negatively selected (2) wild-type allele can have tumor suppressive effect. The first explanation would illustrate the paradigm of oncogene addiction (dependency of tumor on a single activated oncogenic pathway). The second hypothesis has been studied for *RAS* genes, where the wild-type allele seems to have an inhibitory effect on the mutant allele [157, 158].

4.2.1.2 Mutation frequencies

To substantiate this observation, we compared mutation frequencies in samples with a gene loss between gene groups, having a group of random genes as a baseline. The frequency of a mutation in a cell population depends on the dosage of the mutation (number of mutant gene copies), ploidy of the cell, and tumor purity (fraction of tumor cells in a sample).

We corrected Variant Allele Frequency (VAF) estimates to avoid variability driven by changes in tumor purity between samples. VAF estimates

after this procedure, however, were not equal to x/y, where $x \leq y$ and x, y are whole numbers (with the expected values across deleted, neutral or amplified state 0, 1/3, 1/2, 2/3, 1). The observed estimates, however, were not following this pattern, likely because of the noise in the read counts, subclonal mutations, or events of Whole Genome Duplications (WGDs). To account for this, we added a control group of random genes to have a neutral baseline for passenger mutations. Higher frequencies of mutations in the tested set of genes imply a stronger positive selection or an earlier-occurring variant.

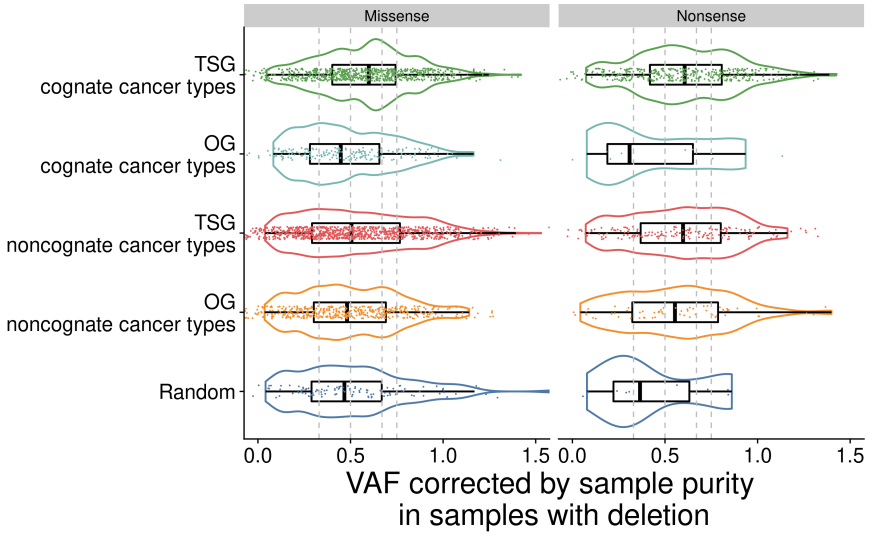

Figure 4.2: Mutation frequencies in samples with a gene loss. One data point corresponds to one adjusted frequency of one missense or nonsense mutation for each tumor sample from The Cancer Genome Atlas Program (TCGA).

Allele frequencies of mutations in tumor suppressor genes were significantly higher than those in random genes (Wilcoxon test p-value $= 4 \cdot 10^{-4}$) with the median for the group of random genes $Q_2 = 0.46$ and TSGs $Q_2 = 0.6$. In particular, for individual genes *STAG2*, *KDM6A*, *AMER1*, *FBXW7*, *VHL* and *TP53*, it was possible to show a significant increase in VAFs compared to the set of random genes at a threshold of FDR ≤ 0.05 (Figure 4.3A).

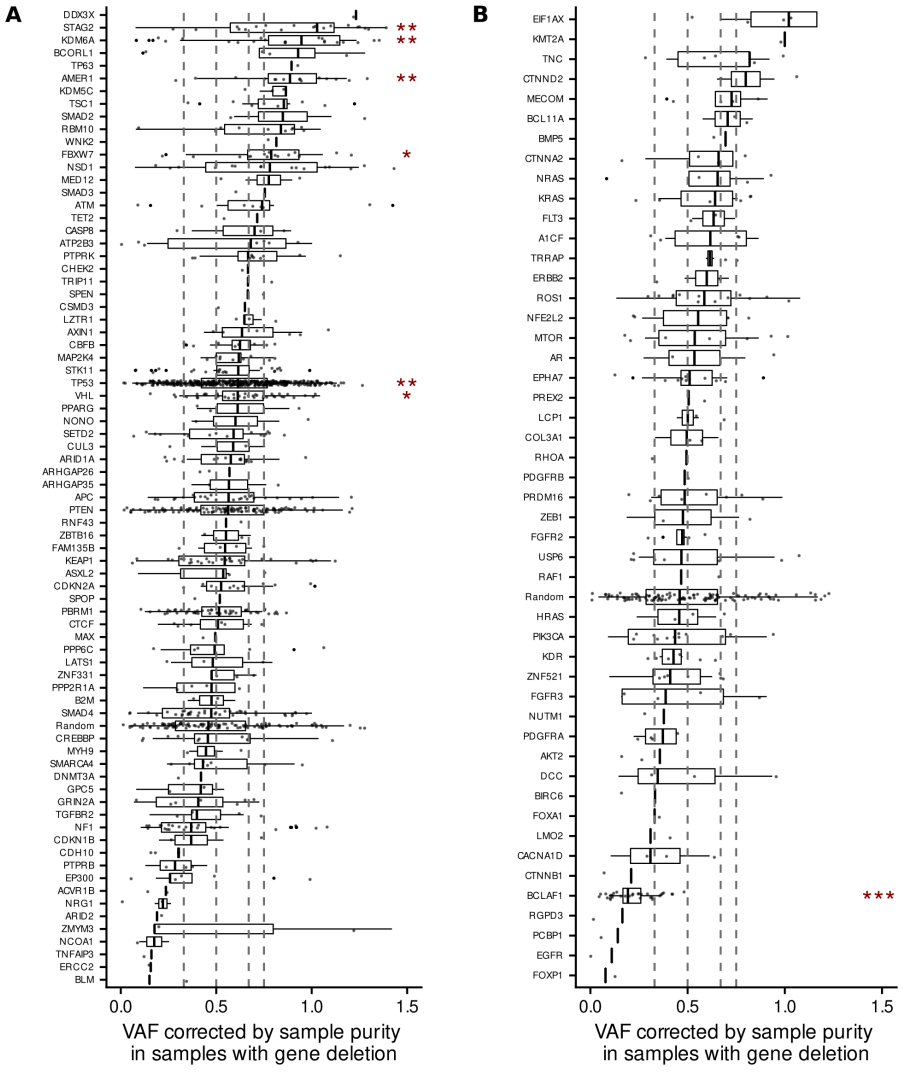

Figure 4.3: Mutation frequencies in samples with a gene loss. A. Variant allele frequencies corrected by the sample purity for TSGs across cognate cancer types (one data point corresponds to one nonsynonymous mutation). Asterisks denote genes with significantly different distributions from the distribution of a group of random genes (Mann-Whitney test with multiple testing corrections). The level of significance of the difference is encoded: '$*$' for FDR ≤ 0.05, '$**$' for FDR $\leq 1 \cdot 10^{-2}$, '$***$' for FDR $\leq 1 \cdot 10^{-3}$, '$****$' for FDR $\leq 1 \cdot 10^{-4}$. **B.** Variant allele frequencies corrected by the sample purity for OGs across cognate cancer types. Asterisks are defined as in panel **A**.

VAFs of oncogenes, on the other hand, were not significantly different from the VAFs of random genes, both for individual genes and all oncogenes together (Figure 4.3, 4.2). For OGs the distribution of allele frequencies of mutations had a median $Q_2 = 0.44$ compared with the $Q_2 = 0.46$ observed in the group of random genes. Restricting analysis only to a set of gene-tissue combinations that showed a mutation enrichment in samples with deletion, did not help to achieve statistical significance ($Q_2 = 0.42$ versus $Q_2 = 0.45$ for the group of random genes).

We checked whether genes that showed a significantly increased VAF of mutations also exhibited a positive selection change on mutations upon deletion (Figure S8.10). Although they all had higher selection estimates than random genes, only selection effects estimated using the neighboring genes baseline in *TP53* gene across cognate cancer types were significantly greater than estimates in the random genes baseline.

4.2.2 Conditional selection upon copy number gain

4.2.2.1 Selection estimates

We estimated conditional selection associated with a copy number gene gain for different cancer genes and mutation types (Figure 4.4). In brief, we consider the interaction term between the selection variable t (separating the tested gene from genes in its neighborhood) and the condition variable c (defining the gene copy number state in a tumor). We observed that driver mutations (nonsense for TSGs and missense for TSGs and OGs) had an increased selection ($\delta > 0$) in samples with copy number gene gain in both TSGs and OGs.

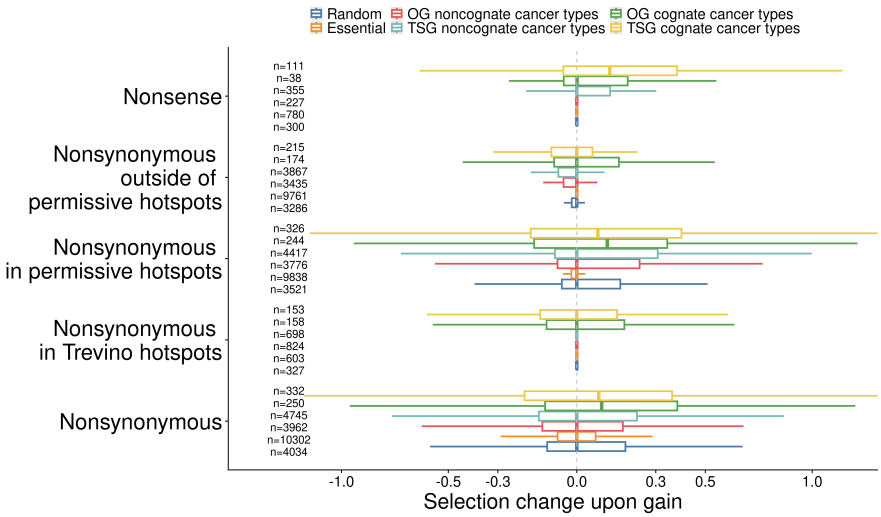

Figure 4.4: Regression coefficient δ on the interaction term between the selection variable t and copy number variable c. Debiased conditional selection estimates obtained for each gene using the different mutation classes of the discovery cohort. The change of selection strength between samples where genes are in the diploid state and where a gene copy was gained was estimated using neighboring genes as a mutational baseline. One data point corresponds to one gene-tumor combination. The number of gene-tumor pairs used to produce each of the boxplots is written in the left part of the plot.

The interpretation of this result is that tumor samples carrying an additional gene copy had an increased mutation rate compared to the expected implying an increased positive selection in tumors with copy number gains. Mutation enrichment in oncogenes was shown previously in papers by Bielski et al. (2018) [108] and Park et al. (2021) [32] in samples with gene amplification. Here we additionally show that this mutation rate increase still stands after controlling for the confounding effect of gene dosage, which increases the apparent mutation rate trivially (via the increased amount of DNA) to a certain extent also in nonselected genes.

On the other hand, additional positive selection pressure on tumor suppressor mutations in samples with gains was not anticipated. We further investigated this by checking which alleles – the wild-type or the mutant

– were the ones that got amplified in TSGs.

4.2.2.2 Mutation frequencies

To collect additional evidence that mutations in oncogenes are positively selected upon the gain of an additional gene copy, we analyzed VAFs of cancer genes mutations in samples with different numbers of gene copies (Figure 4.5).

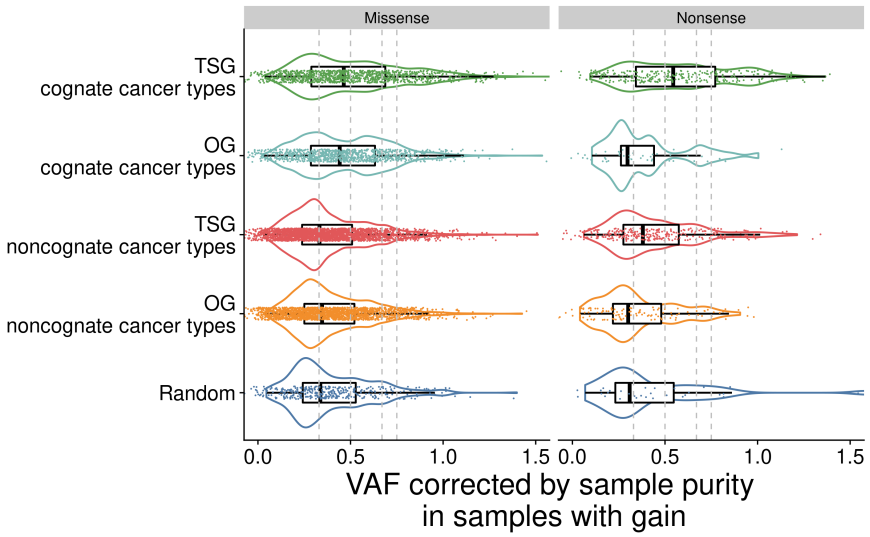

Figure 4.5: Variant allele frequencies of mutations in tumor samples with a gene copy number gain. One data point corresponds to one adjusted frequency of one missense or nonsense mutation for each tumor sample from TCGA.

Compared to the VAFs of random genes, frequencies of missense mutations in oncogenes showed higher values, suggesting that the mutant allele is the one that gets duplicated more often than expected by chance. On the contrary, nonsense mutations had lower frequencies suggesting negative selection on the increasing dosage of truncated proteins.

The allelic imbalance has been previously shown to be selected in can-

cers, with examples of oncogenes under positive and negative selection on increasing the dosage of mutant alleles [108].

Oncogenes *PDGFRB*, *NRAS*, *KRAS*, *FGFR3*, *PIK3CA* individually had significantly higher mutant allele frequencies compared to a random genes baseline at FDR ≤ 0.05. On the contrary, the allelic imbalance was favoring a duplication of a wild-type allele for *BCLAF1* and *EGFR* genes, consistent with a putative signal of negative selection in *EGFR* described above (Figure 4.6B).

Similarly, potentially inactivating mutations (both missense and nonsense) had higher VAFs in tumor suppressor genes compared to random genes baseline. We suggest that preferential amplification of mutant gene copies (or a negative selection on amplification of wild-type alleles) implies the non-dominant-acting character of mutations in tumor suppressor genes. In other words, increasing the proportion of non-functional or less active (hypomorphic) gene products might be beneficial for the fitness of cancer cells, conceivably due to haploinsufficiency or due to a partially dominant-negative effect.

In particular, *PBRM1*, *STK11*, *CDKN1A*, *TP53*, *CDKN2A*, *NSD1*, *PPP2R1A*, *FBXW7*, *NF1*, *ARHGAP35*, *KEAP1*, *PTPRB*, and *APC* genes had significantly higher values of mutant VAFs at FDR ≤ 0.05 (Figure 4.6A), which suggests a dosage-sensitivity in those tumor suppressor genes and non-dominant (or incompletely dominant) character of inactivation of those genes.

We compared the results of the VAF test with the estimates of conditional selection obtained with the MutMatch method. Of all the oncogenes with significantly higher VAFs, only *KRAS* was shown to have a stronger selection in samples with gene gains in cognate cancer types. Other genes, such as *NRAS*, *FGFR3* and *PIK3CA* also had a positive selection change upon gene gain but the difference was not significant at a threshold of FDR $\leq$ 0.05.

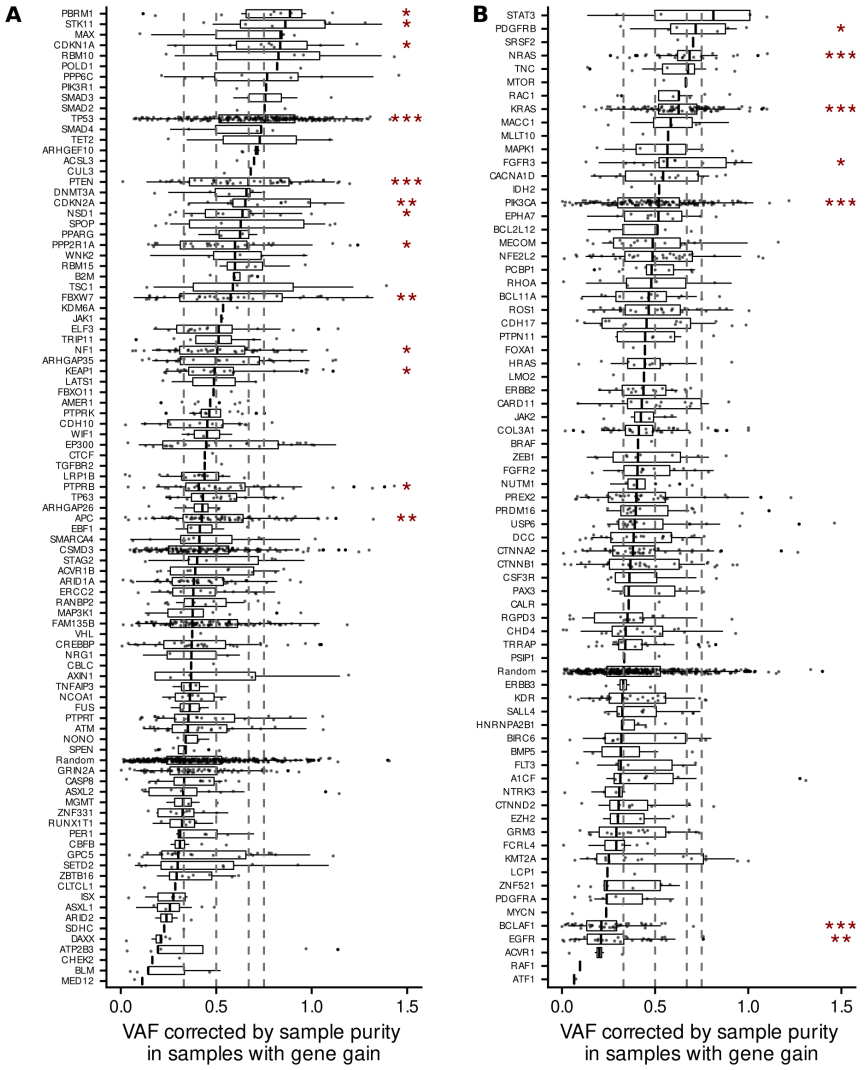

Figure 4.6: Mutation frequencies in samples with a gene copy number gain. A. Variant allele frequencies corrected by the sample purity for TSGs across cognate cancer types (one data point corresponds to one nonsynonymous mutation). Asterisks denote genes with significantly different distributions (Mann-Whitney test with multiple testing corrections). The level of significance of the difference is encoded: '$*$' for FDR ≤ 0.05, '$**$' for FDR $\leq 1 \cdot 10^{-2}$, '$***$' for FDR $\leq 1 \cdot 10^{-3}$, '$****$' for FDR $\leq 1 \cdot 10^{-4}$. **B.** Variant allele frequencies corrected by the sample purity for OGs across cognate cancer types. Asterisks are defined as in panel **A**.

4.2.3 Validation in an independent dataset

To verify the positive selection change upon deletion of a gene or copy number gain in TSG and OG, we estimated conditional selection in an independent Genomics Evidence Neoplasia Information Exchange (GENIE) dataset using low-impact nonsynonymous mutations to estimate a baseline mutation rate (Figure 4.7).

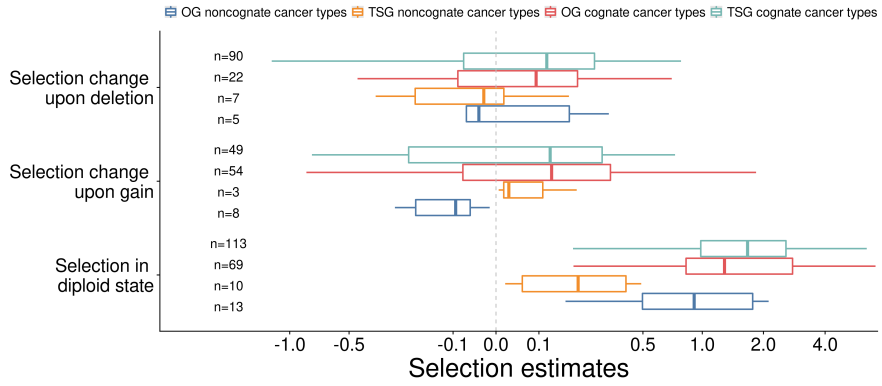

Figure 4.7: Selection on nonsynonymous mutations in diploid state and conditional selection upon CNA events in GENIE study. Debiased regression coefficient ω (estimation of selection pressure in the diploid state and debiased regression coefficients δ on the interaction term between the selection variable t and copy number variable c, for gene deletions, and copy gains. One data point corresponds to one gene-tumor combination. The number of gene-tumor pairs used to produce each of the boxplots is written in the left part of the plot.

The number of genes passing the filtering stage (meaning, with a sufficient number of observed mutations, $n > 4$ in each of the copy number states) was lower compared to the discovery cohort due to the design of the study (panel sequencing of a restricted number of genes, which were all deemed to be cancer driver genes). As a result, the expected behavior of a gene under neutral selection could only be approximated using cancer genes in noncognate cancer types, and a random gene control is not available.

The selection of nonsynonymous mutations for genes in the diploid state

was positive in both groups of cancer types, cognate and noncognate, with a stronger effect in cognate cancer types, expectedly. Low numbers of genes in each group did not allow us to claim a statistically significant difference between the groups. Nevertheless, selection change upon both gain and deletion was positive for both TSGs and OGs, substantiating the overall findings in the discovery cohort.

Although this difference in independent validation data was not significant, it does nonetheless provide additional evidence of positive selection of the cancer genes to increase the dosage of a mutant allele or the proportion of the mutant allele product in a cell by amplifying the mutant allele or removing a wild-type allele.

4.3 Chapter methods

4.3.1 Gene selection models

We estimated selection change upon deletion or single-copy gain for each cancer separately in two datasets: the discovery cohort comprising over 17 000 Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) samples and the validation cohort with ∼90 000 samples with targeted sequencing data. Regression estimates for the interaction term between t and CNA variable were then analyzed, as detailed in Section 2.2.1.

To control for batch effects observed in numbers of mutations and CNAs in the validation dataset, we modified the original formula 3.2 in the following way:

$$\begin{aligned}\log E[Y] = \omega t + \sum_i m_i \mu_i + \sum_j z_j \beta_j + c\gamma + \\ + \delta\, tc + \alpha + \log r,\end{aligned} \tag{4.1}$$

where z is responsible for separating samples from MSK and DFCI cohorts of the GENIE project.

4.3.2 Post-processing of regression estimates of selection

Selection estimates were then debiased using a randomization approach described in Section 2.3.2. For later analysis, we only focused on genes that had at least 4 mutations across all samples where a gene was in a diploid state and at least 4 mutations in samples where a gene copy was deleted or gained, separately for each cancer type.

4.3.3 VAF testing

VAF analysis was performed using the TCGA dataset. We calculated allele frequencies for missense and nonsense mutations located in the same genomic regions that were used for the estimation of selection (including removal of unmappable and conversion-unstable positions). VAF estimates were then adjusted to control for a sample purity (Consensus measurement of Purity Estimation from *TCGAbiolinks* [159, 160, 161]:

$$\frac{n_a}{(n_a + n_r)} \times \frac{1}{P}, \tag{4.2}$$

where n_a and n_r is the number of alternative and reference reads, and P is the sample purity.

Chapter 5

Mechanistic classification of cancer genes based on selection effects

5.1 Overview

In this chapter, we studied the implications that measuring increased or decreased selection on somatic point mutations in non-diploid copy number state has on classifying cancer genes by mechanism of (in)activation. We suggest that the combination of coefficients $\delta_{deletion}$ and δ_{gain} explains the mechanism of action of driver mutations in cancer genes and additionally suggests the contribution of the wild-type allele to tumor fitness. We provide an overview of the possibilities that can generate different combinations of mutant allele imbalance, and characterized the landscape of selection effects on cancer driver genes, summarizing them into "selection signatures" via a dimensionality reduction analysis.

- We characterized 8 possible scenarios of differential selection acting on point mutations in driver genes in tumors that have lost or gained a gene copy.
- Principal Component Analysis (PCA) suggests that genes in tumors

do not form distinct clusters by the mechanism of (in)activation, but rather form a continuous spectrum of cancer-driving potential via multiple mechanisms.

- The most significant trend in variation between genes is correlating with the selection on the diploid state (14.3 % of variance explained). In addition, the interaction of selection on mutations with the copy number state (i.e. conditional selection) can explain at least 24.9 % of variance in the selection effects across cancer genes.

- Genes in tumors do not form discrete clusters according to the selection strength, rather they form a continuum of driving potential.

- Similarly, one-hit and two-hit genes are not clearly separated by the Principal Components (PCs) that constitute "selection signatures". We suggest that mechanisms driving fitness change caused by the mutant allele imbalance in driver genes also form a continuum rather than discrete mechanistic categories.

5.2 Results

5.2.1 Dosage and stoichiometry of the wild-type and mutant alleles affect tumor fitness

Our results (as described in previous chapters) indicate that there is a prevalent interaction between selection strength on point mutations and copy number gains and deletions, in both oncogenes and tumor suppressor genes. Some genes had both increased mutation rates in tumor samples with gene loss and also in tumor samples with an additional copy of the gene; this observation was supported in the group of tumor suppressor genes and oncogenes (Figures 5.1, 5.2). Copy number alterations of any sort correlate with an increased selection of point mutations in the same cancer gene.

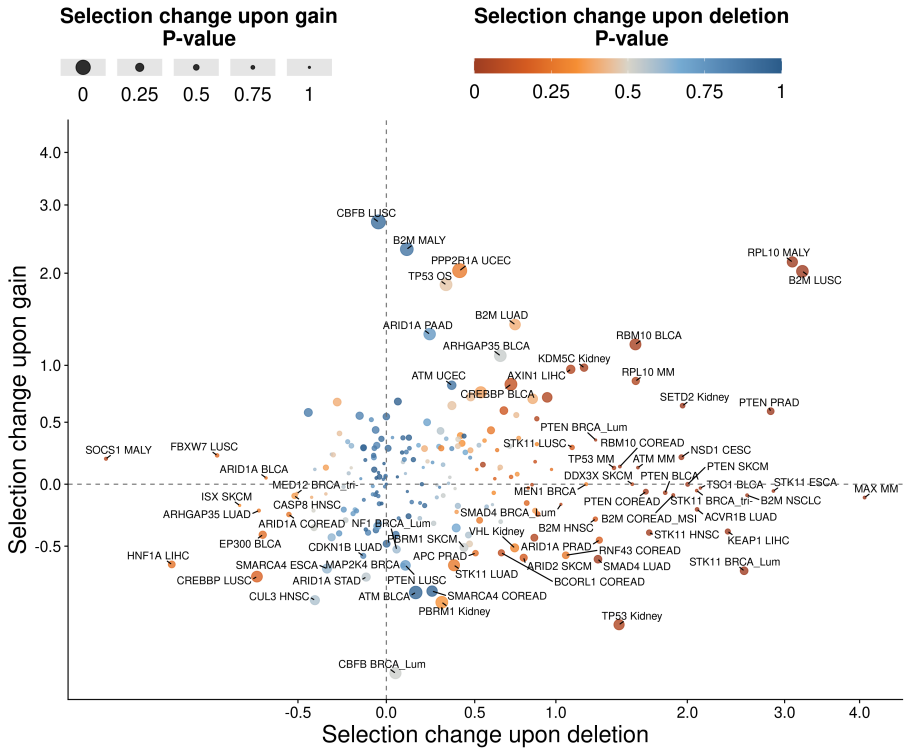

Figure 5.1: Epistatic interaction between selection in a gene and copy number state for tumor suppressors. Debiased regression estimates of conditional selection effects δ obtained for each gene using all nonsynonymous mutations of the discovery cohort. Two analyses were performed to estimate conditional selection: one comparing mutation rates between diploid samples and samples with a hemizygous gene loss (x-axis), and another comparing mutation rates between diploid samples and samples with a copy number gain of the gene (y-axis). Gene name and cancer type are labeled for the strongest effect sizes.

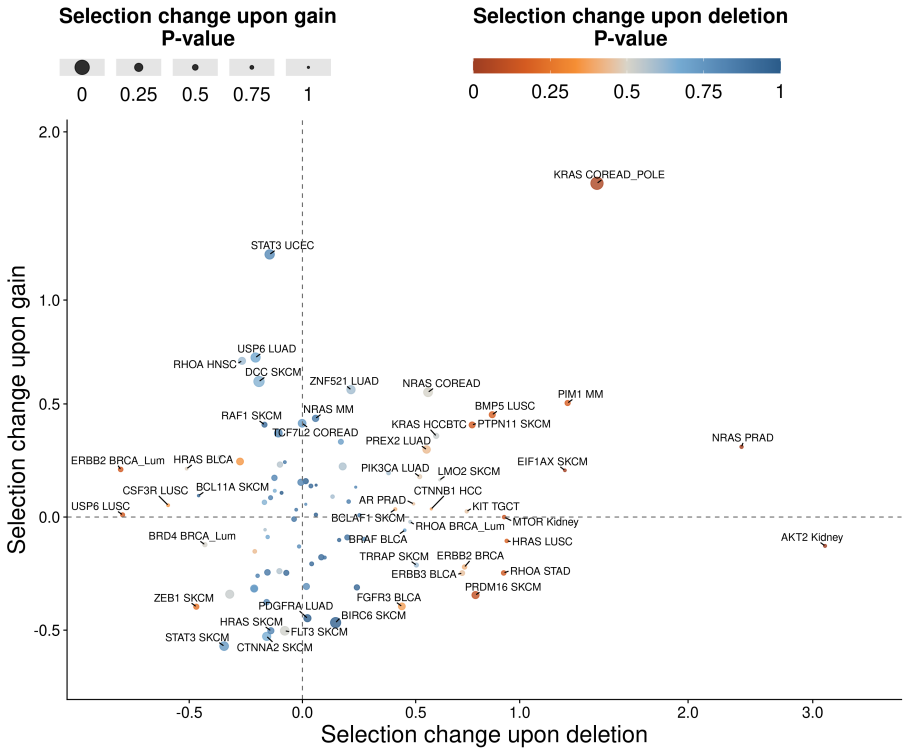

Figure 5.2: Epistatic interaction between selection in a gene and copy number state for oncogenes. Debiased regression estimates of conditional selection effects δ obtained for each gene using all nonsynonymous mutations of the discovery cohort. Two analyses were performed to estimate conditional selection: one comparing mutation rates between diploid samples and samples with a hemizygous gene loss (x-axis), and another comparing mutation rates between diploid samples and samples with a copy number gain of the gene (y-axis). Gene name and cancer type are labeled for the strongest effect sizes.

We suggest that this illustrates a complex interaction where both the relative stoichiometry and the absolute dosage of the mutant allele affect the fitness of cancer cells. Epistasis between a selection of point mutations and Copy Number Alteration (CNA) deletion events (i.e. increased selection upon deletion, $\delta > 0$) can be explained by the increased fitness of cancer cells after the elimination of wild-type alleles, which can play

a tumor-suppressing role (notably including also the wild-type alleles of oncogenes).

Alternatively, a negative difference in selection strength on point mutations between CNA conditions (i.e. negative selection change upon deletion or mutual exclusivity, $\delta < 0$) may signal an essential function of the wild-type allele in a cell. A positive interaction between selection and gene copy gain shows that dosage increases fitness.

Lastly, reduced selection on mutations in tumor samples with CNA gain could be due to the redundancy of two independent gain-of-function genomic alterations (either gene gain/amplification or Gain-of-Function (GoF) mutation is sufficient), or also the toxicity of the increased dosage of the mutant gene product.

We propose a mechanistic classification of cancer driver genes based on a combination of observed selection effects for epistatic interactions involving point mutations and CNA events.

1. **Mutual exclusivity with deletions and with gains**
   $[\delta_{deletion} < 0 \wedge \delta_{gain} < 0]$
   Genes, whose wild-type allele is essential for tumor survival. Gene gain has the same effect as acquiring a mutation. Both GoF mutations are sufficient for the tumorigenesis (no additive effect) and therefore are redundant (for oncogenes). Alternatively, an increased dosage of the mutant allele is toxic to the tumor.

2. **Mutual exclusivity with deletions, cooccurrence with gains**
   $[\delta_{deletion} < 0 \wedge \delta_{gain} > 0]$
   Genes whose wild-type allele is essential for tumor survival. Increasing the amount of mutant allele without removing the wild-type allele is beneficial for a clone.

3. **No interaction with deletions, cooccurrence with gains**
   $[\delta_{deletion} \approx 0 \wedge \delta_{gain} > 0]$
   Wild-type allele does not play a tumor suppressive function nor has an essential role in cancer cell survival. Increasing the dosage of the mutant allele has a positive effect on tumor fitness, possibly due to the weak character of a single mutation.

4. **Cooccurrence with deletions and with gains**
   $[\delta_{deletion} > 0 \wedge \delta_{gain} > 0]$
   The wild-type allele plays a tumor-suppressing role, and so its removal increases tumor fitness. Increasing the dosage (absolute copy number) of the mutant allele, or the proportion (relative copy number compared to the wild-type allele) of the mutant allele results in a similar effect.

5. **Cooccurrence with deletions, mutual exclusivity with gains**
   $[\delta_{deletion} > 0 \wedge \delta_{gain} < 0]$
   The wild-type allele is essential, gains are either redundant with mutations (for oncogenes) or lead to a toxic product of a mutant allele.

6. **No interaction with deletions, mutual exclusivity with gains**
   $[\delta_{deletion} \approx 0 \wedge \delta_{gain} < 0]$
   The presence of the wild-type allele does not change the fitness of tumors with the mutant allele, gains are either redundant with mutations (for oncogenes) or lead to a toxic product of a mutant allele.

7. **Cooccurrence with deletions, no interaction with gains**
   $[\delta_{deletion} > 0 \wedge \delta_{gain} \approx 0]$
   The wild-type allele plays a tumor-suppressive role, and so its removal increases tumor fitness. Increasing the proportion of the mutant allele is not sufficient to inhibit the wild-type allele activity.

8. **No interaction with deletions, cooccurrence with gains**
   $[\delta_{deletion} \approx 0 \wedge \delta_{gain} > 0]$
   The wild-type allele does not play a tumor-suppressive role. Single mutations are hypomorphic and increasing the dosage (absolute number of mutant alleles) rather than the relative proportion of the mutant allele among all alleles benefits tumor fitness.

5.2.2 Inferring gene clusters from the patterns of selection

To reduce the dimensionality of the data and thereby perform an unsupervised analysis to categorize the main trends of variation between cancer

genes in their selection effects by CNA state, we performed PCA.

The estimates for selection in the diploid state (ω), and selection strength change in tumors that have lost ($\delta_{deletion}$) or gained a copy of a gene (δ_{gain}) in various cancer types were summarized using PCA.

The first 3 components were statistically significant (broken stick test [162]) and explained 39.2 % of the variance in the data (Figure S8.11). Although the next principal components were not significant by this particular test, we consider them useful to understand the contribution of different selection effects on the whole dataset (Figure 5.3).

The first Principal Component (PC) had the highest loadings from the selection on all nonsynonymous mutations in the diploid state (including missense and nonsense mutations), and selection interactions with gene gain and gene loss, and the lowest loadings from selection acting on the synonymous mutations (Figure 5.3). Basically, this PC1 can be seen as the strongest overall positive selection observed in a gene, both selection conditional on the gene copy number state and selection acting in the diploid state. As expected, genes in cognate cancer types had higher scores of this PC1 signature (Figures 5.4, S8.13).

The second PC separates the mechanistic groups of one-hit from the two-hit loss genes [31, 32]. High values had gene-tumor pairs with stronger selection in the diploid state ω and a decreased (or not changed) selection in the tumors that lost a gene copy ($\delta <= 0$). The top gene-tumor pairs with the highest score were *SOCS1* in MALY, *PIM1* in DLBC, *BRAF* in THCA, and *KRAS* in UCEC. In contrast, low values of this PC had genes with much stronger selection in tumors with hemizygous gene loss compared to the tumors where the gene is in the diploid state (the top hits: *ACVR1* in GBM, *VHL* in KICH).

The third PC, similarly to the second one, was stratifying two-hit genes versus one-hit genes, however in the case of PC3 it was for gains (in PC2 for losses). Again, genes in cognate cancer types had a significantly shifted distribution of scores in this PC than the rest of the genes, suggesting a broad biological relevance of this PC.

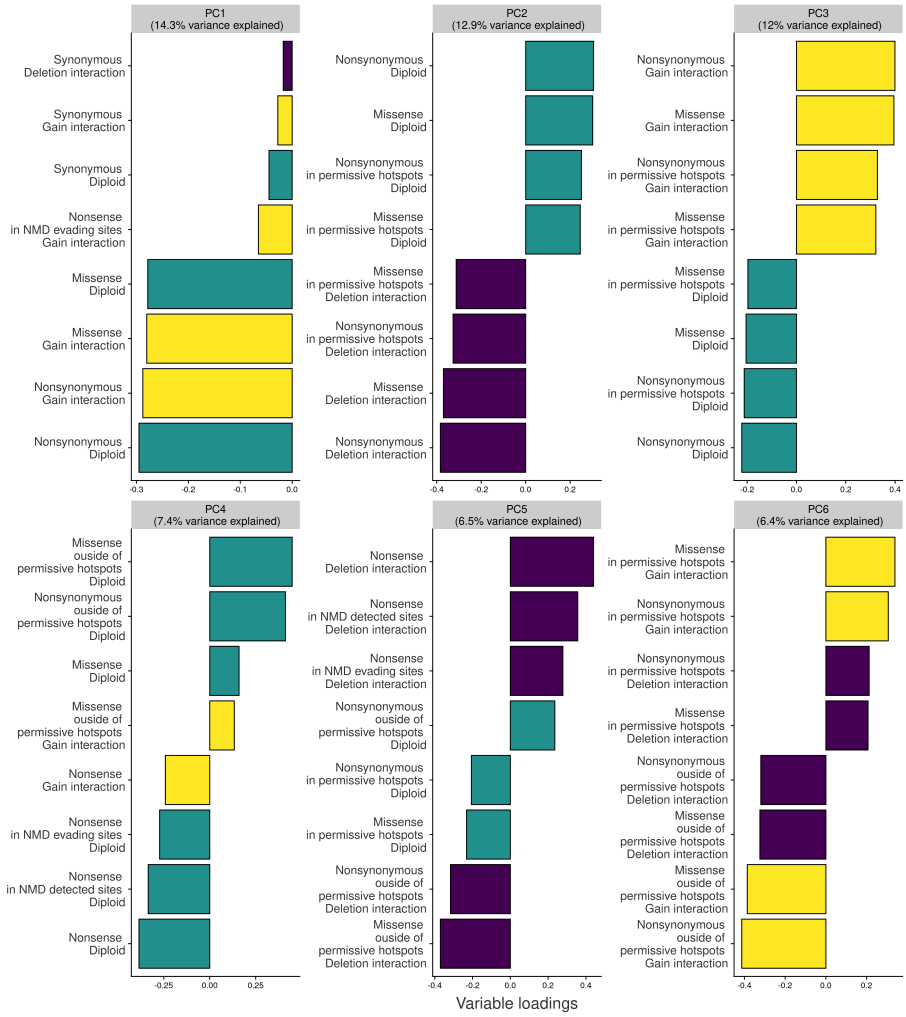

Figure 5.3: Strongest contrubuting features to the first six principal components (PCs). All features were sorted by their loadings to each PC. Here we plot the top 4 features (most positive loading) and bottom 4 features (most negative loading) for each PC.

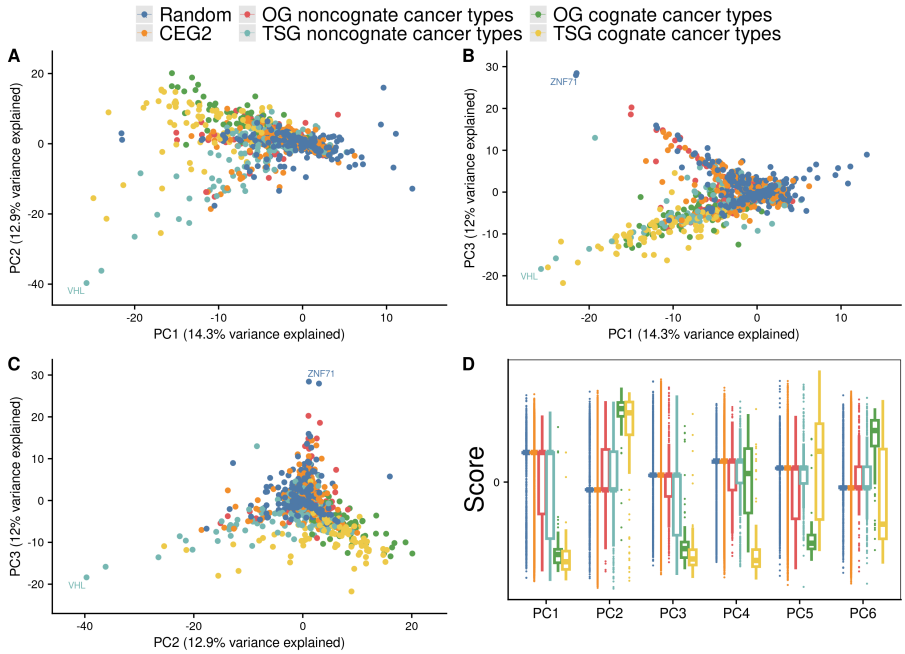

Figure 5.4: PCA analysis of selection effects across cancer types and mutation classes. Principal components 1 to 3 are shown for genes with the 15 most frequently mutated cancer genes: *TP53*, *KRAS*, *APC*, *BRAF*, *PTEN*, *RB1*, *GNAQ*, *PIK3CA*, *VHL*, *GNA11*, *IDH1*, *GTF2I*, *PBRM1*, *ARID1A*, *BAP1*.**A, B C.** Gene groups in the space defined by the three first PCs. Genes with the highest absolute scores are labeled. **D.** Usage of principle components by the gene groups across all cancer types.

Both two-hit loss and two-hit gain genes were rare in the group of cognate gene-tumor pairs, for both Tumor Suppressor Genes (TSGs) and Oncogenes (OGs) (Figure 5.5). However, some of the cognate pairs were indeed two-hit, while, unusually, a bigger number of two-hit genes was observed in the group of noncognate tumors for cancer genes, in cancer types that were neither cognate nor noncognate (unclassified group), and random genes. While some of the genes that were identified as two-hit in PCA were likely to be errors caused by the low mutation rate resulting in noisy estimates (for example, in random genes or essential genes), our data suggest that two-hit can occur in non-cognate cancer types for cancer genes (i.e. those

cancer types where the overall mutation rate for that gene is low thus it was not identified as cognate by our definition).

This points out that some of the selected genes cannot be detected with the standard methods (i.e. searching for the selection across all copy number states jointly) – for example, *VHL* in KICH, *ACVR1* in GBM, *TMSB4X* in THCA and others. In these cases, the known cognate tumor types did not include mentioned cancer types [85, 95, 132]. Using our method, which considers different copy number states separately while estimating selection (even in the case of estimating the overall selection signal) can provide an extended annotation of the putative cognate cancer types for each known cancer gene.

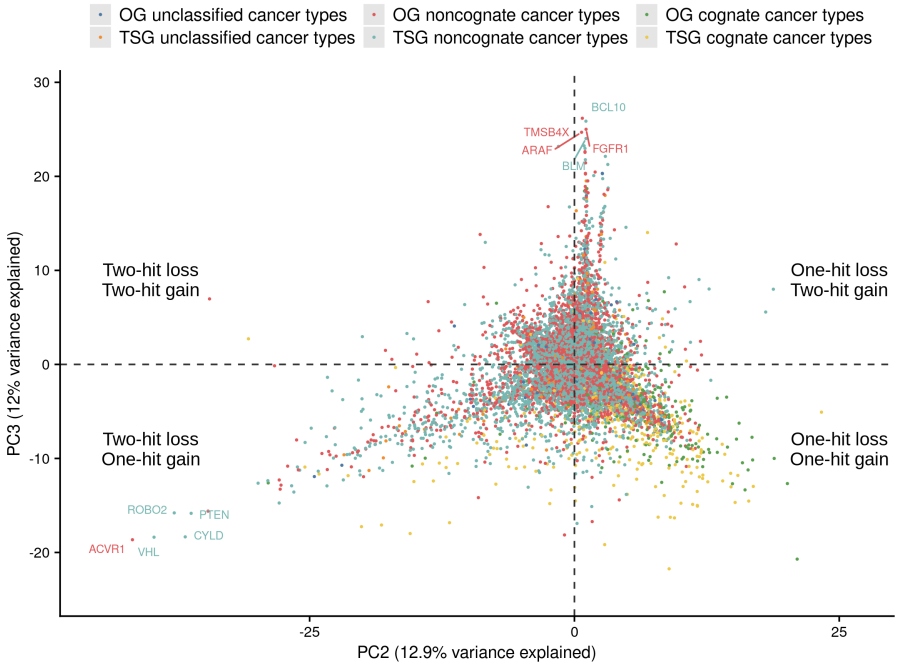

Figure 5.5: Cancer genes form a continuum, rather than distinct groups, between one-hit and two-hit mechanistic classification. All cancer genes in all tumors plotted in the principal component space defined by the PC2 and PC3. Cancer types that could not be defined as cognate or noncognate are separated into the "unclassified" category.

Additionally, we reviewed **the fourth PC**, which represents a stronger selection of nonsense mutations on one extreme, while a stronger selection of missense mutations (outside of hotspots thus presumably mainly Loss-of-Function (LoF) missense) on the other extreme. Expectedly, this PC4 was relevant to the group of TSGs in cognate cancer types, separating them from the rest of gene-tumor pairs (Figures 5.4, S8.12. Thus, certain TSGs appear inactivated more often by missense and others by nonsense mutations.

5.2.3 Selection signatures are differentially active between cancer types

Next, we addressed the question of which tumors were enriched with specific selection signatures summarized by principal components. For the 15 top mutated cancer genes (across all cancer types) [95], we plotted the scores of each PCs in the 30 cancer types with the biggest number of available mutations (Figure 5.6). The strongest selection "overall" determined by PC1 was observed in Kidney, COREAD, LICH, PRAD, and BRCA. Selection on two-hit loss genes (PC2) was the strongest in COREAD, SKCM, PAAD, UCEC. Similarly, two-hit gain genes (PC3) were more often in COREAD, UCEC, HNSC, PRAD.

Interestingly, nonsense mutations (PC4) appeared to be under the strongest selection pressure in SKCM, BLCA, UCEC-POLE, UCEC-MSI, LUSC, COREAD-MSI – tumors with the largest number of available mutations (Pearson's correlation between the median PC4 score in cancer and the total number of nonsynonymous mutations = 0.786).We believe this indicates, at least in part, a technical factor – lack of power to detect selection on nonsense mutations in other cancer types due to the lower number of mutations, rather than necessarily biological reasons. Considering a larger number of cancer genes (for example, all mutated genes from Cancer Gene Census (CGC)) leads to a reduction of this pattern, likely due to the increased proportion of the lowly mutated genes even in cancer types with large mutation numbers.

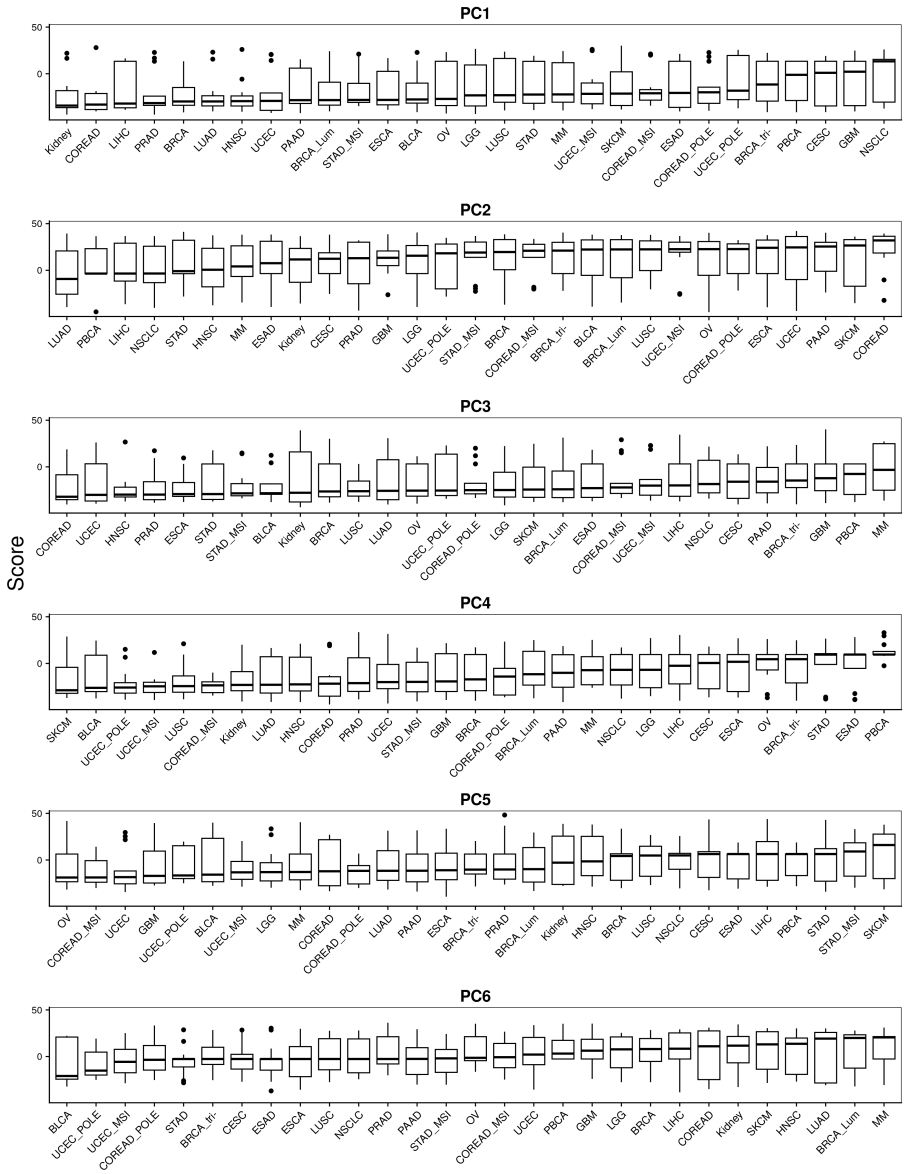

Figure 5.6: Scores of principal components stratified across cancer types with the largest number of mutations. Only 15 top mutated cancer genes were used to produce each boxplot [95]. For PCs, higher values correspond to the stronger selection for glspc2, while for the rest of PCs lower values indicate a stronger selection.

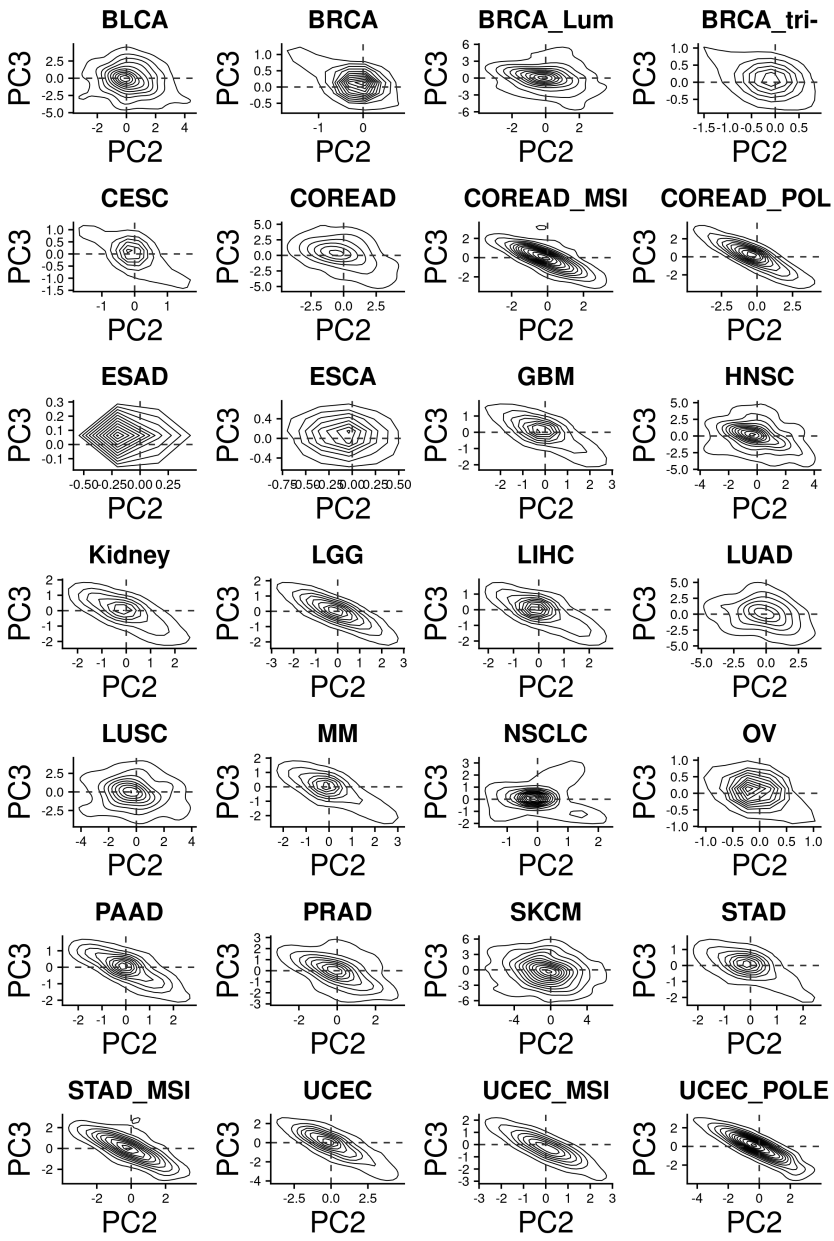

Figure 5.7: 2D-density plots of scores of principal components for each of the cancer types with the largest number of mutations. The correlation between two-hit loss genes (negative values PC2) and two-hit gain genes (positive values of PC3) is supported in many cancer types, with some exceptions (BLCA, LUSC, ESAD, ESCA, HNSC, LUSC, NSCLC, SKCM). Principal component scores from all cancer genes were used in this analysis.

Lastly, we tested if there is a correlation between two-hit gain and two-hit loss genes. For that, using all genes from CGC we plotted their respective scores of PC2 and PC3 separately for each cancer type. Confirming the existence of the proposed correlation, the majority of the cancer types were stretched in a diagonal direction, with genes that were either both two-hit loss and two-hit gain, or one-hit loss and one-hit gain genes (Figure 5.7).

5.3 Chapter methods

5.3.1 PCA on selection estimates

The debiased estimates for (i) selection effect in the diploid state ω, (ii) selection effect change upon deletion $\delta_{deletion}$, and (iii) selection effect change upon gain δ_{gain}, for each gene in each cancer type were used to create an atlas of copy-number dependent selection in the human soma.

To ensure the absence of notable selection in the group of random genes, we excluded from them genes with low Loss-of-function Observed over Expected Upper bound Fraction (LOEUF) score (ranked in the most essential 30 % of the genes) derived from population variants [156, 131], CEG2 essential genes [130]) and cancer genes using MutPanning and CGC gene lists [95, 132]. We also excluded genes that were either very short or very long, allowing only 30 % of difference between the median number of the nucleotides in a gene that was used in the regression model. Finally, we removed genes with very high or very low gene expression using the normalized transcript expression levels summarized per gene in 54 tissues based on transcriptomics data from Human Protein Atlas (HPA) and Genotype-Tissue Expression (GTEx) [163, 164].

For the selection in the diploid state, we averaged the estimates obtained with two different models: one for gene loss and another for gene gain (3.2). We focused on cancer types and genes for which selection effects were estimated in at least 85 % of gene-tumor pairs. Because of the high number of missing values in the case of selection estimates for strict Trevino hotspots, we excluded estimates inside or outside of them from the analysis. For the gene-tumor combination passing these filters, missing values were imputed

using *imputePCA* function from *missMDA* package in R.

PCA was performed on the centered and scaled data. To avoid overplotting, we only show 15 genes from the control group of random genes and 15 genes from the control group of essential genes in the scatter plots above.

Chapter 6

Discussion

In this study, we have developed the MutMatch statistical methodology that can detect and quantify the interaction between selection on mutations and other genomic features (either genetic or non-genetic in nature), thus providing a new framework and tool to study changes in selection effects and epistasis in human cancer genomes.

We next applied this developed MutMatch method to study the interaction between selected mutations and gene copy number changes in cancer genes across cancer types in various datasets. We used the baseline mutation rate derived from the neighboring genes or low-impact nonsynonymous mutations in the same gene, thus rigorously controlling for the confounding effect of gene dosage on the apparent mutation burden.

6.1 Negative selection in cancer genomes

The extent to which negative selection changes mutation rates in cancer genomes has been a much-debated topic in recent years. The common ground is that genomic signatures of negative selection are very subtle [165, 9, 91, 133] and so negative selection does not play a big role in tumor evolution and the majority of the mutations in the genome are not deleterious. According to this view, a somatic genome is accumulating

mutations mostly under neutral or occasionally under positive selection pressure [9]. This goes in agreement that the global selection estimate in the genome aggregated across all genes is neutral or slightly positive [9, 133, 91].

However, several studies were focusing on finding negative selection in the cancer genome [166, 112, 167, 89, 91, 133, 33, 168], since identifying negatively selected (thus tumor-essential) genes would reveal therapeutic targets. For example, Zapata et al. (2018) showed that a state-of-the-art method for the detection of selection in human cancers [9] may overestimate selection in per-gene analyses [133]. As a consequence, the amount of negative selection in cancers is underestimated. They reported a significant negative selection signal in 25 genes, and 668 essential genes with a $dN/dS < 0.5$ cut-off [133]. This set of essential genes was enriched for genes involved in protein synthesis and maturation, as well as genes that participate in molecular transport [133].

Further, parts of genes that produce MHC-exposed native epitope sequences were under negative selection, thus decreasing the number of potential new neo-antigens [133]. Consistently, frameshifting indel mutations that escape Nonsense-Mediated mRNA Decay (NMD) silencing might be under negative selection [169]. This suggests that escaping immune surveillance is an important ability of human tumors that allows them to avoid the immune response. However, as demonstrated by Van den Eynden et al. (2019), the signals of negative selection in the immunopeptidome become weak or absent when considering mutation signatures with respect to the trinucleotide mutation types. They conclude that it can be caused by artifacts acting via differential trinucleotide composition of certain genes and via differential mutational signature activity on tumors, which is particularly relevant in skin cancers [94].

One of the challenges in the identification of tumor-essential genes is that damaging mutations that should, in principle, be negatively selected can act in a recessive manner. In other words, the presence of an intact gene copy in a cell with a wild-type function masks the deleterious effect of the mutation. By searching for purifying selected genes in hemizygous regions, that is those bearing a deletion or a loss-of-heterozygosity, signatures of negative selection may be more evident [9, 34]. As an example, Van den

Eynden et al. (2016) found signals of negative selection in the *POLR2A* gene and genes encoding for protein complex members [89]. This suggests that negative selection may act on the change in the dosage of proteins in the protein complex. Nevertheless, this negative selection might also be explained simply by the essentiality of the protein complexes. 41 % of protein complexes contained at least 70 % of genes that were shown as essential in CRISPR-Cas9 screens [170, 89].

Overall, the presence of negative selection and the strength of the selection against deleterious variants from cancer genomes have been unclear. We addressed this problem by testing for signals of negative selection in known essential genes and oncogenes.

6.1.1 Negative selection on cell-essential genes

Our methodology suggests that mutations in cancer genomes can commonly be both positively and negatively selected. In particular, core essential genes [130] had a depletion in mutation rates compared to the neutral mutation rate baseline derived from neighboring genes. Neighboring genes that are located within the deleted or gained locus rigorously control for any confounding effect due to gene dosage changes, wherein these effects are modeled separately in the regression framework we have implemented. Moreover, to exclude the effect of unknown confounders, we compared the bias-corrected selection effects ω with those from random (putatively non-selected) genes, and also observed a negative shift for essential genes. Notably, this was observed already in the diploid state and it was not necessary to focus only on hemizygous regions as earlier. Our results suggest that heterozygous somatic mutations in essential genes can also sometimes be subject to purifying selection.

The list of 684 reference human core essential genes used in this study was obtained through genome-scale knockout screening performed by CRISPR-Cas9 technology [130] that typically introduces homozygous disruptions. These genes are likely to be essential in all cell types and encode essential functions that are needed for cell survival, for instance, mRNA splicing and protein translation.

Interestingly, the negative selection signal in our analysis was observed

only when comparing mutation rates across all copy number states (while controlling for DNA dosage effects), or in pan-cancer analysis with controlling for copy number state across 13 "elite" cancer types with a high number of available mutations, suggesting that statistical power due to low sample size is limiting to identifying negative selection signatures.

To consider possible cancer type-specific biological effects or technical biases, we have repeated the estimation of selection effects, after the exclusion of every cancer type from the pan-cancer analysis. This was particularly important to control for the case when a cancer type has a high mutation number that can influence the pan-cancer signal, and in which mutation rates may not be modeled accurately by our method.

The negative selection signal remained even after excluding samples with melanoma. Mutational processes in melanoma are mainly induced by UV light and are characterized by the extended mutational signature with a context longer than a trinucleotide [77]. Therefore, trinucleotide mutation spectra might be not sufficient to capture the mutation rate variability between sites in some cases. Inaccuracies in the mutation rate model can create a systematic bias and an illusion of differences in mutation rates between the test and control groups [9]. Future directions of exploring the landscape of selection in somatic cells should control for the extended contexts such as pentanucleotides and heptanucleotides, shown to have relevance to human mutation [77, 78].

Although the exclusion of melanoma samples reduced the difference in selection estimated between essential and random genes, selection effects of essential genes were nonetheless significantly lower than those of random genes (Benjamini-Hochberg correction for multiple testing, FDR $< 1\,\%$). Importantly, the difference in the pan-cancer analysis was observed for selection effects in diploid state ω but was not significant for coefficient δ that quantifies the epistatic interaction between mutations and a gene loss.

We have observed a weak correlation between the cell-essentiality Computational correction of copy-number effect in CRISPR-Cas9 essentiality screens (CERES) score measured from Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) screening assays on cultured cells and the selection effects across all genes in the genome (Pearson's corre-

lation = 0.07) [155]. This demonstrates that cellular essential functions experimentally identified for *in vitro* conditions, do not completely match the essential functions in a tumor, where the environment is different than in cell culture (e.g. pressures from the immune system, and competition for resources, both absent from cultured cells). Certain genes, although having a neutral CERES score (which suggests that their function in cell lines is not essential), demonstrated negative selection in the genomic pan-cancer analysis. Notably, CRISPR inactivates genes homozygously while most somatic mutations do not (and in this analysis, we don't focus on the mutations that do).

These results suggest that although negative selection on essential genes plays a role in cancer genomes, the effect size of selection is relatively small, therefore it can be detected only in very large datasets (such as our pan-cancer analysis with ∼18 000 somatic genomes, or combining mutations from all copy number states together). Additionally, we detected no change in relative selection effect in tumor samples carrying only one gene copy (i.e. the other allele is deleted) indicates that cell-essential genes are often haploinsufficient in tumors. Alternatively, this observation may be a result of the low number of available mutations to properly model the expected local mutation rate upon deletion; this issue would become clearer with larger sample sizes.

6.1.2 Near-neutral selection acting in somatic cells on population-constrained genes

We asked if genes that are under purifying selection in human populations (often referred to as "constraint" in the relevant literature) also are under negative selection in human soma. For that, we calculated the selection effects for all genes in the genome, controlling for copy number state of a gene, across the 13 "elite" cancer types (i.e. those with the highest number of mutations available) in a pan-cancer analysis.

Using ranking of gene essentiality based on intolerance of a gene to a Loss-of-Function (LoF) germline variants in a human population [131, 156], we tested how this gene essentiality score correlates with the somatic selection on that gene in tumors. We found no statistically significant difference in

somatic selection effects in the diploid state between the random group of genes that are not selected in human populations and the top-ranked population essential genes (top-10 %, top-500, top-200, top-100, top-50 of ranked genes were tested).

These results suggest that population-level gene essentiality encodes for a different set of cellular functions than those that are needed for tumor survival and development. Indeed, some population-essential genes may play a key role in the development of the embryo, mutations which cause developmental disorders or fetal anomalies [156]. The essentiality of such genes would presumably be limited to certain stages of organismal development. At the later stages, selection pressure on these genes may relax, and genes accumulate mutations at the same rate as the whole genome on average.

As an additional consideration, the selection of mutations in essential genes is captured at the level of the whole organism or tumor, although the function of the gene might be tissue-specific. Therefore, even if the gene is essential for the tumor with a certain origin, pan-cancer analysis is likely to not detect it if the same gene is not essential for other tissues.

While some cellular functions are essential for the survival of a cell, or a tumor, others depend on the stage of the organismal development, conditions, or cell type. Genes that are essential for clonal tumor growth might be not essential at the organismal level, and vice versa. We observed only a weak correlation between the essentiality metrics for the population and cell-line essentiality metrics, which demonstrates that the number of genes that are simultaneously under a purifying selection in cell lines and a human population is not large.

Our results show that tumor-essential genes have more in common with cell-line essential genes obtained with in vitro experiments rather than with genes that are essential at the organismal level, that were identified by a dearth of LoF variants in the human population.

Further research in this field will provide additional insights into cancer essentiality and vulnerabilities that can be exploited for cancer therapies. An interesting question is to identify genes essential to each tumor and compare them to the cell line-specific essential genes, which would make

good targets for personalized therapy. Additionally, the relationship between different levels of gene essentiality (organismal level, cellular level, and tumor-essential genes) can be studied.

6.1.3 Signatures of negative selection in oncogenes

Oncogenes (OGs) and Tumor Suppressor Genes (TSGs) are the main groups of genes that positively contribute to tumorigenesis. In contrast to TSGs that have to be inactivated, the OGs need to acquire Gain-of-Function (GoF) mutations to increase the fitness of a cell. mutations to increase the fitness of a somatic cell. The dependency of cancer cells on the maintained upregulated activity of certain OGs after they are mutated, called oncogene addiction, has been established by Weinstein et al. (2002) [171]; this principle is already successfully exploited for cancer therapy (for example, for *BRAF*-mutant melanoma by *BRAF* inhibitors). Negative selection against NMD-triggering nonsense mutations in OGs has been demonstrated by Lindeboom et al. (2016) [112]. Similarly, nonsense mutations were depleted in OGs as shown by Bányai et al. (2021) [97].

These results suggest that the function of the OG is vital for the survival and proliferation of cancer cells. We asked the question if the net distribution of mutations in OGs is the consequence of the simultaneous activity of two selective forces: positive selection that increases the frequency of cancer-driving mutations and purifying selection that removes from the mutation pool the cancer-blocking, deleterious mutations.

The majority of mutations in OGs are located in hotspots, which can be selected (increasing the fitness of a cell because of a GoF effect), or in some cases mutational/non-selected mutation accumulation occurs due to other reasons such as high propensity to DNA damage [83]. The regime of selection that acts on the mutations outside of hotspots in OGs has been less studied. We hypothesized that negative selection in OGs might be acting not only on nonsense mutations but also more generally against all nonsynonymous mutations, including the very numerous missense changes, located outside of hotspots of the gene (Figure 6.1).

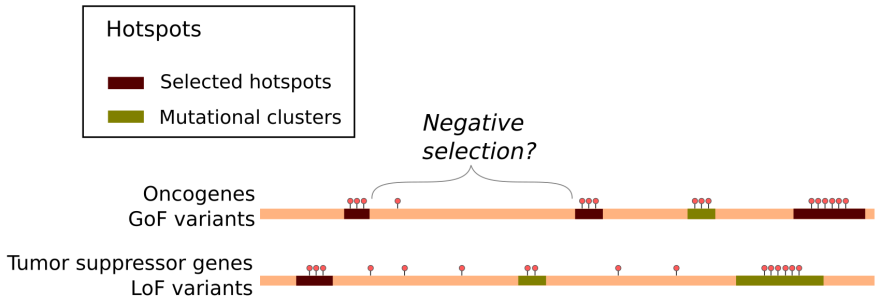

Figure 6.1: Distribution of mutation rates in oncogenes, as opposed to tumor suppressor genes, might reflect both positive and negative selection on mutations.

To check this, we considered mutations within and outside of (1) known selected hotspots identified by Trevino (2020) [82] and (2) in-house detected, permissive set of sites with recurrent mutations (these include the selected and some mutational hotspots; detailed in Section 7.6.1). While the first, strict set of hotspots contains sites under selection, the genic regions that do not include Trevino hotspots might still have an increased mutation rate due to the mutation rate heterogeneity at the sub-gene level (i.e. mutational hotspots). To exclude the influence of such confounders we tested for signals of negative selection in genic regions located outside of our permissive set of hotspots.

Confirming our hypothesis that hotspots-adjacent areas in OGs have depleted mutation rates because of the negative selection, we observed a stronger signal of negative selection in OGs than in random genes or in TSGs. Interestingly, negative selection appeared to be stronger in noncognate cancer types (i.e. those where OGs are not commonly affected by GoF mutations). We have questioned whether those cancer types were indeed noncognate by estimating the selection while limiting to the selected Trevino hotspots.

We found that part of the genes with the negative selection acting on the outside of hotspots was selected in hotspots. This suggested that those cancer types were, in fact, cognate even though our original definition, based on prior databases, did not recognize them as such. This exam-

ple illustrated that mixing signals from positive and negatively selected mutations can lead to underestimation of the selection strength, both for positive and negative.

The possibility of underestimation of negative selection was noted in the paper by Martincorena et al. (2017) if the numbers of mutations that are negatively selected and positively selected are balanced [9]. Nevertheless, they found it unlikely that a large number of genes would have this balance.

Here we show that in the group of oncogenes this type of balance is often present due to the opposing forces that preserve the functionality of essential parts of the gene and simultaneously positively select for gain-of-function mutations. As a consequence of this, we observed that many *de facto* cognate cancer types might be not identified while considering hotspot and non-hotspot sites jointly, as the positive selection signal is diluted. The correct re-annotation of cognate cancer types can help to increase the number of anti-cancer therapies that inhibit oncogenic pathways by repurposing the existing drugs.

As an example, we showed that mutations in *EGFR* gene apart from known cognate cancer types (lung, brain) are also selected for in other, previously unreported cancer types (esophagus, luminal subtype of breast cancer). We also showed that many other genes, including *MYC*, *MYCN*, *MTOR*, *KIT*, have a mutation pattern shaped by both positive and negative selection, which leads to the underestimation of positive selection of these genes in cognate cancer types.

The proportion of the mutations that are removed by the negative selection depends on the cancer type. Because of that, only those cancer types where positive selection fixes more mutations (in hotspots) than the negative selection removes (out of hotspots), appear as selected (or cognate) cancer types. We believe such differences can be driven by the context-dependent negative selection acting on some oncogenes in some tissues. The mechanisms of such cancer-specific dependency on the oncogene function remain to be explored in future work.

6.2 Mutant allele imbalance of cancer genes

The interplay between the selection of carcinogenic mutations and copy number alteration events has been intriguing researchers for decades. The pioneering work of Alfred Knudson on the *RB1* tumor suppressor in heritable retinoblastoma inferred that the inherited variant must be followed by a somatic "second hit" event [31], which is commonly a copy-number deletion. The relationship of how mutations become more or less advantageous depending on the copy number state of a gene in a cell can shed some light on the mechanisms of actions of the mutations, or function of a gene [31, 30, 33]. For example, gene losses lead to increased negative selection in a group of cancer-essential genes [9, 89], which is however not observed in tumor samples that have more than one copy of a gene (as in samples with a gene locus in the diploid state or in samples that have undergone a Whole Genome Duplication (WGD)) [89, 34].

Several studies have been made to address and systematically classify the relationship between Copy Number Alteration (CNA) events and the selection of point mutations at the gene level [108, 32, 172]. In a recent study by Ciani et al. (2022), they found that more than half (56.4 %) of mutant allele imbalance was attributed to the CN-Loss Of Heterozygosity (LOH), which would be classified as wild-type copy number [172]. Nevertheless, the effect of changing the gene dosage on the strength of selection remains understudied, since previous studies of allelic imbalance did not, to our knowledge, implement mutation rate baselines that account for the rate of neutral mutations. Controlling for the differences in the local mutation rate that are caused by CNA events, altering the DNA quantity in a locus, is crucial to decrease the false-positive and false-negative rates.

We addressed this problem by applying the MutMatch method to search for changes in selection strength that are associated with a somatic gene copy number loss or gain event in cancer samples. Independently from the type of the CNA event, we have observed the increased selection for the mutations in OGs and TSGs. While the selection on the allelic balance via gains [108, 32] or heterozygous losses [108] has been previously shown for OGs and TSGs associated with deletions [31], it was to our knowledge not anticipated for the combination of TSGs and copy number gene gains.

6.2.1 Allele imbalance of OGs in cognate cancer types

GoF mutations in OGs – in particular, missense mutations and copy number alterations – have been long considered to act in a dominant manner [122]. For the majority of them with only a few exceptions, OGs that are recurrently mutated in cancers are rarely focally amplified, and those genes which are often focally amplified happen to be not frequently mutated. Therefore, the mutations almost always are present in a heterozygous manner, which implies that they are dominant [122].

Distinct patterns of mutant allele imbalance have been reported by Bielski et al. (2018). By directly estimating the number of copies of the mutant and the wild-type alleles in a large-scale analysis, they showed that these numbers across cancer types were unmatched [108]. The mechanisms of allele imbalance varied depending on the cancer type and the OGs, however, a similar proportion of mutations in studied OGs was attributed to a copy number gain, or, less expectedly to a deletion ($\sim$33 % for each).

Independently, a positive interaction between copy number gene gains and selection of the mutations was shown for some OGs, that were called "two-hit gain "genes, by analogy with the well-established concept "of two-hit" loss TSGs where one allele is mutated and the other deleted or affected by the copy number neutral loss of heterozygosity [32].

In both of these studies, the majority of OGs showed a positive selection of mutations in tumor samples with changes in copy number states. In other words, the mutant allele almost always had a bigger number of copies than the wild-type allele. A few important exceptions that did not follow this pattern were identified by Bielski et al. (2018) [108] and included genes involved in splicing machinery, such as *SF3B1*, *U2AF1*, and *SRSF2*. These OGs were labeled as haplo-essential genes that should retain one copy of a wild-type allele to maintain homeostasis inside a cancer cell; this represents a special case of negative selection acting upon OGs.

The results obtained in our current study are in agreement with this. However, we performed a systematic estimation of selection interaction with copy gains for a larger list of cancer genes and focused on the activating mutations (presumably located in hotspots). This allowed us to get a larger estimate for the proportion of genes that have a higher selection

strength upon copy number change.

More than half of OGs in cognate cancer types had a higher mutation rate in samples with a non-diploid gene state. In contrast, 41 % of OGs exhibited allelic imbalance in the study of Bielski et al. [108] (of 69 examined OGs).

The positive relationship between the selection and the copy number events challenges the common consensus of the fully dominant nature of mutations in OGs. Our data suggest that, while some mutations in OGs might indeed be dominant, others (including the mutations in hotspots) represent a class of weaker, incompletely dominant mutations. This can explain why the increase in the proportion of the mutant allele leads to the increase in selective advantage. In this case, given that mutant allele dosage remains to be the same in case of gene loss, selection of the mutant allele should not be increased in tumor samples with gene deletion. In case when the wild-type allele has an additional inhibitory effect on the activity of the mutant allele, an increased selection of GoF mutations in samples with gene loss is expected.

In our analysis, only 40 % of genes-cancer type combinations with positive interaction between CNA gain and mutation rate, however, did not demonstrate the increase of positive selection in tumor samples with CNA loss, while in the remaining cases (which are the majority) the OG mutations did interact also with CNA loss. This is consistent with a mechanism that in the majority of such cases, even if the mutations are incompletely dominant, there is an additional negative effect of the wild-type allele of the OG on tumor fitness.

On the other hand, the selection change can be explained by the antagonistic, inhibitory action of the wild-type allele of an OG, as shown for *RAS* genes [126, 127, 128, 29]. Similarly, the tumor suppressive effect of the wild-type allele of a TSG can be lowered by decreasing its proportion. Decreasing the proportion of the wild-type allele can, in turn, be achieved either via deletion of the wild-type allele or by acquiring an additional mutant gene copy.

The two described mechanisms are not mutually exclusive. Weak, incompletely dominant GoF mutations that compete against the tumor-

suppressive activity of the wild-type allele of the OG might be a natural mechanism for limiting the opportunity to acquire cancerogenic mutations. Further research to disentangle those models of mutation mode-of-action and estimate what proportion of the tumors is driven by each of the mechanisms in each gene is warranted.

6.2.2 Allele imbalance of TSGs in cognate cancer types

Increased selection of mutations in TSGs in tumor samples with CNA gain compared to the samples in the neutral copy number state possibly indicates the dominant-negative character of these mutations. Dominant negative mutations are characterized by the inhibitory effect of the mutant protein on the function of the wild-type allele [173]. This, for example, can be achieved if the function of the protein is limited by the availability of the substrate, or if the protein acts in a homomeric complex. In these cases, the wild-type activity will be inhibited by forming a non-functional multimer, where mutant subunits poison the whole complex, or due to competitive inhibition [173, 174, 175].

As opposed to dominant negative mutations, for two-hit LoF mutations increasing the fitness of a subclone is expected only after the inactivation of both alleles. The effect of inactivating mutations is not evident in cell fitness when they target only one of the alleles.

Our analysis suggested that only a few TSGs in cognate cancer types have a two-hit mechanism of inactivation, according to this definition. While many of them do show an increased selection on mutation rates in the hemizygous state, the majority of them are also selected in the diploid state without any CNA. Therefore, we can conclude that the increased fitness is explained by an increased proportion of a mutant, dysfunctional protein which leads to a more complete aberration of a tumor-suppressing pathway. In other words, a haploinsufficient TSG – which loses healthy phenotype when one gene copy is inactivated – can be at the same time a two-hit gene. Thus, these two mechanisms are not mutually exclusive.

Dominant-acting mutations can also explain the character of inactivation of one-hit TSGs. The level of dominance (i.e. the efficacy of inhibiting the wild-type allele) varies across different TSGs. For example, selection

on the diploid state for *TP53* mutations was one of the strongest, while in tumor samples with an additional gene copy or with a gene loss the selection did not increase greatly (with exceptions in two cancer types - kidney and multiple myeloma cancers). In contrast, selection on *PTEN* showed a gradient: selection in the samples with two gene copies was much lower than in samples with one gene copy in PRAD, LGG, SKCM, BLCA while the opposite pattern was observed in cancer types GBM, UCEC, and LUSC.

Taking these results into consideration, we suggest that two mechanisms of gene inactivation – through dominant negative-acting mutations or the elimination of both copies of the gene – may be two ends of one spectrum. In other words, because often the efficacy of inhibition of the mutant alleles on the functionality of the wild-type allele can be low, thus inactivation of both alleles is needed to obtain a malignant phenotype. The incomplete dominant-negative effect (which could be considered a kind of hypomorphism) in such mutations causes selection pressure to additionally increase the proportion of the mutant allele. If the inhibition of the wild-type allele by the mutant allele is very effective, increasing the proportion of the mutant allele by selecting for the mutations has a small effect on the fitness.

The relation between the selection in the diploid state and the gain/loss copy number states appears to be cancer-type specific for each gene. As shown in a recent study by Park et al. (2021), switching between one-hit and two-hit genes may depend on the functional pathways active in a cell, which can be specific to the lineage and the cancer type [176, 32].

6.3 Classification of genes based on the selection patterns across copy number states

Using the Principal Component Analysis (PCA) data-driven approach, we have identified at least three components that represent different somatic selection signatures acting on cancer driver genes. With a certain degree of simplification, these components reflect (i) "overall" selection on a gene, (ii) the two-hit loss, and (iii) the two-hit gain selection. Additionally, the

selection of nonsense mutations may be separated into the fourth component.

We have observed that cancer genes do not show well-separated clusters of two-hit genes versus one-hit genes or selected genes versus not selected genes. Rather, they form a continuum of driving potential, and, more surprisingly, seem to form a spectrum between one-hit and two-hit mechanisms of (in)activation meaning that both mechanisms may be operative in the same gene with varying frequencies across tumors.

This can be compared with a recent classification by Park et al. (2021) described four distinct groups of cancer genes genes by the mechanism of action of selected mutations: one-hit genes, two-hit gain genes, two-hit loss genes, and two-hit gain and loss genes [32]. We believe these variable classification schemes might result from the increased size of our dataset, or by rigorously controlling for the confounding effect of CNA on the number of mutations that is not due to selection (but caused trivially by the changes in the amount of available DNA); the latter was enabled by the use of neighboring genes mutation rate baseline in our method. We do recognize that the precision of selection effects estimated with the MutMatch method could be limited because of the high number of variables that are used to control for the heterogeneity of mutation rates using the neighboring genes baseline.

The selection types summarized in some of the principal components may be partially attributed to the difference in the number of mutations in each cancer type dataset. This suggests that statistical power in our data is still limiting. This means that the incorporation of the newly available datasets can further improve the method's sensitivity and precision, thus enabling a more robust classification of cancer genes.

Chapter 7

Methods

7.1 Mutation and copy number data collection and processing

We collected mutation and copy number data for two aggregated datasets in this study: a discovery cohort and a validation cohort.

The **discovery cohort** comprised a mixture of Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) datasets from:

- WES somatic Single Nucleotide Variants (SNVs) from the The Multi-Center Mutation Calling in Multiple Cancers (MC3) Project [177].
- WGS somatic SNVs from the The Cancer Genome Atlas Program (TCGA) consortium [85].
- WGS somatic SNVs from Hartwig Medical Foundation (HMF) project [86, 178].
- WGS somatic SNVs from Pan-Cancer Analysis Of Whole Genomes (PCAWG) dataset [87, 179].
- WES somatic SNVs from PCAWG dataset [87, 179].

- WGS somatic SNVs from Personal Oncogenomics project (POG570) program [88, 180].
- WES somatic SNVs from The Clinical Proteomic Tumor Analysis Consortium (CPTAC)-3 program [181, 182, 183].
- WGS somatic SNVs from MMRF-COMPASS study [184, 183].

The mutations that were called against the human version assembly GRCh38 were were converted to the hg19 reference genome using LiftOver [185]. Variant (nonsense, missense, synonymous) were annotated using *ANNOVAR* software [186]. Highly mutated samples and samples with a high fraction of indels were separated from the rest of the samples for UCEC, CESC, COREAD, ESAD, ESCA, STAD, UCS cancers into separate subtypes denoted as "POLE" or "MSI" (Table 9.1). Cancer types between different datasets were matched to increase the sample size for each of them.

Altogether, we collected the genomes of 23 000 tumor samples with mutations from 117 cancer types; the number of tumor samples per each tumor type was ranging from 1 to over 1000 for PRAD (prostate adenocarcinoma), BRCA-Lum (a luminal subtype of breast cancer), COREAD (colorectal adenocarcinoma), MM (multiple myeloma), kidney, PAAD (pancreatic adenocarcinoma), SKCM (skin cutaneous melanoma). The median number of tumor samples was 57. The average number of SNV mutations in each cancer type was 28 805, and the median number of mutations is 4136. The cancer types with the biggest number of mutations are listed in Table 9.1.

For the **validation dataset**, we downloaded mutation calls for 90 713 tumor samples across 75 cancer types from MSK-IMPACT and the Dana Farber Cancer Institute (DFCI) Oncopanel of the American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange (GENIE) (Release 11.1; syn32309524) [187, 188, 189]. In these studies, only a limited number of cancer genes were sequenced. We determined the list of cancer genes that were targeted in both of these cohorts such that the inter-gene differences of selection effects could not be caused by the differential coverage of the two datasets.

We collected Copy Number Alteration (CNA) data estimated with *GISTIC2* or *purple* programs for the majority of the samples listed above, covering 17 644 samples of the discovery cohort and 89 243 samples of the validation cohort. The estimates of the gene-level copy number status were binarized according to the recommended sample-specific thresholds for the GISTIC2 copy number levels.

For the discovery cohort, only low-level gains or hemizygous deletions were considered for estimation of conditional selection upon CNA alteration event. For the validation cohort, we have aggregated the estimates for the copy number status of the gene. This way, any sample with the number of gene copies greater than 2 was considered to be in a gain state, and similarly for the samples in the deleted state.

7.2 Mutation frequencies

For every nonsynonymous (missense and nonsense) mutation from the MC3 dataset Variant Allele Frequency (VAF) was calculated as F = $\frac{n_a}{(n_a+n_r)}$, where F is the variant allele frequency, n_a and n_r are the numbers of alternative and reference reads.

To control for the purity of the tumor sample (that changes the final number of sequenced reads bearing a mutation) we adjusted the formula in the following way:

$$F_{adj} = \frac{n_a}{(n_a + n_r)} \times \frac{1}{P}, \tag{7.1}$$

where P was the purity estimate from Consensus measurement of Purity Estimation from *TCGAbiolinks* R package [159, 160, 161].

7.3 Annotation of cognate cancer types

To determine which cancer types were cognate (i.e. where cancer genes are positively selected), we used different sources of data: annotation available in the known databases such as CGC or MutPanning [95, 132], and

the selection estimates derived with the neighboring genes baseline in the Discovery cohort.

The latter approach carries fewer risks of missing cognate cancer types due to the different labels between annotation sources and our data, or uncertainty in the case when annotations are not detailed to specify which cancer subtype is cognate. Moreover, this helps to make sure that we do not focus on gene-tumor pairs that are cognate, but have too few mutations in the datasets due to the low number of samples in these cancer types.

We have estimated the selection effects in the discovery cohort for all the cancer genes in the analysis across all copy number states (without controlling for CNA variable). Cancer types where a gene was positively selected ($\omega > 0$ at a threshold of FDR $\leq 25\,\%$) were considered to be cognate cancer types for this gene. To exclude from noncognate cancer types (where cancer genes should not be selected) false-negative hits, we required FDR $\geq 75\,\%$. The rest of the cancer types that did not fall into either of these categories were separated into the "unclassified" group.

7.4 Simulation of bias in regression estimates

We modeled the bias in the estimation of regression coefficients using the simplest neutral model that assumes no correlation between the mutation rate and the genomic locus ($\omega = 0$). The mutation counts are accumulated in the genomic loci t proportionally to its size (length of DNA). The ratio between the DNA lengths of the two genomic loci had one of the next values: $\frac{1}{S}$ or S, where $S \in \{1, 2, 5, 10, 20, 30, 40, 50\}$. $S = 1$ meant that there was no difference in the lengths between genomic sites encoded with the variable t.

The sizes of the genomic loci varied in our simulation, with the CGF values between $1/50$ to 50. The mutation counts Y were simulated using the Poisson distribution with the mean λ:

$$Y \sim \text{Poisson}(\lambda) \tag{7.2}$$

The mutation rate α was the same across genomic sites; the observed

number of mutations λ only depended on the DNA length of the locus r:

$$\lambda = \omega t + \alpha + \log r \tag{7.3}$$

We included the base mutation rate α as the intercept of the model, estimating the model across for different values of α: $-14 \leq \alpha \leq -1$.

Next, we fit the model 7.3 that generated the mutational counts to estimate $\hat{\omega}$. The averaged values of $\hat{\omega}$ from 50 fitting procedures (with random seeds) are presented in Figure 2.3.

7.5 CADD scores

We downloaded a bigWig file that contained pre-computed PHRED-like $-10 \times \log_{10} \frac{rank}{total}$ scaled Combined Annotation-Dependent Depletion (CADD) scores for each genomic position (v.1.4) [142, 190]. The highest CADD score of any 3 possible substitutions is displayed in this file with higher values indicating a higher level of deleteriousness of the variant. Scaled CADD-scores assign value 10 to the top-10 % of all the CADD scores in the reference genome, value 30 to the top-0.1 % and so on.

To separate regions where mutations are likely to have a functional impact versus regions that are evolutionary unconstrained, we used a cut-off of 20. This way, positions in a gene with top-1 % most deleterious mutations were tested for selection using a background mutation rate model from unconstrained gene regions where CADD score was less than 20.

7.6 Genomic filters

7.6.1 Hotspot detection

We define a hotspot based on the codon-specific frequency of mutations in a discovery cohort with the cut-off of 2 mutations per codon (Figure 7.1). This threshold yielded the recovery of 91.6 % of previously identified

selected hotspots [82] and 59.7 % of mutations detected by exploration of mutation clustering in the 3D structure of a protein [191].

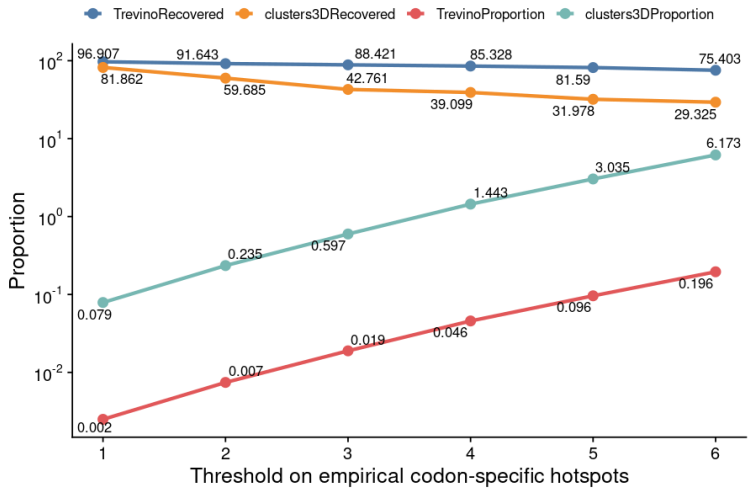

Figure 7.1: Recovery of known hotspots for different hotspot cut-offs in a discovery cohort. Hotspots were defined based on the number of mutations at the codon level (X-axis). For different thresholds we calculated which fraction of known hotspots from Trevino (2020) [82] and Gao et al. (2017) [191] was recovered and what is the fraction of these hotspots in our empirical set of hotspots. The cut-off of 2 mutations per codon was chosen, which corresponds to the 91.6 % of recovery of Trevino hotspots and 59 % of recovery of 3D hotspots (0.235 % and 0.007 % of all pemissive hotspot sites, respecitively.

Hotspots tend to cluster in 3D space as they target a functionally important part of the protein, and often they also target a neighboring amino acid. Some 3D clustered hotspots that were not recovered with this cut-off are located in proximity to a hotspot in the discovery cohort. To avoid missing such sites that are closely located to a detected hotspot, but have fewer mutations (due to the limited number of mutations in the discovery dataset), we lengthened our hotspots to include them. If another mutation was found within 3 codons upstream or downstream from the hotspot, the hotspot was extended to include this mutation. The median length of the hotspots defined this way was 12 nucleotides, or 4 amino acids (Figure 7.2). This yielded in recovery of 75.2 % of 3D clustered hotspots [191] and

95.4 % of Trevino hotspots [82].

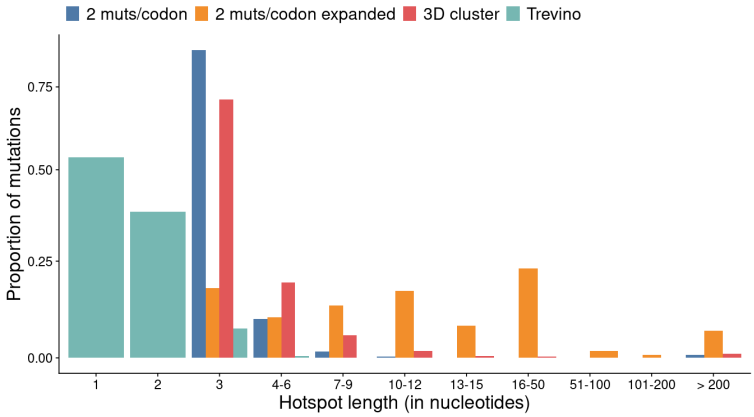

Figure 7.2: Distribution of the hotspot sizes defined by different methods. Median size of the empirical hotspots with expansion was 12 nucleotides, for Trevino hotspots [82] was 1 nucleotide and for 3D cluster hotspots was 3 nucleotides [191].

7.6.2 NMD-detected and NMD-evading regions

Genomic regions were split into those where Premature Termination Codons (PTCs) lead to the degradation of the mRNA in a process of Nonsense-Mediated mRNA Decay (NMD) or those where nonsense mutations lead to a translation of a truncated protein sequence. The efficacy of NMD for PTCs in a human model was predicted using the NMDetective algorithm [192]. The scores of NMDetective-A were obtained from [192, 193]: regions with the score >0.52 were classified as NMD-detected regions, and the rest was classified as NMD-evading regions.

Chapter 8

Supplementary Figures

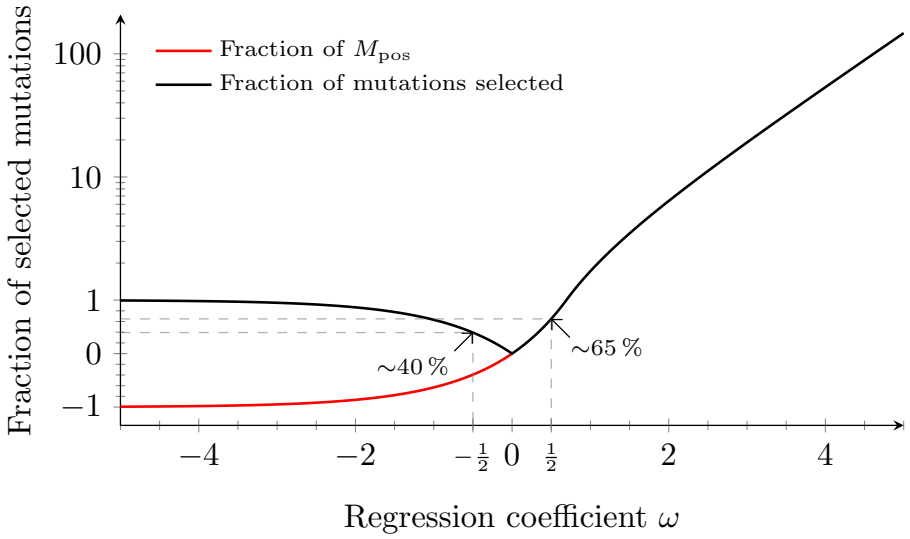

Figure S8.1: Fraction of selected mutations as a function of selection effect ω. The fraction of positively selected mutations in a gene M_{pos} (the difference between the observed number of mutations in a gene M_{observed} and the expected number of mutations M_{expected} relative to the M_{observed}) is shown in red. The fraction of mutations that were either lost or are driver mutations (relative to the M_{observed}) is shown in black.

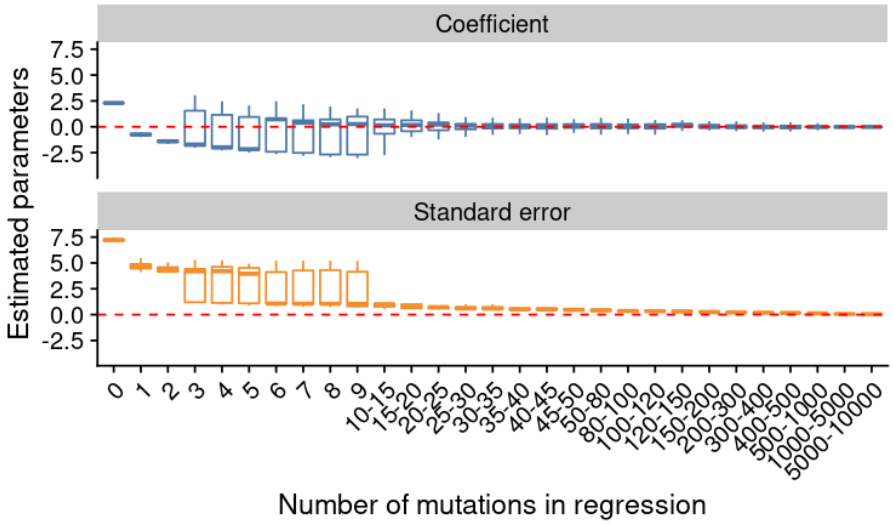

Figure S8.2: Estimation error of regression. The data was simulated using a neutral selection model ($\omega = 0$, red dotted line) with $CGF = 10$. Estimated coefficients $\hat{\omega}$ are shown as a function of the total mutation number that was in the regression.

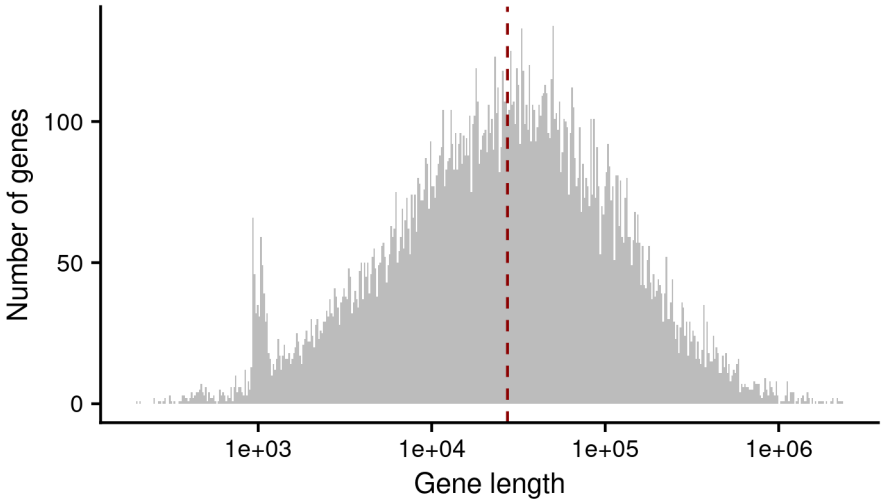

Figure S8.3: Distribution of gene lengths in the human genome. Coordinates of the genes from genome version GRCh37 assembly of the human genome were used to calculate gene lengths. The red dashed line marks the median of the distribution (corresponds to 27 336 bp).

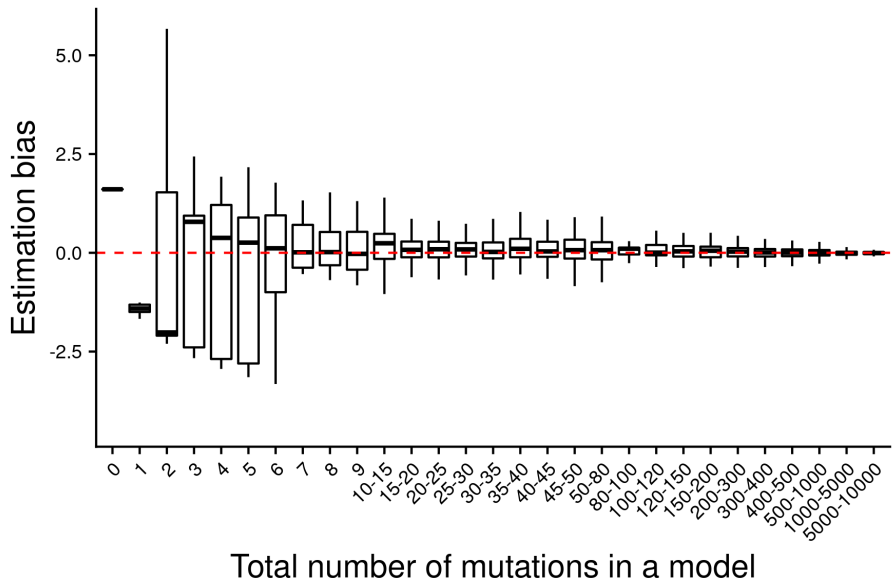

Figure S8.4: Estimation error of regression. The data was simulated using a neutral selection model ($\omega = 0$, red dotted line) with $CGF = 5$. Estimated coefficients $\hat{\omega}$ are shown as a function of the total mutation number in a model (the sum of the mutation counts in the tested and control groups).

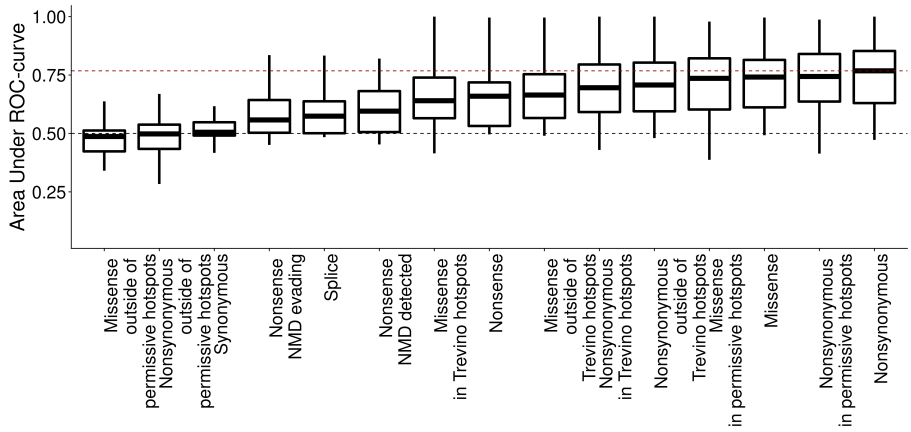

Figure S8.5: Area Under the Receiver Operating Characteristics (AUROC) scores for different mutation sets. Random genes versus mutated Census cancer genes in cognate cancers. The dotted line is the maximal median AUROC score per group in tested sizes.

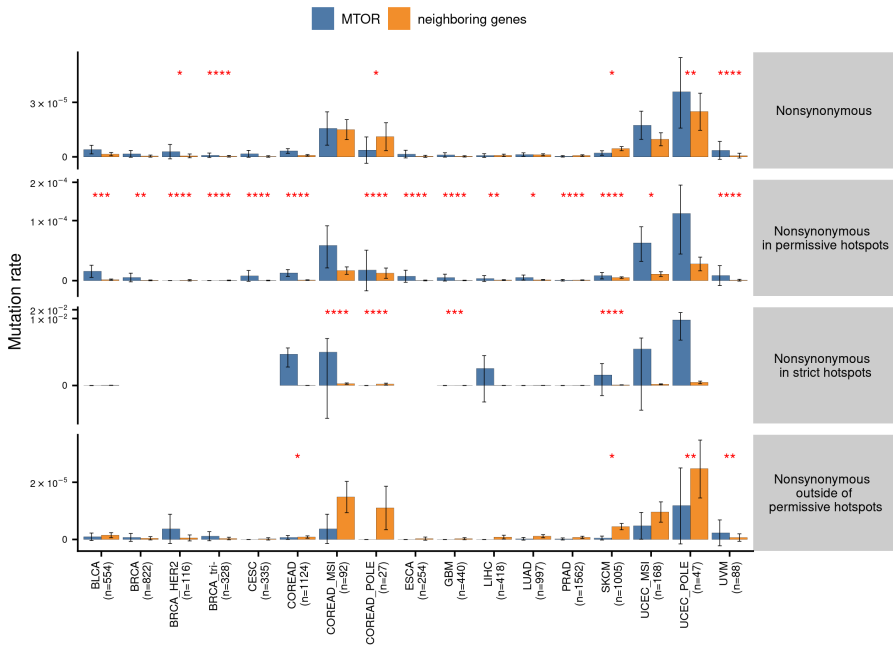

Figure S8.6: Mutation rate of *MTOR* gene and its neighbors across copy number states. Mutation rate is calculated mutation counts divided by the length of the gene without controlling for Trinucleotide Mutation Spectra (MS96) across copy number states; estimated mutation rate is plotted with 95 % confidence interval. Asterisks denote cancer types, where the gene was positively selected. The level of significance of the difference is encoded: '$*$' for FDR ≤ 0.05, '$**$' for FDR $\leq 1 \cdot 10^{-2}$, '$***$' for FDR $\leq 1 \cdot 10^{-3}$, '$****$' for FDR $\leq 1 \cdot 10^{-4}$. Y-axis is log-transformed.

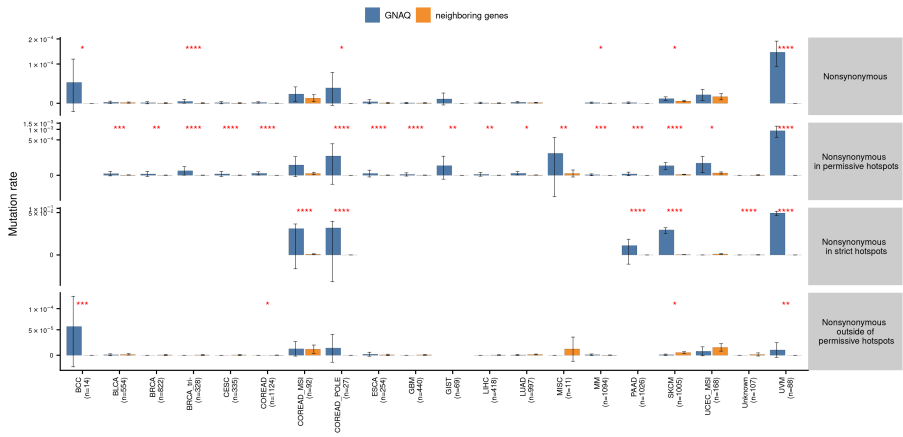

Figure S8.7: Mutation rate of *GNAQ* gene and its neighbors across copy number states. Mutation rate is calculated mutation counts divided by the length of the gene without controlling for MS96 across copy number states; estimated mutation rate is plotted with 95 % confidence interval. Asterisks denote cancer types, where the gene was positively selected. The level of significance of the difference is encoded: '$*$' for FDR ≤ 0.05, '$**$' for FDR $\leq 1 \cdot 10^{-2}$, '$***$' for FDR $\leq 1 \cdot 10^{-3}$, '$****$' for FDR $\leq 1 \cdot 10^{-4}$. Y-axis is log-transformed.

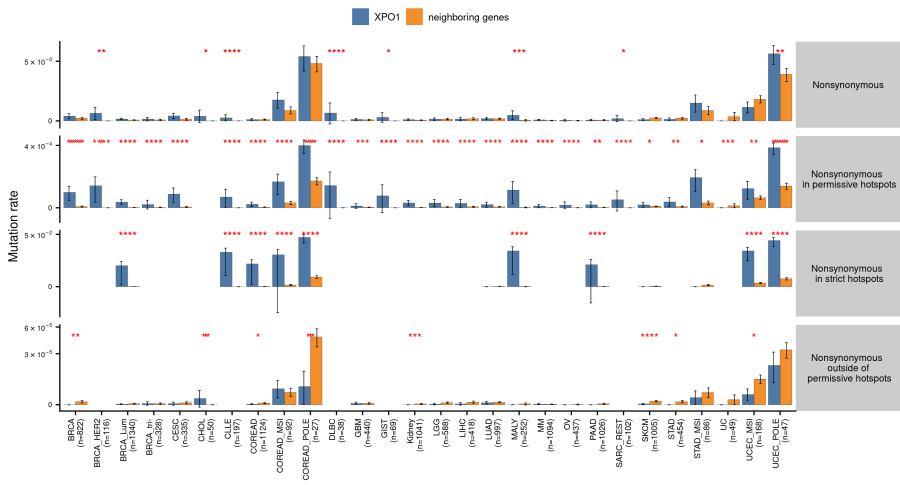

Figure S8.8: Mutation rate of *XPO1* gene and its neighbors across copy number states. Mutation rate is calculated mutation counts divided by the length of the gene without controlling for MS96 across copy number states; estimated mutation rate is plotted with 95 % confidence interval. Asterisks denote cancer types, where the gene was positively selected. The level of significance of the difference is encoded: '$*$' for FDR ≤ 0.05, '$**$' for FDR $\leq 1 \cdot 10^{-2}$, '$***$' for FDR $\leq 1 \cdot 10^{-3}$, '$****$' for FDR $\leq 1 \cdot 10^{-4}$. Y-axis is log-transformed.

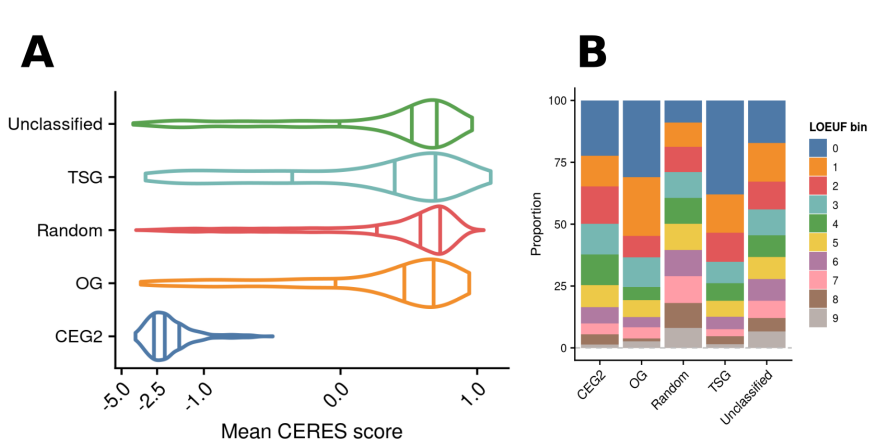

Figure S8.9: Distributions of essentiality scores for main gene classes in this study. A. CERES scores distribution: lower scores correspond to higher cell essentiality in a cell depletion CRISPR–Cas9 essentiality screens. **B.** LOEUF scores distribution: inferred from the population data of depletion of Loss-of-Function (LoF) variants with lower scores correlating with haploinsufficiency.

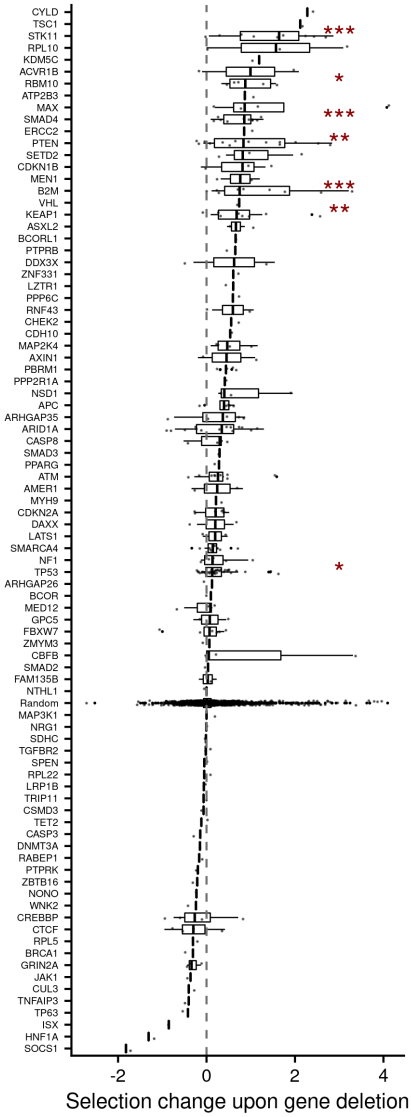

Figure S8.10: The distribution of regression coefficients $\delta_{deletion}$ for Tumor Suppressor Genes (TSGs) in cognate cancer types and the group of random genes. Asterisks denote genes with significantly different distributions from the distribution of a group of random genes (Mann-Whitney test with multiple testing corrections). The level of significance of the difference is encoded: '$*$' for FDR ≤ 0.05, '$**$' for FDR $\leq 1 \cdot 10^{-2}$, '$***$' for FDR $\leq 1 \cdot 10^{-3}$, '$****$' for FDR $\leq 1 \cdot 10^{-4}$

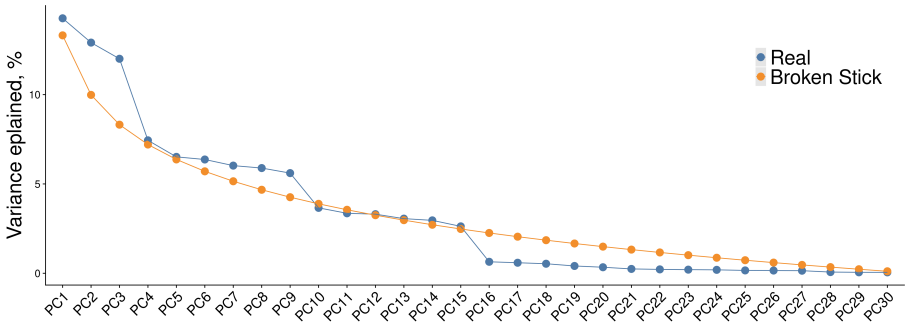

Figure S8.11: Principal components and the variance explained by them The number of significant components (n=3) was decided with the broken stick method [162].

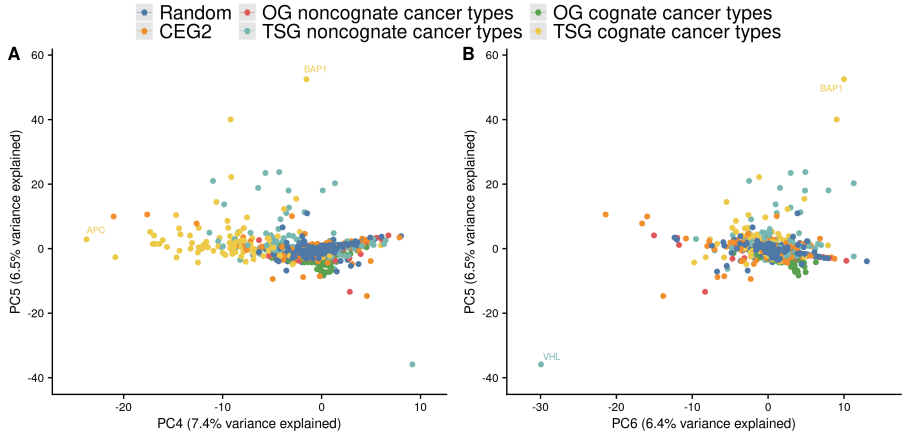

Figure S8.12: PCA analysis of selection effects across cancer types and mutation classes. Principal components 4 to 6 are shown for genes with the 15 most frequently mutated cancer genes: *TP53*, *KRAS*, *APC*, *BRAF*, *PTEN*, *RB1*, *GNAQ*, *PIK3CA*, *VHL*, *GNA11*, *IDH1*, *GTF2I*, *PBRM1*, *ARID1A*, *BAP1*.

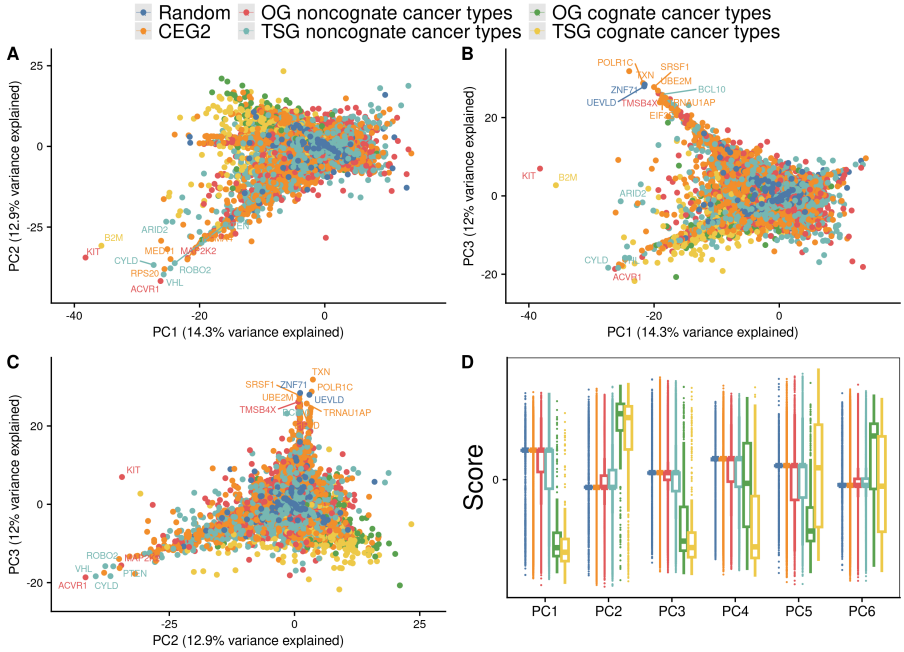

Figure S8.13: PCA analysis of selection effects across cancer types and mutation classes. A, B C. Gene groups in the space defined by the three first Principal Components (PCs). Cancer genes from Cancer Gene Census (CGC) are plotted. Genes with the highest absolute scores are labeled. **D.** Usage of principle components by the gene groups across all cancer types.

Chapter 9

Supplementary Tables

Table 9.1: Cancer types from the discovery cohort with the largest number of nonsynonymous mutations

Cancer	Full cancer name	Number of mutations	Number of samples
SKCM	Skin cutaneous melanoma	499 418	1004
UCEC-MSI	Uterine corpus endometrial carcinoma MSI tumor samples	270 215	168
LUAD	Lung adenocarcinoma	212 630	996
UCEC-POLE	Uterine corpus endometrial carcinoma hypermutated tumor samples	205 066	47
LUSC	Lung squamous cell carcinoma	166 235	700
COREAD	Colorectal adenocarcinoma	128 934	1121
COREAD-POLE	Colorectal adenocarcinoma hypermutated tumor samples	124 671	27
COREAD-MSI	Colorectal adenocarcinoma MSI tumor samples	124 388	92
BLCA	Bladder urothelial xcarcinoma	124 173	554
BRCA-Lum	Breast invasive carcinoma luminal subtype	108 439	1340
HNSC	Head and neck squamous cell carcinoma	99 026	762
STAD-MSI	Stomach adenocarcinoma MSI tumor samples	82 613	86
PRAD	Prostate adenocarcinoma	79 471	1530
PAAD	Pancreatic adenocarcinoma	65 262	1011
BRCA	Breast invasive carcinoma	60 263	810
MM	Multiple myeloma	56 732	1089
LGG	Brain lower grade glioma	53 987	588
Kidney	Kidney cancer	50 870	1041
ESAD	Esophageal adenocarcinoma	50 671	463
NSCLC	Non-small cell lung cancer	49 075	117
STAD	Stomach adenocarcinoma	48 011	450
UCEC	Uterine corpus endometrial carcinoma	46 691	454
GBM	Glioblastoma multiforme	43 163	438
PBCA	Pediatric brain cancer	37 362	617
LIHC	Liver hepatocellular carcinoma	36 784	417
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	36 719	334
SKCA	Skin adenocarcinoma	34 959	100
ESCA	Esophageal Cancer	30 948	253
OV	Ovarian cancer	30 284	436
BRCA-tri-	Breast invasive carcinoma triple-negative subtype	29 983	328
LICA	LICA	26 322	193
MALY	Malignant lymphoma	20 239	252
NET	Neuroendocrine cancer	18 669	263
BRCA-HER2	Breast invasive carcinoma HER2-positive subtype	16 490	116

Bibliography

[1] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. "The cancer genome." In: *Nature* 458.7239 (2009), pp. 719–724. ISSN: 00280836. DOI: 10.1038/nature07943.

[2] American Cancer Society. *Cancer Facts & Figures 2022*. Tech. rep. 2022.

[3] Barbara A. Given and Charles W. Given. "The Burden of Cancer Caregivers." In: *Cancer Caregivers*. Ed. by Allison J. Applebaum. Vol. 2018. Oxford University Press, 2019, pp. 20–33. DOI: 10.1093/med/9780190868567.003.0002. URL: https://academic.oup.com/book/24679/chapter/188086623.

[4] Lauren M.F. Merlo et al. "Cancer as an evolutionary and ecological process." In: *Nature Reviews Cancer* 6.12 (2006), pp. 924–935. ISSN: 1474175X. DOI: 10.1038/nrc2013.

[5] Mel Greaves and Carlo C. Maley. "Clonal evolution in cancer." In: *Nature* 481.7381 (2012), pp. 306–313. ISSN: 00280836. DOI: 10.1038/nature10762.

[6] Antoine M. Dujon et al. "Identifying key questions in the ecology and evolution of cancer." In: *Evolutionary Applications* 14.4 (2021), pp. 877–892. ISSN: 17524571. DOI: 10.1111/eva.13190.

[7] Chris Greenman et al. "Statistical analysis of pathogenicity of somatic mutations in cancer." In: *Genetics* 173.4 (2006), pp. 2187–2198. ISSN: 00166731. DOI: 10.1534/genetics.105.044677.

[8] Iñigo Martincorena and Peter J. Campbell. "Somatic mutation in cancer and normal cells." In: *Science* 349.6255 (2015), pp. 1483–1489. ISSN: 10959203. DOI: 10.1126/science.aab4082.

[9] Iñigo Martincorena et al. "Universal Patterns of Selection in Cancer and Somatic Tissues." In: *Cell* 171.5 (2017), 1029–1041.e21. ISSN: 10974172. DOI: 10.1016/j.cell.2017.09.042.

[10] Marina Salvadores, David Mas-Ponte, and Fran Supek. "Passenger mutations accurately classify human tumors." In: *PLOS Computational Biology* 15.4 (2019). Ed. by Maricel G. Kann, e1006953. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006953. URL: https://dx.plos.org/10.1371/journal.pcbi.1006953.

[11] Susanne Tilk et al. "Most cancers carry a substantial deleterious load due to Hill-Robertson interference." In: *eLife* 11 (2022), pp. 1–22. ISSN: 2050084X. DOI: 10.7554/eLife.67790.

[12] H. J. Muller. "The relation of recombination to mutational advance." In: *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 1.1 (1964), pp. 2–9. ISSN: 00275107. DOI: 10.1016/0027-5107(64)90047-8.

[13] John Haigh. "The accumulation of deleterious genes in a population-Muller's Ratchet." In: *Theoretical Population Biology* 14.2 (1978), pp. 251–267. ISSN: 10960325. DOI: 10.1016/0040-5809(78)90027-8.

[14] Christopher D. McFarland et al. "Impact of deleterious passenger mutations on cancer progression." In: *Proceedings of the National Academy of Sciences of the United States of America* 110.8 (2013), pp. 2910–2915. ISSN: 00278424. DOI: 10.1073/pnas.1213968110.

[15] Douglas Hanahan and Robert A. Weinberg. "The hallmarks of cancer." In: *Cell* 100.1 (2000), pp. 57–70. ISSN: 00928674. DOI: 10.1016/S0092-8674(00)81683-9. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867400816839.

[16] Douglas Hanahan and Robert A. Weinberg. "Hallmarks of cancer: The next generation." In: *Cell* 144.5 (2011), pp. 646–674. ISSN: 00928674. DOI: 10.1016/j.cell.2011.02.013. URL: http://dx.doi.org/10.1016/j.cell.2011.02.013.

[17] Philip J. Stephens et al. "Massive genomic rearrangement acquired in a single catastrophic event during cancer development." In: *Cell* 144.1 (2011), pp. 27–40. ISSN: 00928674. DOI: 10.1016/j.cell.2010.11.055. URL: http://dx.doi.org/10.1016/j.cell.2010.11.055.

[18] Josep V. Forment, Abderrahmane Kaidi, and Stephen P. Jackson. "Chromothripsis and cancer: Causes and consequences of chromosome shattering." In: *Nature Reviews Cancer* 12.10 (2012), pp. 663–670. ISSN: 1474175X. DOI: 10.1038/nrc3352.

[19] Rosina Savisaar and Laurence D. Hurst. "Exonic splice regulation imposes strong selection at synonymous sites." In: *Genome research* 28.10 (2018), pp. 1442–1454. ISSN: 15495469. DOI: 10.1101/gr.233999.117.

[20] Fran Supek et al. "Synonymous mutations frequently act as driver mutations in human cancers." In: *Cell* 156.6 (2014), pp. 1324–1335. ISSN: 10974172. DOI: 10.1016/j.cell.2014.01.051.

[21] Rameen Beroukhim et al. "The landscape of somatic copy-number alteration across human cancers." In: *Nature* 463.7283 (2010), pp. 899–905. ISSN: 00280836. DOI: 10.1038/nature08822.

[22] Smruthy Sivakumar et al. "Pan cancer patterns of allelic imbalance from chromosomal alterations in 33 tumor types." In: *Genetics* 217.1 (2021). ISSN: 19432631. DOI: 10.1093/GENETICS/IYAA021.

[23] Jeffrey R. MacDonald et al. "The Database of Genomic Variants: A curated collection of structural variation in the human genome." In: *Nucleic Acids Research* 42.D1 (2014), pp. 986–992. ISSN: 03051048. DOI: 10.1093/nar/gkt958.

[24] Lukasz P. Gondek et al. "Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML." In: *Blood* 111.3 (2008), pp. 1534–1542. ISSN: 00064971. DOI: 10.1182/blood-2007-05-092304. URL: https://ashpublications.org/blood/article/111/3/1534/25421/Chromosomal-lesions-and-uniparental-disomy.

[25] Cecilia C S Yeung et al. "Impact of copy neutral loss of heterozygosity and total genome aberrations on survival in myelodysplastic syndrome." In: *Nature Publishing Group* 31.4 (2017), pp. 569–580. ISSN: 0893-3952. DOI: 10.1038/modpathol.2017.157. URL: http://dx.doi.org/10.1038/modpathol.2017.157.

[26] Hanna Kryh et al. "Comprehensive SNP array study of frequently used neuroblastoma cell lines; copy neutral loss of heterozygosity is common in the cell lines but uncommon in primary tumors." In: *BMC Genomics* 12.1 (2011), p. 443. ISSN: 1471-2164. DOI: 10.1186/1471-2164-12-443. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-12-443.

[27] James A. Birchler et al. "Dosage balance in gene regulation: Biological implications." In: *Trends in Genetics* 21.4 (2005), pp. 219–226. ISSN: 01689525. DOI: 10.1016/j.tig.2005.02.010.

[28] Silvia Stingele et al. "Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells." In: *Molecular Systems Biology* 8.608 (2012). ISSN: 17444292. DOI: 10.1038/msb.2012.40.

[29] Craig M. Bielski and Barry S. Taylor. "Mutant Allele Imbalance in Cancer." In: *Annual Review of Cancer Biology* 5 (2020), pp. 221–234. ISSN: 24723428. DOI: 10.1146/annurev-cancerbio-051320-124252.

[30] Nicole L Solimini et al. "Recurrent Hemizygous Deletions in Cancers May Optimize Proliferative Potential." In: *Science* 337.6090 (2012), pp. 104–109. ISSN: 0036-8075. DOI: 10.1126/science.1219580. URL: https://www.science.org/doi/10.1126/science.1219580.

[31] A. G. Knudson. "Mutation and cancer: statistical study of retinoblastoma." In: *Proceedings of the National Academy of Sciences of the United States of America* 68.4 (1971), pp. 820–823. ISSN: 00278424. DOI: 10.1073/pnas.68.4.820.

[32] Solip Park, Fran Supek, and Ben Lehner. "Higher order genetic interactions switch cancer genes from two-hit to one-hit drivers." In: *Nature Communications* 12.1 (2021), pp. 1–10. ISSN: 20411723. DOI: 10.1038/s41467-021-27242-3.

[33] Maroulio Pertesi et al. "Essential genes shape cancer genomes through linear limitation of homozygous deletions." In: *Communications Biology* 2.1 (2019), pp. 1–11. ISSN: 23993642. DOI: 10.1038/s42003-019-0517-0. URL: http://dx.doi.org/10.1038/s42003-019-0517-0.

[34] Saioa López et al. "Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution." In: *Nature Genetics* 52.3 (2020), pp. 283–293. ISSN: 1061-4036. DOI: 10.1038/s41588-020-0584-7. URL: http://www.nature.com/articles/s41588-020-0584-7.

[35] P. J. Hastings et al. "Mechanisms of change in gene copy number." In: *Nature Reviews Genetics* 10.8 (2009), pp. 551–564. ISSN: 14710056. DOI: 10.1038/nrg2593.

[36] S. T. Lovett et al. "A sister-strand exchange mechanism for recA-independent deletion of repeated DNA sequences in Escherichia coli." In: *Genetics* 135.3 (1993), pp. 631–642. ISSN: 00166731. DOI: 10.1093/genetics/135.3.631.

[37] Alessandra M. Albertini et al. "On the formation of spontaneous deletions: The importance of short sequence homologies in the generation of large deletions." In: *Cell* 29.2 (1982), pp. 319–328. ISSN: 00928674. DOI: 10.1016/0092-8674(82)90148-9.

[38] Philip J. Farabaugh et al. "Genetic studies of the lac repressor. VII. On the molecular nature of spontaneous hotspots in the lacI gene of Escherichia coli." In: *Journal of Molecular Biology* 126.4 (1978), pp. 847–863. ISSN: 00222836. DOI: 10.1016/0022-2836(78)90023-2.

[39] Mitchell L. Leibowitz, Cheng-Zhong Zhang, and David Pellman. "Chromothripsis: A New Mechanism for Rapid Karyotype Evolution." In: *Annual Review of Genetics* 49.1 (2015), pp. 183–211. ISSN: 0066-4197. DOI: 10.1146/annurev-genet-120213-092228. URL: https://www.annualreviews.org/doi/10.1146/annurev-genet-120213-092228.

[40] Samra Turajlic et al. "Resolving genetic heterogeneity in cancer." In: *Nature Reviews Genetics* 20.July (2019). ISSN: 14710064. DOI: 10.1038/s41576-019-0114-6. URL: http://dx.doi.org/10.1038/s41576-019-0114-6.

[41] A. Bird. "DNA methylation patterns and epigenetic memory." In: *Genes and Development* 16.1 (2002), pp. 6–21. ISSN: 08909369. DOI: 10.1101/gad.947102.

[42] Yong Wang and Frederick C.C. Leung. "An evaluation of new criteria for CpG islands in the human genome as gene markers." In: *Bioinformatics* 20.7 (2004), pp. 1170–1177. ISSN: 13674803. DOI: 10.1093/bioinformatics/bth059.

[43] Manuel Rodríguez-Paredes and Manel Esteller. "Cancer epigenetics reaches mainstream oncology." In: *Nature Medicine* 17.3 (2011), pp. 330–339. ISSN: 10788956. DOI: 10.1038/nm.2305.

[44] Manel Esteller. "Cancer epigenomics: DNA methylomes and histone-modification maps." In: *Nature Reviews Genetics* 8.4 (2007), pp. 286–298. ISSN: 14710056. DOI: 10.1038/nrg2005.

[45] Peter A. Jones and Peter W. Laird. "Cancer epigenetics comes of age." In: *Nature Genetics* 21.2 (1999), pp. 163–167. ISSN: 10614036. DOI: 10.1038/5947.

[46] Sarah J. Aitken et al. "Pervasive lesion segregation shapes cancer genome evolution." In: *Nature* 583.7815 (2020), pp. 265–270. ISSN: 14764687. DOI: 10.1038/s41586-020-2435-1. URL: http://dx.doi.org/10.1038/s41586-020-2435-1.

[47] Philip Stephens et al. "A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer." In: *Nature Genetics* 37.6 (2005), pp. 590–592. ISSN: 10614036. DOI: 10.1038/ng1571.

[48] Steven A. Roberts et al. "Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions." In: *Molecular Cell* 46.4 (2012), pp. 424–435. ISSN: 10972765. DOI: 10.1016/j.molcel.2012.03.030. URL: http://dx.doi.org/10.1016/j.molcel.2012.03.030.

[49] Serena Nik-Zainal et al. "Mutational processes molding the genomes of 21 breast cancers." In: *Cell* 149.5 (2012), pp. 979–993. ISSN: 10974172. DOI: 10.1016/j.cell.2012.04.024.

[50] Hu Fang et al. "Mutational processes of distinct POLE exonuclease domain mutants drive an enrichment of a specific TP53 mutation in colorectal cancer." In: *PLoS genetics* 16.2 (2020), e1008572. ISSN: 15537404. DOI: 10.1371/journal.pgen.1008572.

[51] Donna M. Muzny et al. "Comprehensive molecular characterization of human colon and rectal cancer." In: *Nature* 487.7407 (2012), pp. 330–337. ISSN: 00280836. DOI: 10.1038/nature11252. URL: http://dx.doi.org/10.1038/nature11252.

[52] Fran Supek and Ben Lehner. "Scales and mechanisms of somatic mutation rate variation across the human genome." In: *DNA Repair* 81.July (2019), p. 102647. ISSN: 15687856. DOI: 10.1016/j.dnarep.2019.102647. URL: https://doi.org/10.1016/j.dnarep.2019.102647.

[53] C. Tomasetti, B. Vogelstein, and G. Parmigiani. "Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation." In: *Proceedings of the National Academy of Sciences* 110.6 (2013), pp. 1999–2004. ISSN: 0027-8424. DOI: 10.1073/pnas.1221068110. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1221068110.

[54] Jessica Tang et al. "The genomic landscapes of individual melanocytes from human skin." In: *Nature* 586.7830 (2020), pp. 600–605. ISSN: 14764687. DOI: 10.1038/s41586-020-2785-8. URL: http://dx.doi.org/10.1038/s41586-020-2785-8.

[55] Sophie F. Roerink et al. "Intra-tumour diversification in colorectal cancer at the single-cell level." In: *Nature* 556.7702 (2018), pp. 437–462. ISSN: 14764687. DOI: 10.1038/s41586-018-0024-3.

[56] Yong H. Woo and Wen Hsiung Li. "DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes." In: *Nature Communications* 3 (2012). ISSN: 20411723. DOI: 10.1038/ncomms1982.

[57] Lin Liu, Subhajyoti De, and Franziska Michor. "DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes." In: *Nature Communications* 4 (2013), pp. 1–9. ISSN: 20411723. DOI: 10.1038/ncomms2502.

[58] Fran Supek and Ben Lehner. "Differential DNA mismatch repair underlies mutation rate variation across the human genome." In: *Nature* 521.7550 (2015), pp. 81–84. ISSN: 14764687. DOI: 10.1038/nature14173.

[59] Pablo E García-Nieto et al. "Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis." In: *The EMBO Journal* 36.19 (2017), pp. 2829–2843. ISSN: 0261-4189. DOI: 10.15252/embj.201796717.

[60] Anna R. Poetsch, Simon J. Boulton, and Nicholas M. Luscombe. "APSEQ Genomic landscape of oxidative DNA damage and repair reveals regioselective protection from mutagenesis." In: *Genome Biology* 19.1 (2018), p. 215. ISSN: 1474-760X. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1582-2.

[61] Feng Li et al. "The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα." In: *Cell* 153.3 (2013), pp. 590–600. ISSN: 10974172. DOI: 10.1016/j.cell.2013.03.025. URL: http://dx.doi.org/10.1016/j.cell.2013.03.025.

[62] Eran Hodis et al. "A landscape of driver mutations in melanoma." In: *Cell* 150.2 (2012), pp. 251–263. ISSN: 00928674. DOI: 10.1016/j.cell.2012.06.024.

[63] Laurence D. Hurst and Nizar N. Batada. "Depletion of somatic mutations in splicing-associated sequences in cancer genomes." In: *Genome Biology* 18.1 (2017), pp. 1–12. ISSN: 1474760X. DOI: 10.1186/s13059-017-1337-5.

[64] Nicholas J. Haradhvala et al. "Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair." In: *Cell* 164.3 (2016), pp. 538–549. ISSN: 10974172. DOI: 10.1016/

j.cell.2015.12.050. URL: http://dx.doi.org/10.1016/j.cell.2015.12.050.

[65] Riku Katainen et al. "CTCF/cohesin-binding sites are frequently mutated in cancer." In: *Nature Genetics* 47.7 (2015), pp. 818–821. ISSN: 15461718. DOI: 10.1038/ng.3335. URL: http://dx.doi.org/10.1038/ng.3335.

[66] Rebecca C. Poulos et al. "Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif." In: *Cell Reports* 17.11 (2016), pp. 2865–2872. ISSN: 22111247. DOI: 10.1016/j.celrep.2016.11.055. URL: http://dx.doi.org/10.1016/j.celrep.2016.11.055.

[67] Vera B. Kaiser, Martin S. Taylor, and Colin A. Semple. "Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types." In: *PLoS Genetics* 12.8 (2016), pp. 1–19. ISSN: 15537404. DOI: 10.1371/journal.pgen.1006207.

[68] Kerryn Elliott et al. "Elevated pyrimidine dimer formation at distinct genomic bases underlies promoter mutation hotspots in UV-exposed cancers." In: *PLoS Genetics* 14.12 (2018), pp. 1–15. ISSN: 15537404. DOI: 10.1371/journal.pgen.1007849.

[69] Peng Mao et al. "ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma." In: *Nature Communications* 9.1 (2018). ISSN: 20411723. DOI: 10.1038/s41467-018-05064-0. URL: http://dx.doi.org/10.1038/s41467-018-05064-0.

[70] Tobias Warnecke, Nizar N. Batada, and Laurence D. Hurst. "The impact of the nucleosome code on protein-coding sequence evolution in yeast." In: *PLoS Genetics* 4.11 (2008). ISSN: 15537390. DOI: 10.1371/journal.pgen.1000250.

[71] Alexander J. Brown et al. "Nucleosome positions establish an extended mutation signature in melanoma." In: *PLoS Genetics* 14.11 (2018), pp. 1–21. ISSN: 15537404. DOI: 10.1371/journal.pgen.1007823.

[72] Oriol Pich et al. "Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes." In: *Cell* 175.4 (2018), 1074–1087.e18. ISSN: 10974172. DOI: 10.1016/j.cell.2018.10.004.

[73] Michael Y. Tolstorukov et al. "Impact of chromatin structure on sequence variability in the human genome." In: *Nature Structural and Molecular Biology* 18.4 (2011), pp. 510–516. ISSN: 15459993. DOI: 10.1038/nsmb.2012.

[74] Xiaoshu Chen et al. "Nucleosomes Suppress Spontaneous Mutations Base-Specifically in Eukaryotes." In: *Science* 335.6073 (2012), pp. 1235–1238. ISSN: 0036-8075. DOI: 10.1126/science.1217580. URL: https://www.science.org/doi/10.1126/science.1217580.

[75] Ludmil B. Alexandrov et al. "Signatures of mutational processes in human cancer." In: *Nature* 500.7463 (2013), pp. 415–421. ISSN: 00280836. DOI: 10.1038/nature12477. arXiv: NIHMS150003.

[76] Ludmil B. Alexandrov et al. "Deciphering Signatures of Mutational Processes Operative in Human Cancer." In: *Cell Reports* 3.1 (2013), pp. 246–259. ISSN: 22111247. DOI: 10.1016/j.celrep.2012.12.008. arXiv: arXiv:1408.1149. URL: http://dx.doi.org/10.1016/j.celrep.2012.12.008.

[77] Erin D. Pleasance et al. "A comprehensive catalogue of somatic mutations from a human cancer genome." In: *Nature* 463.7278 (2010), pp. 191–196. ISSN: 00280836. DOI: 10.1038/nature08658.

[78] Varun Aggarwala and Benjamin F. Voight. "An expanded sequence context model broadly explains variability in polymorphism levels across the human genome." In: *Nature Genetics* 48.4 (2016), pp. 349–355. ISSN: 15461718. DOI: 10.1038/ng.3511.

[79] Martin L. Miller et al. "Pan-Cancer Analysis of Mutation Hotspots in Protein Domains." In: *Cell Systems* 1.3 (2015), pp. 197–209. ISSN: 24054720. DOI: 10.1016/j.cels.2015.08.014.

[80] Rehan Akbani et al. "Genomic Classification of Cutaneous Melanoma." In: *Cell* 161.7 (2015), pp. 1681–1696. ISSN: 10974172. DOI: 10.1016/j.cell.2015.05.044.

[81] Eric A. Collisson et al. "Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network." In: *Nature* 511.7511 (2014), pp. 543–550. ISSN: 14764687. DOI: 10.1038/nature13385.

[82] Victor Trevino. "Modeling and analysis of site-specific mutations in cancer identifies known plus putative novel hotspots and bias due to contextual sequences." In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 1664–1675. ISSN: 20010370. DOI: 10.1016/j.csbj.2020.06.022. URL: https://doi.org/10.1016/j.csbj.2020.06.022.

[83] Rémi Buisson et al. "Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features." In: *Science* 364.6447 (2019). ISSN: 10959203. DOI: 10.1126/science.aaw2872.

[84] Ilias Georgakopoulos-Soares et al. "Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis." In: *Genome Research* 28.9 (2018), pp. 1264–1271. ISSN: 15495469. DOI: 10.1101/gr.231688.117.

[85] Matthew H. Bailey et al. "Comprehensive Characterization of Cancer Driver Genes and Mutations." In: *Cell* 174.4 (2018), pp. 1034–1035. ISSN: 10974172. DOI: 10.1016/j.cell.2018.07.034.

[86] Peter Priestley et al. "Pan-cancer whole-genome analyses of metastatic solid tumours." In: *Nature* 575.7781 (2019), pp. 210–216. ISSN: 14764687. DOI: 10.1038/s41586-019-1689-y. URL: http://dx.doi.org/10.1038/s41586-019-1689-y.

[87] Peter J. Campbell et al. "Pan-cancer analysis of whole genomes." In: *Nature* 578.7793 (2020), pp. 82–93. ISSN: 14764687. DOI: 10.1038/s41586-020-1969-6.

[88] Erin Pleasance et al. "Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes." In: *Nature cancer* 1.4 (2020), pp. 452–468. ISSN: 26621347. DOI: 10.1038/s43018-020-0050-6.

[89] Jimmy Van den Eynden, Swaraj Basu, and Erik Larsson. "Somatic Mutation Patterns in Hemizygous Genomic Regions Unveil Purifying Selection during Tumor Evolution." In: *PLoS Genetics* 12.12 (2016), pp. 1–18. ISSN: 15537404. DOI: 10.1371/journal.pgen.1006506.

[90] Luis Zapata et al. "Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes." In: *Scientific Reports* 7.1 (2017), pp. 1–15. ISSN: 20452322. DOI: 10.1038/s41598-017-12888-1.

[91] Donate Weghorn and Shamil Sunyaev. "Bayesian inference of negative and positive selection in human cancers." In: *Nature Genetics* 49.12 (2017), pp. 1785–1788. ISSN: 15461718. DOI: 10.1038/ng.3987. URL: http://dx.doi.org/10.1038/ng.3987.

[92] Michael S. Lawrence et al. "Mutational heterogeneity in cancer and the search for new cancer-associated genes." In: *Nature* 499.7457 (2013), pp. 214–218. ISSN: 00280836. DOI: 10.1038/nature12213. arXiv: 0208024 [gr-qc].

[93] Jimmy Van Den Eynden and Erik Larsson. "Mutational Signatures Are Critical for Proper Estimation of Purifying Selection Pressures in Cancer Somatic Mutation Data When Using the dN/dS Metric." In: *Frontiers in Genetics* 8.June (2017), pp. 1–9. DOI: 10.7908/C11G0KM9.

[94] Jimmy Van den Eynden et al. "Lack of detectable neoantigen depletion signals in the untreated cancer genome." In: *Nature Genetics* 51.12 (2019), pp. 1741–1748. ISSN: 15461718. DOI: 10.1038/s41588-019-0532-6. URL: http://dx.doi.org/10.1038/s41588-019-0532-6.

[95] Felix Dietlein et al. "Identification of cancer driver genes based on nucleotide context." In: *Nature Genetics* 52.2 (2020), pp. 208–218. ISSN: 1061-4036. DOI: 10.1038/s41588-019-0572-y. URL: http://www.nature.com/articles/s41588-019-0572-y.

[96] B Vogelstein et al. "Cancer Genome Landscapes." In: *Science* 339.6127 (2013), pp. 1546–1558. ISSN: 0036-8075. DOI: 10.1126/science.1235122. URL: http://www.sciencemag.org/content/

339/6127/1546.shorthttp://www.sciencemag.org/cgi/doi/10.1126/science.1235122.

[97] László Bányai et al. "Use of signals of positive and negative selection to distinguish cancer genes and passenger genes." In: *eLife* 10 (2021), pp. 1–141. ISSN: 2050084X. DOI: 10.7554/ELIFE.59629.

[98] Matthew T. Chang et al. "Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity." In: *Nature Biotechnology* 34.2 (2016), pp. 155–163. ISSN: 15461696. DOI: 10.1038/nbt.3391.

[99] Eduard Porta-Pardo et al. "Comparison of algorithms for the detection of cancer drivers at subgene resolution." In: *Nature Methods* 14.8 (2017), pp. 782–788. ISSN: 15487105. DOI: 10.1038/nmeth.4364.

[100] Claudia Arnedo-Pac et al. "OncodriveCLUSTL: A sequence-based clustering method to identify cancer drivers." In: *Bioinformatics* 35.22 (2019), pp. 4788–4790. ISSN: 14602059. DOI: 10.1093/bioinformatics/btz501.

[101] Eduard Porta-Pardo and Adam Godzik. "E-Driver: A novel method to identify protein regions driving cancer." In: *Bioinformatics* 30.21 (2014), pp. 3109–3114. ISSN: 14602059. DOI: 10.1093/bioinformatics/btu499.

[102] Eduard Porta-Pardo et al. "A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces." In: *PLoS Computational Biology* 11.10 (2015), pp. 1–18. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004518.

[103] Atanas Kamburov et al. "Comprehensive assessment of cancer missense mutation clustering in protein structures." In: *Proceedings of the National Academy of Sciences of the United States of America* 112.40 (2015), E5486–E5495. ISSN: 10916490. DOI: 10.1073/pnas.1516373112.

[104] Jüri Reimand and Gary D. Bader. "Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers." In: *Molecular Systems Biology* 9.637 (2013). ISSN: 17444292. DOI: 10.1038/msb.2012.68.

[105] Loris Mularoni et al. "OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations." In: *Genome Biology* 17.1 (2016), pp. 1–13. ISSN: 1474760X. DOI: 10.1186/s13059-016-0994-0. URL: http://dx.doi.org/10.1186/s13059-016-0994-0.

[106] Craig H. Mermel et al. "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers." In: *Genome Biology* 12.4 (2011), pp. 1–14. ISSN: 14747596. DOI: 10.1186/gb-2011-12-4-r41.

[107] David Tamborero et al. "Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations." In: *Genome Medicine* 10.1 (2018), pp. 1–8. ISSN: 1756994X. DOI: 10.1186/s13073-018-0531-8.

[108] Craig M. Bielski et al. "Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer." In: *Cancer Cell* 34.5 (2018), 852–862.e4. ISSN: 18783686. DOI: 10.1016/j.ccell.2018.10.003. URL: https://doi.org/10.1016/j.ccell.2018.10.003.

[109] Iñigo Martincorena, Aswin S.N. Seshasayee, and Nicholas M. Luscombe. "Evidence of non-random mutation rates suggests an evolutionary risk management strategy." In: *Nature* 485.7396 (2012), pp. 95–98. ISSN: 00280836. DOI: 10.1038/nature10995.

[110] Sheli L. Ostrow et al. "Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes." In: *PLoS Genetics* 10.3 (2014), pp. 16–20. ISSN: 15537404. DOI: 10.1371/journal.pgen.1004239.

[111] Adam Auton et al. "A global reference for human genetic variation." In: *Nature* 526.7571 (2015), pp. 68–74. ISSN: 14764687. DOI: 10.1038/nature15393.

[112] Rik G H Lindeboom, Fran Supek, and Ben Lehner. "The rules and impact of nonsense-mediated mRNA decay in human cancers." In: *Nature Genetics* 48.10 (2016), pp. 1112–1118. ISSN: 1061-4036. DOI: 10.1038/ng.3664. URL: http://www.nature.com/articles/ng.3664.

[113] Kiyoshi Ezawa, Giddy Landan, and Dan Graur. "Detecting negative selection on recurrent mutations using gene genealogy." In: *BMC Genetics* 14 (2013). ISSN: 14712156. DOI: 10.1186/1471-2156-14-37.

[114] W. Bateson. "Mendel's Principles of Heredity." In: *Nature* 86.2169 (1911), pp. 407–407. ISSN: 0028-0836. DOI: 10.1038/086407a0. URL: https://www.nature.com/articles/086407a0.

[115] R. A. Fisher. "The Correlation Between Relatives on the Supposition of Mendelian Inheritance." In: *Trans. R. Soc. Edinburgh* 52 (1918), pp. 399–433.

[116] Niko Beerenwinkel, Lior Pachter, and Bernd Sturmfels. "Epistasis and shapes of fitness landscapes." In: *Statistica Sinica* 17.4 (2007), pp. 1317–1342. ISSN: 10170405. arXiv: 0603034 [q-bio].

[117] Sewall Wright. "Evolution in mendelian populations." In: *Bulletin of Mathematical Biology* 52.1-2 (1990), pp. 241–295. ISSN: 00928240. DOI: 10.1007/BF02459575.

[118] Dariusz Matlak and Ewa Szczurek. "Epistasis in genomic and survival data of cancer patients." In: *PLoS Computational Biology* 13.7 (2017), pp. 1–16. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005626.

[119] Alan Ashworth, Christopher J. Lord, and Jorge S. Reis-Filho. "Genetic Interactions in Cancer Progression and Treatment." In: *Cell* 145.1 (2011), pp. 30–38. ISSN: 00928674. DOI: 10.1016/j.cell.2011.03.020. URL: http://dx.doi.org/10.1016/j.cell.2011.03.020https://linkinghub.elsevier.com/retrieve/pii/S0092867411002972.

[120] Teresa Davoli et al. "Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome." In: *Cell* 155.4 (2013), pp. 948–962. ISSN: 02315882. DOI: 10.1016/J.CELL.2013.10.011. URL: http://dx.doi.org/10.1016/j.cell.2013.10.011.

[121] Barbara De Kegel and Colm J. Ryan. "Paralog buffering contributes to the variable essentiality of genes in cancer cell lines." In: *bioRxiv* (2019), p. 716043. DOI: 10.1101/716043.

[122] Harold E. Varmus. "THE MOLECULAR GENETICS OF CELLULAR ONCOGENES." In: *Annual Review of Genetics* 18.1 (1984), pp. 553–612. ISSN: 0066-4197. DOI: 10.1146/annurev.ge.18.120184.003005. URL: https://www.annualreviews.org/doi/10.1146/annurev.ge.18.120184.003005.

[123] Li Ding et al. "Somatic mutations affect key pathways in lung adenocarcinoma." In: *Nature* 455.7216 (2008), pp. 1069–1075. ISSN: 14764687. DOI: 10.1038/nature07423.

[124] Toshimi Takano et al. "Epidermal growth factor receptor gene mutations and increased copy numbers predict gefitinib sensitivity in patients with recurrent non-small-cell lung cancer." In: *Journal of Clinical Oncology* 23.28 (2005), pp. 6829–6837. ISSN: 0732183X. DOI: 10.1200/JCO.2005.01.0793.

[125] Sergey Nikolaev et al. "Extrachromosomal driver mutations in glioblastoma and low-grade glioma." In: *Nature Communications* 5 (2014), pp. 1–7. ISSN: 20411723. DOI: 10.1038/ncomms6690.

[126] Xiaolin Nan et al. "Ras-GTP dimers activate the Mitogen-Activated Protein Kinase (MAPK) pathway." In: *Proceedings of the National Academy of Sciences of the United States of America* 112.26 (2015), pp. 7996–8001. ISSN: 10916490. DOI: 10.1073/pnas.1509123112.

[127] Peter M.K. Westcott et al. "The mutational landscapes of genetic and chemical models of Kras-driven lung cancer." In: *Nature* 517.7535 (2015), pp. 489–492. ISSN: 14764687. DOI: 10.1038/nature13898.

[128] Chiara Ambrogio et al. "KRAS Dimerization Impacts MEK Inhibitor Sensitivity and Oncogenic Activity of Mutant KRAS." In: *Cell* 172.4 (2018), 857–868.e15. ISSN: 10974172. DOI: 10.1016/j.cell.2017.12.020.

[129] Martin Boström and Erik Larsson. "Somatic mutation distribution across tumour cohorts provides a signal for positive selection in cancer." In: *Nature Communications* 13.1 (2022), p. 7023. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34746-z. URL: https://www.nature.com/articles/s41467-022-34746-z.

[130] Traver Hart et al. "Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens." In: *G3: Genes, Genomes, Genetics* 7.8 (2017), pp. 2719–2727. ISSN: 21601836. DOI: 10.1534/g3.117.041277.

[131] Konrad J. Karczewski et al. "The mutational constraint spectrum quantified from variation in 141,456 humans." In: *Nature* 581.7809 (2020), pp. 434–443. ISSN: 14764687. DOI: 10.1038/s41586-020-2308-7.

[132] P. Andrew Futreal et al. "A census of human cancer genes." In: *Nature Reviews Cancer* 4.3 (2004), pp. 177–183. ISSN: 1474175X. DOI: 10.1038/nrc1299.

[133] Luis Zapata et al. "Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome." In: *Genome Biology* 19.1 (2018), pp. 1–17. ISSN: 1474760X. DOI: 10.1186/s13059-018-1434-0.

[134] A. Albert and J. A. Anderson. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." In: *Biometrika* 71.1 (1984), p. 1. ISSN: 00063444. DOI: 10.2307/2336390. URL: https://www.jstor.org/stable/2336390?origin=crossref.

[135] Christopher Zorn. "A Solution to Separation in Binary Response Models." In: *Political Analysis* 13.2 (2005), pp. 157–170. ISSN: 1047-1987. DOI: 10.1093/pan/mpi009. URL: https://www.cambridge.org/core/product/identifier/S1047198700001042/type/journal_article.

[136] Andrew Gelman et al. "A weakly informative default prior distribution for logistic and other regression models." In: *Annals of Applied Statistics* 2.4 (2008), pp. 1360–1383. ISSN: 19326157. DOI: 10.1214/08-AOAS191. arXiv: 0901.4011.

[137] Morris H. Hansen and William N. Hurwitz. "On the Theory of Sampling from Finite Populations." In: *The Annals of Mathematical Statistics* 14.4 (1943), pp. 333–362. ISSN: 0003-4851. DOI: 10.1214/aoms/1177731356. URL: http://projecteuclid.org/euclid.aoms/1177731356.

[138] Kuo Feng Tung et al. "Top-ranked expressed gene transcripts of human protein-coding genes investigated with GTEx dataset." In: *Scientific Reports* 10.1 (2020), pp. 1–11. ISSN: 20452322. DOI: 10.1038/s41598-020-73081-5. URL: https://doi.org/10.1038/s41598-020-73081-5.

[139] Marc J. Williams et al. "Identification of neutral tumor evolution across cancer types." In: *Nature Genetics* 48.3 (2016), pp. 238–244. ISSN: 15461718. DOI: 10.1038/ng.3489. URL: http://dx.doi.org/10.1038/ng.3489.

[140] Abel Gonzalez-Perez et al. "Computational approaches to identify functional genetic variants in cancer genomes." In: *Nature Methods* 10.8 (2013), pp. 723–729. ISSN: 15487091. DOI: 10.1038/nmeth.2562.

[141] Sushant Kumar et al. "Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences." In: *Cell* 180.5 (2020), 915–927.e16. ISSN: 10974172. DOI: 10.1016/j.cell.2020.01.032. URL: https://doi.org/10.1016/j.cell.2020.01.032.

[142] Philipp Rentzsch et al. "CADD : predicting the deleteriousness of variants throughout the human genome." In: *Nucleic Acids Research* (2018), pp. 1–9. DOI: 10.1093/nar/gky1016.

[143] Nilah M. Ioannidis et al. "REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants." In: *American Journal of Human Genetics* 99.4 (2016), pp. 877–885. ISSN: 15376605. DOI: 10.1016/j.ajhg.2016.08.016. URL: http://dx.doi.org/10.1016/j.ajhg.2016.08.016.

[144] Cathal Ormond et al. "Converting single nucleotide variants between genome builds: From cautionary tale to solution." In: *Briefings in Bioinformatics* 22.5 (2021), pp. 1–7. ISSN: 14774054. DOI: 10.1093/bib/bbab069.

[145] Weiwei Shi et al. "Reliability of Whole-Exome Sequencing for Assessing Intratumor Genetic Heterogeneity." In: *Cell Reports* 25.6 (2018), pp. 1446–1457. ISSN: 22111247. DOI: 10.1016/j.celrep.2018.10.046. URL: https://doi.org/10.1016/j.celrep.2018.10.046.

[146] Gilad Fuchs et al. "4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells." In: *Genome Biology* 15.5 (2014), R69. ISSN: 1474-760X. DOI: 10.1186/gb-2014-15-5-r69. URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-5-r69.

[147] Paz Polak et al. "Cell-of-origin chromatin organization shapes the mutational landscape of cancer." In: *Nature* 518.7539 (2015), pp. 360–364. ISSN: 14764687. DOI: 10.1038/nature14221.

[148] Alexandra Avgustinova et al. "Loss of G9a preserves mutation patterns but increases chromatin accessibility, genomic instability and aggressiveness in skin tumours." In: *Nature Cell Biology* 20.12 (2018), pp. 1400–1409. ISSN: 14764679. DOI: 10.1038/s41556-018-0233-x. URL: http://dx.doi.org/10.1038/s41556-018-0233-x.

[149] Fran Supek and Ben Lehner. "Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes." In: *Cell* 170.3 (2017), 534–547.e23. ISSN: 10974172. DOI: 10.1016/j.cell.2017.07.003. URL: http://dx.doi.org/10.1016/j.cell.2017.07.003.

[150] Marina Salvadores and Fran Supek. "Redistribution of mutation rates across chromosomal domains in human cancer genomes." In: *bioRxiv* (2022), p. 2022.10.24.513586. URL: https://www.biorxiv.org/content/10.1101/2022.10.24.513586v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.10.24.513586v1.abstract.

[151] Petra Kleiblova et al. "Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint." In: *Journal of Cell Biology* 201.4 (2013), pp. 511–521. ISSN: 00219525. DOI: 10.1083/jcb.201210031.

[152] Adam M. Dinan, John F. Atkins, and Andrew E. Firth. "ASXL gain-of-function truncation mutants: Defective and dysregulated forms of a natural ribosomal frameshifting product?" In: *Biology Direct* 12.1 (2017), pp. 1–16. ISSN: 17456150. DOI: 10.1186/s13062-017-0195-0.

[153] Masaru Katoh. "Fibroblast growth factor receptors as treatment targets in clinical oncology." In: *Nature Reviews Clinical Oncology* 16.2 (2019), pp. 105–122. ISSN: 17594782. DOI: 10.1038/s41571-018-0115-y. URL: http://dx.doi.org/10.1038/s41571-018-0115-y.

[154] Michael S. Lawrence et al. "Discovery and saturation analysis of cancer genes across 21 tumour types." In: *Nature* 505.7484 (2014), pp. 495–501. ISSN: 00280836. DOI: 10.1038/nature12912. URL: http://dx.doi.org/10.1038/nature12912.

[155] Robin M. Meyers et al. "Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells." In: *Nature Genetics* 49.12 (2017), pp. 1779–1784. ISSN: 15461718. DOI: 10.1038/ng.3984. URL: http://dx.doi.org/10.1038/ng.3984.

[156] Pilar Cacheiro et al. "Human and mouse essentiality screens as a resource for disease gene discovery." In: *Nature Communications* 11.1 (2020). ISSN: 20411723. DOI: 10.1038/s41467-020-14284-2.

[157] R. Bremner and A. Balmain. "Genetic changes in skin tumor progression: Correlation between presence of a mutant ras gene and loss of heterozygosity on mouse chromosome 7." In: *Cell* 61.3 (1990), pp. 407–417. ISSN: 00928674. DOI: 10.1016/0092-8674(90)90523-H.

[158] Zhongqiu Zhang et al. "Wildtype Kras2 can inhibit lung carcinogenesis in mice." In: *Nature Genetics* 29.1 (2001), pp. 25–33. ISSN: 10614036. DOI: 10.1038/ng721.

[159] Antonio Colaprico et al. "TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data." In: *Nucleic Acids Research* 44.8 (2016), e71. ISSN: 13624962. DOI: 10.1093/nar/gkv1507.

[160] Tiago C. Silva et al. "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages." In: *F1000Research* 5 (2016). ISSN: 1759796X. DOI: 10.12688/f1000research.8923.2.

[161] Mohamed Mounir et al. "New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEX." In: *PLoS Computational Biology* 15.3 (2019). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006701.

[162] Donald A. Jackson. "Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches." In: *Ecology* 74.8 (1993), pp. 2204–2214. ISSN: 00129658. DOI: 10.2307/1939574. URL: http://doi.wiley.com/10.2307/1939574.

[163] *https://www.proteinatlas.org/about/download*.

[164] Rinki Ratnapriya et al. "Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration." In: *Nature Genetics* 51.4 (2019), pp. 606–610. ISSN: 15461718. DOI: 10.1038/s41588-019-0351-9.

[165] Andrea Sottoriva et al. "A big bang model of human colorectal tumor growth." In: *Nature Genetics* 47.3 (2015), pp. 209–216. ISSN: 15461718. DOI: 10.1038/ng.3214.

[166] Mikhail Pyatnitskiy et al. "Bringing down cancer aircraft: Searching for essential hypomutated proteins in skin melanoma." In: *PLoS ONE* 10.11 (2015), pp. 1–14. ISSN: 19326203. DOI: 10.1371/journal.pone.0142819.

[167] Ilya E. Vorontsov et al. "Negative selection maintains transcription factor binding motifs in human cancer." In: *BMC Genomics* 17.Suppl 2 (2016). ISSN: 14712164. DOI: 10.1186/s12864-016-2728-9. URL: http://dx.doi.org/10.1186/s12864-016-2728-9.

[168] László Bányai et al. "Use of signals of positive and negative selection to distinguish cancer genes and passenger genes." In: *bioRxiv* (2020), p. 2020.06.04.133199. DOI: 10.1101/2020.06.04.133199.

[169] Kevin Litchfield et al. "Escape from nonsense-mediated decay associates with anti-tumor immunogenicity." In: *Nature Communications* 11.1 (2020), pp. 1–11. ISSN: 20411723. DOI: 10.1038/s41467-020-17526-5.

[170] G. Traver Hart, Insuk Lee, and Edward R. Marcotte. "A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality." In: *BMC Bioinformatics* 8 (2007), pp. 8–10. ISSN: 14712105. DOI: 10.1186/1471-2105-8-236.

[171] I. Bernard Weinstein. "Cancer: Addiction to oncogenes - The Achilles heal of cancer." In: *Science* 297.5578 (2002), pp. 63–64. ISSN: 00368075. DOI: 10.1126/science.1073096.

[172] Yari Ciani et al. "Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer." In: *Cell Systems* 13.2 (2022), 183–193.e7. ISSN: 24054720. DOI: 10.1016/j.cels.2021.10.001. URL: https://doi.org/10.1016/j.cels.2021.10.001.

[173] Ira Herskowitz. "dominant negative mutations CS-G-CS." In: *Nature* 329 (1987), pp. 1–4.

[174] L. Therese Bergendahl et al. "The role of protein complexes in human genetic disease." In: *Protein Science* 28.8 (2019), pp. 1400–1411. ISSN: 1469896X. DOI: 10.1002/pro.3667.

[175] Lukas Gerasimavicius, Benjamin J. Livesey, and Joseph A. Marsh. "Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure." In: *Nature Communications* 13.1 (2022), p. 3895. ISSN: 2041-1723. DOI: 10.1038/s41467-022-31686-6. URL: https://www.nature.com/articles/s41467-022-31686-6.

[176] Solip Park and Ben Lehner. "Cancer type-dependent genetic interactions between cancer driver alterations indicate plasticity of epistasis across cell types." In: *Molecular Systems Biology* 11.7 (2015), p. 824. ISSN: 1744-4292. DOI: 10.15252/msb.20156102.

[177] Kyle Ellrott et al. "Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines." In: *Cell Systems* 6.3 (2018), 271–281.e7. ISSN: 24054720. DOI: 10.1016/j.cels.2018.03.002.

[178] *https://www.hartwigmedicalfoundation.nl/en/.*

[179] *https://dcc.icgc.org/pcawg/.*

[180] *https://www.bcgsc.ca/downloads/POG570/.*

[181] Nathan J. Edwards et al. "The CPTAC data portal: A resource for cancer proteomics research." In: *Journal of Proteome Research* 14.6 (2015), pp. 2707–2713. ISSN: 15353907. DOI: 10.1021/pr501254j.

[182] Matthew J. Ellis et al. "Connecting genomic alterations to cancer biology with proteomics: The NCI clinical proteomic tumor analysis consortium." In: *Cancer Discovery* 3.10 (2013), pp. 1108–1112. ISSN: 21598274. DOI: 10.1158/2159-8290.CD-13-0219.

[183] *https://portal.gdc.cancer.gov/.*

[184] Brian A. Walker et al. "A high-risk, Double-Hit, group of newly diagnosed myeloma identified by genomic analysis." In: *Leukemia* 33.1 (2019), pp. 159–170. ISSN: 14765551. DOI: 10.1038/s41375-018-0196-8.

[185] A. S. Hinrichs et al. "The UCSC Genome Browser Database: update 2006." In: *Nucleic acids research* 34.Database issue (2006), pp. 590–598. ISSN: 13624962. DOI: 10.1093/nar/gkj144.

[186] Kai Wang, Mingyao Li, and Hakon Hakonarson. "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data." In: *Nucleic Acids Research* 38.16 (2010), pp. 1–7. ISSN: 03051048. DOI: 10.1093/nar/gkq603.

[187] *https://www.synapse.org/.*

[188] S. M. Sweeney et al. "AACR project genie: Powering precision medicine through an international consortium." In: *Cancer Discovery* 7.8 (2017), pp. 818–831. ISSN: 21598290. DOI: 10.1158/2159-8290.CD-17-0151.

[189] Ryan J. Hartmaier et al. "High-throughput genomic profiling of adult solid tumors reveals novel insights into cancer pathogenesis." In: *Cancer Research* 77.9 (2017), pp. 2464–2475. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-16-2479.

[190] *https://krishna.gs.washington.edu/download/CADD/bigWig/.*

[191] Jianjiong Gao et al. "3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets." In: *Genome Medicine* 9.1 (2017), pp. 1–13. ISSN: 1756994X. DOI: 10.1186/s13073-016-0393-x. URL: http://dx.doi.org/10.1186/s13073-016-0393-x.

[192] Rik G.H. Lindeboom et al. "The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy." In: *Nature Genetics* 51.11 (2019), pp. 1645–1651. ISSN: 15461718. DOI: 10.1038/s41588-019-0517-5. URL: http://dx.doi.org/10.1038/s41588-019-0517-5.

[193] *https://figshare.com/articles/dataset/NMDetective/7803398.*