On the evolution of cancer genomes

Signatures of selection reveal cancer genes across multiple tumor types

Luis Zapata Ortiz

DOCTORAL THESIS / YEAR 2016

THESIS SUPERVISOR
Dr. Stephan Ossowski

Department of Bioinformatics and Genomics, Center for Genomic Regulation

A la memoria de nuestras generaciones: nuestros abuelos, padres, hermanos, hijos y nietos.

Preface

I joined the CRG-UPF program in September of 2011, after my wife, Claire, and I decided to do our PhD in the beautiful city of Barcelona. I must admit that we were surprised by the level and quality of scientific research performed here, as Spain is usually not a typical choice destination to complete a PhD as the USA, UK or Germany are.

However, now that I am turning my thesis in, I have never looked back on this decision we took more than 5 years ago. I have had wonderful time here by the sea, I have met amazing people and I have lived my life to the max, with ups and downs, lefts and rights, 0s and 1s, bottles and diapers, and so on. Altogether, this experience has shaped me as a person, as a scientist, as a husband, and as a father. Moreover, I have worked side by side with one of the most brilliant critical thinkers I have met, my supervisor, Stephan Ossowski. He has always been much more than my PI, as I have always been able to share and talk to him about everything, I consider him a friend. After all these years we have grown together, me as a PhD student and him as a group leader, I cant be anything other than grateful.

This thesis is the reflection of my efforts to better understand the disease with the most impact on modern society, cancer. I have always believed that evolution is one of the most important forces shaping the universe, and in particular, our world. It is not just a concept, it is an ideology, a way of approaching problems. Its dualistic nature underlies the interaction between entropy and environment. Once you have a system that contains some randomness (i.e. atomic particle movement), apply a gradient (i.e. positive and negative electric charges), and multiply such a system an infinite number of times, you will recreate evolution. Life on earth arose from one of these infinite trial and error experiments. But let's go back to a more concrete example: cancer.

Cancer is like a gang of cells that believe they are better than the rest. So they begin to proliferate indiscriminately, taking over resources and eventually leading the organism to death. How does it get to this point? This is what we have been studying for the past decade with the help of next generation sequencing technologies (NGS). Since last century, we know that every cell has an intrinsic code, the DNA sequence, dictating functionality such as the ability to grow and proliferate. This code is far from stable as it is consistently being rewritten due to mutational processes acting on the cell. The beauty of this system, however, is that the cell is also able to detect if a particular change is damaging and can either repair it or commit suicide. On the other hand, if the change is neutral for the cell, it can either remain or be removed. If the change is beneficial, the cell will most likely want to retain it. In summary, evolution works by allowing a constantly

changing system to accumulate a set of beneficial characteristics governed by what Darwin identified as natural selection.
There are approximately 20,000 genes identified in the human genome. A critical function of one of these genes is to recognize an abnormal cell cycle and fix the problem, a gargantuan task performed by the famous TP53. If this gene mutates, it can no longer complete its role, and the cell will eventually become malignant. From an evolutionary perspective, a cell that proliferates more than the rest due to an acquired mutation is considered to be positively selected. The cellular prevalence of a mutation is therefore relevant to tumorigenesis, unless it has hitchhiked with another relevant mutation. In the first part of this work, we learn how to quantify the proportion of cells harboring a mutation. In the second part, we integrate such information to identify genes that impart an advantage if mutated, therefore leaving traces of positive selection in the cancer's evolutionary context. The third part of the thesis is focused on the idea that if positive selection is present, negative selection must also be acting, but how can we detect this effect? In the case of positive selection, in a given cohort of cancer patients, there are many mutations that are shared by patients. If we apply evolutionary theory, we can easily detect these cases because they are more mutated than expected. In the case of negative selection, however, we are searching for the opposite effect: genes that appear less mutated than expected. The large number of patients that we use for these studies gives us the power to observe this effect and assign a significant value for each gene. The impressive amount of data collected through the years, across many tumor types, has allowed us to successfully detect the phenomenon of negative selection in cancer.

Barcelona, Spain
September 2016

Acknowledgements

First, I would like to thanks all my colleagues in the lab, specially to my colleague Hana. One part of this thesis is a great work done together. To my boss for believing on me, on my ideas, and support me to explore them freely. To Korbinian that also put his faith on my work, ending with a beautiful paper on comparative genomics of Arabidopsis. To the CRG for providing me such an amazing place to do science. I doubt that any place in the world is better for combining excellent science and disastrous fun.
I'm really grateful to all the ones that started the PhD journey with me Diego, Alba, Natalia, Pez, Adam. Also, to all the ones that I have met in the process, Aaron, Kadri, JC, my Chilean fellows, Felipe, Pancho y Javi. Rosa, Mata. People from the wednesday/friday futbol. To the secretaries of our programs, Rut, Romina, Imma. I would like to acknowledge the administration of the institute, the PIs, and to everyone that I have shared with in this wonderful institute. To all of you, thanks.
Very important also I'd like to remember my friends outside the science world such as Beta, Kevin, al H, a los GLS, a mis amigos en Chile, a mis amigos repartidos por el mundo. Ellos que estan por ahí, siempre los llevo conmigo.
Quiero recordar a mi abuela Nona, que passed away justo cuando empezé el doctorado; y a mi abuelo Nono, que acaba de dejarnos justo antes de terminar mi doctorado. Quiero darle las gracias a ellos y a mis padres por amarme incondicionalmente desde siempre. Sin su amor y cariño, no sería lo que soy, ni tendría lo que tengo, gracias padres, abuelos y hermanos queridos. Los amo.
Finally, I have to thank my wife Claire for everything. She and Sofia mean to me more than any other person in this world. They bring light to my life like no other person can. Im very proud to be married with such beautiful, smart, good, amazing, dedicated, hardworker, delicious, supportive woman. Besides she is a great wife, she is a great scientist, and mommy. I love you Claire.
Wait, I am missing someone, someone that can not read yet, someone that habla chileno, frances, espanol, catala, ingles, probably a little bit of german and polish and serbian, someone that gives the best hugs and kisses. Someone that if you are hurt will cure you with a petite bisous, or that will ask for caballito, someone that probably knows more people already than any other kid at her age, who is this? Si, Sofia tu! Te amo hija, te amo mi sofi hermosa! Gracias por existir, prometo no romperte y amarte por siempre:P.

Abstract

This Doctoral Thesis aims to identify a comprehensive cancer-related gene landscape using signatures of natural selection. Our work addresses three major gaps in the literature of cancer evolution: on how intratumor heterogeneity fluctuates in a single cancer population, on how heterogeneity can help us predict genes driving cancer, and on how molecular evolution can reveal cancer essential functions. Bridging these gaps is fundamental to develop an effective universal cancer treatment. Using next generation sequencing data, we first quantify the subclonal and clonal heterogeneity within a chronic lymphocytic leukemia patient. Next, we analyze about 7,500 of tumors from different anatomical sites to detect driver genes. And lastly, we test the presence of purifying selection across all cancers. Initially, we found that the clonal composition of the leukemia tumor is responsive to two main selective forces: positive and negative selection. The first Pancancer analysis uncovered 609 tumor type driver gene connections using signatures of positive selection. Our results extend the role of chromatin modifiers to multiple cancers. The second Pancancer analysis revealed a strong purifying selection of 639 cancer essential genes, including two known oncogenes, *ABL1* and *TERT*. Furthermore, we investigated the effect of the identified cancer-related genes on patient prognosis. In conclusion, we have expanded the set of cancer-related genes by detecting drivers using positive selection, and cancer essential genes through unmasking the role of negative selection in cancer evolution.

Resumen

El objetivo de esta tesis doctoral es identificar un catálogo de genes relacionados con cáncer usando evidencia de selección natural. Nuestro trabajo se enfoca en tres problemas descritos en la literatura asociada a la evolución en cancer: el cómo varía la heterogeneidad intratumoral en única población de células cancerosas, el cómo esta heterogeneidad nos puede ayudar a identificar genes ”drivers”, y el cómo la evolución molecular nos puede revelar funciones esenciales para el cáncer. Resolver estos problemas es fundamental para el desarollo de un tratamiento universal contra el cáncer. Usando datos de secuenciación masiva, primero hemos cuantificado la heterogeneidad clonal y subclonal en un paciente con leucemia. Luego, hemos analizado mas de 7,500 casos provenientes de distintos tumores para detectar genes driver. Finalmente, hemos probado la presencia de selección negativa en todos los tumores disponibles. Inicialmente, encontramos que la composición clonal del caso de leucemia responde a dos fuerzas selectivas: selección positiva y negativa. El primer analisis ”Pancancer”nos revela 609 asociaciones entre genes driver y tipos tumorales usando la evidencia de selección positiva. Nuestros resultados destacan la importancia de genes modificadores de la cromatina a multiples tumores. El segundo analysis ”Pancancer”nos revela una fuerte presión selectiva actuando en 639 genes esenciales para el cancer, incluyendo dos oncogenes, *ABL1* y *TERT*. Además, hemos investigado el rol que tienen estos genes en el pronóstico de los pacientes. En conclusión, hemos expandido el número de genes relacionados con el cáncer por medio de la detección de genes driver usando selección positiva, y genes esenciales para el cáncer usando la selección negativa desenmascarada en evolución tumoral.

Contents

List of figures xv

List of tables xvi

I Global Introduction 1

1 Introduction 3
1.1 Of cancer and mutations . . . 3
1.1.1 A brief history . . . 3
1.1.2 Multi-stage theory of cancer . . . 3
1.2 The next-gen sequencing era of cancer research . . . 7
1.2.1 The 21st century . . . 7
1.2.2 Cancer genomics . . . 7
1.3 Tumor heterogeneity . . . 9
1.3.1 Levels of heterogeneity . . . 9
1.3.2 Fitness landscapes and tumor heterogeneity . . . 10
1.3.3 The dynamics of intratumor evolution . . . 12
1.3.4 Heterogeneity and treatment resistance . . . 14
1.3.5 Chronic lymphocytic leukemia . . . 15
1.3.6 The extent of heterogeneity in chronic lymphocytic leukemia 15
1.4 Clonal evolution and cancer gene prediction . . . 16
1.4.1 Estimating cellular prevalence . . . 16
1.4.2 Driver mutations, genes, and pathways . . . 17
1.4.3 Natural selection . . . 20

2 Objectives 23

II Tumor heterogeneity revealed in a case of chronic lymphocytic leukemia 25

3 A longitudinal analysis of a single case of CLL 26

4 Methods 31
4.1 Samples . . . 31

4.2 Whole-genome sequencing . . . 31
4.3 Whole-exome sequencing . . . 32
4.4 Targeted Enrichment Sequencing . . . 32
4.5 Cancer cell fraction (CCF) of cancer subpopulations using WGS and WES data . . . 33
4.5.1 Detection of somatic SNVs . . . 33
4.5.2 Detection of germline SNVs . . . 33
4.5.3 CCF calculation . . . 34
4.5.4 CCF for somatic SNVs . . . 34
4.5.5 CCF for somatic CNAs . . . 35

5 Supplementary tables 37

III Positive selection signatures reveal cancer genes across multiple tumor types 39

6 Introduction 40

7 Results 42
7.1 Evolutionary signatures used by cDriver . . . 42
7.2 Benchmarking cDriver performance . . . 46
7.3 Benchmarking in breast cancer (BRCA) and chronic lymphocytic leukemia (CLL) . . . 47
7.4 cDriver performance in a pooled dataset of 12 tumor types . . . 49
7.5 Tumor driver gene landscape across 21 tumors . . . 52

8 Discussion 55

9 Methods 58
9.1 Data . . . 58
9.1.1 Pancan12 somatic mutation data . . . 58
9.1.2 CLL somatic mutation data . . . 58
9.1.3 Pancan21 somatic mutation data . . . 59
9.2 cDriver package . . . 59
9.2.1 Bayesian inference models . . . 62
9.3 Benchmarking . . . 65
9.3.1 Running competing methods for cancer driver gene identification . . . 65
9.3.2 Gold standard and parameter selection . . . 65
9.4 Defining the landscape of tumor type driver gene connections in Pancan21 . . . 66

9.4.1 Identification of novel TTDG connections by PubMed mining . . . 67
9.4.2 Protein interaction and functional enrichment analysis of novel TTDGs . . . 67
9.4.3 Individual gene analysis . . . 67

10 Supplementary information **68**

IV Purifying selection reveals cancer essential functions **95**

11 Introduction **96**

12 Results **98**
12.1 Purifying selection predominates over positive selection in cancer genomes . . . 98
12.2 Functional role of genes under purifying selection . . . 101
12.3 Natural selection and mutation signatures in cancer . . . 106
12.4 Purified functions reveal tumor specific prognostic markers . . . 106

13 Methods **113**
13.1 Tumor Data . . . 113
13.2 K_n/K_s calculation . . . 113
13.3 Statistical analysis . . . 113
13.4 Postfiltration of cluster of synonymous changes . . . 114
13.5 Damage score of selected genes . . . 115
13.6 Functional enrichment . . . 115
13.7 GO term functional enrichment test and survival analysis . . . 115

14 Discussion **116**

15 Supplementary information **119**

V Global Discussion **133**

VI Appendix **141**

16 List of publications during PhD studies **141**

List of Figures

1.1 History of mutation and cancer 4
1.2 Log-log plot of cancer incidence by age 5
1.3 Stem cell divisions versus lifetime risk of cancer 6
1.4 Spectrum of somatic mutation in cancer genomes 8
1.5 Types of tumor heterogeneity 9
1.6 Clonal evolution of cancer 10
1.7 Schematic 3D fitness landscape 11
1.8 Models of tumor clonal dynamics 13
1.9 Strategies for detecting driver genes 18
1.10 Venn diagram for three lists of published driver genes 19

4.1 Simulation of CCF under a two-population model harboring a heterozygous deletion 35
4.2 Simulation of CCF under a two-population model harboring copy gains 36

7.1 Signatures of positive selection in tumor sequencing data 43
7.2 cDriver proccessing pipeline 45
7.3 Benchmarking of competing methods in breast cancer (BRCA) and chronic lymphocytic leukemia (CLL) 47
7.4 List of gold standard genes identified by competing methods . . . 48
7.5 cDriver results and comparison with other methods for dataset composed of 12 cancers 51
7.6 Novel tumor type - driver gene (TTDG) connections, *CHD4* and *SMARCA4* 54
10.1 F-score measure on filtered versus unfiltered data 68
10.2 Benchmarking five driver identification methods on three different datasets 69
10.3 Evaluation of several measures for five driver identification methods benchmarked on Pancan12 70
10.4 Somatic mutations in FLT3 71
10.5 Somatic mutations in PBRM1 71
10.6 Distribution of genes affecting tumor types 72
10.7 Extended figure of the tumor type driver gene connection landscape 73
10.8 STRING PPI analysis of selected genes 74

10.9 Chromatin modifiers affect a large proportion of individuals with cancer. 75

12.1 Venn diagram for significant genes of three datasets 100
12.2 Damage score mean at different K_a/K_s values 102
12.3 Mean ploidy and expression values for negative and positively selected genes 104
12.4 P2X7 signaling complex 107
12.5 Kaplan-Meier plots comparing wild-type versus mutated P2X7 signaling complex 108
12.6 Distribution of K_n/K_s values across multiple tumor and mutation types. 109

15.1 Models of evolutionary forces acting upon cancer genes 137
15.2 Somatic mutations and selection 138

List of Tables

5.1 Cancer cell fraction estimation for SNVs by WGS/WES sequencing of sample 016-T02 37
5.2 List of copy number altered loci found in case 016 accross multiple time points and the respective fraction of cells harboring the mutation 37
5.3 List of structural variants found in case 016 accross multiple time points 38

10.1 Data used for benchmarking cDriver against competing methods . 75
10.2 List of gold standard genes for breast cancer and chronic lymphocytic leukemia 76
10.3 Comparison of number of significant genes and the best F-score for each method 76
10.4 Data used for driver gene prediction using cDriver accross 21 tumor types 77
10.5 Full list of TTDG connections 92

10.6 Tumor-specific high confident drivers observed across tumor types using cDriver ranks 92
10.8 Fraction of affected patients that have a mutated chromatin modifier 93

12.1 Data used for analysis of purifying selection 99
12.2 Correlation between different selection estimates 99
12.3 Number of significant genes obtained in three datasets 99
12.4 Number of significant genes obtained in each tumor type based on three different tumor types 100
12.5 Negatively selected genes known as cancer genes 103
12.6 Enrichment of GO and Functional pathways of genes under significant purifying selection 105
12.7 Protein interaction members of the P2X7 receptor signaling complex 106
12.8 Number of mutations of *TP53* for each category 110
12.9 GO terms under significant purifying selection 110
12.10 GO terms significance on prognosis 112
15.1 List of significant genes under selection 132

Part I

Global Introduction

> "There is a single light of science, and to brighten it anywhere is to brighten it everywhere"
>
> *Isaac Asimov*

Cancer, as a worldwide disease, affects more than 14 million new patients every year and causes more than 8 million deaths (http://www.cancer.gov/). The risk that a person develops cancer during their lifetime is 42% for males and 37% for females [Howlader et al., 2015], and the chances of cancer-related deaths are 21% and 18%, respectively. Cancer results from an abnormal proliferation of any of the cells in our body. If such proliferation is confined to a specific location and not able to propagate to surrounding cells or tissues, it is denominated a benign tumor [Cooper Geoffrey, 2000]. While most tumors remain benign, some can become malignant and threaten the indivual's life. For several decades, efforts have been focused on understanding the basis of cancer malignancy, as well as improving prevention, early diagnosis and treatment options. So far, many therapeutic options are available, but none holds the promise of universal treatment which could help us permanently defeat the disease. The reason successful treatment options are hard to come by is that tumors adapt rapidly due to their heterogeneity, which results in drug resistance and relapse.

This thesis aims to explore three fundamental and interrelated concepts underlying cancer etiology and development: **tumor heterogeneity**, **positive selection**, and **purifying selection**, in other words, cancer evolution. Before delving into each specific project, we aim to concisely review the topic of tumor heterogeneity and evolution. The amount of cancer literature is vast, therefore this introduction

represents our effort to condense information relevant for the prediction of cancer genes.

First, **tumor heterogeneity**, as a product of random mutations, is driven by environment-dependent selective pressures. If the acquisition of a novel function (i.e. increased proliferation rate or the suppression of cell surveillance mechanisms), increases fitness and is, therefore, under **positive selection**, a tumor clone can arise. Subsequently, other alterations can occur and transform the tumor clone into a malignant one, ultimately, causing cancer. The selection of advantageous and the purification of deleterious mutations shapes the fitness landscape of a tumor, therefore giving rise to tumor heterogeneity. Not all cancer patients share the same genetic alterations, and the same can be said for cancer cells from a single patient.

Lastly, although it is hard to imagine the presence of positive selection without **negative selection**, most current cancer studies aiming at identifying cancer-related genes have focused extensively on the former. The latter has largely been neglected for various reasons, including limited cohort size and lack of statistical power. The advent of next generation sequencing technologies and the resulting availability of thousands of sequenced tumor samples, has marked the start of a new era for exploring cancer genomes using evolutionary concepts, and these new technologies will help us shed light onto a universal strategy to treat cancer patients.

Chapter 1

Introduction

1.1 Of cancer and mutations

1.1.1 A brief history

Cancer begins with the formation of a tumor, usually described as an accumulation of cells with an increased growth capacity. In the early 20th century, the first evidence of chromosomal alterations linked to cancer was discovered [Balmain, 2001][Stratton et al., 2009]. In 1914, Theodor Boveri postulated that tumors might arise from an abnormal set of chromosomes passed on to daughter cells [Boveri, 1914]. Cancer research scientists spent half a century trying to refine Boveri's initial theory (see Fig 1.1 adapted from [Knudson, 2001]). Almost 50 years passed until the theory was finally proven, by the discovery of the "Philadelphia chromosome" [Nowell and Hungerford, 1960] and the first oncogenes (i.e. genes that when amplified or overexpressed stimulate cell proliferation). In parallel, the work of Knudson statistically demonstrated that the acquisition of two inactivating mutational events were sufficient to develop Retinoblastoma [Knudson, 1971]. A decade later, the first tumor suppressor gene involved in Retinoblastoma was cloned, the *RB* gene [Friend et al., 1986]. Later, Fearon and Vogelstein reviewed a "multi-stage" model for the genetic progression of colorectal tumors from adenoma to carcinoma. They reported that a minimum of four to five mutations, including activation of an oncogene and inactivation of a tumor suppressor, were needed to develop a malignant tumor[Fearon and Vogelstein, 1990].

1.1.2 Multi-stage theory of cancer

Supporting evidence for the multi-stage theory of cancer progression came from the analysis of age-specific curves of cancer incidence. The incidence tends to increase logarithmically with age [Armitage and Doll, 1954] [Nordling, 1953] indicating a progressive accumulation of mutations over time, six to seven cellular changes underlying carcinogenesis co-occur within the same cell (See Fig. 1.2). Although such a conclusion seems to be valid for most tumors, others did not fol-

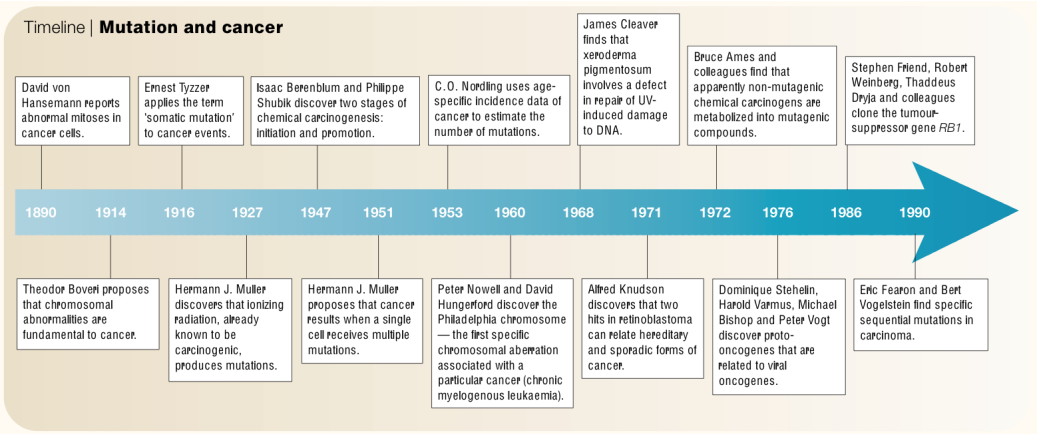

Figure 1.1: **History of mutation and cancer.** Timeline describing the historical landmarks on the discovery of mutations and their relationship to cancer. Reproduced from [Knudson, 2001].

low the expected behavior[Armitage and Doll, 1954]. For instance, (1) tumors exhibiting a cancer-predisposing *de novo* or inherited germline mutation [Knudson, 1971], (2) tumors exhibiting an intermediate growth advantage [Ashley, 1969], and (3) tumors strongly influenced by environmental factors such as tobacco exposure [Proctor, 2001], show an increased incidence. As a result of these studies, it was evident that more complex models including the division rate, the mutation rate, mutagen exposure, and the selective processes acting on the cell type were needed to explain tumor-specific incidence curves.

Another exception to the multi-hit theory came from the cytogenetic analysis of tumors. Nowell and Hungerford observed a recurrently small chromosome 22 when examining chronic myelogenous leukemia (CML) cells [Nowell and Hungerford, 1960]. This chromosome was named the Philadelphia chromosome (Ph) and consisted of a translocation between two chromosomes, 9 and 22. The fusion of these two chromosomes activated the Abelson gene (*ABL*), creating a chimeric BCR-ABL transcript. Activation of BCR-ABL increases tyrosine kinase activity [Konopka et al., 1985], interfering with cell cycle regulation and apoptosis [Skorski et al., 1997]. Recently, the *ABL* gene was identified as cancer cell essential in a CRISPR-based screening of a CML cell line [Wang et al., 2015]. Inactivating the gene is lethal for the cell, a fact consistent with our finding that the *ABL* gene is under strong purifying selection. An extended revision of the roles and the potential implication of *ABL* in leukemias and solid tumors has been provided in [Greuber et al., 2013].

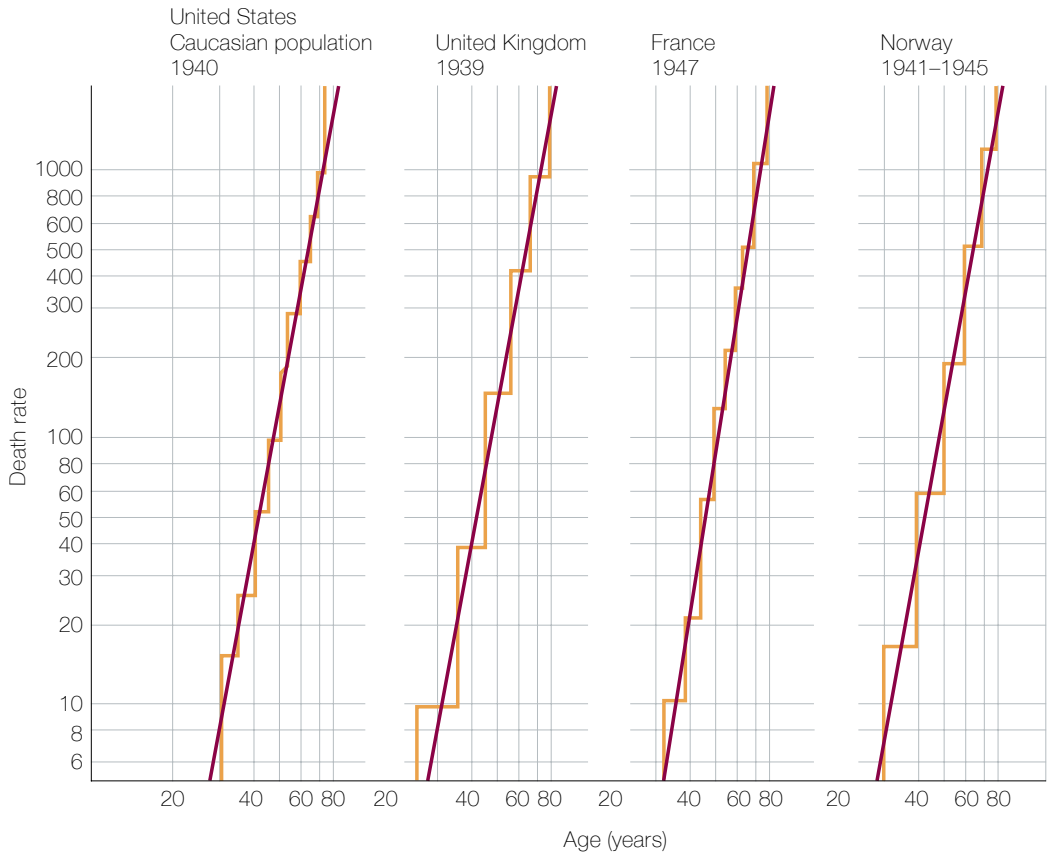

Figure 1.2: **Log-log plot of cancer incidence by age.** Log-log plots of cancer death rates in males versus age. Reproduced from [Knudson, 2001].

The body is composed of a large number of different cell types. Every proliferating cell has to make the decision of either continuing to proliferate or becoming quiescent [Malumbres and Barbacid, 2001]. Cell division rates vary drastically between different cell types: a hematopoietic stem cell divides every 30 days, a brain cell does not divide, and a human fibroblast is able to perform between 60 to 80 cell divisions *in vitro* [Mathon and Lloyd, 2001]. Interestingly, Tomasetti and Vogelstein, as well as Hao and colleagues, have recently shown that several tumor types arise simply as a function of the number of stem cell divisions [Tomasetti and Vogelstein, 2015](Fig. 1.3, [Hao et al., 2016]), demonstrating that some tumor types are mostly produced by pure ”bad luck”. Nonetheless, the amount of endogenous ”bad luck” versus exogenous factors remains controversial [Wu et al., 2016]. In any case, during every cell division, telomeres shrink, eventually driv-

ing the cell to senescence (i.e. a barrier acting as a tumor-suppressor mechanism) [Campisi and d'Adda di Fagagna, 2007]. To escape from such senescent or apoptotic signal produced by telomere attrition, the tumor cell has to activate *TERT*,a telomerase reverse transcriptase gene [Zhang et al., 1999].

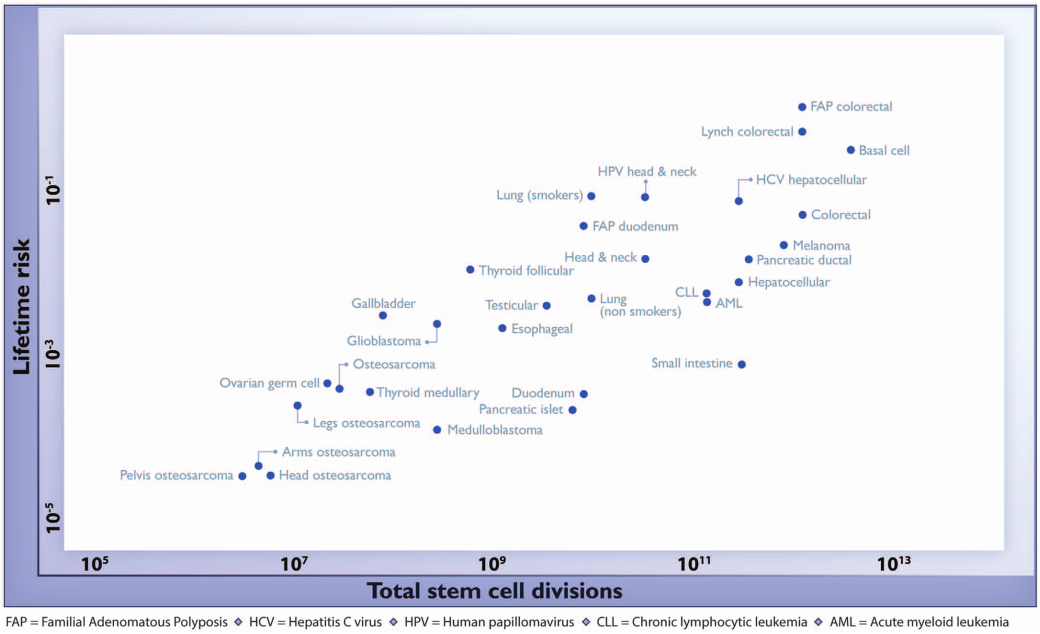

Figure 1.3: **The relationship between stem cell divisions and lifetime risk of cancer.** Reproduced from [Tomasetti and Vogelstein, 2015].

If the number of somatic changes underlying cancer etiology is tumor-type specific, then how many events are needed for the establishment of the malignant phenotype? Which of these events are relevant? Can we therapeutically inhibit these events? Hereditary retinoblastoma and CML just needed one-hit whereas colorectal tumors required up to six hits. In addition, it was observed that chromosome (CIN) and microsatellite instability (MIN) can further change the number of hits needed for tumor development [Nowak et al., 2002][Cahill et al., 1999], therefore complicating the identification of mutations in genes which are causally implicated in oncogenesis. Although great advances were made in the 1990s in the discovery of oncogenes and tumor suppressors [Weinberg, 1996] [Weir et al., 2004], it was not until the establishment of next generation sequencing (NGS) technologies that the full landscape of cancer genomes could be explored [TCGANetwork, 2008].

1.2 The next-gen sequencing era of cancer research

1.2.1 The 21st century

Two key historical scientific achievements have paved the way for our current understanding of cancer. One was the release of the first draft of the human genome in 2001 after a frenetic competition between public and private efforts [Lander et al., 2001]. The other was the development of novel high throughput sequencing technologies (NGS), including pyrosequencing [Margulies et al., 2005] and sequencing by synthesis [Bentley et al., 2008]. In contrast to the three-decades used Sanger sequencing, these new technologies were able to read the DNA in parallel and output billions of sequencing reads [Mardis, 2008]. The current cost for re-sequencing a whole human genome at 30X coverage is about 1,000 USD [Hayden, 2014]. The most widely used sequencing machine is from Illumina, but others such as PacBio or Oxford Nanopore, provide complementary sequencing chemistries and novel features. The Illumina HiSeq, MiSeq and NextSeq machine families use the sequencing-by-synthesis method to produce short read fragments of up to 300 base pairs. Novel technologies, such as PacBio or Nanopore are reaching read lengths of up to 100kb, which will allow us to observe genomic variation at an unprecedented resolution. Nevertheless, some limitations remain akin the amount of single nucleotide or insertion/deletion errors reported, which sometimes can be solved by raising the coverage of the region.

The massive amounts of NGS data has allowed us to explore the variability present in hundreds of individuals and their relationship to diseases and traits. Particularly for cancer studies, sequencing the tumor tissue and the germline DNA (normal tissue) of a patient enabled the identification of somatic variants that could underlie cancer development. Transcriptome sequencing allows for reconstruction of the expression patterns observed in a cancer sample. Whole-Exome-Sequencing (WES), on the other hand, allows the identification of all possible single nucleotide variants or short insertion/deletions in coding regions, as well as the copy number state (ploidy level). Lastly, Whole-Genome-Sequencing (WGS) allows for the exploration of the non-coding part of the genome. Other ”omics” approaches, such as ChIP or DNA methylation sequencing, reveal different features of the chromatin of a cancer cell and are reviewed in [Reuter et al., 2015].

1.2.2 Cancer genomics

The Cancer Genome Atlas Initiative (TCGA) was launched in the mid 2000s with the goal of characterizing the universe of genomic changes involved in all types

of human cancer `http://cancergenome.nih.gov/`. The first publication revealing the landscape of genomic alterations in a single cancer type, the brain tumor glioblastoma multiforme [TCGANetwork, 2008], initiated a massive release of genomic data covering several tumor types. The first analyses of different tumor tissues revealed the extent of genetic inter-tumor heterogeneity: different tumor tissues harbor a variable number of somatic mutations of characteristic mutation signatures (Fig. 1.4). In addition, studies on malignancies in the lung [TCGANetwork, 2012a], brain [Parker et al., 2016], ovarian [TCGANetwork, 2011], colon and rectum [TCGANetwork, 2012b], breast [TCGANetwork, 2012c], pancreas [Bailey et al., 2016] among others, revealed an extensive intratumor heterogeneity [Vogelstein et al., 2013], that ultimately hampers global cancer treatment strategies and hinting at the urgent need for personalized medicine.

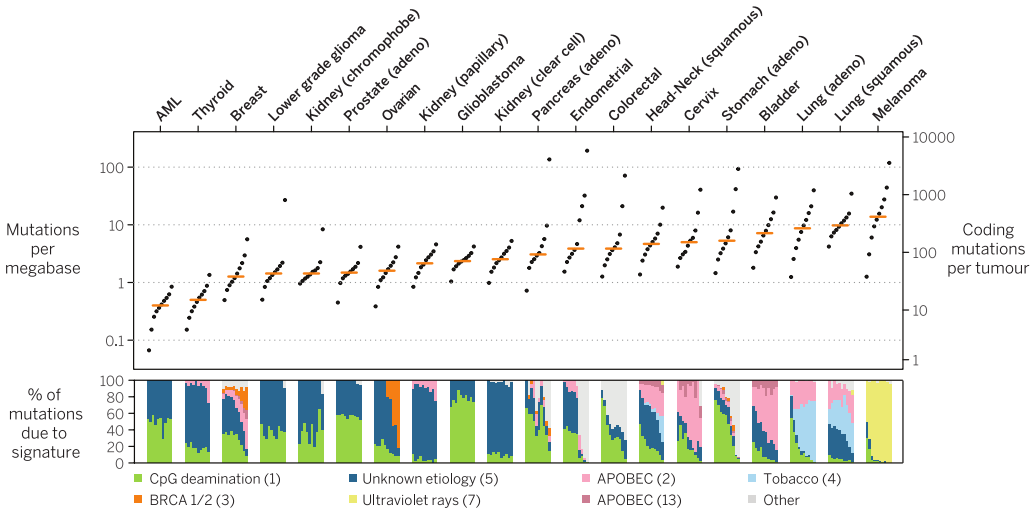

Figure 1.4: **Spectrum of somatic mutation in cancer genomes.** Number of somatic mutations per megabase across 20 tumor types. (Top) The median of deciles is shown as a dot, the orange bar is the median of all samples. (Bottom) The contribution of different mutation signatures as stacked bars. Reproduced from [Martincorena et al., 2015a].

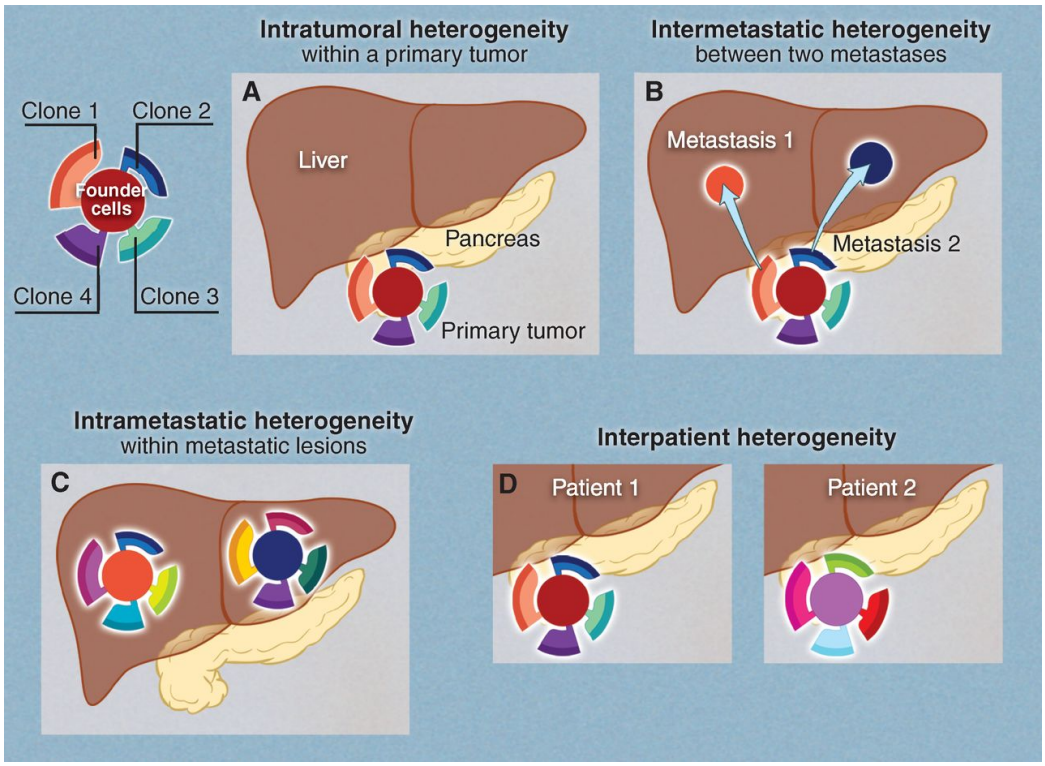

Figure 1.5: **Types of tumor heterogeneity.** Reproduced from [Vogelstein et al., 2013].

1.3 Tumor heterogeneity

1.3.1 Levels of heterogeneity

Cancer is an heterogeneous disease on different levels, including inter- and intra-tumor heterogeneity, as well as differences between primary and metastatic tumor cells (Fig. 1.5). As previously described, the variability between tumor types, or inter-tumor heterogeneity, was recognized by age-cancer incidence patterns, cell division dynamics, and the number of mutations needed to develop the malignancy. Also, UV light and tobacco smoke produce particular mutation signatures on the genome revealing the importance of environmental exposure to different carcinogens [Alexandrov et al., 2013]. Therefore, it was clear that the genetic composition (genotype) of the patient, multiple cell-type-specific factors, and the environment can lead to a patient-tissue-specific set of genetic abnormalities.

Intratumor heterogeneity was recognized in the 1970s by observing different biological properties, including antigen resistance, immunogenicity, and growth rate.

These discoveries led to the hypothesis of tumor evolution based on the acquisition of alterations conferring a selective advantage to a neoplasic cell over surrounding cells [Nowell, 1976]. Under this scheme, further alterations can lead to new better-adapted clones that can become the next predominant cancer subpopulation in a specific microenvironment. In addition, there are multiple mutational trajectories but only some of them are viable. For instance, in renal carcinomas a gene regulating apoptosis, *VHL*, needs to be deactivated before other mutations influencing proliferation rates become acceptable [Gerlinger et al., 2012]. While mutations conferring a selective advantage to the cancer cell will remain and increase their frequency in the cancer cell population, reduced-fitness or mildly deleterious mutations are removed from the genetic pool by eliminating the cells carrying them (See fig. 1.6).

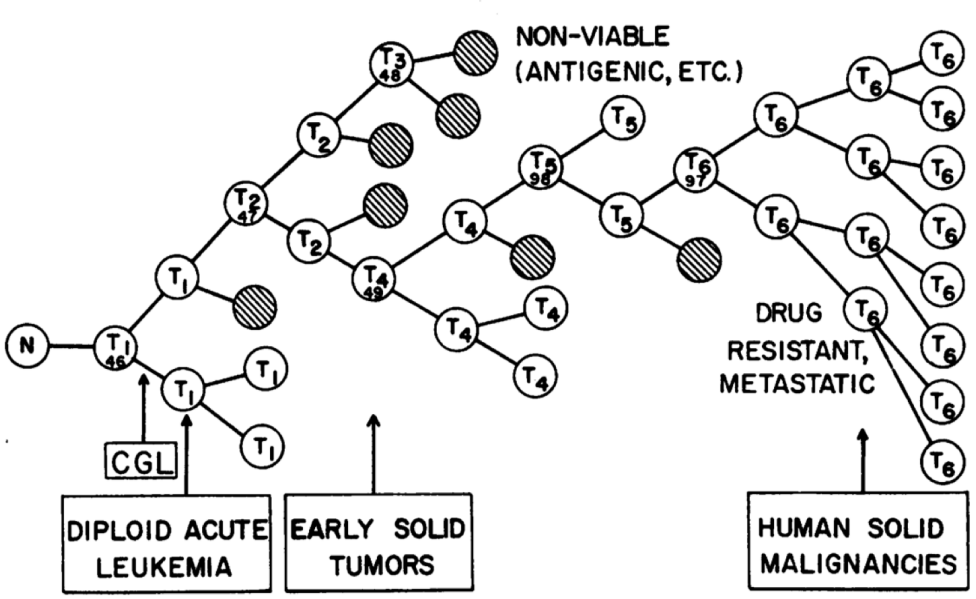

Figure 1.6: **The process of clonal evolution of cancer.** Adapted from [Nowell, 1976].

1.3.2 Fitness landscapes and tumor heterogeneity

To understand the basis of cancer heterogeneity, we first need to understand the concept of fitness landscape proposed by Sewall Wright in the 1930s [Wright, 1932]. A fitness landscape is defined as the "relationship between selection, structure, and adaptation" [Simpson, 1944], and is visualized as a hypothetical terrain

composed of peaks and valleys (Fig. 1.7), where each particular genotype has an associated fitness value. In general, empirical fitness can be defined as any trait of interest such as infectivity [Hayashi et al., 2006], fluorescence intensity [Sarkisyan et al., 2016], or proliferation. In cancer, somatically acquired mutations alter fitness by modifying the birth-death rate of the cell [Merlo et al., 2006]. Mutations conferring an increased fitness (i.e. higher proliferation) are under positive selection [Babenko et al., 2006] and have been termed driver mutations [Greenman et al., 2007]. The rest of mutations that do not contribute to tumorigenesis are called passengers [Stratton et al., 2009]. Many passengers can accumulate before malignant transformation occurs, and then hitchhike along the fitness landscape together with the positively selected mutation [Smith and Haigh, 1974]. Most passenger mutations are random mutations with no selective advantage, however they may modulate the mutation rate [Lynch, 2016], or become important later through co-option [Billaud and Santoro, 2011].

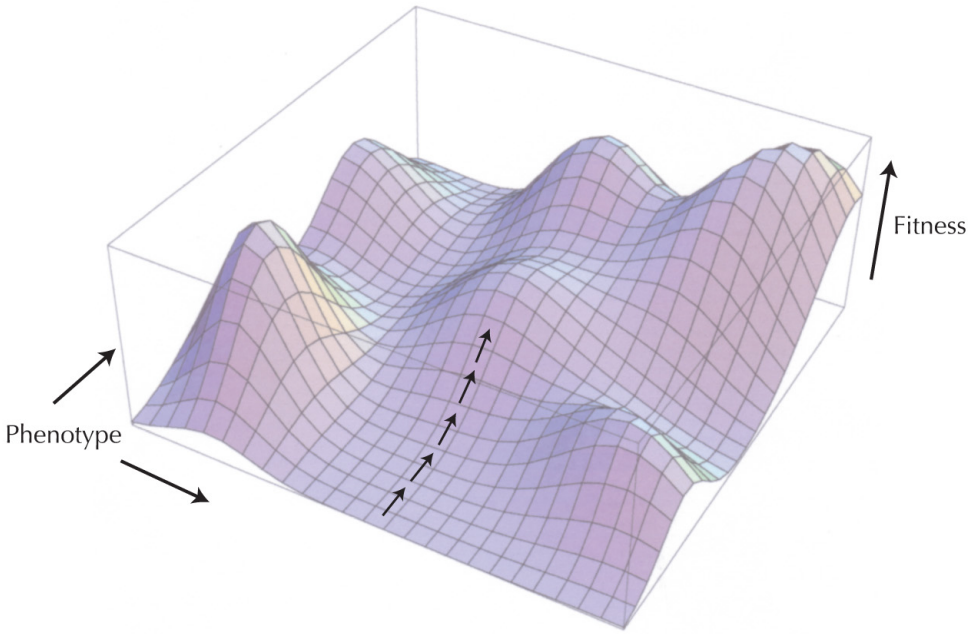

Figure 1.7: **Schematic 3D fitness landscape.**

The accumulation of somatic changes moves the cell along the fitness landscape. Several major driver mutations have been identified in genes such as *APC, KRAS* and *TP53*. Nonetheless, many cancer patients do not harbor mutations in any of these major drivers. It has therefore been proposed that weak tumor-promoting

mutations accumulate in "mini-driver" genes[Castro-Giner et al., 2015]. This is consistent with a polygenic model of small selective advantages, which slowly move the cell towards a fitness peak. If we imagine the multiple paths for random mutations to accumulate, it is not surprising to observe a substantial amount of heterogeneity within and between tumor types. In addition, genes perform multiple functions (pleiotropy) and, at the same time, specific functions are performed by multiple genes (polygenic). Consequently, multiple genetic combinations can give rise to the same phenotype and thus have the same or very similar fitness. Hanahan described that cancer develops through the alteration of six main hallmarks [Hanahan and Weinberg, 2011]. Therefore it is likely that during tumorigenesis, two cancer cells acquire a somatic change in different genes affecting one of these hallmarks, fueling the observed genetic heterogeneity.

1.3.3 The dynamics of intratumor evolution

A considerable problem for cancer treatment is the pervasive presence of tumor heterogeneity [Michor and Polyak, 2010]. One of the key aspects of tumor progression is the appearance of a large number of individual clones carrying common and clone-specific mutations [Torres et al., 2007] [Navin et al., 2010] [Yachida et al., 2010]. This observation was also evident from early cytogenetic studies showing distinct karyotypes within the same tumor [Wolman, 1986]. As many mutations accumulate over time, some seemingly neutral mutations may be adaptive in a novel environment (i.e. chemotherapy) [Podlaha et al., 2012], conferring resistance and ultimately leading to a relapse of the disease (Fig. 1.8). Several studies using next generation sequencing (NGS) technologies have attempted to reconstruct the evolutionary history of single solid tumors such as renal-cell carcinoma [Gerlinger et al., 2012], breast [Yates et al., 2015], prostate [Cooper et al., 2015], brain [Sottoriva et al., 2013], lung [Zhang et al., 2014], among others. Following up on these findings, it has also been demonstrated that the tumor clonal composition changes after treatment [Almendro et al., 2014] and that knowing the composition of the clonal architecture can predict treatment success [Zhao et al., 2014b].

One of the first genomic studies revealing the great extent of intratumor heterogeneity came from the analysis of multiple regions of 20 breast carcinomas [Navin et al., 2010]. These 20 carcinomas were classified as either monogenomic or polygenomic based on the pattern of copy number alterations present in the sections. Monogenomic tumors consisted of a homogenous population of tumor cells, whereas polygenomic tumors carried multiple subclonal populations harboring variable genetic abnormalities. A similar study was performed using flow-sorted

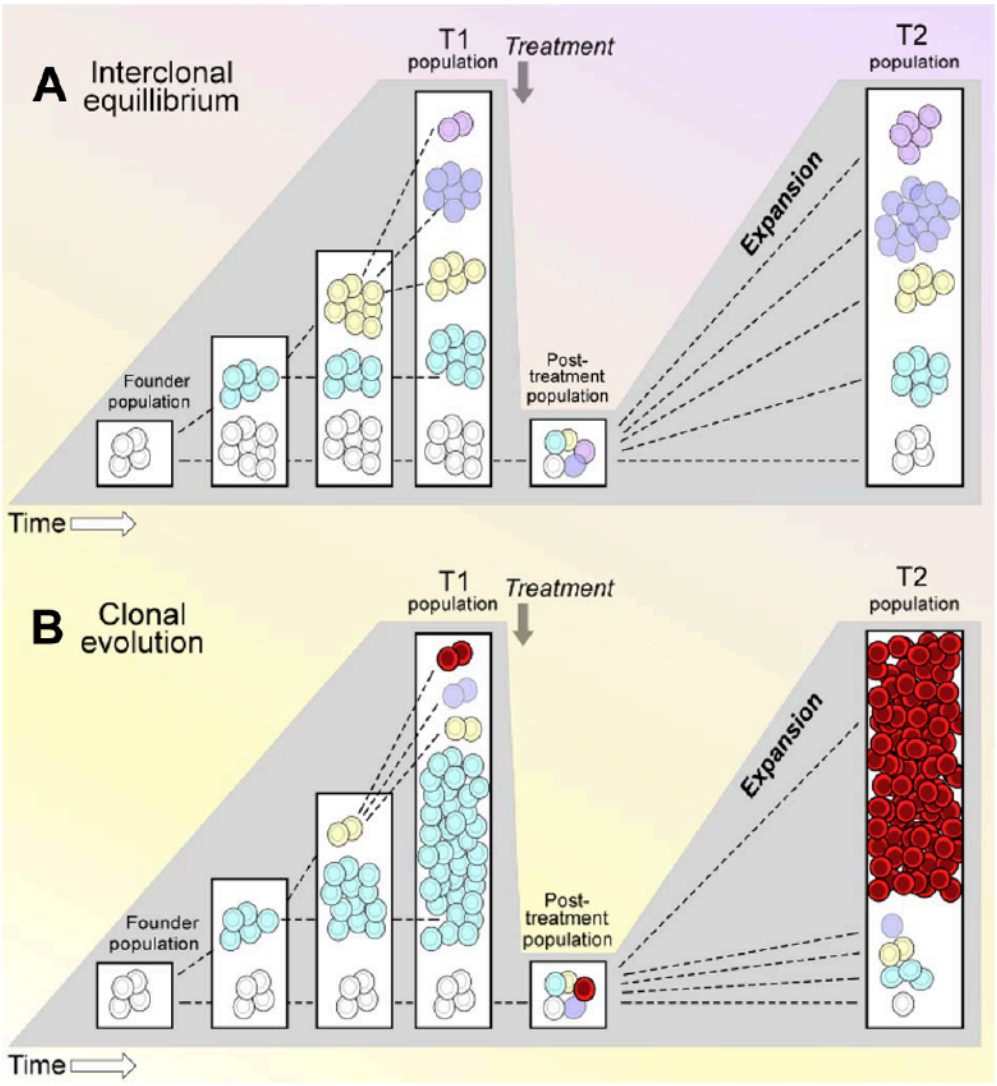

Figure 1.8: **Models of tumor clones progression in a blood tumor.** Adapted from [Wu, 2012].

nuclei revealing that in a single tumor biopsy, a clonal expansion formed the primary tumor and then seeded the metastatic sample, and many subpopulation were present [Navin et al., 2011]. Another example from primary renal carcinomas revealed a pattern of branched evolution where few mutations were shared by all sections, others were present only in the primary tumor, others were metastasis specific, and others were private to each individual sample site [Gerlinger et al., 2012]. In the latter study, mutations in *VHL* were ubiquitous while *PTEN* muta-

tions were lineage-specific, unmasking a clonal evolution pattern that started in the primary site and "traveled" to other regions.

These findings sparked the development of mathematical models of tumor progression [Altrock et al., 2015]. A recent analysis of colorectal tumors concluded that most region-specific mutations arose in an early clone which had already achieved an optimum fitness [Sottoriva et al., 2015]. Therefore, the total fraction of cancer cells harboring these private alterations is directly related to the time of appearance when the major cancer clone was expanding. Such model points to a neutral accumulation of low frequency variants without a major selective advantage [Williams et al., 2016]. Interestingly, ultra-deep sequencing of different sections of a normal epithelial tissue has revealed an extreme accumulation of somatic mutations in driver genes, without driving a malignant phenotype [Martincorena et al., 2015a]. These findings raised interesting questions that still remain to be answered: Is selection a driving force in cancer? Is it only important for tumor initiation but not tumor progression? Or vice versa? Do these observations hold for non-solid tumors, such as hematological malignancies?

1.3.4 Heterogeneity and treatment resistance

It is widely known that a cancer treatment will often fail to remove all cancer cells from a patient, suggesting an acquired resistance to the drug. The source of resistance remains to be elucidated: does resistance come from somatic mutations or from an already resistant cell previously not previously detected? [Podlaha et al., 2012]. For instance, methotrexate (MTX) is a classical cytotoxic drug used in cancer therapy that inhibits the DHFR enzyme. If a cancer subclone shows enhanced *DHFR* activity due to extra copies of the gene, the treatment will eliminate the rest of the clones but positively select this one [Rosowsky et al., 1985],[Morales et al., 2005].Therefore, an alternative treatment would consist of targeting all possible relapsing clones in combination with the major clones [Zhao et al., 2014a],[Zhao et al., 2016]. Further studies on MTX have proved valuable to understand resistance mechanisms involving clonal heterogeneity [Chabner and Roberts, 2005]. Recent technological advances can help us to dissect the full landscape of genomic mutations that may cause treatment resistance. To achieve this, we need better strategies that quantify tumor heterogeneity and assess how it will affect the response [Griffith et al., 2015] [Tabassum and Polyak, 2015].

1.3.5 Chronic lymphocytic leukemia

Intratumor heterogeneity has also been observed in hematological malignancies such as acute myelogenous leukemia (AML) [Ding et al., 2012] and chronic lymphocytic leukemia (CLL) [Schuh et al., 2012]. The latter, CLL, is the predominant leukemia in the western world with an incidence of 4.5 per 100,000 individuals [Fabbri and Dalla-Favera, 2016]. The malignancy is characterized by an abnormal accumulation of mature B-lymphocytes in the blood and lymph nodes, and two major subtypes have been identified, an agressive form defined by an unmutated Ig heavy chain variable locus (IGHV) and a non-aggressive form defined by a mutated IGHV locus [Hamblin et al., 1999]. Additionally, genomic studies have reported a high degree of genetic and molecular heterogeneity inter- and intra-CLL patients [Landau and Wu, 2013].

CLL is an interesting model in which to analyze tumor evolution and progression since sampling is quite accessible (blood extraction), tumor purity is very high (>98%), and all possible subclones are mixed together. A study describing the underlying genomic alterations driving CLL has recently been published [Puente et al., 2015]. A deletion on the q arm of chromosome 13 (Del13q14) is the most recurrent alteration across CLL patients followed by a deletion of chromosome arm 11q (Del11q), and single point mutations in the NOTCH receptor. Chromosomal aberrations and copy number alterations are observed in more than 80% of CLL cases, a percentage which is similar in other tumors [Knight et al., 2012]. About 2-3% carry chromothripsis, a catastrophic event where multiple focal copy number alterations appear in one chromosome [Stephens et al., 2011]. Nevertheless, it is important to note that there is no common genetic event explaining most cases and that multiple subclonal alterations are present [Landau et al., 2013].

1.3.6 The extent of heterogeneity in chronic lymphocytic leukemia

In 2011, a longitudinal analysis of CLL samples using whole-genome sequencing revealed a heterogeneous clonal architecture [Schuh et al., 2012]. This study analyzed the effect of chemotherapeutic therapy on the population dynamics of observed subclones. After treatment, a massive extinction of white blood cells occurred, then for every case the tumor clones re-expanded with different temporal-spacial dynamics. In one case, clones show an equilibrium similar to the pre-treatment stage (see Fig 1.8a). In another case, a minor subclone in the pre-treatment stage replaced the entire population (see Fig 1.8b). Further studies have also suggested that multiple coexisting subclones carrying specific mutations may be positively selected in the case of an environmental stress [Wu et al., 2012]

[Braggio et al., 2012], underscoring the importance of subclonal quantification for proper diagnosis and treatment selection. In the second part of this thesis, we analyzed a single CLL case and developed a mathematical formula to quantify subclonal mutations based on the allele frequencies obtained from sequencing data.

1.4 Clonal evolution and cancer gene prediction

1.4.1 Estimating cellular prevalence

Exome sequencing allows the identification of all possible mutations in coding regions for genes using algorithms such as Mutect [Cibulskis et al., 2013] and GATK[McKenna et al., 2010], as well as the copy number state (ploidy level) per gene based on the coverage. Notably, it can be scaled to several samples allowing for identification of recurrently mutated genes and common CNAs. Algorithms optimized for detection of somatic variants using tumor-normal pairs facilitate the sensitive estimation of allelic variants in a low fraction of the tumor population [Cibulskis et al., 2013]. Whole-exome and whole-genome sequencing allow for the characterization of the clonal subpopulation structure of thousands of tumor samples.

Several methods have been developed to quantify the cellular prevalence of a mutation (i.e cancer cell fraction) and to reconstruct the clonal architecture of a single tumor sample [Ding et al., 2014a][Li et al., 2014]. Initially, methods were focused on determining the allelic copy-number from an admixture of aberrant and non-aberrant cells using SNP arrays[Van Loo et al., 2010]. Soon after, similar methods were adapted to use whole-exome and whole-genome sequencing data [Su et al., 2012], [Oesper et al., 2013], [Lönnstedt et al., 2014] [Fischer et al., 2014]. Later, these methods adopted more sophisticated statistical strategies including clustering of probabilities [Andor et al., 2013], karyotype likelihoods [Carter et al., 2012], hierarchical Bayesian clustering [Lee et al., 2014][Roth et al., 2014], and Bayesian mixture modeling [Miller et al., 2014]. Nevertheless, two important steps were common to most of the methods: first, the somatic allele count per SNV should be transformed into a cell fraction based on parameters such as purity and locus ploidy, and second, the data from multiple SNVs should be deconvoluted into the number of individual clones. The main assumption here is that if two variants have different allele counts, they must belong to two different clones.

These methods have been widely applied to infer the tumor clonal structure of individual tumor types. Most have used sequencing data from multi-regional [Gerlinger et al., 2012] data or longitudinal analyses [Schuh et al., 2012] to determine the presence of multiple subpopulations. The number of clones present in a single sample varies from one to five in AML [Miller et al., 2014], one to 16 in glioblastoma [Andor et al., 2013], and only one or two in melanoma [Ding et al., 2014b]. Another study has tested the value of single cell sequencing to reconstruct tumor heterogeneity in breast cancer [Wang et al., 2014a]. Nevertheless, the question remains: how representative is the biopsy of the whole tumor sample[Yates and Campbell, 2012]. Interestingly, despite all the efforts to reconstruct the clonal structure of a tumor sample, and the importance of tumor heterogeneity for therapeutic management, cellular prevalence has remained unused for prediction of driving events.

1.4.2 Driver mutations, genes, and pathways

Understanding the effect of somatically acquired mutations is a central goal in cancer studies [Greenman et al., 2006][Greenman et al., 2007]. Whether these mutations are single nucleotide variants, insertions/deletions, copy number alterations, epigenetic modifications, or large scale variants, their relationship to tumorigenesis is far from completely understood. A complicating issue is the distinction between ”driver mutation” and ”driver gene” [Vogelstein et al., 2013]. Many genes labeled as drivers do not harbor point mutations since their relationship to tumorigenesis is due to activating translocations [Ren, 2005], or overexpression induced by regulatory variants [Horn et al., 2013] . To avoid such confusion, genes specifically harboring driver mutations have been named ”Mut-driver” or mutational driver genes. Moreover, recent approaches are trying to uncover mutated ”driver” pathways [Leiserson et al., 2013], further stressing the need to revisit the ”driver versus passenger” concept. It is still not clear whether a driver gene is specific to one or to many tumor types. In the second part of this thesis, we use ”driver” to refer to a mutational driver.

Several methods have been developed to identify driver genes [Marx, 2014]. Statistical approaches have used recurrence [Dees et al., 2012][Hua et al., 2013], genetic context [Lawrence et al., 2013], functional impact [Gonzalez-Perez and Lopez-Bigas, 2012] [Bertrand et al., 2015], and spacial features [Tamborero et al., 2013b] of somatic mutations observed across multiple patients (Fig. 1.9). On one hand, recurrence methods have focused on signals coming from a large number of patients, benefiting mainly from positive selection at the individual level. Functional impact methods rely on the prediction of a damage score, which represents

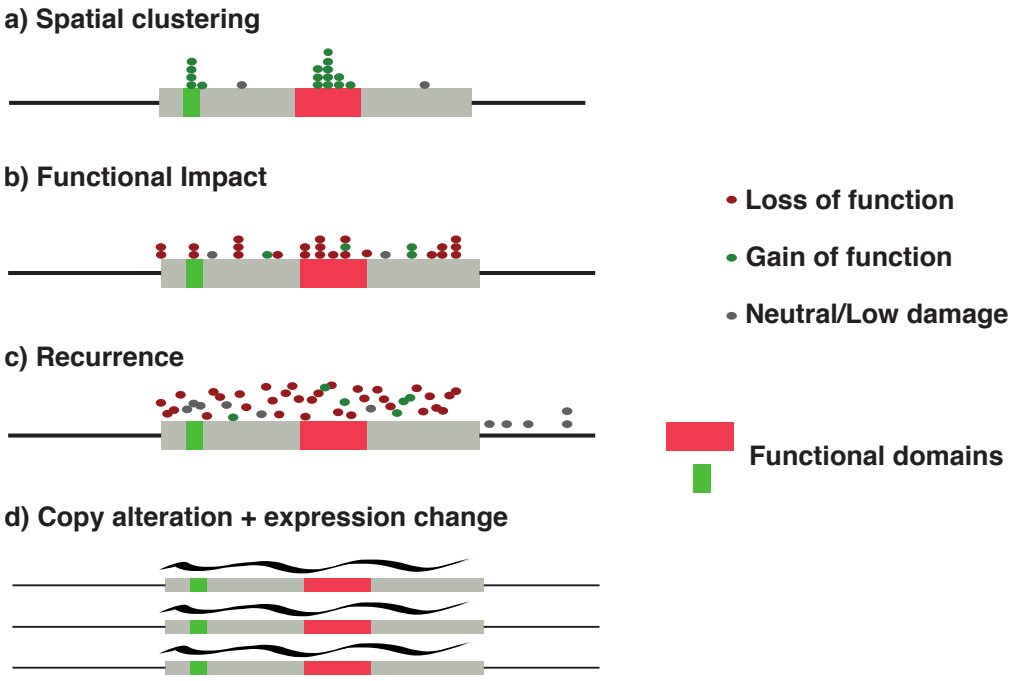

Figure 1.9: **Strategies for detecting driver genes.** a) Spatial clustering. Tools such as oncodriveCLUST [Tamborero et al., 2013b] and ActiveDriver [Reimand and Bader, 2013] identify genes with a clustering bias towards particular sites. b) Functional impact. OncodriveFM [Gonzalez-Perez and Lopez-Bigas, 2012] identify genes by exploiting pre-calculated functional scores of the substitution at that position. c) Recurrence. MuSiC [Dees et al., 2012] identifies genes with a significantly high observed-to-expected ratio of affected patients. MutSig-CV [Lawrence et al., 2013] uses the same principle but includes other features for calculating the background mutation rate. d) CONEXIC [Akavia et al., 2010] and oncodriveCIS [Tamborero et al., 2013d] identifies genes with a differential expression and copy-number alteration bias.

molecular selection. Other methods have tried to differentiate between oncogenes and tumor suppressors by the spacial distribution of activating mutations [Schroeder et al., 2014] [Davoli et al., 2013], while others have tried to identify mutated pathways underlying cancer initiation [Leiserson et al., 2013]. To date, a combination of methods to capture driver genes with different underlying principles is the most successful strategy [Tamborero et al., 2013c] [Youn and Simon, 2011]. Surprisingly, despite the lessons learned from tumor heterogeneity (driver genes are significantly more clonal than others [McGranahan et al., 2015]), cellular prevalence of somatic mutations has not been used as a signature of positive selection to identify driver genes.

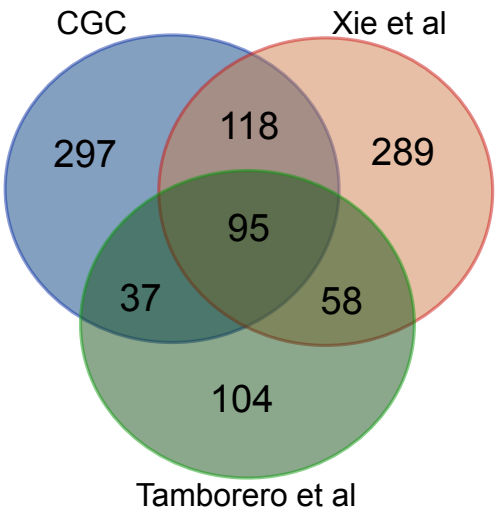

Figure 1.10: **Venn diagram for three lists of published driver genes.** CGC, refers to the cancer gene census database [Futreal et al., 2004]. Tamborero et al and Xie et al refer to the list of driver genes published in [Tamborero et al., 2013c] and [Xie et al., 2014], respectively.

The cancer gene census (CGC) was the first catalog of cancer genes released in 2004 [Futreal et al., 2004]. The CGC list is regularly updated by a curator team solely based on publication evidence. Currently, there are 595 genes, and of these, 55 are tumor suppressors (TS), 84 oncogenes (OG), 17 TS/OG, and 439 are uncharacterized. Interestingly, 326 of the 595 have been found only due to translocation evidence, whereas less than 100 have been found due to single point mutations. Results from computational genomic methods to predict driver genes are not included. Thus, the overlap between the results obtained with current computational algorithms and the CGC is low. For instance, [Xie et al., 2014] published a list of 560 driver genes and only 213 overlap between both sets (Fig. 1.10). A similar observation arises when looking at the list of genes published in [Tamborero et al., 2013c]. Such inconsistency between gene lists has motivated the second and last part of this thesis. There are around 600 genes associated to cancer and computational models that predict cancer genes achieve a recall of 35%. Can the integration of multiple levels of evolutionary signatures improve the recall of cancer genes? How much can we extract from genomic data? Clearly, exome sequencing cannot identify mutations in regulatory regions such as the one affecting *TERT* expression, or the translocation activation in *ABL*. Can whole-genome sequencing capture more signals? Interestingly, we have massively exploited cancer genome data looking for positive selection signals, but naively we have forgotten about a major force of natural selection, purifying selection. This important evo-

lutionary force is the last topic that I will address in this doctoral thesis.

1.4.3 Natural selection

Charles Darwin postulated that the survival of the fittest was the norm among competing populations [Darwin, 1872], setting the foundation of evolutionary biology. It is not within our scope to discuss the vast amount of literature with respect to evolution, but two key Darwinian insights are relevant for the evolution of cancer. First, Darwin observed that traits were heritable and second, he observed that individuals had different chances of survival and reproduction depending on their environment: better adapted individuals are more likely to remain and pass their heritable traits to the next generation, evoking the process of natural selection. Similarly, a cell divides and passes on its genetic material to daughter cells. Changes in the genetic material create variability that may affect the chances of survival and proliferation, fueling the action of natural selection to preserve the best adapted cells. On one hand, a change that increases the proliferation rate of a cell is under positive selection, whereas a change that decreases cell survival is under negative or purifying selection.

The effect of positive selection of somatic mutations driving tumorigenesis has been widely studied [Stratton et al., 2009] [Vogelstein et al., 2013] [Tamborero et al., 2013b]. Similarly to the evolution of microorganisms, mutations conferring a selective advantage will increase in frequency in the population, eventually achieving fixation (clonal). Although some mutations may not confer any selective advantage, they can hitchhike with a positively selected clone [Smith and Haigh, 1974]. The classical evidence for positive selection of driver mutations comes from the observation of a gene mutated more often than expected randomly across multiple patients [Dees et al., 2012][Martincorena et al., 2015a]. We denominate such form of evidence, selection at the individual level. Another piece of evidence for positive selection comes from the functional impact at the molecular level. Multiple selection signatures are hidden in the genome at multiple levels, therefore, we present an effort to uncover different levels of selection for the identification of cancer-related genes.

In a 2009 review, Michael Stratton wrote that ”some somatic mutations may actually impair cell survival. These will usually be subject to negative selection and hence be absent from the cancer genome.” [Stratton et al., 2009]. Furthermore, McFarland has modeled the acquisition of mutations with different selective advantages in a tumor scenario [McFarland et al., 2013]: cells acquiring strongly beneficial mutations (driver) will sweep and become fixed quickly, whereas mildly

deleterious mutations can also be spread by random drift. Nonetheless, a strong deleterious mutation has almost no possibility of becoming fixed by drift [Charlesworth, 2009].

The amount of literature about the role of negative selection in cancer is sparse. Notwithstanding the fact that two forms of selection were evident from Nowell's work in 1976, three decades passed before the first reports exploring the effect of negative selection in cancer were published. Whereas one study identified evidence for strong purifying selection when comparing cancer versus non-cancer genes [Thomas et al., 2003], another study rejected the importance of negative clonal selection simulated on a mutator phenotype [Beckman and Loeb, 2005]. In addition, an early study emphasized the importance of compartment size in selection strength [Michor et al., 2003]: a small compartment size favors random drift over selection. In 2012, [Ovens and Naugler, 2012], in their self-proclaimed preliminary work, obtained a set of cancer-related genes, aligned them against a non-cancer version of the same genes, and calculated the deviation from selection using the Nei-Gojobori statistic. From 46 cancer-related genes, they reported nine genes under purifying selection, although they recognized several limitations in their study.

Despite [Ovens and Naugler, 2012] description of purifying selection, some recent studies reject a strong effect of negative selection on cancer genomes [Ostrow et al., 2014][Tao et al., 2015][Ling et al., 2015]. However, in 2015, [Pyatnitskiy et al., 2015] identified 91 cancer essential genes by combining negative selection bias, expression, and evidence of low-impact mutations. They calculated the ratio of non-synonymous substitutions rate to synonymous substitutions rate (dN/dS), a classical method to detect evidence of selection using comparative data [Nielsen, 2005]. Interestingly, they found that membrane proteins appeared functionally enriched, thereby linking negative selection to immune surveillance mechanisms. Nevertheless, the extent of negative selection on the cancer genome remains unexplored. In this thesis, for the first time, we exploit the massive amount of sequencing data from 26 tumor types to uncover global patterns of negative selection in cancer.

In part II, we develop a mathematical model for estimating the fraction of cells harboring a copy number alteration using sequencing data. Then, we apply our model to a single case of chronic lymphocytic leukemia to determine the coexistence of clonal and sub-clonal variation along the lifetime of the patient. In the subsequent parts, we have chosen to focus on both classical forces of natural selection: positive and negative selection. In part III, we describe a novel Bayesian algorithm that uses signatures of positive selection at the population, cellular, and

molecular level to identify cancer driver genes. And lastly, in part IV, we estimate the ratio of non-synonymous substitutions to synonymous substitutions, dN/dS or K_n/K_s, to detect genes and functions under negative selection using more than 7,000 exomes spanning 26 tumor types. While the results of the first study were published in the journal *Leukemia*, the results of the second are under revision and the results of the last project are currently being prepared for publication.

Chapter 2

Objectives

The relationship between cancer and the underlying evolution of tumor etiology gives rise to three main questions discussed in this thesis. This work will thus be structured into three parts: (i) quantification of mutation-carrying clones determining intratumor heterogeneity, (ii) the importance of positive selection at multiple levels by estimating the allelic fraction of mutations in a cancer cohort, and (iii) the evidence of purifying selection of cancer essential functions detected using genomic data.

In the first part, we answer the following questions: In a single tumor case, is it possible to quantify the proportion of competing clones coexisting with each other? Do catastrophic events, such as chromothripsis, lead always to the settlement of a tumor phenotype? Can tumor evolutionary dynamics be used to predict treatment outcome?

In the second part, we explore the following questions: Are positive selection signatures imprinted in tumor genomes? How can we exploit them to identify cancer driver genes?

In the last part, we raise the following questions: Is negative selection a driving force shaping the cancer genome? How can we identify cancer essential genes using such a signal?

Part II

Tumor heterogeneity revealed in a case of chronic lymphocytic leukemia

Adapted from published manuscript:

Sporadic and Reversible Chromothripsis in Chronic Lymphocytic Leukemia (CLL) Revealed by Longitudinal Genomic Analysis. Bassaganyas L., Beà S., Escaramís G., Tornador C., Salaverria I., **Zapata L.**, Drechsel O., Ferreira P. G., Rodriguez-Santiago B., Tubio J. M., Navarro A., Martín-García D., López C., Martínez-Trillos A., López-Guillermo A., Gut M., Ossowski S., López-Otín C., Campo E. and Estivill X. *Leukemia* (2013) 27, 23762379; doi:10.1038/leu.2013.127

Chapter 3

A longitudinal analysis of a single case of CLL

This chapter covers the analysis of a chronic lymphocytic leukemia patient over the period of 11 years. The main contribution in this publication was the analysis of subclonal mutations using different genomic technologies. Here, we developed a mathematical formula that quantifies the fraction of clones carrying a somatic mutation based on the deviation of perfect heterozygosity of germline variants. In theory, the number of reads is proportional to the original number of DNA molecules from the sample. In a diploid locus, the simplest case, a heterozygous single nucleotide variant (SNV) present in the whole cancer population has a variant allele frequency (VAF) of 0.5 (half of the reads). Then, we simulated different ploidy scenarios (SNV in a deletion or amplification), and mixed different proportions to obtain a deterministic formula. Such a formula allows us to directly calculate the proportion of cells harboring a CNA from the variation on the allele frequency of the germline SNVs present. Since several deletions and copy gains were detected in this patient, our goal was to calculate the proportion of the somatic variants (i.e. the clonal architecture of the sample).

Bassaganyas L, Beà S, Escaramís G, Tornador C, Salaverria I, Zapata L, Drechsel O, Ferreira PG, Rodriguez-Santiago B, Tubio JM, Navarro A, Martín-García D, López C, Martínez-Trillos A, López-Guillermo A, Gut M, Ossowski S, López-Otín C, Campo E, Estivill X. Sporadic and reversible chromothripsis in chronic lymphocytic leukemia revealed by longitudinal genomic analysis. Leukemia. 2013 Dec;27(12):2376-9. doi: 10.1038/leu.2013.127.
Erratum in:
Leukemia. 2015 Mar;29(3):758

Chapter 4

Methods [1]

4.1 Samples

Collection and processing of samples 016-T02 (B-lymphocytes) and 016-N09 (normal mononuclear cells), as well as biological characteristics of the tumor sample, has been previously published [Quesada et al., 2012]. To perform genomic longitudinal analyses, we used the same tumor samples studied by cytogenetics and FISH, with the exception of sample 016-T06, which was replaced by a tumor sample from year 2005 (016-T05), obtained between the first treatment and the first relapse. Samples that underwent molecular analysis were obtained by single cell sorting of cells to achieve purities of over 95% of either tumor or normal cells, with the exception of sample 016-T05, which was obtained from non-purified DNA (79% of tumor cells).

4.2 Whole-genome sequencing

Whole genome sequencing of sample 016 has been performed within the framework of CLL-Genome Project. Sequencing libraries were constructed with two insert sizes 430-bp and 460-bp according to the TruSeq DNA sample preparation protocol with minor modifications, in particular double size selection. Two ug of genomic DNA were fragmented with a Covaris E210 and size-selected to 300-700 bp. The fragmented and size-selected DNA was end-repaired, adenylated and ligated to Illumina paired-end adaptors and size-selected to very tight size distribution using an E-Gel (Life Technologies, Carlsbad, CA, USA). Size-selected adapter-insert fragments were amplified with 10 polymerase chain reaction (PCR) cycles and sequenced on an Illumina HiSeq 2000 platform with paired end run of 2x100 bp. Sequenced reads from each library were mapped to the human reference genome (GRCh37) using Burrows-Wheeler aligner (BWA) [Li and Durbin, 2009], generating BAM files. From each BAM file, read-pairs corresponding to a

[1] In order to specifically address my contribution to this project, I have omitted the methodological sections not relevant to the purpose.

single paired-end library were extracted using SAMtools [Li et al., 2009a]. Statistics for the number of mapped reads and depth of coverage (RD) for each sample are shown in Table S3 (See publication). For the identification of somatic SVs, we used the PeSV-Fisher algorithm, as described in [Puente et al., 2011]

4.3 Whole-exome sequencing

Detailed information about the collection processing of samples has been previously described [Quesada et al., 2012]. Three ug of genomic DNA from each sample was sheared and used for the construction of a paired-end sequencing library as described in the protocol provided by Illumina [Bentley et al., 2008]. Enrichment of exonic sequences was then performed for each library using the SureSelect Human All Exon 50-Mb kit (Agilent Technologies, Santa Clara, CA) following the manufacturers instructions. Exon-enriched DNA was precipitated with magnetic beads coated with streptavidin (Invitrogen, Life Technologies, Carlsbad, CA, USA) and was washed and eluted. An additional 18 cycles of amplification were then performed on the captured library. Exon enrichment was validated by real-time PCR in 7300 Real-Time PCR System (Applied Biosystems, Life Technologies, Carlsbad, CA, USA) using a set of two pairs of primers to amplify exons and one pair to amplify an intron. Enriched libraries were sequenced in one lane of an Illumina Genome Analyzer IIx sequencer, using the standard protocol. Sequenced reads from each library were mapped to the human reference genome (GRCh37) using BWA with the sampe option, and a BAM file was generated using SAMtools. Reads from the same paired-end libraries were merged, and optical PCR duplicates were removed using Picard. Statistics for the number of mapped reads and depth of coverage for each sample, as well as information about the Sidron algorithm (for the identification of somatic substitution) have been previously described [Quesada et al., 2012].

4.4 Targeted Enrichment Sequencing

Targeted DNA enrichment was performed with the SureSelect (Agilent Technologies, Santa Clara, CA) custom capture system, according to the SureSelect Target Enrichment protocol and sequenced on an Illumina HiSeq2000 instrument, following manufacturers protocols. Briefly, three ug of human genomic DNA were sheared using a Covaris E220 to a size of roughly 150 bp. The purified samples were end repaired, adaptor ligated and 6 cycles of PCR were applied. In the fol-

lowing step the PCR product was hybridized to the Agilent bait capture kit for 24 h for the enrichment and subsequently amplified by 12 cycles of PCR. The resulting libraries were sequenced on Illumina HiSeq2000 flow-cells in pools of three samples per sequencing lane, generating 50-80 million paired end sequence reads of 76-bp per sample. Image analysis, base calling, and base call quality were generated during the run with the Illumina HiSeq Real Time Analysis (RTA 1.13.48) software with default parameters. FASTQ files containing sequence information and quality scores for each base call were exported for further analysis. Sequenced reads were mapped to the human reference genome (GRCh37/hg19) using the BWA aligner. Structural variant detection by PeSV-Fisher was performed using alignments created by BWA, while split-read analysis was performed using alignments created by GEM algorithm [Marco-Sola et al., 2012]. The efficiency of the capture for each sample is described on Table S3 (See publication). Somatic single-nucleotide variants (SNVs) detection was done by the additional in-house pipeline for the detection of SNVs described above, using BWA alignment.

4.5 Cancer cell fraction (CCF) of cancer subpopulations using WGS and WES data

4.5.1 Detection of somatic SNVs

We used MuTect [Cibulskis et al., 2013] to detect somatic SNVs (sSNVs) at a broad frequency spectrum including low frequency mutations down to 0.05 B-Allele Frequency in samples 016-T02 and 016-N09 (also denoted as Variant allele or BAF). MuTect identified 13 out of 14 validated sSNVs (Table 5.1) and the observed allele frequency estimations were corroborated by visual inspection of the alignment at the sSNV position using Integrative Genomic Viewer tool (IGV, http://www.broadinstitute.org/igv/home). The list of detected sSNVs was used to directly estimate the clonality of the population carrying the alternative allele, knowing the ploidy of the overlapping locus and the tumor purity of the sample following the method described in the following section.

4.5.2 Detection of germline SNVs

We used GATK [McKenna et al., 2010] to detect SNVs in samples 016-T02 and 016-N09 using a combination of WGS and WES data. The combined average coverage of $> 120x$ for each tumor and normal samples allowed us to obtain more precise values for SNV allele frequencies. We chose this since using WES

and WGs data separately lead to variation in the calculated BAF up to 20%. Germline SNVs (gSNVs) found at near perfect heterozygosity in the normal sample ($0.45 < BAF < 0.55$) were selected for further steps. In order to estimate the cancer cell fraction of chromothripsis, the BAF for each gSNV within regions of the chromothripsis was calculated in the tumor sample and the median deviation from 0.5 for each region was obtained (Average distance in Table 5.2). Additionally, the same procedure was performed for whole genome using all heterozygous gSNVs in 100-kb segments.

4.5.3 CCF calculation

To infer the CCF of each somatic SNV and the chomotrypsis regions, we adapted a mathematical formula described elsewhere [Van Loo et al., 2010]:

$$BAF_i = \sum_{j=1}^{L}(n_{ij}X_{ij}) / \sum_{j=1}^{L}(N_{ij}X_{ij}) \tag{4.1}$$

where for each variant (i), the allelic fraction observed (BAF_i) depends on the integer allelic copy-number of population j (n_{ij}), the population frequency harboring that variant (X_{ij}), and the integer ploidy of that variant(N_{ij}). The most simple case is a somatic mutation in the whole cancer population ($L = 1$), having an heterozygous allele ($n_i = 1$) in a diploid locus ($N = 2$). The resulting BAF is then 0.5.

This formula allows for simulation of allele frequencies observed under any number of populations (L) with specific ploidy levels (N_{ij}).

4.5.4 CCF for somatic SNVs

For sSNVs in diploid regions, the distance to 0.5 (Bdev) indicates linearly the clonality (i.e. X_i or CCF) of a specific subpopulation (assuming tumor purity of 100%). Therefore, a higher Bdev indicates a low CCF (subclonality). This model has been applied to most sSNVs since they were present in a diploid locus, except for *NFKBIE* and *DCAF12L2*. The same model was applied to the estimation of CCF of sSNVs from targeted resequencing data across the different time-points, given no copy number alterations were found.

4.5.5 CCF for somatic CNAs

CCF estimations for somatic copy-number alterations (CNA) are based on a more complex model. The Bdev for a germline variant (gSNV) refers to the distance from 0.5 in the tumor sample and correlates with the cancer cell fraction of a CNA. The median Bdev of all gSNVs found in the tumor sample was calculated in a 100-kb sliding-window. Median Bdev is transformed into CCF assuming a two-populations model in our equation 4.1 and shown in figures 4.1 and 4.2. The allele frequency of alleles present in the same deleted/duplicated locus follows a similar trend when Bdev varies but with different signs. The simulation fits a polynominal function of degree 2 (quadratic), with an R^2 close to 1 and intercepting the y-axis on the expected 0.5. Thus, modifying the quadatric formula obtained ($ax2 + bx + c = y$ by $ax2 + bx + (c - 0.5) = 0$), allows to solve the equation to estimate x, since the value of c is the observed BAF, then *c-0.5* is the Bdev or the deviation of BAF from 0.5 (perfect heterozygosis of the germline variant), and x is the CCF of the CNA.

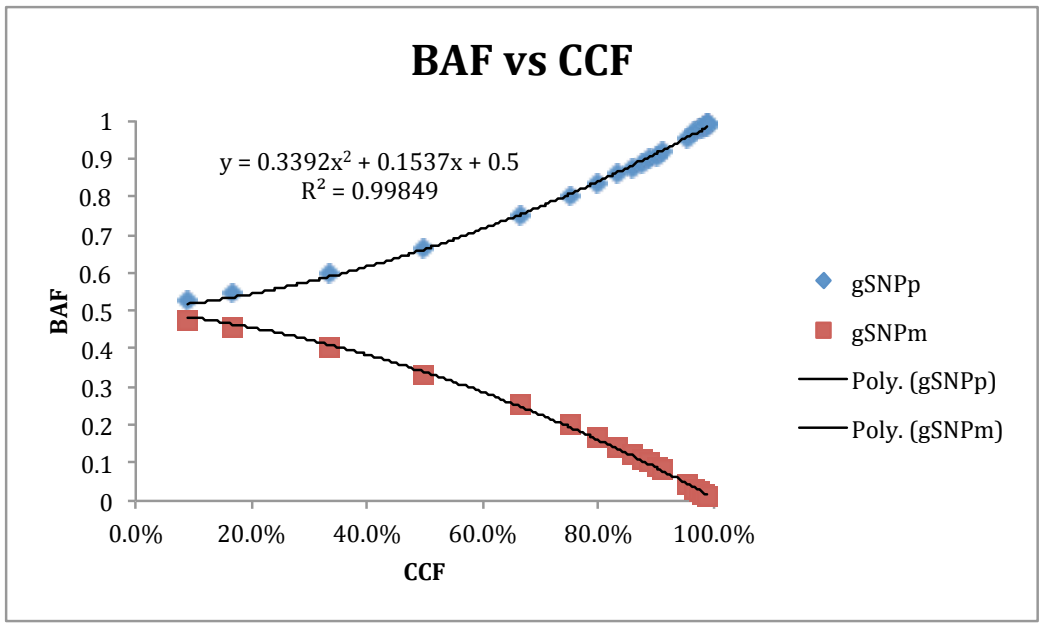

Figure 4.1: **Simulation of CCF under a two-population model harboring a heterozygous deletion.** Relationship between expected BAF and CCF for the heterozygous deletion model, assuming ploidy of 1n. The quadratic formula obtained with this simulation can be used to directly estimate the clonality in these types of regions.

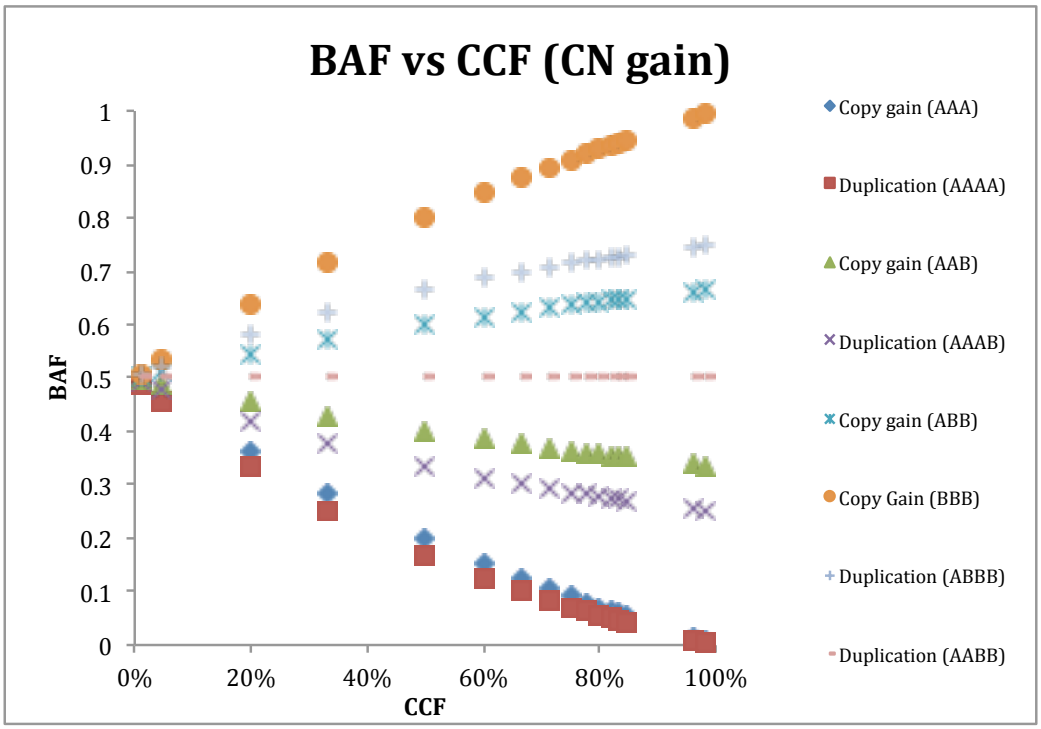

Figure 4.2: **Simulation of CCF under a two-population model harboring copy gains.** Relationship between expected BAF and CCF for different copy number gain scenarios. Here we simulated the correlation of BAF and CCF assuming a ploidy of 3n and 4n, following the SNP array results.

The estimation of CCF for the CNAs by targeted re-sequencing was not possible because probes for targeted re-sequencing were designed for 1-kb fragments around breakpoints, which allowed the validation of targeted structural variants breakpoints but not included enough SNVs to calculate CCF in tumor cells.

Chapter 5

Supplementary tables

Chr	Pos	Ref	Obs	Gene	WGS+WES(BAF mut)	WGS+WES(BAF obs)	Ploidy	predicted CCF
X	125299321	C	A	DCAF12L2	1	1	1	100%
15	45409873	A	G	DUOXA1	0,48	0,5	2	100%
16	745751	T	C	FBXL16	0,63	0,5	2	100%
1	78425887	G	A	FUBP1	0,409	0,44	2	88%
17	42463002	G	A	ITGA2B	0,371	0,37	2	74%
19	19431945	G	A	MAU2	-	0,5	2	100%
6	44230329	C	A	NFKBIE	0,346	0,36	1	48%
1	186277552	A	G	PRG4	0,468	0,52	2	100%
6	146264389	T	C	SHPRH	0,531	0,494	2	99%
6	132892234	A	C	TAAR6	0,4434	0,45	2	90%
19	56173950	A	T	U2AF2	0,439	0,45	2	90%
3	141163154	C	T	ZBTB38	0,437	0,45	2	90%
12	21358921	C	A	SLCO1B1	0,477	0,48	2	96%
11	108114749	G	C	ATM	0,5	0,34	2	100%

Table 5.1: Cancer cell fraction estimation for SNVs by WGS/WES sequencing of sample 016-T02.

Chr	Pos	Pos	Obs	Locus	Number SNVs	Average distance	Ploidy	CCF	T00	T02*	T05	T11
1	158916616	162228238	DEL	1q23	1391	0,15	1	47%	-	+	-	-
1	204884807	207098441	DEL	1q32	651	0,14	1	46%	-	+	-	-
2	57538505	57546240	DEL	2p16	1	0,00	1	0%	-	+	-	-
3	66307700	66848190	DEL	3p14	238	0,14	1	46%	-	+	-	-
4	92032232	92035673	DEL	4q22	1	0,41	1	90%	+	+	-	-
6	21165446	21202889	DEL	6p22	14	0,13	1	43%	-	+	-	-
6	42655943	44314131	DEL	6p21	276	0,32	1	77%	+	+	+	+
6	87787340	87937492	DEL	6q14	1991	0,15	1	48%	-	+	-	-
6	87831293	92372455	DEL	6q14-q15	-	-	-	-	-	+	-	-
6	98726379	100945032	DEL	6q16	682	0,15	1	48%	-	+	-	-
6	105121236	107322576	DEL	6q16-q21	716	0,15	1	48%	-	+	-	-
6	135571668	137893724	DEL	6q23	514	0,15	1	48%	-	+	-	-
7	142493998	142495773	DEL	7q34	-	-	-	-	+	+	+	+
9	19557108	19559274	DEL	9p22	1	0,45	1	95%	+	+	+	+
10	103329084	105060989	DEL	10q24	238	0,14	1	45%	-	+	-	-
12	6174700	8452800	DEL	12p13	736	0,14	1	46%	-	+	-	-
12	10781231	13107215	DEL	12p13	1255	0,14	1	46%	-	+	-	-
14	106330475	107178831	DEL	14q32	675	0,452	1	95%	+	+	+	+
18	118560	14978275	DEL	18p	3907	0,456	1	95%	+	+	+	+
4	55259599	190916819	GAIN	4q	52089	0,16	3	94%	+	+	+	+

Table 5.2: List of copy number altered loci found in case 016 accross multiple time points and the respective fraction of cells harboring the mutation.

Strategy	SVtype	locus	chrA	brk1	chrB	brk2	size	016-T00	016-T02	016-T05	016-T11
WGS/Re-seq	TRANS-INV	t(1;12)(q23;p13)	1	158895420	12	9009852	-	-	+	-	-
WGS/Re-seq	INV	1q23-q32	1	158895935		207100978	48204807	-	+	-	-
WGS/Re-seq	INTRA-TRANS	1q23-q32	1	158908565		204868609	45960044	-	+	-	-
WGS/Affy	DEL	1q23	1	158916616		162228238	3311622	-	+	-	-
WGS/Re-seq	TRANS-INV	t(1;10)(q23;q24)	1	162273578	10	103313969	-	-	+	-	-
WGS/Re-seq	INTRA-TRANS	1q23-q32	1	162299645		207110393	44810748	-	+	-	-
WGS/Re-seq	TRANS-INV	t(1;12)(q32;p13)	1	204873185	12	13128829	-	-	+	-	-
WGS/Affy	DEL	1q32	1	204884807		207098441	2213634	-	+	-	-
WGS/Re-seq	TRANS-INV	t(1;6)(q32;q14)	1	204884852	6	87790817	-	-	+	-	-
WGS/Re-seq	TRANS-INV	t(1;6)(q32;q24)	1	204885286	10	103311319	-	-	+	-	-
WGS/Re-seq	TRANS-INV	t(1;6)(q32-q16)	1	207109744	6	98720478	-	-	+	-	-
WGS/Re-seq	TANDEM-DUP	1q44	1	245750841		245754210	3369	-	+	-	-
WGS/Re-seq	DEL	2p16	2	57538505		57546240	7735	-	+	-	-
WGS/Affy	DEL	3p14	3	66307700		66848190	540490	-	+	-	-
WGS/Re-seq	DEL	4q22	4	92032232		92035673	3441	+	+	-	-
Affy/WGS	GAIN	4q	4	55259599		190916819	135657220	+	+	+	+
Cyto/WGS	TRANS	t(4;18)(p11;p11)	4	49275540	18	14983384	-	+	+	+	+
WGS/Re-seq	DEL	6p22	6	21165446		21202889	37443	-	+	-	-
WGS/Affy	DEL	6p21	6	42655943		44314131	1658188	+	+	+	+
WGS/Re-seq	DEL	6q14	6	87787340		87937492	150152	-	+	-	-
WGS/Re-seq	INTRA-TRANS	6q14-q16	6	87823275		98688205	10864930	-	+	-	-
WGS/Affy	DEL	6q14-q15	6	87831293		92372455	4541162	-	+	-	-
WGS/Re-seq	INV	6q15-q16	6	92369212		100987026	8617784	-	+	-	-
WGS/Re-seq	TRANS-DIR	t(1;10)(q15;q24)	6	92405452	10	103302215	-	-	+	-	-
WGS/Re-seq	INV	6q15-q23	6	92406400		137889638	45483157	-	+	-	-
WGS/Re-seq	INV	6q15-q16	6	92407820		100973954	8566039	-	+	-	-
WGS/Affy	DEL	6q16	6	98726379		100945032	2218653	-	+	-	-
WGS/Affy	DEL	6q16-q21	6	105121236		107322576	2201340	-	+	-	-
WGS/Affy	DEL	6q23	6	135571668		137893724	2322056	-	+	-	-
WGS/Re-seq	TRANS-DIR	t(6;12)(q23;p13)	6	137871508	12	10749843	-	-	+	-	-
WGS/Re-seq	DEL	7q34	7	142493998		142495773	1775	+	+	+	+
WGS/Re-seq	DEL	9p22	9	19557108		19559274	2166	+	+	+	+
WGS/Re-seq	TRANS-INV	t(3;10)(q26;q21)	10	70172296	3	168885793	-	+	+	-	-
WGS/Re-seq	TRANS-INV	t(2;10)(p11;q24)	10	101290756	2	89132330	-	+	+	+	+
WGS/Re-seq	TRANS-INV	t(6;10)(q16;q24)	10	103308280	6	98688129	-	-	+	-	-
WGS/Re-seq	TRANS-INV	t(6;10)(q16;q24)	10	103310475	6	100949176	-	-	+	-	-
WGS/Re-seq	TRANS-DIR	t(6;10)(q23;q24)	10	103328997	6	100987263	-	-	+	-	-
WGS/Affy	DEL	10q24	10	103329084		105060989	1731905	-	+	-	-
WGS/Re-seq	TRANS-INV	t(6;10)(q16;q24)	10	105080595	6	137868980	-	-	+	-	-
WGS/Re-seq	TRANS-INV	t(6;12)(q15;p13)	12	5988079	6	92374031	-	-	+	-	-
WGS/Re-seq	TRANS-DIR	t(6;12)(q14;p13)	12	5998576	6	87786886	-	-	+	-	-
WGS/Affy	DEL	12p13	12	6174700		8452800	2278100	-	+	-	-
WGS/Re-seq	TRANS-DIR	t(10;12)(q24;p13)	12	10749697	10	105073574	-	-	+	-	-
WGS/Re-seq	TRANS-INV	t(6;12)(q14;p13)	12	10766354	6	87823973	-	-	+	-	-
WGS/Re-seq	DEL	12p13	12	10781231		13107215	2325984	-	+	-	-
WGS/Re-seq	INV	12p13	12	13102811		13123781	20857	-	+	-	-
WGS/Re-seq	TRANS-DIR	t(1;12)(q23;p13)	12	13127322	1	162299794	-	-	+	-	-
WGS/Re-seq	DEL	14q32	14	106330475		107178831	848356	+	+	+	+
WGS/Re-seq	TRANS-INV	t(5;17)(q21;q23)	17	61565839	5	100389381	-	-	+	-	-
Affy/WGS	DEL	18p	18	118560		14978275	14859715	+	+	+	+

Table 5.3: List of structural variants found in case 016 accross multiple time points.

Part III

Positive selection signatures reveal cancer genes across multiple tumor types

Adapted from manuscript under revision:

Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. Zapata L., Susak H., Drechsel O., Friedlander MR., Estivill X., Ossowski S.

Tumors are composed of an evolving population of cells subjected to tissue-specific selection, which fuels tumor heterogeneity and ultimately complicates cancer driver gene identification. Here, we integrate cellular prevalence, population recurrence, and functional impact of somatic mutations as signatures of positive selection into a Bayesian model for driver prediction. We demonstrate that our model, cDriver, outperforms competing methods when analyzing solid tumors, hematological malignancies, and pan-cancer datasets. Applying cDriver to exome sequencing data of 21 cancer types from 6,870 individuals revealed 123 unreported tumor type-driver gene connections. These novel connections are highly enriched for chromatin-modifying proteins, hinting at a universal role of chromatin regulation in cancer etiology. Although infrequently mutated as single genes, we show that chromatin modifiers are altered in a large fraction of cancer patients. In summary, we demonstrate that integration of evolutionary signatures is key for identifying mutational driver genes, thereby facilitating the discovery of novel therapeutic targets for cancer treatment.

Chapter 6

Introduction

Since the 1970s, tumors have been considered the product of evolutionary forces such as positive selection of highly proliferative cancer genotypes or negative selection of non-adaptive cancer genotypes [Nowell, 1976]. Analogous to the evolution of multi-cellular organisms, random somatic mutations in cancer cells interplay with natural selection, creating phenotypic diversity [Stratton et al., 2009] and allowing for adaptation [Vogelstein et al., 2013]. It has been shown that this process of clonal evolution follows different paths depending on the background genotype of the patient [Fearon and Vogelstein, 1990, Sakoparnig et al., 2015], the tissue microenvironment [Bissell and Hines, 2011], and the functional redundancy of somatic mutations acquired [McLendon et al., 2008, Szczurek and Beerenwinkel, 2014]. This leads to increased diversity, ultimately contributing to intra- and inter- tumor heterogeneity [Vogelstein et al., 2013]. This tumor heterogeneity has been reported in multiple tissues [Wang et al., 2014b, Landau and Wu, 2013, Gerlinger et al., 2012, Bolli et al., 2014, Stransky et al., 2011, Zhao et al., 2014a], effectively hampering the identification of driver genes and limiting the effectiveness of therapeutic targets [Marusyk et al., 2012].

Next generation sequencing (NGS) technologies have allowed mutational screening across thousands of tumors, uncovering the extent of tumor heterogeneity. [Lawrence et al., 2013, Sottoriva et al., 2013, Lee et al., 2015]. The standard analysis involves aligning NGS reads from tumor-normal sample pairs to the reference genome and identifying allelic variants present only in the tumor (somatic mutations) [Mwenifumbo and Marra, 2013]. Recently developed methods have transformed somatic allele counts from single tumor sequencing data to cancer cell fraction (CCF) revealing complex tumor architecture [Oesper et al., 2013, Fischer et al., 2014, Roth et al., 2014, Miller et al., 2014, Li and Li, 2014]. In addition, various studies have sequenced multiple locations of single solid tumors, and observed common mutations coexisting with region-specific mutations [Campbell et al., 2010, Ding et al., 2010, Gerlinger et al., 2012, Gerstung et al., 2012, Lee et al., 2015]. In non-solid or circulating tumors, clones mixed within the sample have also revealed mutations at different cellular fractions (or CCF)[Landau and Wu, 2013, Schuh et al., 2012, Bassaganyas et al., 2013]. Although all the studies

recognize the importance of subclonal genetic variation and the dynamics of tumor evolution, little clinical impact has been achieved in terms of the identification of resistant genotypes [Yates et al., 2015]. Thereby, highlighting the importance of CCF as measure of positive selection and thus, an indicator of driver gene status [Landau and Wu, 2013, Landau et al., 2015].

Current solutions for identifying driver genes rely on the mutation frequency across a large number of cancer patients [Dees et al., 2012], the genomic context where they occur [Lawrence et al., 2013], the functional impact of the mutation, [Gonzalez-Perez and Lopez-Bigas, 2012] and the clustering of mutations within active sites [Tamborero et al., 2013a]. However, statistical methods based on mutation recurrence and context alone have not been able to classify infrequently mutated genes as drivers [Vogelstein et al., 2013]. Methods based on molecular selection signatures, such as functional impact and clustering, have been recently applied to identify infrequently mutated genes [Tamborero et al., 2013b] but without considering cancer cell fraction as a key measure of positive selection. Moreover, most of them are based on frequentist statistics, explicitly stating the need for a Bayesian approach to rank driver genes using somatic single nucleotide polymorphism data. A large number of tumor samples will continue to be sequenced, ultimately requiring new models to identify the full driver gene landscape. Knowledge of this landscape is key to improve diagnosis, monitor progression and select treatment options [Lawrence et al., 2014].

Here, we present cDriver, a novel Bayesian inference approach to identify and rank driver genes using multiple measures of selection. We benchmark our results against standard tools on public tumor datasets. Finally, we apply cDriver to more than 6,000 cancer exomes to uncover unreported associations between known driver genes and tumor types, expanding the set of possible therapeutic targets.

Chapter 7

Results

7.1 Evolutionary signatures used by cDriver

To identify driver genes, our Bayesian inference model, cDriver, integrates three measures of positive selection for each gene: the proportion of affected cases (recurrence), the fraction of cancer cells (CCF), and the functional impact of the mutated allele (Damage score)(Fig. 7.1). cDriver uses the frequency of the mutated allele in the cohort as evidence of positive selection at the population level (Fig. 7.1 a). Given that not all driver genes are frequently mutated, cDriver uses the frequency of the mutated allele in an individual cancer cell population as evidence of positive selection at the cellular level (Fig. 7.1 b). Thus, somatic mutations that increase cellular fitness can be present at the root of the tumor evolutionary tree, or lead to tumor clonal expansions, displaying a high cancer cell fraction (CCF). It is possible that most of the somatic mutations observed at high CCF are actually passenger events that occurred in a premalignant clone before the initiating driver event. Nevertheless, driver mutations alter the protein function drastically, so cDriver also includes functional scores as evidence of positive selection at the molecular level to downgrade these mutations (Fig. 7.1 c).). Summarizing, cDriver benefits from organismal, cellular, and molecular information given by recurrence, CCF, and functional impact values.

To account for the variability of background mutation rate (bmr) between genes, cDriver uses silent mutations to locally estimate the expected number of non-silent mutations. Under neutral evolution, the ratio of non-silent mutations per non-silent sites and silent mutations per silent sites is equal to one (K_a/K_s = 1, see methods). The observed number of non-silent mutations exceeds this value for genes under positive selection ($K_a >> K_s$) and fall below this value for genes under negative or purifying selection ($K_a << K_s$). However, somatic mutations are rare for most cancers and many genes do not have silent mutations, restricting computability of K_a/K_s. Therefore, cDriver calculates an average bmr using the K_a/K_s formula and a precalculated bmr taken from literature [Lawrence et al., 2013].

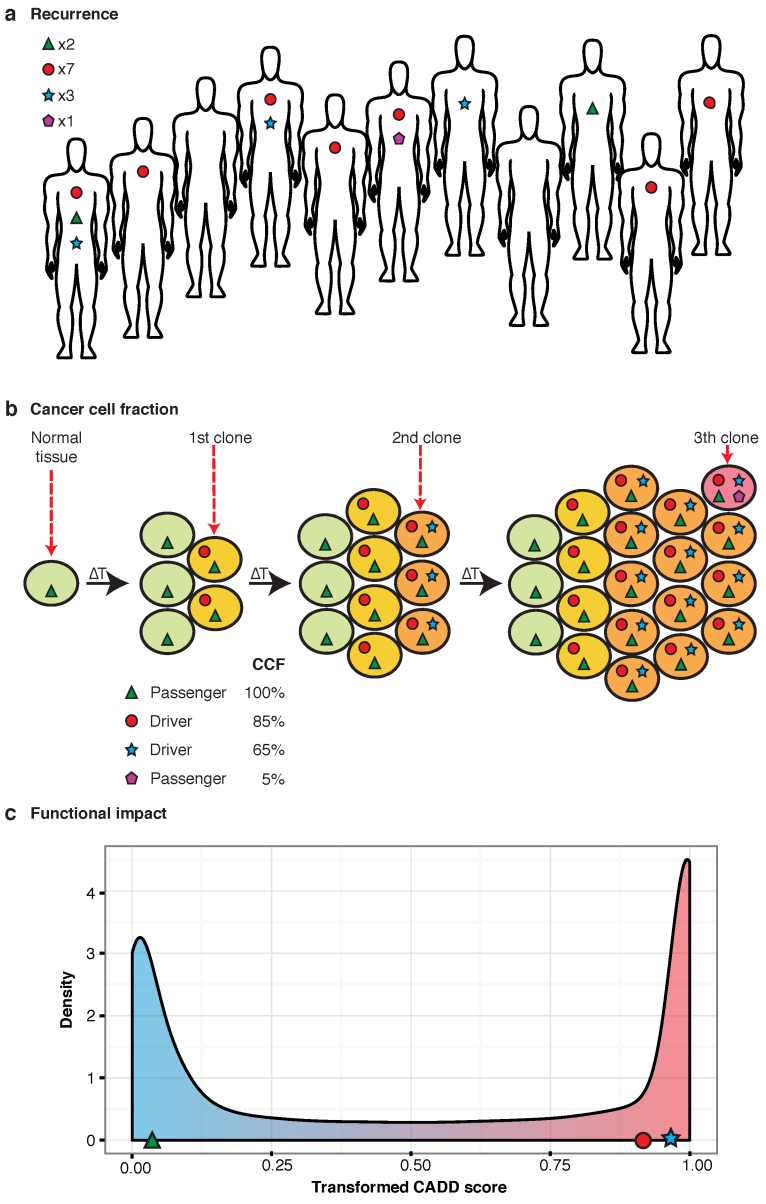

Figure 7.1: (Continued in the following page.)

Figure 7.1: **Signatures of positive selection in tumor sequencing data.** (a) Large scale sequencing experiments of patient cohorts reveal the mutational landscape of a cancer across a population. Somatic mutations under positive selection (circle and star) are expected to be more frequent than somatic mutations that confer no selective advantage (triangle and pentagon). Therefore, most of the current algorithms consider recurrently mutated genes as drivers and randomly mutated genes as passengers. (b) Illustrative model of clonal evolution showing four time points. Each clone is represented by a unique genotype, and is depicted as a group of cells (ellipsoids) with the same background color. Shapes inside the cell represent mutations. Two types of mutations under positive selection are illustrated: a tumor-initiating driver (red circle) and a late-driver causing clonal expansion (blue star). The initial driver mutation causes the emergence of the first malignant clone (last onko-common ancestor, LOCA) and it propagates to all daughter cells, thus having a high cancer cell fraction (CCF) at any time point. The second driver mutation confers a selective advantage over the rest of the clones and it generates a selective sweep in the last time point. Two types of passenger mutations are shown: early passengers or hitchhikers (green triangle) present at a high CCF since they appeared before the emergence of the LOCA and late passenger mutations (purple pentagon) present only in a small fraction of cancer cells. The CCF value describes the fraction for each mutation observed if the sample was sequenced at the last time point. (c) Highly damaging mutations are expected to be under selection given they disrupt the normal protein function. In contrast, passenger mutations are mostly neutral and are not expected to bias towards high functional damage. In this study we integrate signals depicted in a-c in one model for driver gene identification.

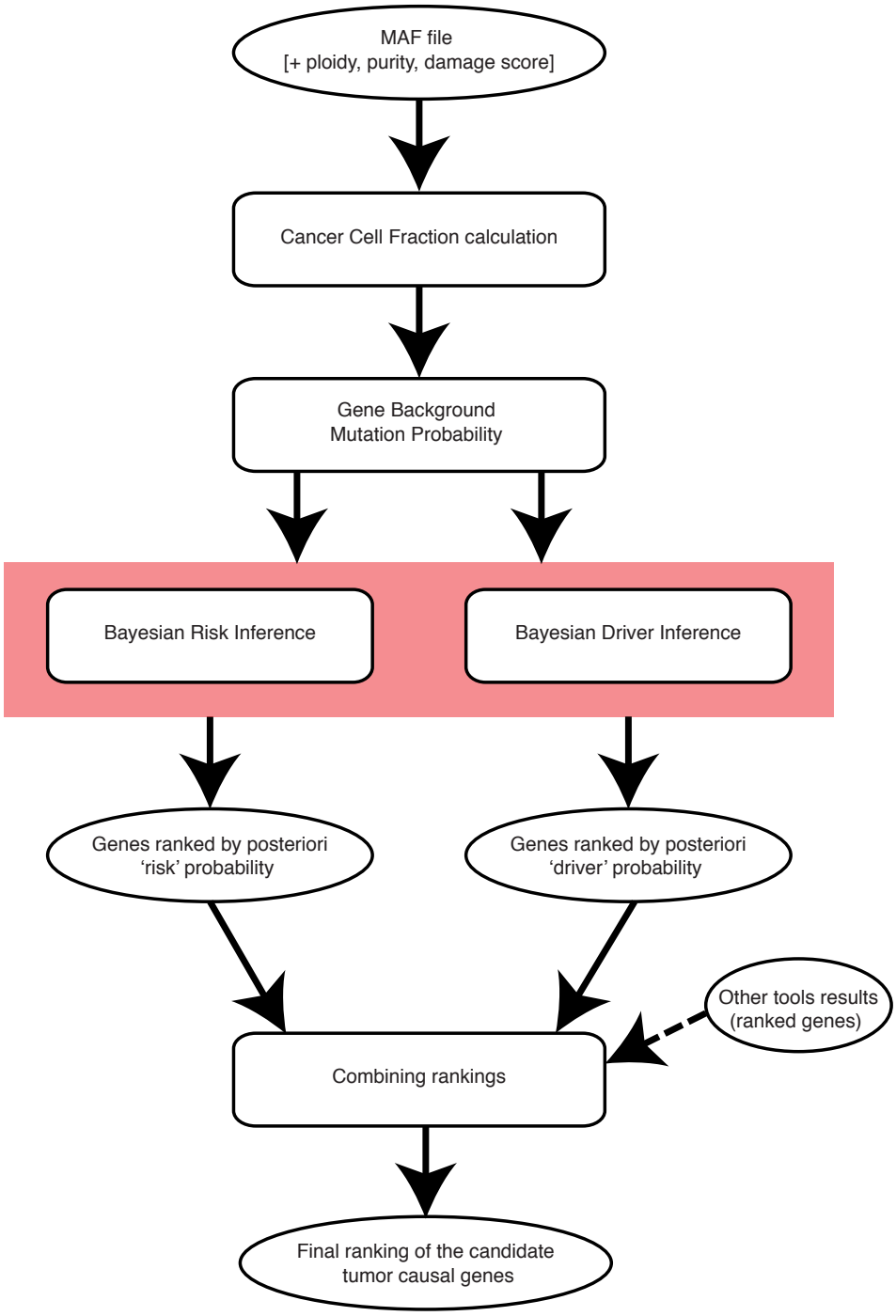

Figure 7.2: (Continued in the following page.)

Figure 7.2: **cDriver proccessing pipeline.** Schematic overview of cDriver processing pipeline. Input is a standardized MAF file with optional columns: ploidy, purity and functional impact score. In this diagram, ellipsoids represent data or files. Rectangles represent functions or operations. The first step is the calculation of cancer cell fraction. The second step calculates the background mutation probability using the model described in methods. The third step calculates the posterior probabilities per gene using two Bayesian models. The final output is a ranking of all genes given by the combination of the previously obtained rankings.

Next, cDriver calculates posterior probabilities per gene using two Bayesian models, (i) the risk of having cancer and (ii) the probability of being a driver gene. The first model requires the incidence of the tumor in the population, while the second requires the frequency of mutations in driver genes. In summary, cDriver combines recurrence, CCF, and functional impact as a foreground measure, and the averaged bmr as a background measure to calculate posterior probabilities for each gene. This provides the user with a per gene-rank as an agreement between the two probabilities given by each model (Fig. 7.2).

7.2 Benchmarking cDriver performance

To evaluate the performance of cDriver, we benchmarked precision, recall and F-score against four frequently used driver gene identification methods (see methods). These were OncodriveFM [Gonzalez-Perez and Lopez-Bigas, 2012], OncodriveCLUST [Tamborero et al., 2013a], MuSiC [Dees et al., 2012], and MutsigCV [Lawrence et al., 2013] as they are canonical methods based on different underlying principles. In parallel, we selected three public cancer datasets that differed in the number of samples, the number of mutations, the tissue-of-origin, and the purity of the tumor. These datasets consisted of 762 cases of breast cancer (BRCA, a solid tumor type) [Kandoth et al., 2013], 385 cases of chronic lymphocytic leukemia (CLL, a non-solid tumor type) [Puente et al., 2015], and 3,205 cases from a combined set of 12 tumor types (Pancan12) [Kandoth et al., 2013] (Table 10.1). Results of each method for each dataset were sorted by p-values or posterior probabilities to compare ranked driver genes.

7.3 Benchmarking in breast cancer (BRCA) and chronic lymphocytic leukemia (CLL)

To evaluate the performance of cDriver and the competing methods when analyzing solid and circulating tumor data, we assembled a list of 33 and 22 gold standard (GS) genes for BRCA and CLL, respectively (Table 10.2). For comparison we only considered genes that were ranked in the top 66 for BRCA and the top 44 for CLL (twice the number of GS genes), given no further improvement was achieved by any tool beyond these thresholds. We found that cDriver outperforms all other methods in both BRCA (Fig. 7.3 a) and CLL (Fig. 7.3 b), showing the highest F-score and precision as well as similar recall as the second best method (Fig. 10.2).

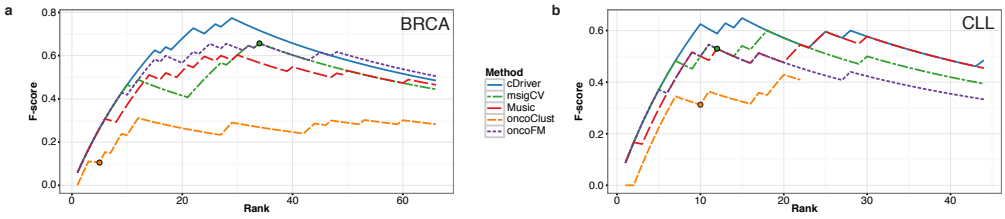

Figure 7.3: **Benchmarking of competing methods in breast cancer (BRCA) and chronic lymphocytic leukemia (CLL).** F-score for cDriver (solid blue line) and four other driver identification algorithms using BRCA **(a)** and CLL **(b)** datasets. Results of each method were transformed to ranks by ordering p-values or posterior probabilities. For comparison, F-score is shown to rank 66 for BRCA and 44 for CLL (twice the number of genes in the gold standards), since all methods reach the F-score peak before these ranks.

Five and four GS genes were missed by all methods in BRCA (Fig. 7.4 a) and CLL (Fig. 7.4 b), respectively, suggesting that these genes are likely affected by types of variation not detectable by single nucleotide variant analysis (e.g. CNVs, fusion genes, expression or epigenomomic imbalance). On the other hand, cDriver identified nine genes in BRCA and eight genes in CLL not present in the GS (Fig. 7.4, bottom part). Most of these genes were also predicted by at least one other approach, except *KDM6A* and *EGR1*. The former was reported to play a role in rare aggressive breast cancer during the preparation of this manuscript [Dieci et al., 2016], while the latter was found in CLL using gene expression and network analysis [Álvarez Silva et al., 2015]. Other driver genes not present in the GS, such as *MYH14, MED23 and ZFP36L1* in BRCA and *ZNF292, FUBP1* and *DTX1* in

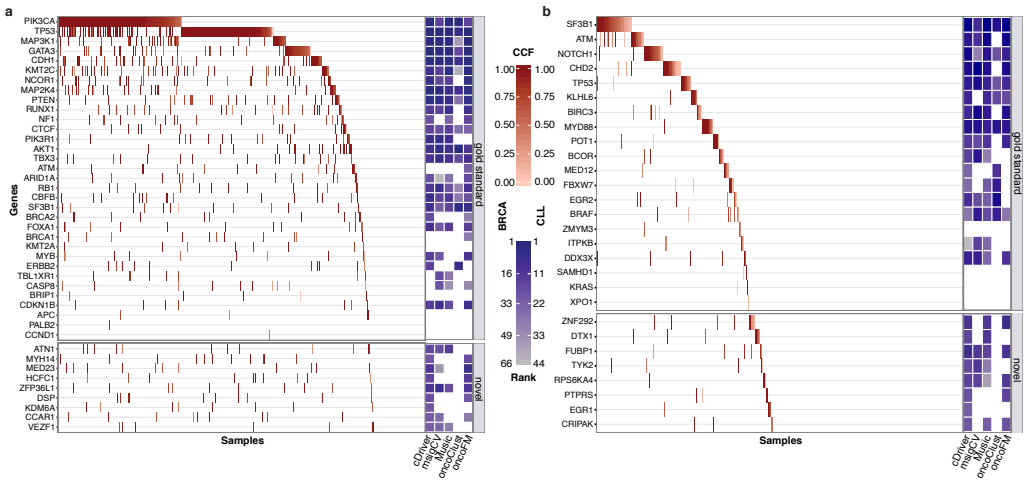

Figure 7.4: **List of gold standard and novel genes identified by competing methods.** We compared the results for all methods irrespective of the p-value using only the ranking for BRCA (a) and CLL (b). Gold standard genes were ordered by mutation frequency and samples were ordered by cancer cell fraction (CCF). The CCF of each mutation in each gene-patient pair is indicated by the red color gradient. On the right, gene rankings of each algorithm are indicated by the blue color gradient. White means that this gene was not ranked under 66 for BRCA and 44 for CLL. At the bottom of figures c and d results for genes not present in the gold standard but highly ranked by cDriver are shown. 15 out of 17 of these genes were identified as drivers by at least one other method.

CLL, had also recently been implicated in tumor development [Nik-Zainal et al., 2016, Surcel et al., 2015, Puente et al., 2015, Landau et al., 2015]. As previously reported, less than half of CLL cases harbor a somatic point mutation in a driver gene [Puente et al., 2015], in contrast to BRCA where more than 80% of cases are affected.

7.4 cDriver performance in a pooled dataset of 12 tumor types

Capitalizing on the large number of cancer sequencing studies recently published by TCGA, we benchmarked all methods on a publicly available compilation of 12 cancers (Pancan12)[Kandoth et al., 2013] using five published gold standards[Futreal et al., 2004],[Kandoth et al., 2013],[Xie et al., 2014],[Lawrence et al., 2014],[Tamborero et al., 2013b] (Methods). Across the 3,200 whole exome sequencing (WES) samples, cDriver and oncodriveFM performed best in F-score using Cancer Gene Census (Fig. 7.5a) with and without filtration of non-expressed genes (Fig. 10.1). Noteworthy, MuSiC benefited extensively from this post-filtration while all other methods benefited marginally. In addition, cDriver showed the highest precision and outperformed all other methods amongst the top 50 ranked genes across all gold standards (Fig. 10.3). We also noticed that significance thresholds used by different methods often do not coincide with their respective F-score peak, e.g. MutsigCV identified 100 significantly mutated genes, while the peaks of the F-score using the five GS varied from 126 to 228.

To assess whether cDriver contributes to combinations of complementary methods [Tamborero et al., 2013b], we calculated two ensemble F-score curves using all methods with and without cDriver (Methods). Inclusion of cDriver increased the F-scores especially in the long tail between ranks 150 and 800 (Fig. 7.5b). Likewise, to evaluate the contribution of CCF integration to the performance of our method, we benchmarked it using only recurrence, recurrence and functional impact, recurrence and CCF, and the combination of all signals (Methods). We observed that CCF and functional impact independently improve F-score and show best performance in combination, corroborating the importance of cancer cell fraction for the identification of mutational drivers (Fig. 7.5c).

The top 30 genes predicted by cDriver showed a high median CCF, although with a large variance (Fig. 7.5d). Despite all of these genes were present in the gold standard, several of them are missed (Fig. 7.5e). For example, oncodriveFM

missed *FLT3* due to a cluster of medium impact mutations (Fig. 10.4) and oncodriveCLUST missed several tumor suppressor genes, since loss of function mutations in these genes do not necessarily cluster (e.g. *PBRM1*, Fig. 10.5). MutsigCV missed *KEAP1* and *MTOR* in this dataset, but it was able to find these genes within a larger sample size [Lawrence et al., 2013].

In summary, cDriver achieved a superior precision in the 'short' and in the 'long tail' of ranked methods, allowing us to explore an extended landscape of driver genes across multiple tumor types.

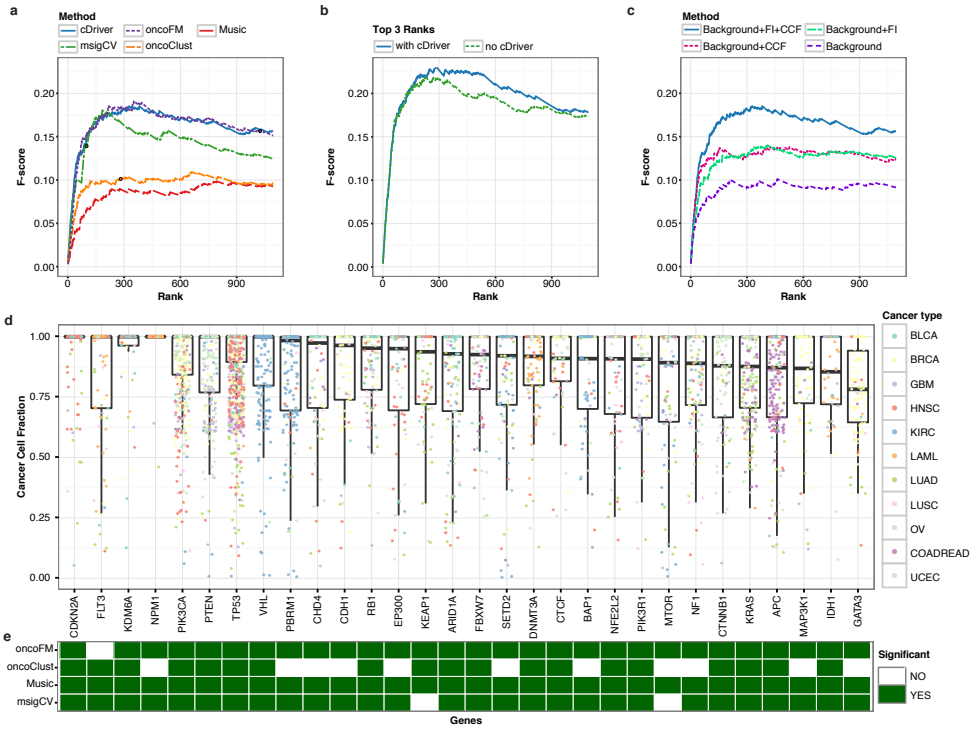

Figure 7.5: cDriver results and comparison with other methods for dataset composed of 12 cancers. (a) F-score for cDriver (solid blue line) and four other driver identification methods using the Pancan12 dataset (b) F-score for an ensemble approach of all tools with and without our Bayesian model, cDriver (blue and green lines respectively). (c) F-score for cDriver using: (i) only a published background model, (ii) including functional impact (FI), (iii) including cancer cell fraction, CCF, and (iv) a combination of all signals. (d) We ordered the top 30 cDriver-ranked genes on Pancan12 by their median CCF. (d) Matrix showing whether these top 30 genes were predicted as significant by the other four algorithms (Q value or FDR less than 0.1).

7.5 Tumor type driver gene landscape across 21 tumors

To obtain a list of high confident driver genes, we run cDriver on each and on a pooled set of 21 tumor types comprised of 6,870 samples (Pancan21, Table 10.4). By combining the top 10 ranked genes from each tumor type with the top 200 genes from the Pancan21 pooled analysis, we identified 234 high confidence driver genes. These driver genes were connected to each tumor type, ultimately defining a tumor type-driver gene (TTDG) landscape composed of 609 TTDG connections (Table 10.5). We investigated whether this landscape reveals associations for genes thought to be specific to one tumor type (Figure 10.6). We found that 56 out of 95 genes present only in the top 10 of one tumor type are also found in the long tail (top 100 ranked genes) of other tumor types (Table 10.6). Interestingly, genes formerly known to be tumor specific were also present at medium frequencies in other tumor types, such as *APC* in 13% of patients with stomach adenocarcinoma (STAD) and *FLT3* in 10% of patients with melanoma.

Consequently, we explored the entire set of 609 TTDG connections to identify novel candidate therapeutic targets. First, we assessed the number of PubMed records obtained when querying each of the 234 genes using the MeSH term neoplasm, resulting in 203 (86%) genes with at least one and 152 (65%) genes with at least five neoplasm-related PubMed records. Next, we queried the gene name in combination with the TTDG specific tumor term (see Methods). We identified 123 (20%) novel TTDG connections consisting of 73 genes with no or only one publication associated to the tumor type (Table 10.5). Furthermore, the network of these genes had significantly more interactions than expected in the STRING database (Adj. P value=1.47e-9, Fig. 10.8). Surprisingly, 22 of these interacting genes were annotated as chromatin modifiers in the gene ontology database (Adj. P value=2.55e-14), revealing an overlooked role of chromatin modification and chromatin organization in several tumor types. Importantly, we found that across most tumor types a large fraction of patients (up to 80%) is affected by a mutation in one of these chromatin-modifying proteins (Fig. 10.9) and at least one of these chromatin modifiers had a significant prognostic impact in one tumor type (Table 10.8).

Finally, we investigated the TTDG connections of known therapeutic targets, chromodomain helicase 4, *CHD4* (Fig. 7.6a), and chromatin regulator, *SMARCA4* (Fig. 7.6b). We found that *CHD4* acts as a driver for seven tumor types, while initially it was only associated to endometrial [Le Gallo et al., 2012] and ovarian carcinoma (Fig. 7.6c). *CHD4* is a tumor suppressor and core member of the nucle-

osome remodeling and deacetylase (NuRD) complex [Cai et al., 2014], which has been linked to multiple cellular processes including cell cycle regulation, DNA damage repair, and chromatin stability [Lai and Wade, 2011],[Chudnovsky et al., 2014]. Survival analysis showed that bladder carcinoma patients with mutations in *CHD4* have better prognosis than patients with other mutated drivers (Fig. 7.6e). *SMARCA4* has a known role in lung cancer and esophageal carcinoma, and it was found recurrently mutated in pancreatic, breast, lung, and prostate cancer cell lines [Roberts and Orkin, 2004]. However, the importance of this gene as a driver in tumorigenesis has been neglected in others cancers such as head and neck and liver carcinomas (Fig. 7.6d). It is the core subunit of a SWI/SNF complex and has several binding motifs to other tumor suppressors proteins [Orvis et al., 2014],[Biegel et al., 2014]. Most of the mutations fall in the active domains SNF2 (Fig. 7.6b) involved in the unwinding of the DNA. Additionally, we observed that liver carcinoma patients carrying a mutation in *SMARCA4* have a poor prognosis (Fig. 7.6f). In summary, all novel TTDG connections could be exploited as potentially therapeutic targets ultimately increasing the number of options for cancer prognosis.

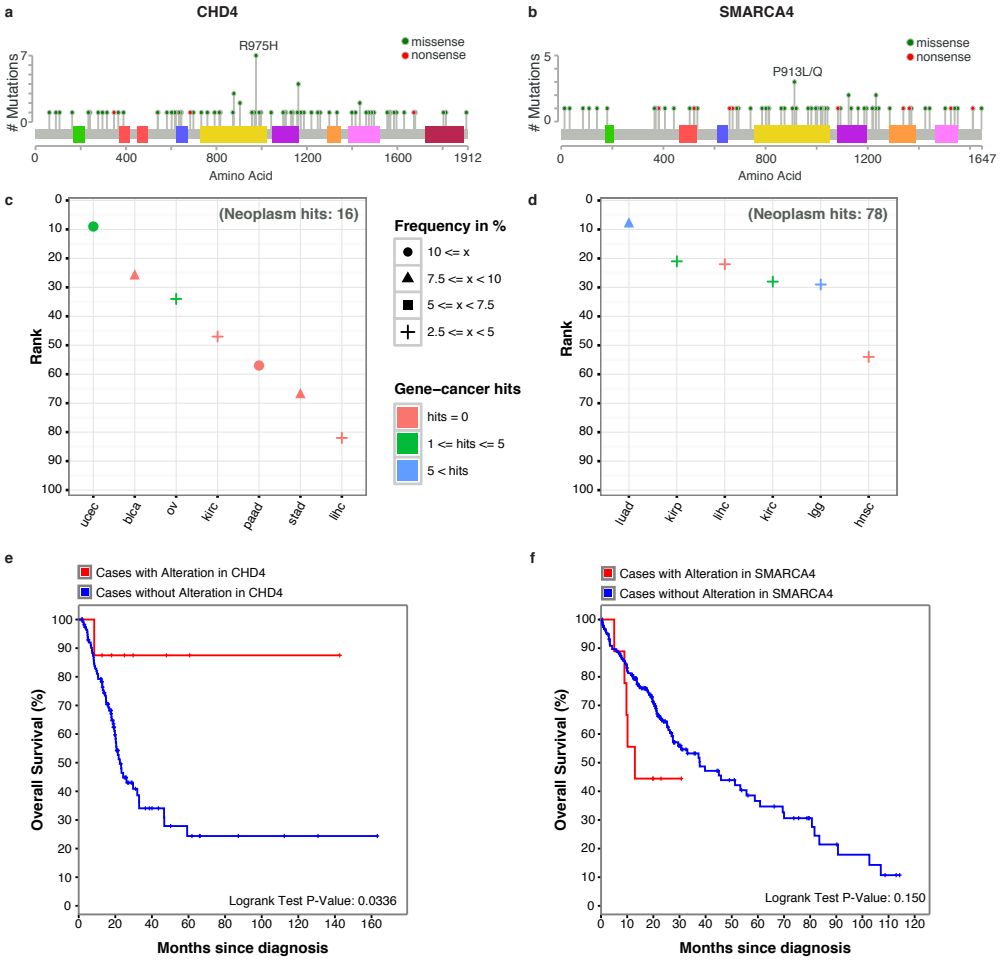

Figure 7.6: **Novel tumor type - driver gene (TTDG) connections, *CHD4* and *SMARCA4***. Distribution of somatic mutations found in (a) *CHD4* and (b) *SMARCA4*. The domains are colored following the cBioPortal color scheme. Most of the mutations are evenly distributed in *CHD4*, except for two small clusters at the beginning of the protein. In the case of *SMARCA4*, mutations tend to accumulate in the domains for ATP hydrolysis or DNA unwinding. TTDG connection landscape for (c) *CHD4* and (d) *SMARCA4*: the color indicates the number of pubmed hits related to each MeSH term. The shape indicates the frequency of patients affected by a mutation in the gene. Survival curves for (e) *CHD4* in bladder carcinoma and for (f) *SMARCA4* in liver hepatocellular carcinoma. Patients affected by a mutation are plotted in red.

Chapter 8

Discussion

Evolutionary signatures are imprinted in tumor genomes, and cDriver leverages them at the population, cellular and molecular level to identify cancer driver genes. For the first time, we integrate these measures into a Bayesian framework to detect known driver genes in three different tumor datasets, and we discover 123 unreported tumor type-driver gene connections across 21 tumor types. We show that these novel connections are strongly enriched for chromatin modifying proteins and have prognostic relevance, revealing an unexplored landscape of therapeutic targets.

Tumor heterogeneity complicates the discovery of cancer driver genes. Somatic mutations in these genes are subject to different selective pressures leading to complex tissue- and patient- specific clonal structures. Previous studies have shown that functional impact bias and recurrence represent evidence of positive selection at the molecular and population level, respectively [Tamborero et al., 2013b]. On one hand, the number of patients carrying a mutation in a gene hints at the importance of this gene for cancer etiology. On the other hand, mutations severely affecting protein function are more likely to be relevant for tumor formation [Ostrow et al., 2014]. Remarkably, while former studies have demonstrated that driver mutations rise in frequency within the tumor cell population [Schuh et al., 2012],[McGranahan et al., 2015] and passenger mutations accumulate neutrally following a power-law distribution [Williams et al., 2016], cellular prevalence (i.e. cancer cell fraction) of mutations has been neglected as a feature for cancer driver prediction.

We found that integrating cancer cell fraction, CCF, in our model increases the number of true driver genes detected. This makes intuitive sense because positively selected mutations will be present in a large fraction of cancer cells. Indeed, our results in Pancan12 agree with previous studies showing that most driver genes are affected by clonal mutations [McGranahan et al., 2015]. Furthermore, a CCF-adjusted mutation count allows estimation of background mutation rates at a higher resolution (i.e. at the cellular level). We speculate that CCF reduces the impact of technical artifacts arising from low allele fraction of false positive mu-

tations. However, the accuracy of CCF calculation depends on correct estimates of tumor purity and tumor ploidy, as well as adequate coverage of mutated positions. We expect that ultra-deep and single cell sequencing will further improve the power to detect mutations in small fractions of the tumor, making CCF an indispensible feature for accurate driver gene prediction.

Different types of driver genes are functionally constrained by different evolutionary pressures. Therefore, in concordance with the idea of combining complementary methods to improve the accuracy of capturing driver genes [Tamborero et al., 2013b], we show that functional impact and CCF equally improve performance, and their combination outperforms the use of each independently. Our results indicate that selection signatures at a molecular, cell, and population level are complementary, because they address different biological principles. Consequently, we demonstrate that cDriver improves performance when added to an ensemble of commonly used methods, and specifically contributes to detect infrequently mutated driver genes missed by the other approaches.

The total number of cancer genes driving tumorigenesis is still incomplete. Multiple gold standard datasets have been assembled in the literature (from 100 to 600 genes), none constituting a definite set of cancer driver genes. Although this is a limitation when benchmarking different methods, we show that the performance order is consistent for our method across all applied gold standards. Moreover, in this study none of the methods achieved a recall higher than 30% against Cancer Gene Census, suggesting that many genes have a role in tumorigenesis not related to positive selection of non-silent point mutations. Indeed, all methods tested here neglect other types of complex variation that may be driving cancer malignancy. These events are also under selection such as, positive selection of copy number alterations, fusion genes, regulatory, and synonymous mutations, as well as negative selection of cancer essential genes. In future work, we will extend our method to include these signatures to comprehensively catalog genes involved in tumorigenesis.

Our study also highlights the importance of prior information on driver gene prediction. A gene that is highly mutated (known driver) in one tumor type is probably a driver in other tumor types, even if it is infrequently mutated. We show that 73 genes highly ranked in one tumor type have been neglected in other cancers, despite a low, but substantial number of affected patients. Interestingly, this list of novel tumor type driver gene connections includes 22 chromatin-modifying proteins, extending the well-known and important role of chromatin remodeling in cancer [Roberts and Orkin, 2004]. These genes are global regulators of transcription activity and often act in a tissue-specific manner. According to a

network analysis, all 22 genes are interacting or co-localized, suggesting that a single hit is needed to drive tumorigenesis. Indeed, we found that mutations in the 22 chromatin-modifying proteins affect a large fraction of cancer patients and that in every analyzed tumor type at least one of these genes significantly affects prognosis. The genes *CHD4* and *SMARCA4* demonstrate how the landscape of "tumor type driver gene" connections can be exploited to identify novel therapeutic targets, especially for patients without a canonical driver mutation.

In conclusion, we show that an extensive landscape of therapeutic targets awaits exploration. We demonstrate that integrating cellular prevalence of somatic mutations as part of multiple signatures of tumor evolution allows for improved discovery of driver genes. As a result, it facilitates identification of novel "tumor type - driver gene" connections, which are key for improved cancer diagnosis, monitoring, and targeted treatment selection.

Chapter 9

Methods

9.1 Data

9.1.1 Pancan12 somatic mutation data

Filtered MAF files from 12 tumor types were obtained from synapse (syn1729383). Allele counts, ploidy status and purity estimates were merged into a single MAF file containing information for 3,276 samples and 617,354 mutations described elsewhere [Kandoth et al., 2013]. Allele counts for 782 samples not available from synapse were obtained from DCC-Firehose MAF files. Damage probability scores were added by applying a sigmoid transformation to CADD scores [Kircher et al., 2014] with mean $u = 15$ and scale factor of 2. Individual MAF files for each cancer were produced in order to perform downstream analysis. Expression values for this dataset were also obtained from synapse (syn1729383).

9.1.2 CLL somatic mutation data

385 CLL tumor-normal pairs sequenced by WES were analyzed using an in-house pipeline. Reads were aligned to hg19 using BWA-mem [Li et al., 2009b] and BAM files were post-processed (indel realignment, base quality recalibration) using GATK (https://www.broadinstitute.org/gatk/). Mutect [Cibulskis et al., 2013], Indelocator (https://www.broadinstitute.org/cancer/cga/indelocator) and ClinDel (unpublished) were used to produce a set of somatic SNVs and Indels. 27,625 mutations were annotated using eDiVA (www.ediva.crg.eu) to obtain several measures including allele counts, CADD damage score and population allele frequencies. Somatic SNVs that have a high number of occurrences in all paired normal samples, i.e. are likely germline variants (ND occurrence > 10), or a high rate of exclusion by MuTect across all samples (more than five times excluded) were excluded from the analysis. Indels that fall into a repeat masked region of the genome within 30 bp were also removed. Additionally, to reduce common false positive in detecting indels, we excluded indels that were reported in exons not typically expressed in B-lymphocytes. We considered exons not to be expressed

if they had an average or median fragment per kilobase per million (FPKM) < 1. For the expression analysis we calculated FPKMs for 270 CLL RNA-seq samples using tophat and cufflinks.

Finally, we produced a MAF file excluding variants in segmental duplications, common in the population (AF in EVS or 1000GP > 1%) or with alternative allele fraction (VAF) < 0.05. CADD damage score and VAF were added to each mutation in the final data. Ploidy values and cancer cell fraction of CNVs were obtained using the in-house developed tool clinCNV (unpublished).

9.1.3 Pancan21 somatic mutation data

The MAF files for 6,485 exome samples from 20 tumor types were downloaded from DCC-Firehose and combined with 385 CLL cases to obtain a large dataset of 21 tumor types (Table 10.4). Allele counts were transformed to VAF and CADD scores were added for each mutation. We removed duplicated samples and updated the gene symbols using the Hugo Gene Symbol database. Colon and rectal tumors were merged into one tumor type giving us a final set of 20 tumor types. All curated MAF files used in this study were uploaded to synapse (syn5593040).

9.2 cDriver package

We have developed cDriver (R package) to identify driver genes using next generation sequencing data from cancer genome studies (Fig 7.2). cDriver uses a MAF file (v2.4) as input data with the optional columns: (i) variant allele frequency (VAF), (ii) damage score, (iii) ploidy, (iv) purity, and (v) cellular fraction of the CNV. These measures can be obtained from current cancer genome or exome sequencing studies and public genome annotation databases.

One of the conceptual advances of our method is the inclusion of cancer cell fraction (CCF). Therefore, we developed a function to estimate CCF based on VAF, ploidy, CCF of the CNV, and tumor purity. However, cDriver also accepts clonality estimates from any other method.

To account for the variability of the background mutation rate (bmr) between genes, cDriver uses silent mutations to locally estimate the expected number of non-silent mutations. To this end, we applied a classical formula (K_a/K_s ratio) to detect selection bias in comparative genome analysis to incorporate CCF of silent mutations. However, somatic mutations are rare for most cancers and

many genes do not harbor silent mutations, restricting the usefulness of K_a/K_s. Consequently, cDriver calculates an average bmr using the CCF-adjusted Ka/Ks formula and a pre-calculated bmr taken from the literature [Lawrence et al., 2013].

Next, cDriver calculates posterior probabilities per gene using two Bayesian models, (i) the cancer hazard model and (ii) the driver model. The first model requires the incidence of the tumor in the population as prior probability, while the second requires a list of known driver genes (e.g. any gold standard used in this study) to estimate prior and likelihood values. In summary, cDriver combines recurrence, CCF, and functional impact as a foreground signal, and the averaged bmr as a background measure to calculate posterior probabilities for each gene.

CCF calculation

To calculate Cancer Cell Fraction (CCF) we developed a function as part of the cDriver package. Intuitively, it is easy to deduce that the variant allele should be observed in approximately half of the reads if it represents a heterozygous variant in a diploid cell. Therefore, CCF calculation in a diploid locus is the variant allele frequency (VAF) multiplied by two and corrected by the purity of the cancer sample. More general for diploid and non-diploid loci, for any SNV or Indel, the general equation to calculate CCF is:

$$CCF_{SNV} = \frac{VAF_{SNV} * (2 + (ploidy_{SNV} - 2) * CCF_{SNV})}{purity} \tag{9.1}$$

where VAF_{SNV} is the observed variant allele frequency, $ploidy_{SNV}$ is the ploidy of the locus, CCF_{SNV} is the CCF for the copy number change, and $purity$ is the fraction of cancer cells in the sequenced tumor sample.

Background mutation probability using CCF-adjusted K_a/K_s

We adjusted the classic formula for detecting selection from comparative data [Nielsen, 2005], the ratio between the rate of nonsynonymous substitutions (n_a) per nonsynonynous sites (N_a) and the rate of synonymous substitutions (n_s) per synonymous sites (N_s):

$$\frac{K_a}{K_s} = \frac{(n_a/N_a)}{(n_s/N_s)} \tag{9.2}$$

to estimate the expected number of non-silent mutations under no selective pressures (i.e. neutral evolution). First, we adapted the formula to take into account

cancer cell fraction of mutations by calculating n_s and n_a as the sum of CCF of silent and non-silent mutations, respectively, resulting in n_s^{ccf} and n_a^{ccf}. The CCF-adjusted $\frac{K_a}{K_s}$ formula is:

$$\frac{K_a^{ccf}}{K_s^{ccf}} = \frac{(n_a^{ccf}/N_a)}{(n_s^{ccf}/N_s)} \tag{9.3}$$

Next, we estimated the expected n_a^{ccf} for each gene in a cancer cohort-specific based on the observed number of CCF-adjusted silent mutations in coding regions (n_s^{ccf}) within the provided cohort (e.g. WES data from BRCA, CLL, Pancan12, Pancan21). The total number of sites (N_a and N_s) was taken from Lawrence et al [Lawrence et al., 2013]. Under the assumption of neutral selection ($\frac{K_a^{ccf}}{K_s^{ccf}} = 1$), we estimated n_a^{ccf} as:

$$n_a^{ccf} = \frac{n_s^{ccf} * N_a}{N_s} \tag{9.4}$$

To avoid zero expected non-silent mutations for genes where zero silent mutations were observed, we defined a minimum n_s^{ccf}. This is based on the assumption that one non-silent mutation per gene in the cohort can occur by chance and should not be considered a positive selection signal. Hence, the minimum n_s^{ccf} is defined such that $\frac{K_a^{ccf}}{K_s^{ccf}} = 1$ (neutral) under the assumption that a single non-silent mutation at 100% CCF and zero silent mutations are observed. To avoid similar issues of arbitrarily small expected n_a^{ccf} when silent mutations at very low CCF are found, any observed n_s^{ccf} smaller than the minimum defined previously is adjusted.

After obtaining the expected n_a^{ccf} for all genes we calculated the probability that a patient has at least one non-silent mutation in gene X. To this end, we approximated the average number of somatic non-silent mutations in a healthy cohort (r) using the cancer cohort. This is feasible based on the assumption that the majority of clonal mutations ($CCF_{snv} >= 0.85$) are hitchhiking passengers present before tumor initiation. Following this assumption, we estimated the probability $P(X >= 1)$ using a binomial distribution:

$$P(X >= 1) = 1 - P(X = 0) = 1 - (1 - \frac{n_a^{ccf}}{\sum_{i}^{genes} n_{a_i}^{ccf}})^r \tag{9.5}$$

Integration of multiple background mutation rate estimates

To compensate for the lack of power of the CCF-adjusted K_a/K_s model for genes with few or zero mutations in coding regions, we additionally incorporated the non-coding mutation rate (ncmr) provided by Lawrence et al [Lawrence et al., 2013]. For each gene the ncmr and the gene length were used to calculate the number of total expected mutations (silent and non-silent). Under the assumption of neutral evolution ($K_a/K_s = 1$), we determined the expected number of non-silent mutations following a rearrangement of the classical formula into:

$$n_a = \frac{n_t}{1 + \frac{N_s}{N_a}} \tag{9.6}$$

Finally, we calculated the probability that a gene has at least one non-silent mutation using the aforementioned binomial distribution formula but without CCF.

Similarly, cDriver could integrate any measure of background mutation rate. Here, we used as final background mutation probability (bmp) the average bmp obtained by the two methods described above (CCF-adjusted K_a/K_s, and ncmr). Note that both probabilities are upper bound estimates of bmp for two reasons: i) silent and non-coding mutations might be under positive selection and, ii) cancer essential genes might never be observed with non-silent mutations, as these might be lethal for the cell, while silent and non-coding mutations are mostly neutral.

9.2.1 Bayesian inference models

To identify and rank driver genes we developed two Bayesian models, called cancer-hazard inference model and driver inference model. The cancer-hazard inference model estimates the posterior probability of developing cancer if the focal gene is mutated, given evidence from the data, i.e. somatic mutations of the gene in a cohort of cancer patients. The driver inference model estimates the posterior probability that a gene is a true cancer driver given evidence from the data.

The cancer-hazard model

In the first model, we adapted Bayes formula to calculate a posterior probability of developing cancer given that a focal gene is mutated as

$$P(cancer|nsmut) = \frac{P(nsmut|cancer) * P(cancer)}{P(nsmut|cancer) * P(cancer) + P(nsmut|\neg cancer) * P(\neg cancer)} \quad (9.7)$$

where the prior probability for developing cancer, $P(cancer)$, is the incidence of the cancer type in the population. The likelihood, $P(nsmut|cancer)$,that a cancer patient carries a non-silent mutation in a gene of interest is estimated from the cohort. To this end, we used the sum of CCF times the adjusted-CADD damage probability per gene across all patients:

$$P(nsmut|cancer) = \frac{\sum_{i=1}^{n} CCF_i * CADD_i}{n} \quad (9.8)$$

where i is the index of the patient and n the total size of the cohort. If a patient did not have any non-silent mutation, then CCF was equal to zero. If two non-silent mutations were found in a patient in the same gene of interest, we used the mutation with the highest CCF.

We defined the marginal probability of having a non-silent mutation as the sum of the numerator of eq. 9.8, plus the conditional probability of having a non-silent mutation in a healthy population times the probability of a healthy individual. We denoted $P(nsmut|\neg cancer)$ as the somatic background mutation probability (bmp). To our knowledge there is no large enough cohort of healthy people examined for tissue specific somatic mutations, therefore direct estimation from data is not possible. However, we estimated an upper bound of the bmp as described in the previous section.

The driver inference model

In the second model, we calculated the posterior probability that a gene is a cancer driver given the mutation data in the studied cohort using the formula:

$$P(driver|D) = \frac{P(D|driver)^m * P(\neg D|driver)^{n-m} * P(driver)}{P(D|driver)^m * P(\neg D|driver)^{n-m} * P(driver) + P(D|passenger)^m * P(\neg D|passenger)^{n-m} * P(passenger)} \quad (9.9)$$

where now $m = \sum_{i=1}^{n} CCF_i * CADD_i$ and n is the total size of the cohort.

To estimate a prior probability of a gene being a driver we need to consider that most tumor types can be caused by mutations in a different set of genes. Depending on which tumor type or group of cancers (pan-cancer) we were analyzing, the number of known driver genes differs and hence the prior probabilities change (e.g. ovarian cancers are in most cases caused by a mutation in *TP53*, while the number of published genes involved in CLL ranges from 20 to 40). We estimated the prior probability that a random gene is a driver as equal to the ratio between the number of known driver genes of the cancer type and the total number of protein coding genes:

$$P(driver) = \frac{\#drivergenes}{\#genes}$$

The number of driver genes can be approximated as the number of published driver genes for a particular cancer type. If the cancer has not been studied yet, or if we deal with pan-cancer sets of multiple cancer types, the prior can be approximated using any gold standard list of cancer driver genes.

Because of inter-tumor heterogeneity genes that are known to be cancer drivers in a given tumor type will not necessarily be mutated in all patients. The probability that a gene is mutated given that it is a known driver can be estimated as:

$$P(D|driver) = \frac{\#mutationsindrivers}{\#genes * \#drivers}$$

where we assume that all drivers have the same chance to be mutated. As this assumption is weak the cDriver package allows the user to define better estimates for this likelihood. The probability that a gene is mutated given that it is not a driver is estimated from the background mutation rate as described previously.

The other terms in the equation 9.9, $P(\neg D|driver)$ and $P(\neg D|passenger)$ were calculated as the complementary events of $P(D|driver)$ and $P(D|passenger)$ described above.

9.3 Benchmarking

9.3.1 Running competing methods for cancer driver gene identification

(i) MuSiC on Pancan12 and BRCA: gene coverage files and results from the MuSiC suite for the Pancan12 dataset (including all BRCA cases used here) were obtained from synapse (syn819550, syn1713813, syn1734155). For this set of results, we only considered genes that were less than 0.05 FDR in at least two out of three measures. The rank order was based on the P-values given by the CT test, followed by the LRT test, followed by the number of cases affected. MuSiC on CLL dataset: gene coverage files for 385 samples were generated from tumor-normal BAM files using the function calc-bmr available in the MuSiC suite. The region of interest files (ROI) were downloaded from synapse and merged to avoid duplicates. For comparative analysis we used the sorted list returned by the tool. (ii) MutSigCV on all datasets: we ran MutSigCV using default parameters on all datasets, assuming full coverage and using the example covariate space provided in the source code. (iii) OncodriveClust and (iv) oncodriveFM: The analysis was performed by the group of Nuria Lopez-Bigas. (v) cDriver on all dataset: We ran cDriver using default parameters. Prior values used for the cancer-hazard model are shown in table 10.4.

9.3.2 Gold standard and parameter selection

For comparative analysis of several competing methods we assembled different gold standard datasets. We downloaded published lists of significantly mutated genes (SMGs): 1) [Kandoth et al., 2013],consisting of 127 significantly mutated genes across 12 tumor types; 2) [Lawrence et al., 2014], 261 cancer genes predicted from 21 tumor types; 3) [Tamborero et al., 2013b], 435 cancer genes predicted by a combination of multiple algorithms. For benchmarking purposes, we only used the 291 genes labeled as high confident; 4) Cancer Gene Census [Futreal et al., 2004], list of 547 manually curated cancer driver genes; 5) [Xie et al., 2014], list of 556 cancer-associated genes; 6) [Landau and Wu, 2013], a list of 21 CLL specific genes and 7) TCGA breast [TCGANetwork, 2012c], 35 breast cancer genes.

The gold standard genes for breast cancer consisted of the union of dataset (7), and breast cancer genes found in (2) and (4) plus the top 20 genes identified in COSMIC. For CLL we merged dataset (6) and CLL genes found in (2) and (4)

plus the top 20 genes identified in COSMIC. Subsequently, we manually curated this list by checking the number of PubMed records found by querying the HUGO gene name and the corresponding MeSH term for breast cancer and CLL. We excluded histone genes that have no relevant publication associating them to recurrent somatic mutations. Results for Pancan12 were benchmarked using datasets 1-5. Single tumor types (i.e. CLL and breast cancer) were benchmarked against the tumor specific gold standards assembled as described above. In addition, we compared the performance of the methods on Pancan12 with and without filtration of non-expressed genes.

Furthermore, we benchmarked our method cDriver under several scenarios and parameter settings: the F-score curve for (i) cDriver using a simple recurrence model where no CCF or functional impact was used and the background model did not include CCF-adjusted Ka/Ks, (ii) cDriver using only functional impact, (iii) cDriver using CCF adjusted mutation counts and the CCF-adjusted Ka/Ks background model, and (iv) cDriver using all signatures of positive selection. Lastly, we benchmarked an ensemble of complementary methods (MuSic, MutSigCV, OncodriveFM, OncodriveClust) including and not including cDriver. For this, we calculated a combined rank and calculate the F-score. We used Borda count ranking method with truncated ranks (up to two times the gold standard size) but using ranks of only the three best methods.

9.4 Defining the landscape of tumor type driver gene connections in Pancan21

We ran cDriver on each of the 21 tumor types separately (syn5593040, Table 10.4) and in the pooled Pancan21 dataset. Next, we used the union of the top 10 genes for each tumor type and the top 200 genes of the pooled Pancan21 analysis to create a list of high confidence driver genes. For each of these genes, we noted their presence among the top 100 ranked genes of each tumor type to define a tumor type driver gene (TTDG) connection. We defined genes found in the top 10 of only one tumor type and not in the top 100 of any other tumor type as highly tumor specific.

9.4.1 Identification of novel TTDG connections by PubMed mining

For each high confidence gene, we queried the HUGO symbol together with the MeSH term neoplasm (i.e. ATM[TIAB] AND neoplasm[MH]) against the PubMed database in order to test if the gene have been associated to any cancer type. The HUGO symbol had to be found in the title and/or abstract. Next, for each TTDG connection we queried the HUGO symbol together with the MeSH term of the associated tumor type. Based on the PubMed mining results, we used the following criteria to detect novel TTDG connections: (i) the tested gene was among the top 200 of the Pancan21 analysis, (ii) the gene had at least 5 neoplasms-related PubMed records, (iii) the gene was among the top 100 of the corresponding tumor type (defined by its TTDG connection), (iv) the gene had zero or one TTDG specific PubMed record, and, (v) if one TTDG specific PubMed entry was retrieved, we required that the publication did not report recurrent somatic mutations for that gene in the tumor type of interest.

9.4.2 Protein interaction and functional enrichment analysis of novel TTDGs

We used STRING v10.0 (www.string-db.org) find the connectivity among the novel TTDGs reported. We input the list of genes into the webserver and retrieved the network using all STRING features except text-mining and database evidence. STRING provides built-in analysis functions to detect protein-protein interactions and to perform GO term enrichment analysis. For the latter, STRING performs a Hypergeometric test and corrects for multiple testing using Benjamini and Hochberg.

9.4.3 Individual gene analysis

To visualize the somatic mutations on the gene structure we used MutationMapper [Vohra and Biggin, 2013]. We input our list of non-silent mutations for genes individual genes from the Pancan12 dataset. The clinical data (defined as processed data, level 2, by TCGA) for the patients was downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). Kaplan-Meier curves and log-rank p-values for the selected genes were calculated using the R package Surv.

Chapter 10

Supplementary information

Figures

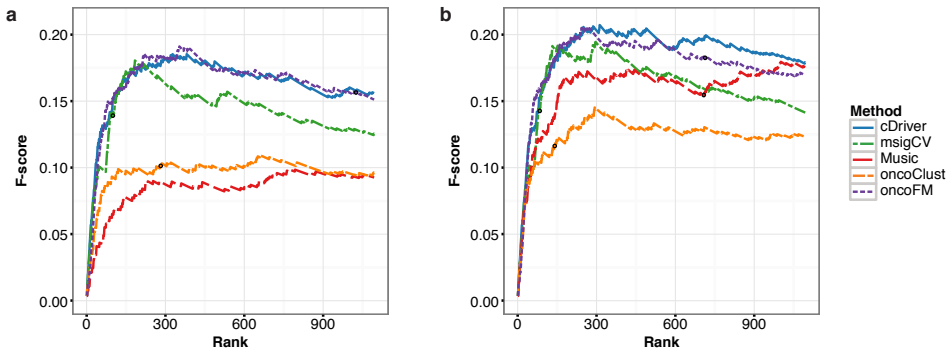

Figure 10.1: **F-score measure on filtered versus unfiltered data**. F-score curves for competing methods with and without post-filtration of non-expressed genes in the Pancan12 dataset. All methods are shown, and their corresponding significance threshold ranking using Q value < 0.1.

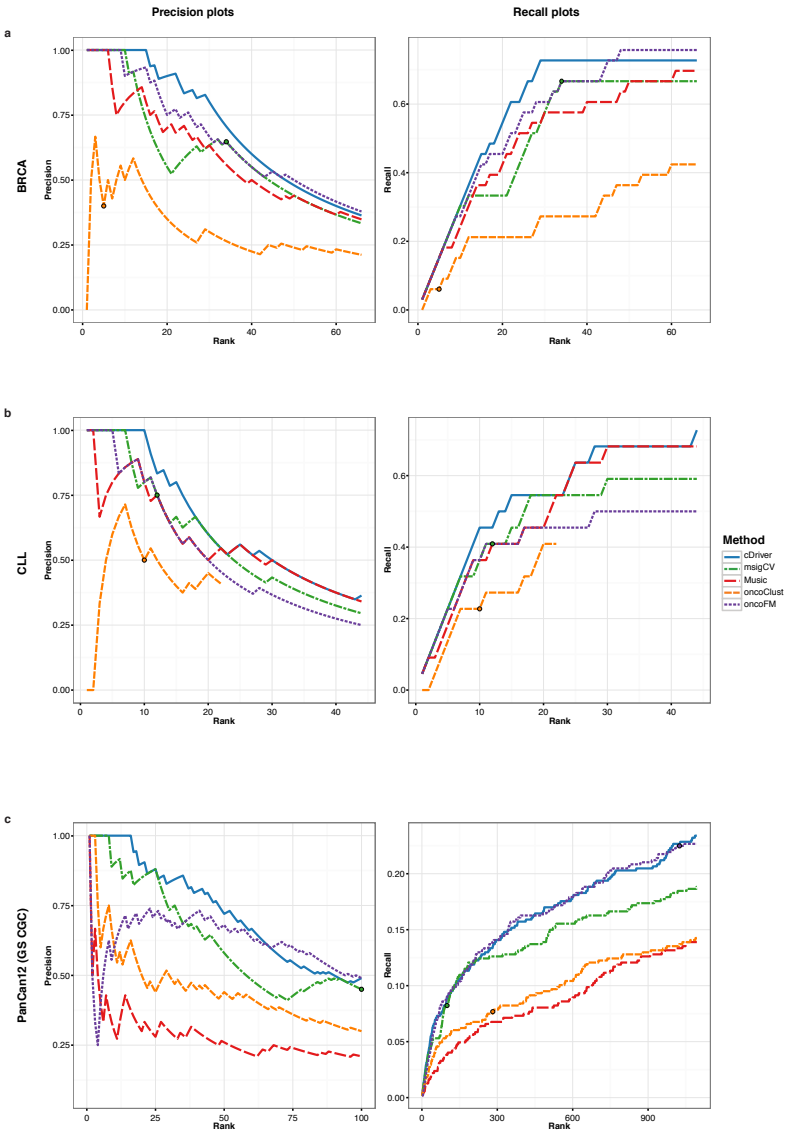

Figure 10.2: **Precision and recall for five driver identification methods benchmarked on three different datasets.** Precision and recall plots for BRCA (a), CLL (b) and Pancan12 (c) cohorts. Precision and recall are shown for methods: cDriver (blue), MutsigCV (misgCV, green), MuSiC(red), OncodriveFM (oncoFM, purple) and OncodriveCLUST(oncoClust, orange). As gold standard, manually compiled lists of 44 genes for BRCA and 22 genes for CLL were used, while Cancer Gene Census was used for Pancan12.

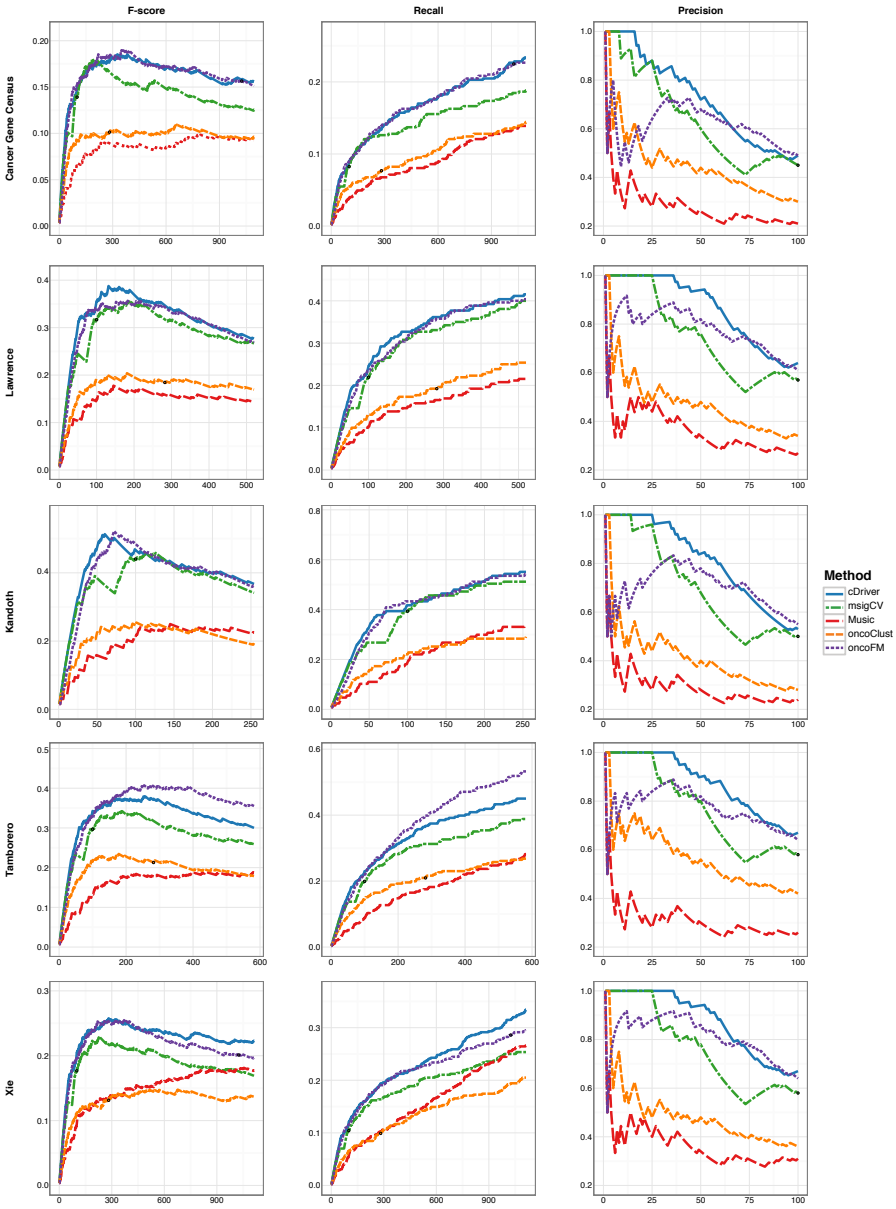

Figure 10.3: **Evaluation of several measures for five driver identification methods benchmarked on Pancan12.** Benchmarking of F-score, precision and recall measures for five driver identification methods benchmarked on Pancan12 across five gold standard datasets. X-axis shows the ranked list of genes for each tool. Y-axis shows F-score, Precision, and Recall according to the header.

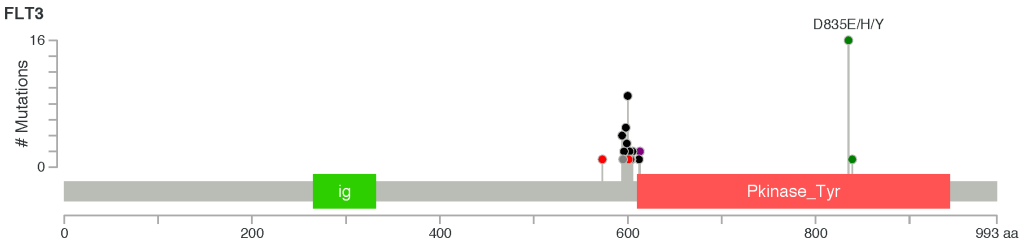

Figure 10.4: **Somatic mutations in *FLT3***.

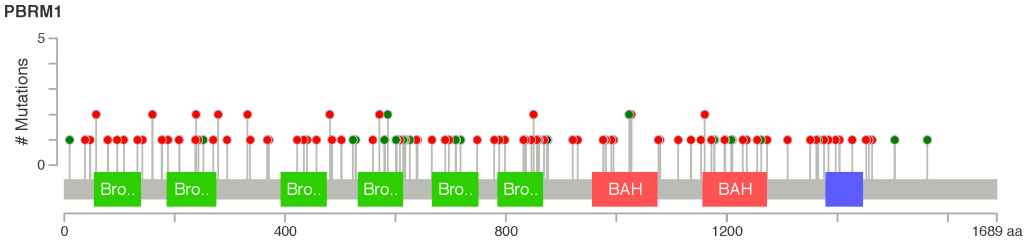

Figure 10.5: **Somatic mutations in *PBRM1***.

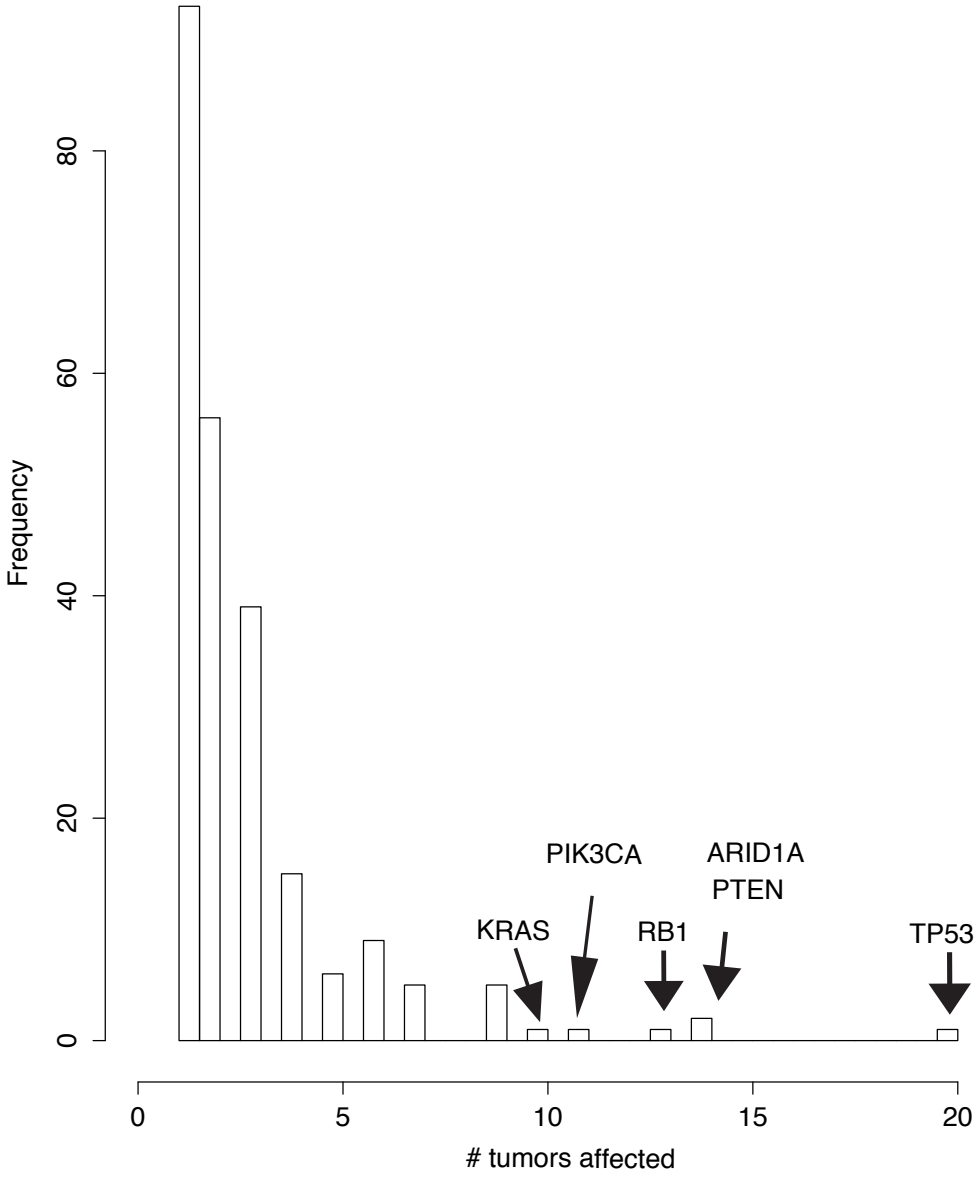

Figure 10.6: **Distribution of genes affecting tumor types.** Histogram of high confidence driver genes and the number of tumor types affected by them.

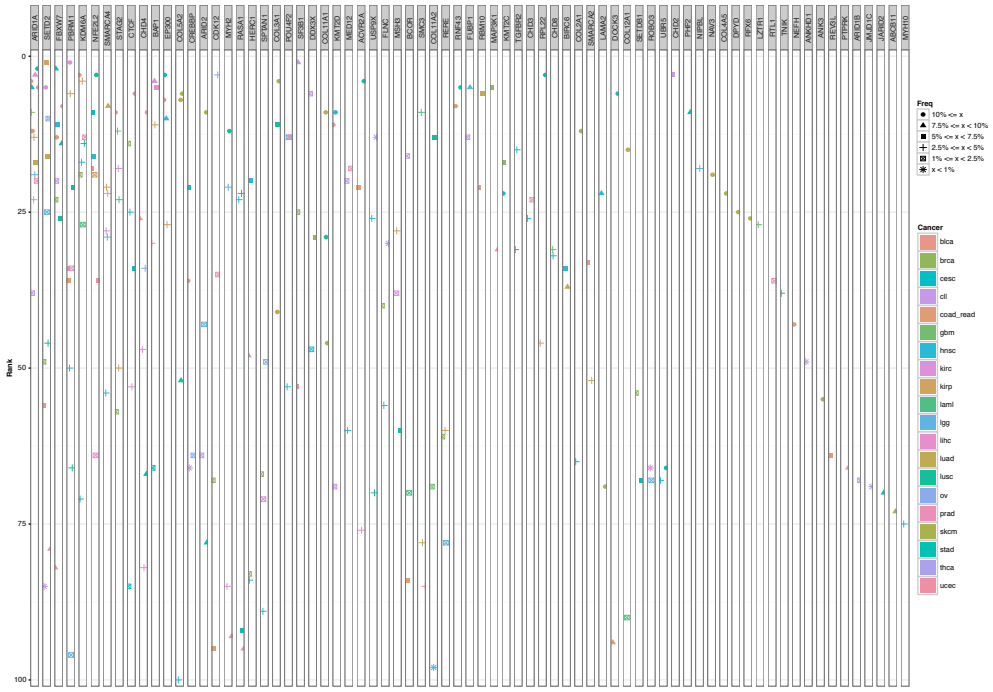

Figure 10.7: **Extended figure of the tumor type driver gene connection landscape.** Thirty selected genes are shown together with the tumors affected by them, the ranking in that tumor type, and the frequency of patients having the gene mutated.

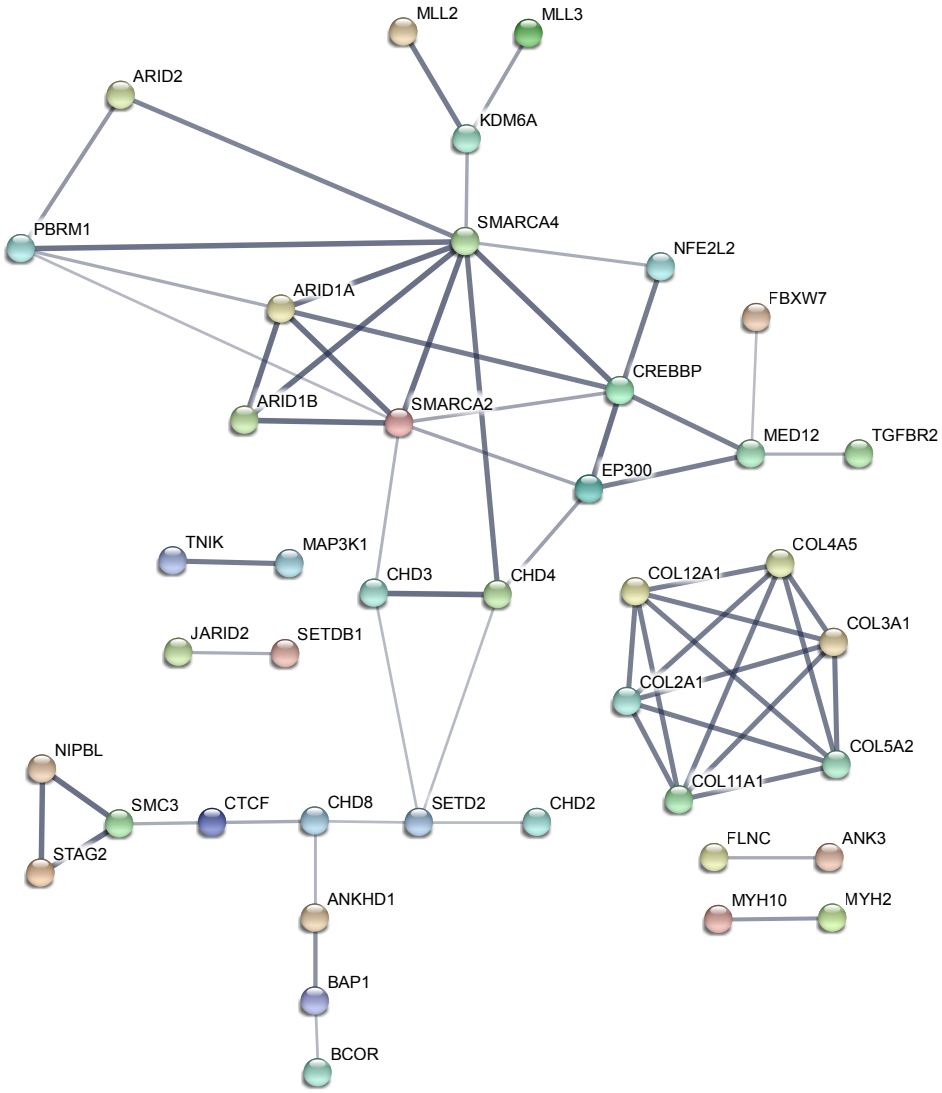

Figure 10.8: **STRING PPI analysis of selected genes.** STRING enrichment analysis using all functions except text mining shows a significant enrichment for interactions in the unreported TTDG dataset. The main function revealed in these genes is chromatin modification.

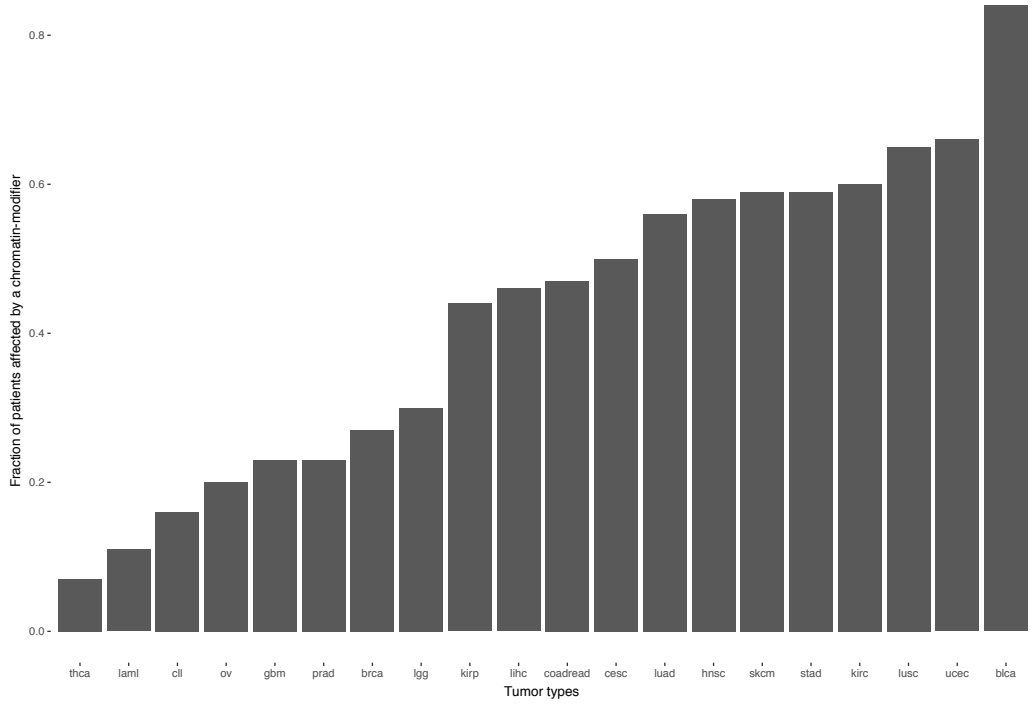

Figure 10.9: **Chromatin modifiers affect a large proportion of individuals with cancer.** Proportion of individuals harboring a non-silent mutation in at least one of the novel chromatin modifiers described in the text.

Tables

Tumor type	Patients	Non-silent	Silent	Incidence	Source	Gold standard genes
Breast invasive carcinoma	762	29929	8612	0,00125	Kandoth et al 2012	33
Chronic Lymphocytic Leukemia	385	8145	3015	0,00005	ICGC	22
Pancancer12	3205	291129	90884	0,0045	Kandoth et al 2012	CGC

Table 10.1: **Data used for benchmarking cDriver against competing methods.**

BRCA	CLL
AKT1	ATM
APC	BCOR
ARID1A	BRAF
ATM	CHD2
BRCA1	DDX3X
BRCA2	EGR2
BRIP1	FBXW7
CASP8	ITPKB
CBFB	KLHL6
CCND1	KRAS
CDH1	MED12
CDKN1B	MYD88
CTCF	NOTCH1
ERBB2	NRAS
FOXA1	POT1
GATA3	SAMHD1
KMT2A	SF3B1
KMT2C	TP53
MAP2K4	XPO1
MAP3K1	ZMYM3
MYB	BIRC3
NCOR1	MYC
NF1	
PALB2	
PIK3CA	
PIK3R1	
PTEN	
RB1	
RUNX1	
SF3B1	
TBL1XR1	
TBX3	
TP53	

Table 10.2: **List of gold standard genes for Breast cancer and Chronic Lymphocytic Leukemia.**

Method	Significance ($Qvalue < 0.05$)	Significance ($Qvalue < 0.1$)	Best F-score rank	Best F-score
cDriver	-	-	327	0.19
msigCV	99	100	184	0.18
MuSiC	2175	2782	793	0.10
oncoCLUST	181	282	660	0.11
oncoFM	798	1025	353	0.19

Table 10.3: **Comparison of number of significant genes and the best F-score for each method.**

Tumor Code	Total samples	Total mutations	MESH term	Incidence	Mutation/sample
blca	238	84839	Urinary Bladder Neoplasms[MH]	0.00011	356
brca	982	89105	breast neoplasms[MH]	0.00125	91
cesc	194	44441	Uterine Cervical Neoplasms[MH]	0.000089	229
cll	385	18791	Leukemia, Lymphocytic, Chronic, B-Cell[MH]	0.00007	49
coadread	298	129686	Colorectal Neoplasms[MH]	0.00047	435
gbm	290	21075	Glioblastoma[MH]	0.00003	73
hnsc	509	120520	Head and Neck Neoplasms[MH]	0.00003	237
kirc	213	21716	Kidney Neoplasms[MH]	0.00012	102
kirp	168	16210	Kidney Neoplasms[MH]	0.00001845	96
laml	197	2420	Leukemia, myeloid, Acute[MH]	0.00003	12
lgg	467	42943	Glioma[MH]	0.000125	92
lihc	198	27865	Liver Neoplasms[MH]	0.00005	141
luad	543	235233	Lung Neoplasms[MH]	0.0006	433
lusc	178	65218	Lung Neoplasms[MH]	0.0006	366
ov	316	18576	Ovarian Neoplasms[MH]	0.00022	59
prad	300	21206	Prostatic Neoplasms[MH]	0.001047	71
skcm	369	324163	Skin Neoplasms[MH]	0.000174	878
stad	372	247156	Stomach Neoplasms[MH]	0.000076	664
thca	405	7443	Thyroid Neoplasms[MH]	0.000129	18
ucec	248	180680	Uterine Neoplasms[MH]	0.0002	729

Table 10.4: **Data used for driver gene prediction using cDriver accross 21 tumor types.**

							MESH term	
Tumor Type	Hugo Symbol	Frequency	Samples	Rank	TOP10	TOP200	Tumor Specific	Neoplasms
blca	TP53	0,50	118	1	Y	Y	195	5236
blca	RB1	0,17	41	2	Y	Y	41	1394
blca	KDM6A	0,24	57	3	Y	Y	5	43
blca	ARID1A	0,26	61	4	Y	Y	4	179
blca	ELF3	0,12	28	5	Y	N	0	37
blca	CDKN1A	0,10	23	6	Y	N	6	423
blca	EP300	0,14	34	7	Y	Y	3	66
blca	PIK3CA	0,19	46	8	Y	Y	26	1547
blca	STAG2	0,14	33	9	Y	Y	6	24
blca	CDKN2A	0,06	15	10	Y	Y	44	1963
blca	KMT2D	0,29	70	11	Y	Y	1	12
blca	ARHGAP35	0,08	20	12	N	Y	0	1
blca	ERBB3	0,11	25	17	N	Y	12	584
blca	FGFR3	0,12	29	18	N	Y	197	568
blca	NFE2L2	0,06	15	19	Y	Y	0	54
blca	RBM10	0,07	16	20	Y	Y	0	10
blca	FNBP4	0,05	11	23	Y	Y	0	0
blca	FOXA1	0,05	11	24	Y	N	2	201
blca	CHD4	0,08	19	25	Y	Y	0	16
blca	HRAS	0,04	10	26	Y	N	38	504
blca	SMARCA2	0,07	16	30	N	Y	0	22

blca	BAP1	0,04	9	31	Y	Y	1	151
blca	PBRM1	0,07	16	34	Y	Y	0	46
blca	DMXL2	0,09	21	35	N	Y	0	0
blca	CREBBP	0,13	30	36	N	Y	2	78
blca	SCAF4	0,06	15	39	N	Y	0	0
blca	PTEN	0,04	9	42	Y	Y	74	5794
blca	GBF1	0,04	9	45	Y	Y	0	2
blca	HERC1	0,08	20	48	N	Y	0	5
blca	SF3B1	0,06	14	54	Y	Y	1	119
blca	SETD2	0,05	13	56	Y	Y	1	49
blca	ADGRL2	0,06	14	59	N	Y	0	0
blca	MYH7	0,07	16	61	N	Y	0	2
blca	ADGRV1	0,12	28	62	Y	Y	0	0
blca	CNOT1	0,06	15	68	N	Y	0	2
blca	MYH8	0,05	11	70	N	Y	0	3
blca	FBXW7	0,09	21	82	Y	Y	0	176
blca	KRAS	0,03	6	94	Y	Y	13	4562
brca	TP53	0,31	300	1	Y	Y	704	5236
brca	PIK3CA	0,32	318	2	Y	Y	279	1547
brca	CDH1	0,11	111	3	Y	Y	179	925
brca	GATA3	0,10	97	4	Y	Y	103	242
brca	MAP3K1	0,07	70	5	Y	Y	54	69
brca	PTEN	0,04	35	6	Y	Y	667	5794
brca	MAP2K4	0,03	32	7	Y	N	9	35
brca	ARID1A	0,03	28	8	Y	Y	10	179
brca	CBFB	0,02	23	9	Y	N	2	182
brca	FOXA1	0,02	23	10	Y	N	106	201
brca	NCOR1	0,04	41	12	N	Y	15	47
brca	MYH7	0,02	18	13	N	Y	0	2
brca	CTCF	0,02	17	14	Y	Y	36	163
brca	ADGRV1	0,03	30	15	Y	Y	0	0
brca	RB1	0,02	19	16	Y	Y	100	1394
brca	KMT2C	0,07	69	17	N	Y	0	5
brca	KDM6A	0,02	16	22	Y	Y	5	43
brca	HCFC1	0,02	17	23	N	Y	0	1
brca	FBXW7	0,02	15	24	Y	Y	14	176
brca	SF3B1	0,02	16	26	Y	Y	3	119
brca	CCAR1	0,01	11	27	Y	N	0	8
brca	CAD	0,01	14	33	N	Y	569	1556
brca	ZSWIM8	0,02	17	36	N	Y	0	0
brca	FLNC	0,02	21	43	N	Y	0	19

brca	ERBB3	0,02	18	47	N	Y	183	584
brca	SETD2	0,02	17	48	Y	Y	4	49
brca	LRBA	0,02	24	49	N	Y	4	4
brca	ADGRL2	0,01	11	50	N	Y	0	0
brca	SETDB1	0,02	16	55	N	Y	2	22
brca	STAG2	0,01	13	59	Y	Y	0	24
brca	COL14A1	0,02	18	61	N	Y	1	4
brca	SRRM2	0,01	14	63	N	Y	0	0
brca	RERE	0,01	11	64	N	Y	0	8
brca	SPTAN1	0,01	12	69	N	Y	0	7
brca	CDK12	0,01	14	72	Y	Y	1	13
brca	BRCA2	0,02	16	80	Y	N	3148	4282
brca	NF1	0,03	28	84	Y	Y	42	2707
cesc	PIK3CA	0,27	53	1	Y	Y	19	1547
cesc	FBXW7	0,10	19	2	Y	Y	3	176
cesc	EP300	0,11	21	3	Y	Y	2	66
cesc	PTEN	0,08	16	4	Y	Y	44	5794
cesc	ADGRV1	0,13	26	5	Y	Y	0	0
cesc	ADGRV1	0,13	26	5	Y	Y	0	0
cesc	ARID1A	0,08	15	6	Y	Y	4	179
cesc	HLA-A	0,08	16	7	Y	N	55	1714
cesc	ZNF750	0,05	10	8	Y	N	0	2
cesc	KRAS	0,06	11	9	Y	Y	12	4562
cesc	NFE2L2	0,06	12	10	Y	Y	1	54
cesc	ADGRL3	0,04	8	12	N	Y	0	0
cesc	RB1	0,05	9	14	Y	Y	13	1394
cesc	MT-CO1	0,04	8	17	Y	N	0	5
cesc	DOCK11	0,05	9	18	N	Y	0	4
cesc	ERBB3	0,06	11	19	N	Y	3	584
cesc	CREBBP	0,07	14	22	N	Y	0	78
cesc	KMT2C	0,15	29	23	N	Y	0	5
cesc	CHD3	0,05	10	25	N	Y	0	6
cesc	EPHA2	0,04	8	28	N	Y	3	329
cesc	CASP8	0,05	9	30	Y	Y	3	231
cesc	TGFBR2	0,03	5	33	N	Y	0	184
cesc	FAM193A	0,03	6	36	N	Y	0	0
cesc	TNIK	0,05	10	42	N	Y	0	12
cesc	MT-ND5	0,10	19	43	N	Y	0	0
cesc	YLPM1	0,05	9	48	N	Y	0	0
cesc	ARHGAP35	0,04	8	54	N	Y	0	1
cesc	CCAR1	0,03	5	57	Y	N	0	8

cesc	TP53	0,05	9	61	Y	Y	57	5236
cesc	NOTCH1	0,06	11	63	Y	Y	21	1195
cesc	UBR5	0,05	9	68	N	Y	0	7
cesc	COL2A1	0,03	6	70	N	Y	0	31
cesc	BAP1	0,02	4	72	Y	Y	0	151
cesc	KDM6A	0,03	6	74	Y	Y	2	43
cesc	CTCF	0,02	4	91	Y	Y	2	163
cesc	ADGRB3	0,06	11	92	Y	Y	0	0
cll	SF3B1	0,10	38	1	Y	Y	55	119
cll	NOTCH1	0,08	29	2	Y	Y	87	1195
cll	ATM	0,07	26	3	Y	Y	167	2197
cll	CHD2	0,06	25	4	Y	Y	1	6
cll	MYD88	0,03	12	5	Y	N	18	360
cll	TP53	0,03	12	6	Y	Y	259	5236
cll	FUBP1	0,03	13	7	Y	Y	0	25
cll	ZNF292	0,03	10	8	Y	N	1	4
cll	KLHL6	0,03	10	9	Y	N	2	6
cll	DDX3X	0,02	6	10	Y	Y	3	23
cll	BCOR	0,02	8	13	N	Y	0	42
cll	MED12	0,03	10	16	N	Y	0	56
cll	FBXW7	0,02	7	21	Y	Y	3	176
cll	BRAF	0,02	8	22	Y	Y	8	4834
cll	POLR3B	0,01	4	23	Y	Y	0	0
cll	CARD11	0,03	10	26	Y	N	2	46
cll	BRCA1	0,01	3	27	Y	N	4	7290
cll	MGA	0,02	9	28	N	Y	2	81
cll	CREBBP	0,03	11	31	N	Y	1	78
cll	IFT172	0,02	8	39	N	Y	0	1
cll	ARID1A	0,02	6	47	Y	Y	1	179
cll	ANKHD1	0,01	4	52	N	Y	0	6
cll	UBR4	0,03	10	59	N	Y	0	1
cll	KMT2D	0,02	9	63	Y	Y	1	12
cll	ZMYM3	0,01	4	70	Y	N	2	7
cll	ARID2	0,01	5	74	Y	Y	0	19
coadread	APC	0,76	225	2	Y	Y	2953	6015
coadread	TP53	0,62	185	1	Y	Y	466	5236
coadread	KRAS	0,41	122	3	Y	Y	1840	4562
coadread	NEFH	0,34	101	42	N	Y	1	11
coadread	PIK3CA	0,22	67	6	Y	Y	337	1547
coadread	TMPRSS13	0,20	60	5	Y	N	0	0
coadread	KRTAP4-5	0,20	59	4	Y	N	0	0

coadread	SMAD4	0,14	42	10	Y	Y	254	1128
coadread	FBXW7	0,13	39	13	Y	Y	22	176
coadread	ATM	0,12	37	53	Y	Y	108	2197
coadread	BRAF	0,12	37	15	Y	Y	1158	4834
coadread	RNF43	0,11	33	8	Y	Y	18	38
coadread	KRT1	0,11	32	36	N	Y	0	4
coadread	ARID1A	0,10	31	12	Y	Y	11	179
coadread	LURAP1L	0,10	31	7	Y	Y	0	0
coadread	ZNF787	0,08	25	9	Y	N	0	0
coadread	DOCK3	0,08	24	96	Y	Y	0	8
coadread	PTEN	0,08	23	23	Y	Y	361	5794
coadread	ACVR2A	0,07	20	21	Y	Y	2	11
coadread	CTNNB1	0,07	20	57	Y	Y	114	679
coadread	REV3L	0,07	20	68	N	Y	0	7
coadread	MTOR	0,06	18	51	Y	Y	247	6439
coadread	NRAS	0,06	17	18	Y	Y	123	1053
coadread	PLEKHA6	0,06	17	64	N	Y	0	1
coadread	PBRM1	0,05	15	38	Y	Y	0	46
coadread	SMARCA1	0,05	15	39	Y	Y	0	2
coadread	RPL22	0,03	9	43	Y	Y	1	8
gbm	PTEN	0,31	89	1	Y	Y	374	5794
gbm	TP53	0,28	82	2	Y	Y	204	5236
gbm	PIK3R1	0,11	33	3	Y	Y	9	83
gbm	PIK3CA	0,11	32	4	Y	Y	23	1547
gbm	IDH1	0,05	15	5	Y	Y	167	914
gbm	OR5AR1	0,02	7	6	Y	N	0	0
gbm	RB1	0,08	24	7	Y	Y	26	1394
gbm	NF1	0,11	32	8	Y	Y	45	2707
gbm	PRB2	0,02	6	9	Y	N	1	92
gbm	PROKR2	0,02	6	10	Y	N	0	2
gbm	STAG2	0,04	12	11	Y	Y	3	24
gbm	FRG1B	0,06	16	12	Y	N	0	1
gbm	ABCB1	0,03	10	26	N	Y	11	785
gbm	LZTR1	0,03	10	27	N	Y	1	5
gbm	CHD8	0,03	10	31	N	Y	0	10
gbm	MYH8	0,03	8	41	N	Y	0	3
gbm	CARD11	0,02	5	47	Y	N	0	46
gbm	LRP2	0,06	18	54	N	Y	0	2
gbm	CHD9	0,03	10	63	N	Y	0	3
hnsc	CDKN2A	0,22	113	1	Y	Y	161	1963
hnsc	TP53	0,71	360	2	Y	Y	438	5236

hnsc	NSD1	0,12	63	3	Y	Y	2	50
hnsc	PIK3CA	0,18	94	4	Y	Y	176	1547
hnsc	NOTCH1	0,17	88	5	Y	Y	89	1195
hnsc	AJUBA	0,06	32	6	Y	N	3	16
hnsc	FAT1	0,23	117	7	Y	Y	9	30
hnsc	CASP8	0,11	55	8	Y	Y	21	231
hnsc	KMT2D	0,16	82	9	Y	Y	9	12
hnsc	EP300	0,08	40	10	Y	Y	4	66
hnsc	FBXW7	0,06	33	11	Y	Y	8	176
hnsc	EPHA2	0,05	25	12	N	Y	21	329
hnsc	ZNF750	0,04	22	13	Y	N	2	2
hnsc	HRAS	0,06	29	14	Y	N	92	504
hnsc	TGFBR2	0,05	23	15	N	Y	14	184
hnsc	NFE2L2	0,05	26	16	Y	Y	7	54
hnsc	KDM6A	0,03	17	17	Y	Y	3	43
hnsc	HERC1	0,05	27	18	N	Y	0	5
hnsc	LAMA2	0,08	43	19	N	Y	0	9
hnsc	RB1	0,04	18	20	Y	Y	71	1394
hnsc	PTEN	0,03	14	21	Y	Y	376	5794
hnsc	USP9X	0,05	25	22	N	Y	3	30
hnsc	ADGRB3	0,06	30	23	Y	Y	0	0
hnsc	KEAP1	0,04	22	24	Y	Y	11	248
hnsc	RASA1	0,03	17	25	N	Y	0	30
hnsc	CTCF	0,03	17	28	Y	Y	5	163
hnsc	MICALCL	0,03	13	29	Y	Y	0	0
hnsc	TLN1	0,04	22	32	N	Y	4	15
hnsc	BIRC6	0,06	29	33	N	Y	0	16
hnsc	CHD8	0,03	16	35	N	Y	0	10
hnsc	FGFR3	0,03	13	37	N	Y	13	568
hnsc	NUDT11	0,02	8	39	Y	Y	0	4
hnsc	POLDIP2	0,02	8	43	Y	Y	0	1
hnsc	DOCK11	0,03	17	46	N	Y	0	4
hnsc	FNBP4	0,02	12	47	Y	Y	0	0
hnsc	DDX3X	0,02	8	48	Y	Y	3	23
hnsc	SMARCA4	0,05	23	51	Y	Y	0	78
hnsc	PBRM1	0,03	16	52	Y	Y	0	46
hnsc	POU4F2	0,03	13	53	N	Y	0	10
hnsc	FLNC	0,03	14	57	N	Y	1	19
hnsc	MED12	0,04	18	61	N	Y	0	56
hnsc	MYCBP2	0,06	28	65	N	Y	0	2
hnsc	MYH10	0,04	20	69	N	Y	0	5

hnsc	ALB	0,02	12	89	Y	N	14	313
hnsc	NF2	0,02	8	97	Y	N	37	1162
kirc	PBRM1	0,43	91	1	Y	Y	32	46
kirc	VHL	0,56	120	2	Y	Y	984	2040
kirc	MICALCL	0,08	18	3	Y	Y	0	0
kirc	SETD2	0,11	23	4	Y	Y	26	49
kirc	BAP1	0,10	21	5	Y	Y	27	151
kirc	KDM5C	0,07	14	6	Y	N	10	14
kirc	MT-ND1	0,04	8	7	Y	N	0	4
kirc	MT-CYB	0,06	12	8	Y	N	0	3
kirc	MTOR	0,08	18	9	Y	Y	534	6439
kirc	MT-CO1	0,03	7	10	Y	N	0	5
kirc	POLDIP2	0,02	5	11	Y	Y	0	1
kirc	LURAP1L	0,02	5	12	Y	Y	0	0
kirc	MT-ND4	0,02	5	13	Y	N	0	3
kirc	PTEN	0,05	10	14	Y	Y	95	5794
kirc	STAG2	0,03	7	15	Y	Y	0	24
kirc	CNOT1	0,03	6	16	N	Y	0	2
kirc	KRT1	0,02	5	17	N	Y	0	4
kirc	NF2	0,02	5	17	Y	N	4	1162
kirc	NUDT11	0,02	4	20	Y	Y	0	4
kirc	ATM	0,06	13	21	Y	Y	7	2197
kirc	CDC27	0,09	19	22	Y	N	0	19
kirc	ARID1A	0,04	8	24	Y	Y	8	179
kirc	SMARCA4	0,04	8	25	Y	Y	5	78
kirc	TP53	0,02	5	33	Y	Y	59	5236
kirc	SCAF4	0,04	8	34	N	Y	0	0
kirc	MSH3	0,01	3	43	N	Y	1	118
kirc	CHD4	0,04	8	49	Y	Y	0	16
kirc	MYH8	0,02	5	59	N	Y	0	3
kirc	ROBO3	0,01	2	65	N	Y	0	8
kirc	SPTAN1	0,02	4	68	N	Y	0	7
kirc	SMC3	0,02	5	93	Y	Y	0	19
kirc	ADGRV1	0,04	9	98	Y	Y	0	0
kirp	SETD2	0,06	10	1	Y	Y	26	49
kirp	MET	0,07	12	2	Y	N	374	14537
kirp	ANKLE1	0,04	6	3	Y	N	0	2
kirp	NF2	0,05	9	4	Y	N	4	1162
kirp	KDM6A	0,04	6	5	Y	Y	5	43
kirp	PBRM1	0,04	6	7	Y	Y	32	46
kirp	FOXE1	0,02	4	8	Y	N	0	40

kirp	EBLN1	0,02	3	9	Y	N	0	0
kirp	BHMT	0,02	4	10	Y	N	1	26
kirp	BAP1	0,03	5	11	Y	Y	27	151
kirp	ARID1A	0,05	8	13	Y	Y	8	179
kirp	GBF1	0,03	5	17	Y	Y	0	2
kirp	SMARCA4	0,04	7	18	Y	Y	5	78
kirp	NFE2L2	0,02	4	21	Y	Y	1	54
kirp	KRTAP4-5	0,02	3	25	Y	N	0	0
kirp	EP300	0,03	5	26	Y	Y	0	66
kirp	MSH3	0,04	6	28	N	Y	1	118
kirp	UBR4	0,04	7	29	N	Y	0	1
kirp	KRT1	0,01	2	30	N	Y	0	4
kirp	DSCAML1	0,03	5	35	N	Y	0	1
kirp	HRCT1	0,02	3	46	Y	N	0	0
kirp	STAG2	0,04	6	49	Y	Y	0	24
kirp	PLEKHA6	0,02	4	54	N	Y	0	1
kirp	RERE	0,04	6	57	N	Y	0	8
kirp	DNMT3A	0,02	4	75	Y	Y	5	436
laml	DNMT3A	0,26	51	1	Y	Y	132	436
laml	IDH2	0,10	20	2	Y	N	113	387
laml	IDH1	0,10	19	3	Y	Y	149	914
laml	FLT3	0,28	56	4	Y	Y	1234	2038
laml	NRAS	0,08	15	5	Y	Y	98	1053
laml	CEBPA	0,07	13	6	Y	N	254	336
laml	TET2	0,09	17	7	Y	N	81	239
laml	TP53	0,08	15	8	Y	Y	111	5236
laml	SMC3	0,04	7	9	Y	Y	6	19
laml	WT1	0,06	12	10	Y	N	285	1921
laml	BRINP3	0,03	5	13	N	Y	0	1
laml	KRAS	0,04	8	16	Y	Y	44	4562
laml	MT-ND5	0,02	4	16	N	Y	0	0
laml	MT-CYB	0,03	6	20	Y	N	0	3
laml	STAG2	0,03	6	23	Y	Y	4	24
laml	KDM6A	0,02	3	27	Y	Y	3	43
laml	BCOR	0,01	2	72	N	Y	11	42
laml	MT-CO1	0,01	2	87	Y	N	0	5
laml	CBFB	0,01	2	90	Y	N	97	182
lgg	IDH1	0,77	360	1	Y	Y	568	914
lgg	TP53	0,48	226	2	Y	Y	484	5236
lgg	CIC	0,22	103	3	Y	Y	32	355
lgg	ATRX	0,39	181	4	Y	Y	42	102

lgg	FUBP1	0,09	42	5	Y	Y	20	25
lgg	PIK3CA	0,09	40	6	Y	Y	37	1547
lgg	PTEN	0,05	23	7	Y	Y	685	5794
lgg	NOTCH1	0,09	41	8	Y	Y	56	1195
lgg	PIK3R1	0,05	22	9	Y	Y	10	83
lgg	POLDIP2	0,02	9	10	Y	Y	0	1
lgg	NUDT11	0,02	11	11	Y	Y	0	4
lgg	NF1	0,06	30	13	Y	Y	316	2707
lgg	EGFR	0,07	33	14	Y	Y	1512	17418
lgg	POU4F2	0,02	10	15	N	Y	0	10
lgg	IDH2	0,04	19	17	Y	N	153	387
lgg	NIPBL	0,03	16	18	N	Y	0	6
lgg	ARID1A	0,04	21	19	Y	Y	2	179
lgg	MYH2	0,03	15	20	N	Y	0	5
lgg	DNMT3A	0,02	8	21	Y	Y	6	436
lgg	KRT1	0,01	6	22	N	Y	0	4
lgg	SMARCA4	0,04	20	25	Y	Y	8	78
lgg	SETD2	0,02	11	27	Y	Y	2	49
lgg	HEATR5B	0,02	8	29	N	Y	0	0
lgg	CDC27	0,03	15	31	Y	N	1	19
lgg	MYH8	0,03	14	37	N	Y	0	3
lgg	ARID2	0,02	11	38	Y	Y	0	19
lgg	RB1	0,01	6	39	Y	Y	57	1394
lgg	HCFC1	0,01	7	50	N	Y	1	1
lgg	CAD	0,01	4	51	N	Y	12	1556
lgg	MYH1	0,02	8	52	N	Y	0	4
lgg	ROBO3	0,01	7	69	N	Y	0	8
lgg	MTOR	0,01	5	70	Y	Y	376	6439
lgg	PBRM1	0,01	6	89	Y	Y	1	46
lgg	ZNF292	0,03	13	96	Y	N	0	4
lihc	TP53	0,31	62	1	Y	Y	182	5236
lihc	CTNNB1	0,26	51	2	Y	Y	102	679
lihc	ALB	0,09	18	6	Y	N	141	313
lihc	ADGRV1	0,08	16	78	Y	Y	0	0
lihc	ARID1A	0,08	16	3	Y	Y	11	179
lihc	RB1	0,08	15	3	Y	Y	43	1394
lihc	BAP1	0,06	11	5	Y	Y	10	151
lihc	GCN1L1	0,06	11	8	Y	N	0	1
lihc	KIF19	0,05	10	10	Y	N	0	0
lihc	SMARCA4	0,05	9	21	Y	Y	1	78
lihc	CARD11	0,04	7	9	Y	N	0	46

lihc	KEAP1	0,04	7	11	Y	Y	28	248
lihc	PTEN	0,04	7	15	Y	Y	236	5794
lihc	ANO4	0,03	6	57	N	Y	0	0
lihc	ACVR2A	0,03	5	77	Y	Y	1	11
lihc	CHD4	0,03	5	84	Y	Y	1	16
lihc	CTCF	0,03	5	50	Y	Y	5	163
lihc	CDKN1A	0,02	4	7	Y	N	23	423
lihc	NFE2L2	0,02	4	63	Y	Y	9	54
lihc	PBRM1	0,02	4	33	Y	Y	2	46
lihc	IDH1	0,02	3	75	Y	Y	8	914
luad	TP53	0,54	295	1	Y	Y	367	5236
luad	KRAS	0,30	164	2	Y	Y	1067	4562
luad	KEAP1	0,17	95	3	Y	Y	43	248
luad	STK11	0,15	80	4	Y	Y	36	300
luad	EGFR	0,14	74	5	Y	Y	5215	17418
luad	RBM10	0,06	35	6	Y	Y	4	10
luad	COL5A2	0,10	55	7	Y	Y	0	11
luad	SMARCA4	0,08	43	8	Y	Y	17	78
luad	COL11A1	0,21	116	9	Y	Y	4	29
luad	BRAF	0,08	41	10	Y	Y	259	4834
luad	MGA	0,08	42	11	N	Y	3	81
luad	NF1	0,11	62	12	Y	Y	22	2707
luad	RB1	0,06	31	13	Y	Y	82	1394
luad	ATM	0,08	46	14	Y	Y	122	2197
luad	COL12A1	0,10	56	15	N	Y	0	13
luad	SETD2	0,06	33	16	Y	Y	3	49
luad	ARID1A	0,07	36	17	Y	Y	7	179
luad	NAV3	0,21	114	19	N	Y	1	19
luad	MYH7	0,10	56	20	N	Y	0	2
luad	MYH8	0,12	65	22	N	Y	0	3
luad	CDKN2A	0,04	22	23	Y	Y	105	1963
luad	RGS7	0,06	31	27	N	Y	1	4
luad	BIRC6	0,08	46	29	N	Y	1	16
luad	COL3A1	0,11	58	30	Y	Y	4	23
luad	POLDIP2	0,02	13	31	Y	Y	0	1
luad	MICALCL	0,03	17	34	Y	Y	0	0
luad	MT-CO1	0,03	14	37	Y	N	0	5
luad	NALCN	0,15	83	39	N	Y	1	2
luad	MT-ND1	0,02	13	41	Y	N	0	4
luad	MT-CYB	0,03	15	42	Y	N	0	3
luad	ADAMTS19	0,06	33	44	N	Y	0	0

luad	MT-ND4	0,03	16	46	Y	N	0	3
luad	CDH10	0,17	90	47	N	Y	0	2
luad	SLC4A3	0,04	22	48	N	Y	0	1
luad	MYH1	0,11	62	53	N	Y	0	4
luad	CTNNB1	0,03	18	54	Y	Y	21	679
luad	PTPRU	0,03	19	55	N	Y	1	4
luad	SMARCA2	0,03	19	56	N	Y	3	22
luad	KCNH8	0,08	42	59	N	Y	1	2
luad	HEATR5B	0,04	20	60	N	Y	0	0
luad	DNMT3A	0,04	23	61	Y	Y	20	436
luad	MT-ND5	0,03	15	65	N	Y	0	0
luad	PIK3CA	0,05	29	67	Y	Y	166	1547
luad	SORCS3	0,09	51	70	N	Y	0	1
luad	CAD	0,05	25	73	N	Y	288	1556
luad	SRRM2	0,05	29	74	N	Y	0	0
luad	ADGRV1	0,11	58	90	Y	Y	0	0
lusc	TP53	0,79	141	1	Y	Y	367	5236
lusc	CDKN2A	0,15	26	2	Y	Y	105	1963
lusc	NFE2L2	0,15	27	3	Y	Y	15	54
lusc	KEAP1	0,12	22	4	Y	Y	43	248
lusc	PTEN	0,08	14	5	Y	Y	313	5794
lusc	RB1	0,07	12	6	Y	Y	82	1394
lusc	TDRD5	0,08	14	7	Y	N	0	1
lusc	PIK3CA	0,15	27	8	Y	Y	166	1547
lusc	COL3A1	0,07	13	9	Y	Y	4	23
lusc	ADGRB3	0,12	22	10	Y	Y	0	0
lusc	BRINP3	0,15	27	11	N	Y	0	1
lusc	MYH2	0,16	28	12	N	Y	4	5
lusc	KDM6A	0,04	7	13	Y	Y	1	43
lusc	COL11A2	0,07	13	14	N	Y	0	7
lusc	ABCB1	0,08	14	18	N	Y	54	785
lusc	MYH8	0,12	21	19	N	Y	0	3
lusc	KMT2D	0,20	36	20	Y	Y	3	12
lusc	NOTCH1	0,08	14	21	Y	Y	48	1195
lusc	RASA1	0,04	8	22	N	Y	1	30
lusc	MYH1	0,12	21	23	N	Y	0	4
lusc	SCN11A	0,06	11	24	N	Y	0	1
lusc	FBXW7	0,06	11	27	Y	Y	7	176
lusc	NRAP	0,07	12	28	N	Y	0	2
lusc	COL11A1	0,20	35	30	Y	Y	4	29
lusc	DNAH8	0,17	31	40	N	Y	0	3

lusc	SETD2	0,03	5	47	Y	Y	3	49
lusc	ADGRV1	0,17	31	48	Y	Y	0	0
lusc	ANKHD1-EIF4EBP3	0,05	9	49	N	Y	0	1
lusc	MYH9	0,06	10	51	N	Y	2	27
lusc	CNKSR2	0,07	12	52	N	Y	0	0
lusc	COL5A2	0,08	14	53	Y	Y	0	11
lusc	PBRM1	0,04	7	67	Y	Y	1	46
lusc	USP9X	0,04	8	71	N	Y	4	30
ov	TP53	0,95	299	1	Y	Y	315	5236
ov	BRCA1	0,03	11	2	Y	N	2496	7290
ov	ADGRB3	0,03	9	3	Y	Y	0	0
ov	CDK12	0,03	9	4	Y	Y	6	13
ov	CCAR1	0,02	6	5	Y	N	0	8
ov	RB1	0,02	6	5	Y	Y	31	1394
ov	PCDHB13	0,02	5	7	Y	N	0	0
ov	PSMC3	0,01	4	8	Y	N	0	4
ov	NF1	0,04	12	9	Y	Y	14	2707
ov	BRCA2	0,03	10	9	Y	N	1520	4282
ov	SETD2	0,02	6	11	Y	Y	0	49
ov	LRP2	0,04	14	17	N	Y	0	2
ov	MADD	0,02	6	20	N	Y	4	22
ov	MTOR	0,02	6	29	Y	Y	165	6439
ov	DSCAML1	0,02	6	33	N	Y	0	1
ov	DOCK11	0,01	4	35	N	Y	0	4
ov	CHD4	0,03	8	36	Y	Y	2	16
ov	BRINP3	0,01	3	47	N	Y	0	1
ov	SPTAN1	0,02	5	50	N	Y	1	7
ov	CREBBP	0,02	7	69	N	Y	2	78
ov	KDM5C	0,02	6	78	Y	N	0	14
prad	TP53	0,10	29	1	Y	Y	89	5236
prad	SPOP	0,10	29	2	Y	N	21	33
prad	PTEN	0,04	13	3	Y	Y	709	5794
prad	ZMYM3	0,03	9	4	Y	N	0	7
prad	FOXA1	0,04	13	5	Y	N	55	201
prad	FNBP4	0,02	7	6	Y	Y	0	0
prad	NUDT11	0,02	6	7	Y	Y	3	4
prad	MAP3K9	0,03	8	8	Y	N	0	7
prad	MT-ND4	0,02	5	9	Y	N	2	3
prad	SMARCA1	0,02	6	10	Y	Y	0	2
prad	KDM6A	0,02	6	11	Y	Y	3	43
prad	POU4F2	0,02	5	12	N	Y	0	10

prad	MED12	0,02	6	17	N	Y	4	56
prad	ARID1A	0,02	5	17	Y	Y	1	179
prad	POLDIP2	0,01	4	19	Y	Y	0	1
prad	LURAP1L	0,01	3	21	Y	Y	0	0
prad	ATM	0,04	13	23	Y	Y	81	2197
prad	CHD3	0,02	5	24	N	Y	0	6
prad	CTNNB1	0,03	9	27	Y	Y	5	679
prad	CDK12	0,02	7	32	Y	Y	0	13
prad	RTL1	0,01	4	37	N	Y	0	6
prad	ANO4	0,02	6	50	N	Y	0	0
prad	MT-CO1	0,01	3	57	Y	N	0	5
prad	MT-ND1	0,01	2	71	Y	N	0	4
prad	NCOR1	0,02	7	72	N	Y	11	47
prad	FRG1B	0,17	50	73	Y	N	0	1
prad	ADGRB3	0,02	7	78	Y	Y	0	0
prad	KMT2D	0,04	13	84	Y	Y	3	12
prad	SMAD4	0,01	4	84	Y	Y	45	1128
skcm	BRAF	0,51	189	1	Y	Y	781	4834
skcm	NRAS	0,27	98	2	Y	Y	225	1053
skcm	TP53	0,15	56	3	Y	Y	138	5236
skcm	COL3A1	0,19	71	4	Y	Y	0	23
skcm	CDKN2A	0,12	46	5	Y	Y	339	1963
skcm	COL5A2	0,14	51	6	Y	Y	0	11
skcm	KCNQ5	0,13	48	7	Y	Y	0	1
skcm	PTEN	0,09	35	8	Y	Y	148	5794
skcm	ADGRV1	0,34	127	9	Y	Y	0	0
skcm	ARID2	0,14	50	10	Y	Y	0	19
skcm	BRINP3	0,16	60	11	N	Y	0	1
skcm	COL2A1	0,10	37	13	N	Y	0	31
skcm	COL1A1	0,12	46	18	N	Y	76	181
skcm	SNAP91	0,10	37	19	N	Y	0	0
skcm	RGS7	0,12	44	20	N	Y	0	4
skcm	COL4A5	0,17	63	23	N	Y	0	36
skcm	DPYD	0,17	64	26	N	Y	0	133
skcm	RFX6	0,11	41	28	N	Y	0	7
skcm	ADGRB3	0,17	61	29	Y	Y	0	0
skcm	COL5A3	0,15	57	31	N	Y	0	0
skcm	DDX3X	0,06	22	33	Y	Y	0	23
skcm	NF1	0,13	47	34	Y	Y	91	2707
skcm	NDST4	0,12	46	36	N	Y	0	2
skcm	CTNNB1	0,05	19	40	Y	Y	25	679

skcm	KIAA1109	0,15	54	42	N	Y	0	1
skcm	ANO4	0,15	57	44	N	Y	0	0
skcm	COL11A1	0,19	69	46	Y	Y	0	29
skcm	ANK3	0,32	118	55	N	Y	0	5
skcm	IDH1	0,05	20	56	Y	Y	3	914
skcm	AMPD1	0,08	31	59	N	Y	0	4
skcm	PPFIA2	0,12	46	60	N	Y	0	0
skcm	GRIK1	0,08	30	61	N	Y	0	2
skcm	MYH7	0,14	51	64	N	Y	0	2
skcm	GRID2	0,15	55	70	N	Y	0	4
skcm	LAMA2	0,15	57	72	N	Y	0	9
skcm	ABCB11	0,10	36	74	N	Y	0	18
skcm	FLT3	0,09	34	83	Y	Y	8	2038
skcm	POLR3B	0,05	20	88	Y	Y	0	0
stad	TP53	0,48	180	1	Y	Y	121	5236
stad	ARID1A	0,28	105	2	Y	Y	22	179
stad	KMT2D	0,19	69	85	Y	Y	2	12
stad	PIK3CA	0,17	63	10	Y	Y	49	1547
stad	DOCK3	0,14	51	6	Y	Y	0	8
stad	APC	0,13	49	40	Y	Y	234	6015
stad	UBR5	0,13	48	28	N	Y	1	7
stad	RNF43	0,12	45	5	Y	Y	1	38
stad	ACVR2A	0,12	44	3	Y	Y	0	11
stad	RPL22	0,12	43	4	Y	Y	0	8
stad	TTK	0,11	42	9	Y	N	1	63
stad	CHD4	0,10	36	34	Y	Y	1	16
stad	KRAS	0,10	36	7	Y	Y	72	4562
stad	PHF2	0,09	35	8	Y	Y	0	5
stad	ARID2	0,09	33	38	Y	Y	0	19
stad	FBXW7	0,09	33	12	Y	Y	5	176
stad	CIC	0,09	32	49	Y	Y	9	355
stad	CEP290	0,08	31	30	N	Y	0	1
stad	SCAF4	0,08	31	35	N	Y	0	0
stad	SMAD4	0,08	31	17	Y	Y	56	1128
stad	JARID2	0,08	30	44	N	Y	0	10
stad	PBRM1	0,07	27	21	Y	Y	0	46
stad	SETDB1	0,07	27	41	N	Y	0	22
stad	CTCF	0,06	24	36	Y	Y	1	163
stad	PLEKHA6	0,06	23	20	N	Y	0	1
stad	MSH3	0,06	22	54	N	Y	4	118
stad	MT-CO1	0,05	20	80	Y	N	0	5

stad	RB1	0,05	18	88	Y	Y	10	1394
stad	CDKN2A	0,05	17	62	Y	Y	41	1963
stad	POLDIP2	0,03	11	67	Y	Y	0	1
thca	BRAF	0,60	242	1	Y	Y	921	4834
thca	NRAS	0,08	34	2	Y	Y	74	1053
thca	HRAS	0,03	14	3	Y	N	59	504
thca	EIF1AX	0,01	6	4	Y	N	1	8
thca	PPM1D	0,01	5	5	Y	N	1	66
thca	MSI1	0,01	3	6	Y	N	0	31
thca	KRAS	0,01	4	7	Y	Y	74	4562
thca	CHEK2	0,01	5	8	Y	N	5	362
thca	GBF1	0,01	4	9	Y	Y	0	2
thca	SLA	0,01	3	10	Y	N	1	55
thca	USP9X	0,01	4	13	N	Y	0	30
thca	PTEN	0,00	2	13	Y	Y	159	5794
thca	ZSWIM8	0,00	2	23	N	Y	0	0
thca	FLNC	0,01	3	31	N	Y	0	19
thca	ATM	0,01	5	36	Y	Y	18	2197
thca	ARID1B	0,01	5	68	N	Y	1	18
thca	JMJD1C	0,01	4	69	N	Y	0	5
thca	RB1	0,00	2	78	Y	Y	16	1394
thca	ADGRV1	0,01	5	91	Y	Y	0	0
thca	TP53	0,01	3	94	Y	Y	41	5236
thca	KMT2D	0,00	2	99	Y	Y	0	12
ucec	PTEN	0,65	161	1	Y	Y	499	5794
ucec	TP53	0,28	69	2	Y	Y	164	5236
ucec	PIK3CA	0,53	132	3	Y	Y	117	1547
ucec	CTNNB1	0,30	74	4	Y	Y	57	679
ucec	ARID1A	0,33	83	5	Y	Y	38	179
ucec	CTCF	0,18	45	6	Y	Y	6	163
ucec	KRAS	0,21	53	7	Y	Y	97	4562
ucec	CHD4	0,14	35	8	Y	Y	2	16
ucec	FBXW7	0,16	39	9	Y	Y	10	176
ucec	PIK3R1	0,33	83	10	Y	Y	8	83
ucec	FGFR2	0,13	31	11	N	Y	29	465
ucec	TAF1	0,14	35	12	N	Y	2	13
ucec	SPOP	0,08	21	14	Y	N	2	33
ucec	HEATR5B	0,06	16	15	N	Y	0	0
ucec	MTOR	0,10	26	17	Y	Y	176	6439
ucec	ANKHD1-EIF4EBP3	0,08	19	19	N	Y	0	1
ucec	RBM27	0,06	16	20	N	Y	0	0

ucec	AMPD1	0,06	14	24	N	Y	0	4
ucec	MAP3K1	0,08	21	30	Y	Y	1	69
ucec	ERBB3	0,07	17	33	N	Y	10	584
ucec	NFE2L2	0,06	15	36	Y	Y	1	54
ucec	NRAP	0,08	19	40	N	Y	0	2
ucec	CNOT1	0,09	22	47	N	Y	0	2
ucec	ADGRL2	0,07	17	55	N	Y	0	0
ucec	MYH1	0,09	23	58	N	Y	0	4
ucec	POLR3B	0,05	13	66	Y	Y	0	0
ucec	PTPRK	0,08	19	68	N	Y	0	20
ucec	SETD2	0,09	22	81	Y	Y	0	49
ucec	SMC3	0,04	9	85	Y	Y	0	19
ucec	RB1	0,08	20	92	Y	Y	21	1394
ucec	CHEK2	0,05	13	97	Y	N	6	362

Table 10.5: **Full list of TTDG connections.**

Tumor	Top200	Top10-Top100	Total Drivers in top 100	Total Affected	Total Patients	Percent altered	Specific genes
blca	2	1	26	232	238	97%	ELF3
brca	5	3	22	834	993	84%	CDH1,GATA3,MAP2K4
cesc	5	1	20	163	194	84%	HLA-A
cll	5	3	19	186	385	48%	CHD2,MYD88,KLHL6
coadread	4	2	23	294	301	98%	TMPRSS13,ZNF787
gbm	5	3	13	253	290	87%	OR5AR1,PRB2,PROKR2
hnsc	6	3	29	491	509	96%	FAT1,NSD1,AJUBA
kirc	1	1	25	199	235	85%	VHL
kirp	6	4	19	132	171	77%	MET,ANKLE1,BHMT,EBLN1
laml	3	3	16	152	197	77%	TET2,CEBPA,WT1
lgg	2	1	24	456	468	97%	ATRX
lihc	2	2	20	177	198	89%	GCN1L1,KIF19
luad	10	1	27	526	561	94%	STK11
lusc	6	1	19	176	178	99%	TDRD5
ov	3	2	14	313	316	99%	PDCHB13,PSMC3
prad	2	1	23	211	300	70%	MAP3K9
skcm	14	1	18	356	369	96%	KCNQ5
stad	4	2	23	358	372	96%	TTK,PHF2
thca	6	4	16	328	429	76%	EIF1AX,PPM1D,MSI1,SLA
ucec	4	0	19	245	248	99%	-

Table 10.6: **Tumor-specific high confident drivers observed across tumor types using cDriver ranks.**

Gene/Tumor	BLCA	BRCA	CESC	CRC	GBM	HNSC	KIRC	KIRP	LGG	LIHC	LUAD	LUSC	OV	PRAD	SKCM	STAD	THCA	UCEC
ARID1A	26%	3%	8%	10%		4%	3%	5%	4%	8%	6%	7%		**2%**	5%	28%		33%
ARID1B	5%	2%		5%		4%	3%		2%	5%	5%	4%			7%	8%	**1%**	4%
ARID2	8%	1%	3%	5%		4%		2%	**2%**	3%	5%	5%	2%	2%	**14%**	**9%**		6%
BCOR	3%	1%	3%	5%	3%	2%			3%		3%	4%			4%	7%		12%
CHD2	8%	1%		4%		3%	2%		1%	2%	2%	2%		2%	5%	5%		6%
CHD3	5%	2%	**5%**	5%		3%					4%			**2%**	4%	8%		12%
CHD4	**8%**	2%	5%	4%		3%	**3%**		1%	**3%**	3%	3%	3%	1%	7%	**10%**		14%
CHD8	3%	2%	4%	4%	**3%**	**3%**					5%	4%	2%		7%	6%		8%
CREBBP	13%	1%	**7%**	6%	2%	7%		2%	1%	3%	5%	8%	2%	1%	7%	11%		9%
CTCF	3%	2%	2%	3%		3%				3%	1%	0%			0%	**6%**		18%
EP300	14%	1%	11%	5%		8%	3%	**3%**			2%	4%		1%	5%	6%		8%
JMJD1C	6%	1%	2%	3%		4%		2%			4%	4%	2%		5%	8%	**1%**	7%
KDM6A	24%	2%	3%	2%		3%		4%			2%	**4%**		2%	1%	3%		4%
KMT2C	20%	7%	15%	8%	3%	9%	6%	8%	3%	7%	16%	17%	2%	5%	14%	14%	1%	10%
KMT2D	29%	2%	11%	12%	2%	16%	2%	8%	2%	6%	8%	20%		4%	14%	19%		13%
MSH3	3%	1%		3%		2%	**1%**	**4%**			2%	2%			3%	6%		3%
NIPBL	5%	2%	5%	6%		5%			**3%**	2%	5%	4%	2%	1%	5%	10%		9%
PBRM1	**7%**	1%		**5%**		**3%**	39%	4%	**1%**	2%	2%	**4%**			4%	**7%**		4%
PHF2				17%		1%					2%	2%			3%	**9%**		4%
RERE	5%	1%		3%	2%	1%		4%	1%		2%	0%			7%	8%		6%
SETD2	**5%**	2%	3%	4%	2%	2%	10%	6%	2%	4%	6%	3%	**2%**	1%	5%	4%		**9%**
SETDB1	2%	2%	3%	4%		2%			1%		2%				2%	**7%**		6%
SMARCA2	**7%**	1%		4%		3%					3%	2%			4%	7%		7%
SMARCA4	6%	1%	3%	5%		**5%**	3%	4%	4%	**5%**	8%	4%	1%		7%	7%		6%
SMC3	2%	1%		2%	1%	1%	**2%**				3%	3%			2%	4%		**4%**

Table 10.8: **Fraction of affected patients that have a mutated chromatin modifier.** Bold numbers are novel tumor type - driver gene connections. Underlined numbers refer to a significant prognosis impact in that tumor type ($QValue < 0.2$).

Part IV

Purifying selection reveals cancer essential functions

Adapted from manuscript under preparation:

Purifying selection reveals cancer essential functions. Zapata L., Schaefer M., Pich O., Vlasov P., Serrano L., Kondrashov F., Ossowski S.

Cancer is an evolutionary process. Tumors are composed of rapidly proliferating cells that become malignant under selection of biological functions needed for cancer development. Here, we explore how purifying selection shapes the mutational landscape of cancer genomes. Previous work has focused on the adaptive genetic landscape imparted by cancer driver mutations, while neglecting the importance of essential cellular functions constrained by the new metabolic requirements of the malignant cell. We analyzed the exome of 7,546 individuals, spanning 26 tumor types, and identified 645 genes under strong purifying selection. The main enriched function among these genes is membrane plasma protein regulating ion transport, a putative non-cell-autonomous mechanisms of cancer progression. In summary, we demonstrate that purifying selection is a major force shaping the evolution of cancer genomes, and that by identifying essential genes and functions we expand the number of targets for cancer treatment.

Chapter 11

Introduction

Similarly to unicellular microorganisms, cancer cells are subjected to selective pressures where adaptive genotypes outcompete less fit genotypes, therefore removing deleterious alleles from the population [Nowell, 1976]. Since early 1970s, several studies have explored this evolutionary model of tumor development to determine which genes are relevant for malignant transformation and tumor progression. Somatic mutations conferring a selective advantage are most likely affecting a cancer hallmark [Hanahan and Weinberg, 2011]. Accordingly, described cancer hallmarks such as increased proliferative capacity, suppression of cell cycle control, and escape from immune surveillance have been the focus of such studies, thereby enabling the identification of a large number of cancer driver genes.

Recent cancer studies have identified driver genes by detecting signals of positive selection [Tamborero et al., 2013b]. The International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas initiatives (TCGA) have made available thousands of cancer genomes [TCGANetwork, 2008]. Consequently, several methods have exploited sequencing data to reveal an extensive landscape of cancer genes across more than 30 tumor types. Our group has previously identified driver genes affecting tumor development from different anatomical sites by integrating measures of positive selection at multiple levels (See part III). We proposed an extensive landscape of tumor type driver gene connections by including known driver genes that are neglected in tumors where the proportion of affected patients is low. Our strategy has expanded the set of therapeutic targets for multiple patients without a causative explanation for the cancer.

Although the idea that purifying selection acts on cancer genomes is not new [Stratton et al., 2009], most studies focused on identifying the neutral accumulation of somatic mutations [Williams et al., 2016] or demonstrating the presence of pervasive positive selection [Ostrow et al., 2014]. While some studies failed to identify strong negative selection [Ovens and Naugler, 2012] or rejected its presence under mutator phenotypes [Beckman and Loeb, 2005], others have recently revealed its action on transcription factor binding motifs [Vorontsov et al., 2015] and membrane proteins in melanoma [Pyatnitskiy et al., 2015], sparking

an interesting controversy in the field. To our knowledge, this is the first study that addresses the extent and characteristics of purifying selection across several tumor types. In addition, we looked for enriched pathways, and demonstrate the prognostic relevance of said functions.

We adapted a previously described statistical method to detect genes or functions under significant selection [Greenman et al., 2006]. We identify 10-fold more genes under negative than positive selection, suggesting that a substantial portion of the cancer genome is under functional constraints. Interestingly, we also identified two known cancer driver genes, *ABL1* and *TERT*, not related to somatic point mutations at the coding level, and highly conserved across 25 tumor types. In addition, we observe that extracellular signaling processes are under strong purifying selection suggesting that non-cell-autonomous selection is a driving force in cancer progression. We expand the analysis to functional terms in order to assess the effect of selection at the phenotypic level. Finally, we describe two functional examples, the P2X7 signaling complex and the GO term translational elongation, that have an impact on the prognosis of cancer patients.

Chapter 12

Results

12.1 Purifying selection predominates over positive selection in cancer genomes

To characterize genes under purifying selection, we analyzed 7,546 individual samples across 26 tumor types (Table 12.1) using the ratio of non-synonymous substitutions per non-synonymous sites to synonymous substitutions per synonymous sites, K_n/K_s. In addition, to account for the impact of other types of protein-affecting mutations, such as nonsense and splicing substitutions, we calculated a similar ratio for each of these mutation types (K_{non}/K_s and K_{spl}/K_s) and for all mutation types combined (K_a/K_s, Methods). We observed that these estimates are significantly correlated to each other (Table 12.2), thus allowing us to use the K_a/K_s as a surrogate for all three estimates. Although we did not observe large differences by using K_a/K_s, some chromatin-modifying proteins disappeared when only using K_n/K_s (not shown), suggesting an enrichment towards nonsense and/or splicing mutations for this functional group.

The mutation rate across tumor types varied considerably [Lawrence et al., 2013]. Skin carcinoma (SKCM), for example, is well known for its extreme mutation rate, which in principle should increase the power for detecting selection. We observed that the number of significant genes increased 3-fold when using a pooled set of 26 tumor types (Pancan26) compared to a pooled set of all tumors except skin carcinoma (Pancan25, Table 12.3). Moreover, the number of significant genes using only SKCM data was more than 1,600, whereas the number of significant genes for other individual tumors ranged from 0 to 37 (Table 12.4). We observed that 144 genes are shared by all three datasets, more than 500 between the pooled datasets, and 625 (30%) are specifically shared by Pancan26 and SKCM (Fig. 12.1). To reduce the anticipated number of false positives, we performed subsequent analysis using the Pancan25 dataset, which has the smallest number of significant genes, while analyzing SKCM as single tumor.

We identified 685 genes under significant selection bias using all mutation types,

Tumor Type	Tumor Type Abbreviation	Num Samples	Num Mutations	Filtered Num of Mutations
Adrenocortical carcinoma	ACC	91	13130	8145
Bladder Urothelial Carcinoma	BLCA	395	133351	93726
Breast invasive carcinoma	BRCA	976	86769	70037
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	194	45894	35521
Chronic Lymphocityc Leukemia	CLL	385	18791	8396
Colon adenocarcinoma and Rectum adenocarcinoma	COADREAD	298	139583	103879
Esophageal carcinoma	ESCA	183	40765	25571
Head and Neck squamous cell carcinoma	HNSC	508	109877	73159
Kidney renal clear cell carcinoma	KIRC	213	21716	12408
Kidney renal papillary cell carcinoma	KIRP	161	15585	9899
Brain Lower Grade Glioma	LGG	513	45096	25408
Liver hepatocellular carcinoma	LIHC	198	27891	19352
Lung adenocarcinoma	LUAD	542	216292	131982
Lung squamous cell carcinoma	LUSC	178	65306	44375
Pheochromocytoma and Paraganglioma	PCPG	178	4432	1415
Prostate adenocarcinoma	PRAD	425	24911	11038
Sarcoma	SARC	255	31906	16475
Skin Cutaneous Melanoma	SKCM	367	298328	220898
Stomach adenocarcinoma	STAD	428	238694	152275
Testicular Germ Cell Tumors	TGCT	149	14127	2456
Thyroid carcinoma	THCA	402	7061	3209
Thymoma	THYM	123	6673	1867
Uterine Carcinosarcoma	UCEC	248	184862	156187
Uterine Corpus Endometrial Carcinoma	UCS	56	5886	3798
Uveal Melanoma	UVM	80	2529	1351

Table 12.1: Data used for analysis of purifying selection.

Measure	K_n/K_s	K_{non}/K_s	K_{spl}/K_s	K_a/K_s
K_n/K_s	1	0.41	0.27	0.99
K_{non}/K_s		1	0.53	0.49
K_{spl}/K_s			1	0.33

Table 12.2: Correlation between different selection estimates.

Dataset	Number of significant	Positive selection	Negative Selection
PC26	1951	60	1891
PC25	685	42	643
SKCM	1645	6	1639

Table 12.3: Number of significant genes obtained in three datasets.

K_a/K_s (Table 15.1). Four genes, including *TP53*, had a silent mutation cluster, indicating that silent sites could be under selection and confound the interpretation of negative selection. Hence, these genes have been removed from negative selection analysis. 42 out of the 681 genes were under positive selection (PS) and included driver genes such as *TP53, IDH1, PIK3CA, BRAF*, among others. The remaining 639 genes (90% of the total set), displayed a purifying selection signal (NS) suggesting that natural selection is removing deleterious alleles rather than promoting novel functions in cancer. 32 out of the 42 PS genes are known cancer genes, which demonstrates that our approach has higher precision (75%) than previously reported using a similar measure [Ostrow et al., 2014]. On the other hand, 35 out of the 639 NS genes are known cancer genes. Three of them, *ALK, EPHB2, and MYH11*, are annotated as cancer drivers due to activating point

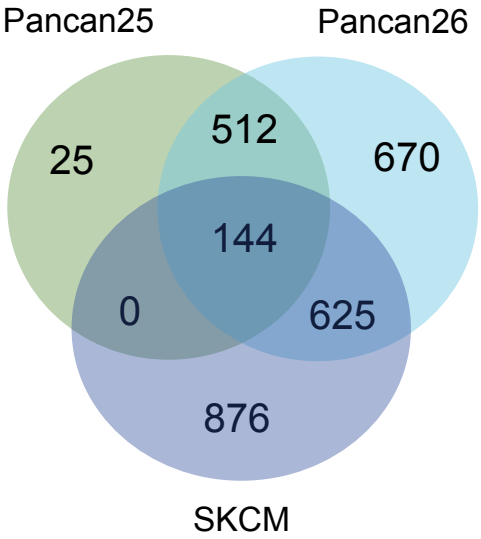

Figure 12.1: **Venn diagram for significant genes of three datasets.** Venn diagram for the significant selected genes of all tumor types (Pancan26), all tumor types except skin carcinoma (Pancan25), and only skin carcinoma (SKCM).

Tumor Type	Tumor Abbreviation	$K_a = (K_n + K_{non} + K_{spl})$	$K_n + K_{non}$	K_n
Adrenocortical carcinoma	ACC	0	0	0
Bladder Urothelial Carcinoma	BLCA	11	10	7
Breast invasive carcinoma	BRCA	1	1	2
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	0	0	0
Chronic Lymphocityc Leukemia	CLL	1	1	0
Colon adenocarcinoma and Rectum adenocarcinoma	COADREAD	36	29	16
Esophageal carcinoma	ESCA	2	2	1
Head and Neck squamous cell carcinoma	HNSC	4	3	2
Kidney renal clear cell carcinoma	KIRC	0	0	0
Kidney renal papillary cell carcinoma	KIRP	0	0	0
Brain Lower Grade Glioma	LGG	2	2	2
Liver hepatocellular carcinoma	LIHC	0	0	0
Lung adenocarcinoma	LUAD	7	6	2
Lung squamous cell carcinoma	LUSC	1	1	1
Pheochromocytoma and Paraganglioma	PCPG	0	0	0
Prostate adenocarcinoma	PRAD	0	0	0
Sarcoma	SARC	1	1	1
Skin Cutaneous Melanoma	SKCM	1645	1552	1419
Stomach adenocarcinoma	STAD	27	20	19
Testicular Germ Cell Tumors	TGCT	0	0	0
Thyroid carcinoma	THCA	1	1	1
Thymoma	THYM	0	0	0
Uterine Carcinosarcoma	UCEC	37	32	30
Uterine Corpus Endometrial Carcinoma	UCS	0	0	0
Uveal Melanoma	UVM	0	0	0

Table 12.4: Number of significant genes obtained in each tumor type based on three different mutation types.

mutations (Intogen) and 21 of them are linked to cancer due to translocation/other events in CGC (Table 12.5). Five of these 35 NS known cancer genes, *ABL1, CARS, IGF1R, KCNJ5, PARP1,* and *PTRPRB*, as well as 91 of the total 639 NS

genes, were associated to decreased viability of cancer cell lines [Cheung et al., 2011] using genome-wide screening of shRNA. In another study, nine out of the 35 NS known cancer genes (P value < 0.05), *ABL1, CARS, DNMT1, ERCC3, FLI1, IKZF1, IGF1R, TSC2,* and *TYMS* were cataloged as essential for proliferation and survival using the CRISPR genome editing technology in cancer cell lines. [Wang et al., 2015]. Similar to activating mutations, *TERT* has been implicated in cancer development not because of an accumulation of loss or gain of function somatic point mutations, but because of mutations in its promoter, thereby effectively increasing gene expression [Horn et al., 2013]. Thus, we hypothesize that oncogenes that contribute to cancer by increasing copy number and/or expression show signs of negative selection at the protein level.

To validate global properties of genes under selection, we compared damage scores of mutations in genes under negative, neutral, and positive selection. We observed that genes displaying a low K_a/K_s (i.e. under strong negative selection) harbored less damaging mutations than neutral and positively selected genes (Fig. 12.2). Moreover, a higher threshold of K_a/K_s increased the mean damage score, revealing a dependence between selection strength and functional impact. Our results confirmed the importance of the functional impact of point mutations as a signature of positive and negative selection, which are instrumental for the detection of cancer driver and essential genes, respectively.

Next, we explored the average ploidy and expression levels of negative versus positively selected genes (Fig. 12.3). Genes under positive selection displayed (a) a lower copy number status and (b) an increased mean expression value, when compared to negatively and neutral genes which showed no differences. In summary, we show that positively selected genes harbor highly damaging mutations when compared to random genes, as previously described [Gonzalez-Perez and Lopez-Bigas, 2012]. In addition, PS are highly expressed (oncogenes) and are more prone to harbor deletions (Tumor suppressors). We confirmed that negatively selected genes harbor mostly less damaging somatic mutations, which are less likely to affect protein function.

12.2 Functional role of genes under purifying selection

To address which biological functions and processes are under strong negative selection, we performed a gene functional enrichment analysis using STRING and

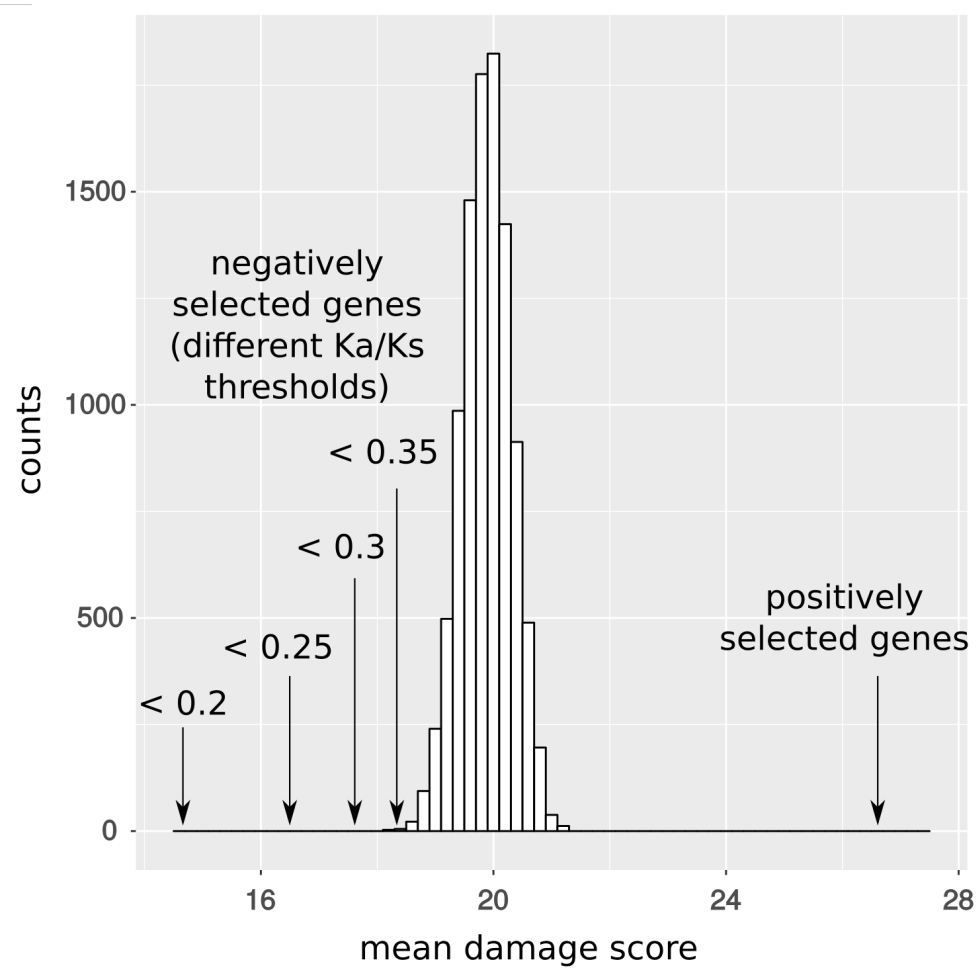

Figure 12.2: **Damage score mean at different K_a/K_s values.** Distribution of functional impact of mutations in the cohort. Genes under negative selection harbor less damaging mutations than positively selected genes.

Gene Name	Intogen Info	Cancer Gene Census
ABL1	No	oncogene,T
ALK	Yes,Activating	oncogene,T,Mis,A
ATP2B3	No	O
BCL11B	No	T
CAMTA1	No	T
CANT1	No	T
CARS	No	T
CDH11	No	T
CYP3A5	No	-
DNMT1	No	-
EPHB2	Yes,Activating	-
EPPK1	No	-
ERCC3	No	Mis,S
ETV1	No	T
FAT3	No	-
FGF12	No	-
FLI1	No	T
FZD1	No	-
GOPC	No	T
GRIN2A	No	Mis,N,F,O
HIST1H1C	No	-
IGF1R	No	-
IKZF1	No	TSG,D,T
KCNJ5	No	M
MYH11	Yes,Activating	T
NAB2	No	oncogene,T
PARP1	No	-
PRDM1	No	D,N,Mis,F,S
PTPRB	No	N,Mis,F,S
RNF213	No	T
SLC22A2	No	-
TERT	No	Promoter
TSC2	No	TSG,D,Mis,F,S
TSHZ3	No	-
TYMS	No	-

Table 12.5: Negatively selected genes known as cancer genes. Abbreviations: T, Translocation. TSG, Tumor supresor gene. Mis, Missense. S, Splice. F, Frameshift. O, Other. D, Deletion. N, Nonsense.

gProfiler (Table 12.6). In STRING, global terms such as cell differentiation and development, cell projection organization, neurogenesis, and cell morphogenesis were the most significant biological processes affected. Interestingly, enriched molecular functions and cellular localization of negatively selected genes were related to non-cell-autonomous mechanisms, such as regulation of transporter activ-

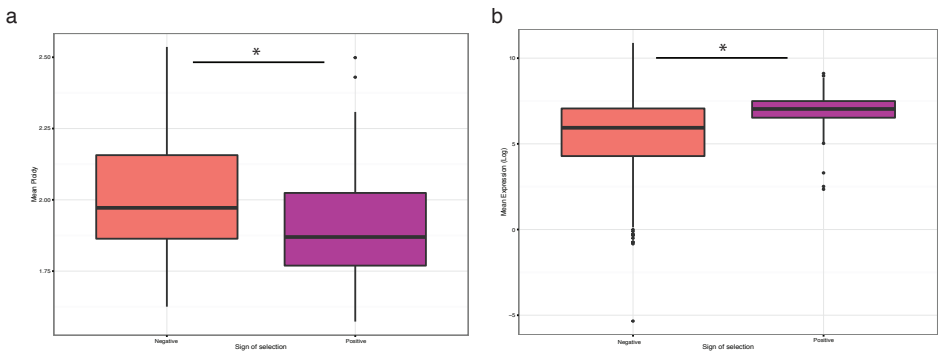

Figure 12.3: **Mean ploidy and expression values for negative and positively selected genes.** Comparison between negative and positive selected genes in mean ploidy level and mean expression value. (*P value < 0.005)

ity, substrate-specific binding, cell projection, and plasma membrane, which suggests a microenvironment-specific selective pressure. Furthermore, we observed that the number of significant terms produced by gProfiler was substantially lower than by STRING. The enriched GO terms were related to ion transport, neuron projection, axon extension, regulation of cell size, and regulation of cell growth. The last four terms coincided in several genes which were principally members of the plexins and semaphorins families. The overlapping number of genes between axon guidance, neurogenesis, and regulation of cell growth reaffirms a common functional role for these genes in cancer etiology [Pyatnitskiy et al., 2015].

Semaphorin and Plexin protein families

Semaphorin and plexin protein families have been recognized for a long time as scatter factors in cancer progression [Trusolino and Comoglio, 2002]. More specifically, Ch'ng et al. Have reviewed the role of *SEMA4D* and Plexin-B1 in tumor progression [Ch'ng and Kumanogoh, 2010], revealing their implication in tumor angiogenesis, regulation of tumor-associated macrophages, and control of invasive growth. Although their interaction has been widely recognized in immune and nervous systems, our study ratifies the importance of other semaphorins and plexins in tumor progression. Moreover, evidence of overexpression of *SEMA4D* supports our hypothesis of the relationship between increased oncogenic activity and negative selection of protein-altering mutations.

Feature name	ID	P Value	Genes in term
Sodium ion transport	GO:0006814	0,0455	14
Neuron projection extension involved in neuron projection guidance	GO:1902284	0,021	8
Regulation of cell size	GO:0008361	0,0071	13
Regulation of extent of cell growth	GO:0061387	0,0197	10
Axon extension	GO:0048675	0,00302	12
Axon extension involved in axon guidance	GO:0048846	0,021	8
Regulation of axon extension	GO:0030516	0,0116	10
Regulation of axon extension involved in axon guidance	GO:0048841	0,0495	7
Developmental growth involved in morphogenesis	GO:0060560	0,00634	15
Regulation of GTPase activity	GO:0043087	0,0387	23
P2X7 receptor signalling complex	CORUM:725	0,00819	4
Brachydactyly syndrome	HP:0001156	0,043	26
Glycosphingolipid biosynthesis - globo series	KEGG:00603	0,036	3
Insulin secretion	KEGG:04911	0,00881	11
Circadian entrainment	KEGG:04713	0,0294	11
Calcium signaling pathway	KEGG:04020	0,00296	18
Muscle contraction	REAC:397014	0,00327	20
Cardiac conduction	REAC:5576891	0,00453	16
Platelet calcium homeostasis	REAC:418360	0,00801	7
Other semaphorin interactions	REAC:416700	0,00788	6
Factor: ZEB; motif: NNNCAGGTGNSN; match class: 1	TF:M07371_1	0,0193	10

Table 12.6: Enrichment of GO and Functional pathways of genes under significant purifying selection.

P2X7 signaling complex

Another functional group under strong purifying selection was the P2X7 signaling complex. The main member of this complex, *P2X7R*, has been implicated in ATP-mediated cell death, and is responsible for the regulation of cation influx into the cell [Kim et al., 2001]. Various P2 receptor subtypes have been linked to different cancer types, in both human tissue samples and cell lines [Di Virgilio and Adinolfi, 2016]. Our results suggest that the P2X7 complex plays a universal role in the modulation of pathophysiological functions such as proliferation, differentiation, and apoptosis. We observed that *P2X7R* and its direct interactors, which include *PTPRB, ACTN4, IGTB2,* and *MPP3*, were under significant negative selection. However, complex members that do not directly interact with *P2X7R* were not under negative selection 12.4. The role of *P2X7R* and ATP in cancer is controversial. High ATP concentrations had a cytotoxic effect in all cell lines, eventually leading to cell necrosis, while in basal conditions, low ATP concentrations prevented cell death. Furthermore, ATP concentration exerted an effect on cancer cell migration and invasion [Giannuzzo et al., 2015]. Although not statistically significant, patients affected by a mutation in one of the members of the P2X7 complex have better prognosis in most tumor types (Fig. 12.5). We believe that in most tumors, wild-type *P2X7R* is favorable for cancer progression, based on the prognostic effect across individual tumor types.

Member protein	Gene Name	Protein atlas	Localization	Main function	K_a/K_s PC25
Actin, cytoplasmic 1	ACTB	All	Intracellular	Structural	-
Alpha-actinin-4	ACTN4	All	Intracellular	Structural	0,45*
Heat shock protein HSP 90-beta,	HSP90AB1	All	Intracellular	Structural	0,31
Heat shock 70 kDa protein 1	HSPA1A	All	Intracellular	Structural	-
Heat shock cognate 71 kDa protein	HSPA8	All	Intracellular	Structural	-
Integrin beta-2	ITGB2	All	Intracellular,Membrane	Structural	0,4*
Laminin subunit alpha-3	LAMA3	Mixed	Intracellular,Membrane	Structural	1
MAGUK p55 subfamily member 3	MPP3	Mixed	Intracellular	Structural	0,41*
P2X purinoceptor 7	P2RX7	All	Membrane	Cation pore	0,3*
Phosphatidylinositol 4-kinase alpha	PI4KA	Mixed	Intracellular	Signalling	0,82
Receptor-type tyrosine-protein phosphatase beta	PTPRB	Mixed	Intracellular,Membrane,Secreted	Signalling	0,63*
Supervillin	SVIL	All	Intracellular	Structural	0,88

Table 12.7: Protein interaction members of the P2X7 receptor signaling complex. Localization and tissue expression as annotated in the Protein Atlas, and calculated KaKs for each member on the Pancancer25 dataset. (* Q Value < 0.1)

12.3 Natural selection and mutation signatures in cancer

To reveal nucleotide context-specific selection, we calculated the K_n/K_s per mutation type across the Pancan25 dataset (Fig. 12.6). First, we found that K_n/K_s for all tumor and mutation types is normally distributed around one, revealing that the majority of genes in cancer genomes were under neutral selection. Second, statistical analysis only using C-to-T changes, the most common type of mutation, revealed similar functional enrichment results (axon guidance, neurogenesis, cation transport). Nonetheless, the number of significant PS genes is drastically reduced to only one, *TP53*, revealing that by restricting the sites to only C-to-T changes, the positive selection signal was more affected than the one for negative selection (P value ¡0.05, χ^2). In addition, we calculated the K_n/K_s value for *TP53* for each mutation category across every tumor type, demonstrating that *TP53* is under positive selection irrespective of the nucleotide context of mutations. (Table 12.8).

12.4 Purified functions reveal tumor specific prognostic markers

To assess the effect of selection over global functions instead of individual genes, we calculated K_a/K_s for each GO term and calculated a P value using the Pan-

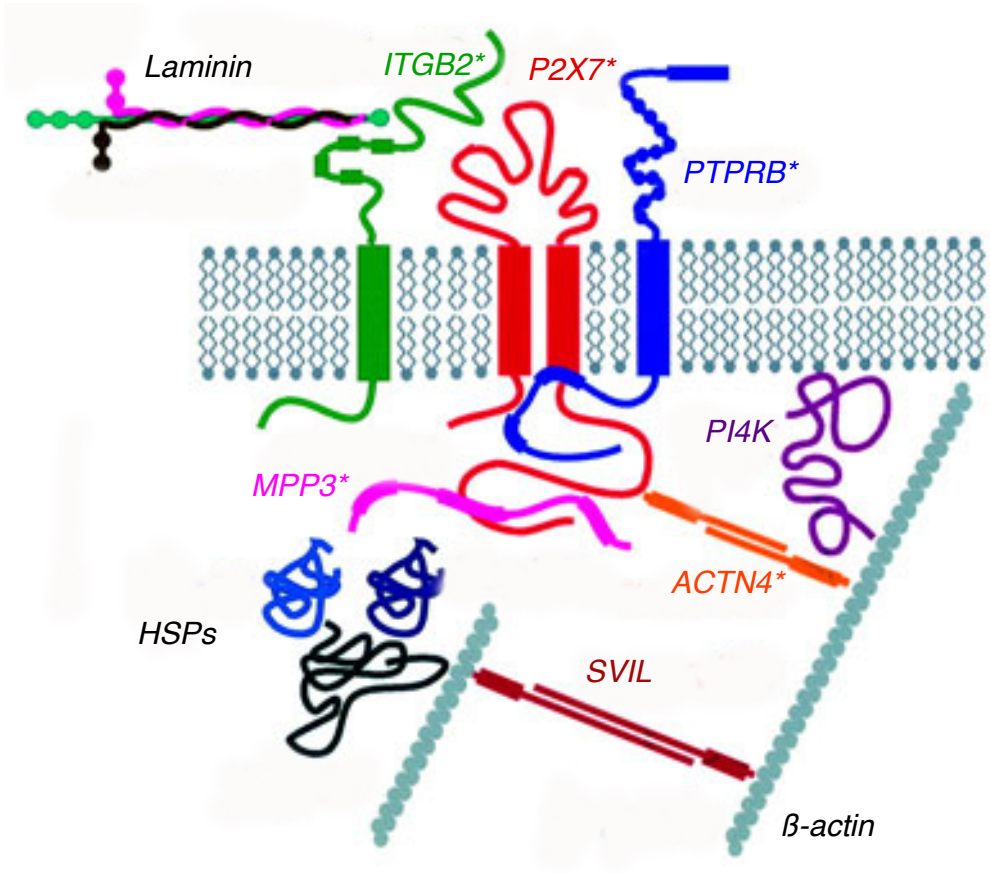

Figure 12.4: **P2X7 signalling complex.** Proteins interacting with *P2X7R*. Figure adapted from [Kim et al., 2001]. * indicates a significant selection on the gene.

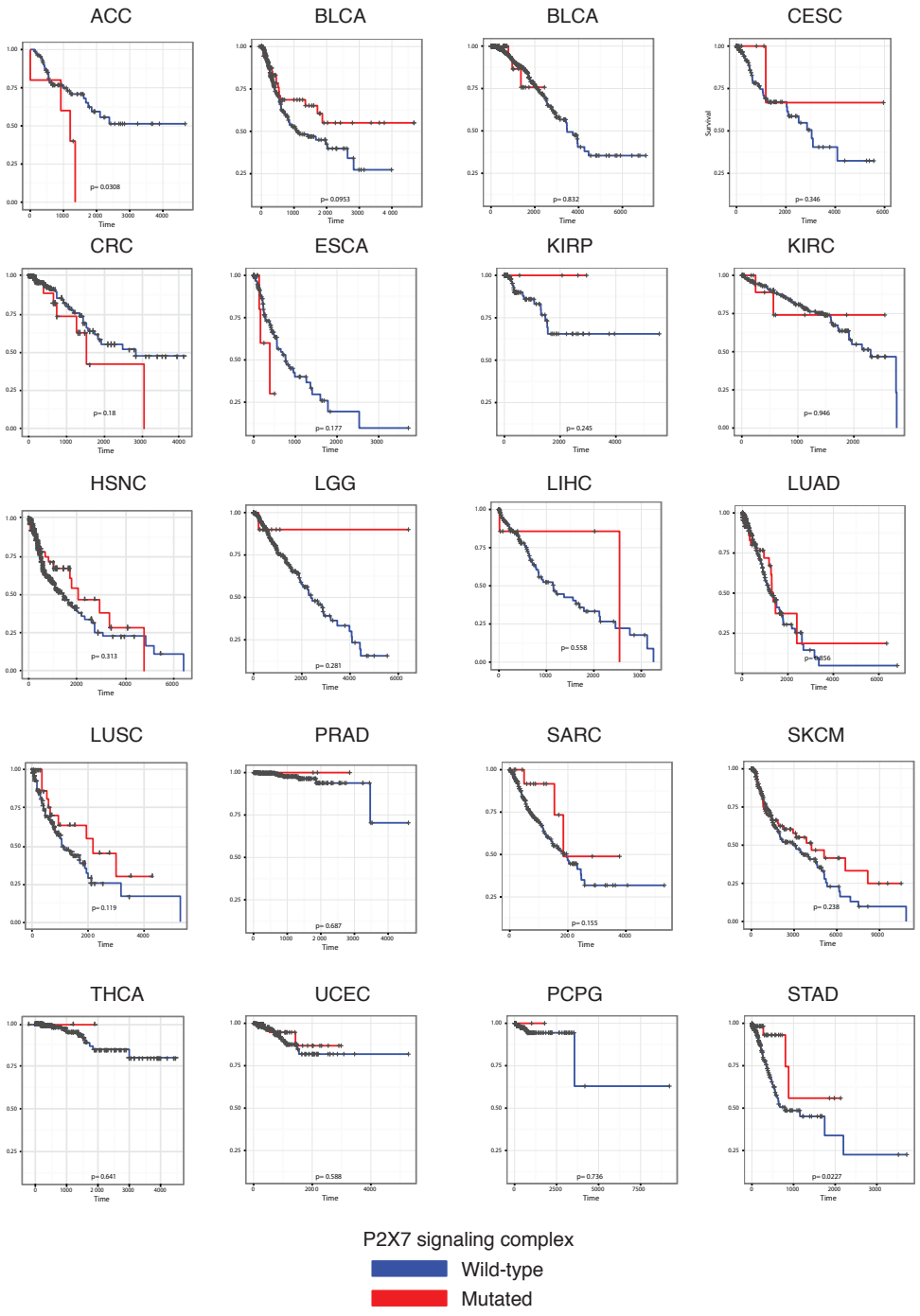

Figure 12.5: **Kaplan-Meier plots comparing wild-type versus mutated P2X7 signaling complex.**

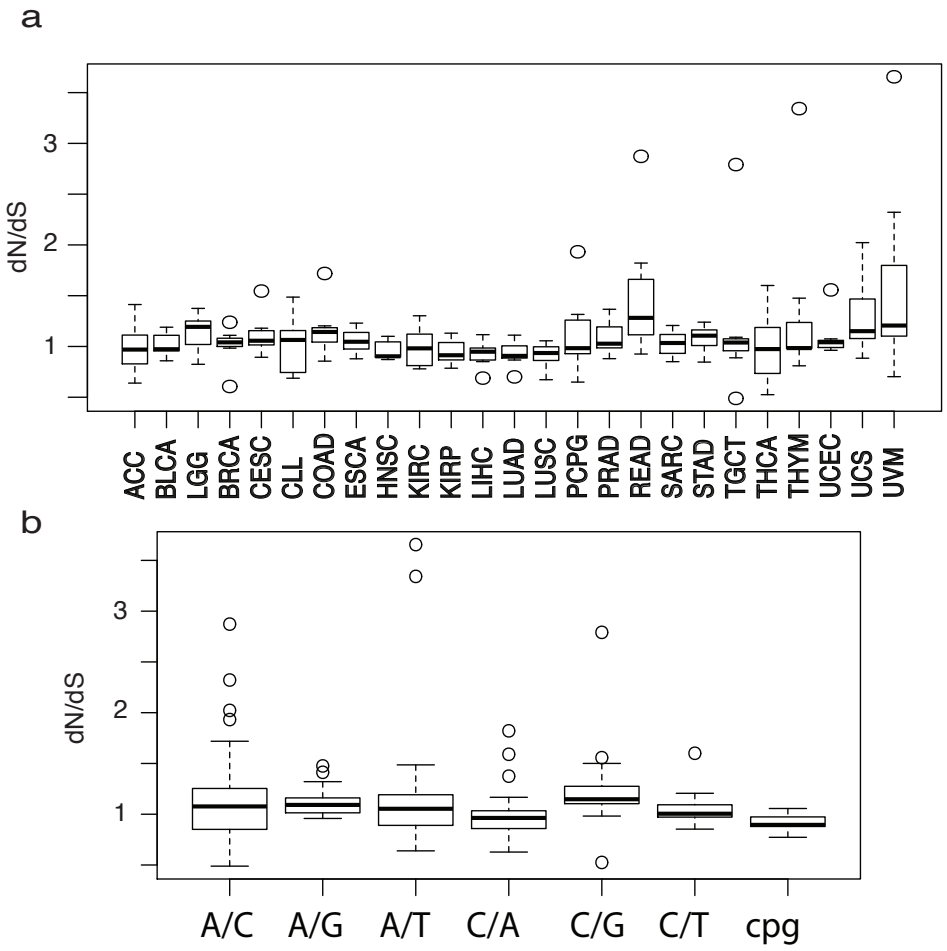

Figure 12.6: **Distribution of K_n/K_s values accross multiple tumor and mutation types.** a) Per tumor type and b) per mutation type.

Mutation type	Number of nonsynonymous	Synonymous	Nonsynonynomous Sites	Synonymous Sites	KnKs
A/C or T/G	100	2	417	93	11.15
A/G or T/C	256	1	357	153	109.7
A/T or T/A	85	1	422	88	17.72
C/A or G/T	206	4	479	109	11.71
C/G or G/C	104	3	486	102	7.27
C/T or G/A	290	23	361	227	7.92
CpG	665	26	193	59	7.81

Table 12.8: Number of somatic substitutions, total number of sites and calculated K_n/K_s of *TP53* (ENST00000269305) for each mutation category.

can25 dataset. We identified 14 GO terms under purifying selection (Table 12.9. The main functions identified were related to translation (GO:0006414, GO:0006415, GO:0022625, GO:0022627), metabolic rewiring (GO:0004129), and two described processes of cancer development: the first related to angiogenesis (GO:0030948) and the second to cell invasion (GO:0010719). These functions are highly relevant for the cancer cell to adapt to the new environment where metabolic needs are altered given low oxygen accessibility, or where intense competition for extracellular resources is present. Protein translation is a basic cellular function that must be perfectly conserved in order to comply with the increased burden of protein expression in cancer. Furthermore, angiogenesis and cancer invasion are under negative selection, which hints at the fact that the cell is able to perform these functions without the acquisition of novel genotypes.

p value	q	size	median	SD	term name	term id
0	0	15	0,506077348	0,497024295	cytochrome-c oxidase activity	GO:0004129
0	0	69	0,50711275	0,651667806	translational elongation	GO:0006414
0	0	58	0,519043101	0,91819403	translational termination	GO:0006415
0	0	21	0,570232558	0,309018972	negative regulation of epithelial to mesenchymal transition	GO:0010719
0	0	72	0,60246247	0,889821573	SRP-dependent cotranslational protein targeting to membrane	GO:0006614
0	0	9	0,446779262	0,697112268	cellular response to zinc ion	GO:0071294
0	0	23	0,533536585	0,249749002	cochlea morphogenesis	GO:0090103
0	0	42	0,592872318	0,456751917	keratin filament	GO:0045095
0	0	63	0,639360051	0,484612667	peptidase activity	GO:0008233
0	0	25	0,571914532	0,705044453	phospholipid transport	GO:0015914
0	0	28	0,518144981	0,997794893	cytosolic small ribosomal subunit	GO:0022627
0	0	35	0,530973451	0,492937493	cytosolic large ribosomal subunit	GO:0022625
0	0	7	0,391316527	0,538280809	negative regulation of vascular endothelial growth factor receptor signaling pathway	GO:0030948
0	0	85	0,631090487	0,791458445	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	GO:0000184

Table 12.9: GO terms under significant purifying selection.

In an attempt to demonstrate the impact of negatively selected functions in cancer treatment instead of negatively selected genes, we tested the prognostic relevance of the significant GO terms. We plotted Kaplan-Meier curves for each GO term by pooling all patients with at least one mutated gene into the ”knockout” category and compared them to the ”wild type” category for each tumor type separately.

We observed that most tumor types, except colorectal carcinoma, uterine carcinoma, and uveal melanoma, had a significant difference in prognosis for at least one term (Table 12.10).

GO Term	acc	blca	brca	cesc	coadread	esca	hnsc	kirc	kirp	lgg	lihc	luad	lusc	pcpg	prad	sarc	skcm	stad	tgct	thca	thym	ucec	uvm	Min P Valuel
GO:0000184	0,00	0,02		0,07							0,15		0,14				0,00			0,02	0,00			0,00
GO:0004129		0,01		0,08		0,08		0,15																0,01
GO:0006414	0,20	0,02	0,06	0,17		0,08					0,03										0,00			0,00
GO:0006415	0,00	0,10		0,04		0,10				0,07	0,08		0,04											0,00
GO:0006614	0,08	0,07	0,04	0,09		0,05	0,05			0,17			0,18								0,00			0,00
GO:0008233		0,08													0,09	0,03	0,01				0,04			0,01
GO:0010719			0,00	0,07				0,20										0,07						0,00
GO:0015914			0,02					0,00									0,05	0,18			0,01			0,00
GO:0022625	0,01	0,00		0,11							0,00	0,17	0,18				0,09							0,00
GO:0022627													0,10				0,18	0,14						0,10
GO:0030948		0,17					0,18	0,03								0,19								0,03
GO:0045095		0,02		0,04								0,08			0,08		0,14		0,00		0,00			0,00
GO:0071294	0,17									0,18											0,00			0,00
GO:0090103			0,01				0,07	0,14		0,12	0,20	0,16		0,00		0,10	0,06							0,00
Min P Value	0,00	0,00	0,00	0,04	-	0,05	0,05	0,00	0,35	0,07	0,00	0,08	0,04	0,00	0,08	0,03	0,00	0,07	0,00	0,02	0,00	-	-	0,00

Table 12.10: GO terms significance on prognosis.

Chapter 13

Methods

13.1 Tumor Data

Initial tumor dataset was downloaded from the following link `https://dl.dropboxusercontent.com/u/8680991/mafs/tcga_pancancer_dcc_mafs_082115.tar.gz`. Details on how the MAF files were assembled are given in a separate document within the tar file. CLL dataset was obtained as stated in the methods section of part III. The 26 selected tumor types, including CLL, are described in 12.1. Variant allele frequency (VAF) and CADD damage score were added to each mutation. Mutations were excluded based on the following criteria: (i) VAF < 0.1, (ii) Number of alternative allele supporting reads <= 4, (iii) EVS freq >= 1%, (iv) segmental duplication score > 0.5, and (v) repeat overlapping the mutation. In addition, we compiled a list of blacklisted genes and positions that we removed from the final MAF file. These files were uploaded to synapse (syn6115413).

13.2 K_n/K_s calculation

All mutations were reannotated using VEP to provide consistent ensembl transcript IDs. The number of non-synonymous, nonsense, splicing, and context specific sites for each transcript was calculated by counting all possible substitutions for each site in the transcript.

13.3 Statistical analysis

To assess the significance of selection acting on genes, we adapted a previously published exact statistical test [Greenman et al., 2006]. This test is based on the principle that silent (synonymous) somatic mutations are passenger mutations, which allow us to estimate an expected number of non-silent mutations and test this as a null hypothesis. We calculated P-values for non-synonymous

only, non-synonymous plus nonsense, and non-synonymous plus nonsense plus splice mutations, always considering the total number of sites for each category. In addition, we also applied the test using different base substitution categories ($A > T, A > C, A > G, C > A, C > T, C > G,\ and\ CpG > X$). Multiple test correction was performed using Benjamini and Hochberg.

We applied the Greenman method to each individual tumor, to all cancers (Pancan26), to all cancers except one (Leave-one-out Pancan). For all genes, we defined K_n as the classical non-synonymous substitutions per non-synonymous sites, K_{non} as the nonsense substitutions per nonsense sites, K_{spl}, as the ratio of splicing substitutions per splicing sites and K_a as the ratio of all nonsilent mutations per nonsilent sites. Then, each of these measures was divided by the ratio of synonymous mutations per synonymous sites. We tested the correlation of these measures using Pearson's product-moment correlation in R.

We compared the results from three datasets: (i) Pancan26, consisting of all tumor types, (ii) Pancan25, consisting of all tumor types except Skin Carcinoma, and (iii) only SKCM. Significant genes were selected based on the adjusted p-value considering all three mutation types: non-synonymous, nonsense, and splicing ($Adj.Pvalue < 0.1$). Significantly positive and negative genes were selected based on the K_a/K_s measure ($> 1positive, < 1negative$). For comparison of K_a/K_s in the whole dataset, we removed genes with 0 silent mutations. In addition, we calculated K_n/K_s for each mutation category in the Pancan25 dataset to compare the effect of selection bias in different contexts. In the case of this dataset, we removed genes with less than 10 silent mutations prior to the statistical analysis to increase power.

13.4 Postfiltration of cluster of synonymous changes

To check for genes that appear as negatively selected due to positive selection of synonymous changes, we ran OncodriveCLUST, replacing the annotation label of synonymous by non-synonymous. We compiled 60 genes showing a strong clustering bias in at least one of the 21 tumor types (syn5593040) ($Qvalue < 0.1$) and at least one cluster harboring more than 2 mutations). For skin cancer, we used all significant genes reported in oncoCLUST ($Qvalue < 0.1$).

13.5 Damage score of selected genes

We retrieved the CADD PHRED-scaled scores for all mutations used to compute the K_a/K_s ratios. Then, we considered selected genes as those that had a q value below 0.1, and neutral genes as those with a p-value above 0.8. We computed the mean damage score (among non-silent mutations in all genes in the respective sets) for different K_a/K_s cutoffs for negatively selected genes, for random samples from the neutral genes and for the positively selected genes ($K_a/K_s > 1$).

13.6 Functional enrichment

We used STRING v10.0 to find the connectivity for the set of purified genes. STRING provides built-in functions to detect protein-protein interactions and to perform GO term/KEGG pathway enrichment analysis. It performs a Hypergeometric test and corrects for multiple testing using Benjamini and Hochberg. In addition, we performed the enrichment analysis using gprofiler with the following parameters: ordered query, hierarchical sorting, no electronic GO, and setting 5 as the minimum and 350 as the maximum number of genes [Reimand et al., 2016]. The enrichment analysis was performed by first sorting the K_n/K_s of significant genes.

13.7 GO term functional enrichment test and survival analysis

A list of genes for each selected GO term was compiled. The genes from each term were concatenated, in other words, the number of substitutions and total sites were added, and a median K_a/K_s value was calculated for each of them. Then, patients were categorized into mutated and wild-type by the presence of at least one gene of the mutated GO list. Kaplan-Meier curves were calculated for each term using the two categories. The log-rank test function was used to asses significance difference in prognosis using the R package surv.

Chapter 14

Discussion

Here, we present evidence of selection in cancer genes. Most of the selection identified is in the form of negative selection, demonstrating an important role for cancer essential functions. We discuss some examples of known cancer genes under strong purifying selection, observations which challenge current theories stating the absence of negative selection in cancer [Martincorena et al., 2015b], [Ostrow et al., 2014].

Purifying selection is a major evolutionary force shaping the genomes of cells, individuals, and species. We exploited massive cancer genomic data from 25 tumor types to uncover a set of 639 genes under negative selection (NS) and 42 genes under positive selection (PS). A remarkable finding is that more than 70% of PS genes is present in the cancer gene census database, a much larger value than previously reported using a similar strategy [Ostrow et al., 2014]. Also, we demonstrate that negatively selected genes harbor less damaging mutations than neutral or positive selected. Intuitively, only lowly damaging mutations can accumulate on essential cancer genes, and these do not affect protein function. Moreover, we observe that the K_n/K_s ratio correlates with the mean functional impact of the somatic mutations, uncovering a relationship between gene selection and molecular selection. As previously noted, longitudinal analysis of allele fraction of somatic mutations will allow testing the effect of purifying selection at the cellular level.

Similarly to our cDriver approach, but in the opposite direction, low cellular prevalence of somatic mutations hints at negative selection. Nonetheless, many mildly deleterious mutations show a high CCF if they have hitchhiked along with a driver mutation during a clonal expansion [McFarland et al., 2013]. Under this scenario, most mildly deleterious mutations evade negative selection, first, through drift, and then by hitchhiking. If such somatic mutations reduce cancer cell fitness, it is possible that cancer progression is slowed, effectively improving a patient's prognosis. We identified one such example: the P2X7 signaling complex.

Five out of nine members of the P2X7 signaling complex displayed a K_n/K_s significantly lower than one across 25 tumor types. All direct interactors of the *P2X7*

receptor were under significant purifying selection, suggesting a coordinated essential role in oncogenesis. Indeed, studies have demonstrated the importance of P2X7 for cancer progression [Adinolfi et al., 2002],[Salaro et al., 2016],[Roger and Pelegrin, 2011] but not for all the interactors. *P2X7R* is ubiquitously expressed and when activated up regulates most of the glycolysis-related proteins linked to the "Warburg effect" [Di Virgilio and Adinolfi, 2016], a known hallmark of tumor progression [Warburg et al., 1924]. Additionally, *P2X7R* has been described as an immune modulator that releases immune-suppressive factors modifying the tumor micro-environment. Our findings confirms previous literature stating the essentiality of this complex for tumor progression and that selectively targeting this complex might be a powerful anticancer therapy.

Human cancers are subjected to different mutational processes [Alexandrov et al., 2013]. We divide our statistical calculation of K_n/K_s into seven mutation types including the most common across cancers, C-to-T. Although such analysis limits the statistical power of detection of selection for single genes, K_n/K_s values of *TP53* were always substantially higher than one, suggesting a mutation type-independent mechanism of selection. We also identified that tumor types show little variation when split into these categories, with the exception of A-to-T changes in Uveal Melanoma and Thymoma, A-to-C in Rectum adenocarcinoma, and C-to-G in Testicular cancer (all show strong positive selection). In contrast, C-to-G changes in Breast cancer and A-to-C in Testicular cancer were negatively selected. On one hand, such variation could be due to a low number of sites in the category. On the other hand, these results reveal slight differences on selective pressures at these sites in a tumor tissue - specific manner.

Our study also explores whether negatively selected genes are identified as essential genes in recent CRISPR-based screening studies on cancer cell lines [Wang et al., 2015]. Despite a low overlap between both datasets, we found the top essential gene of the chronic myelogenous leukemia cell line, *ABL1*, under strong purifying selection across 25 tumors. A reason for the low overlap between studies could be due to the fact that cultured cell lines are not affected by the tumor micro-environment. The environment in vitro is under tight control and environmental variables may not resemble realistic conditions for the survival of the cell. In fact, many cellular processes detected in [Wang et al., 2015] as essentials were fundamental pathways for basic cell functioning. Another explanation could be the large amount of heterogeneity observed when mixing different tumor types, therefore occluding tissue-specific selection of essential functions. In summary, our strategy reveals key differences between the "*in vivo*" and "*in vitro*" data.

One study reported absence of negative selection on a mutator phenotype [Beck-

man and Loeb, 2005]. Their model tested negative selection in isolation of other selective forces, however, we found that positive and negative selection are hardly divisible. We suggest that some observed negative selection is due to an accumulation of hidden oncogenic signals resulting in a recurrent conservation of the primary sequence of the protein. We present two examples of this, *ABL* and *TERT*. Both of them are oncogenes known to be "activated" in cancer because of translocations or regulatory variants, and we identified them under strong negative selection.

One of our findings points to a non-cell-autonomous selection process. It is therefore possible that mutations that we observe on our sequencing data may be somatic mutations acquired in the normal tissue fraction. It will be interesting to see whether somatic mutations in "healthy" cells have such non-cell-autonomous effect on cancer cells. As it was shown recently that normal cells also harbor somatic mutations under positive selection [Martincorena et al., 2015a], it may be possible that metastatic and angiogenic hallmarks are positively selected in normal cells supporting tumor growth and invasion.

In summary, we have identified a list of 643 putative novel targets for cancer treatment. Some of these genes overlap with current known cancer genes. The method presented here is the first step to uncover global essential functions related to tumor maintenance. In the future, more sophisticated statistical methods will help us to reveal tissue-specific traces of negative selection, therefore, allowing us to design global and personalized strategies for cancer treatment.

Chapter 15

Supplementary information

Ensembl id	K_n/K_s	K_{non}/K_s	K_{spl}/K_s	K_a/K_s	P Value Adj	Hugo symbol
ENST00000269305	17,28	68,71	8,46	19,11	0,0000	TP53*
ENST00000415913	31,10	2,28	0,00	28,16	0,0000	IDH1*
ENST00000256078	22,51	0,84	0,00	20,32	0,0000	KRAS*
ENST00000288602	7,03	3,15	2,51	6,55	0,0000	BRAF*
ENST00000263967	6,68	0,88	9,70	6,46	0,0000	PIK3CA*
ENST00000457016	1,26	25,42	1,04	2,78	0,0000	APC*
ENST00000371953	3,90	30,68	2,22	5,42	0,0000	PTEN*
ENST00000324856	1,50	36,49	1,87	3,18	0,0000	ARID1A*
ENST00000392793	0,46	0,57	0,10	0,46	0,0000	TECTA
ENST00000281708	5,69	20,51	0,83	6,42	0,0000	FBXW7*
ENST00000357586	0,25	0,19	0,14	0,25	0,0000	AP1B1
ENST00000435607	0,43	0,08	0,57	0,42	0,0000	SCN4A
ENST00000414716	0,34	0,00	0,23	0,32	0,0000	CEP170B
ENST00000221347	0,47	0,44	0,61	0,48	0,0000	FCGBP
ENST00000355272	0,37	0,12	0,35	0,36	0,0000	AP3D1
ENST00000457091	0,24	0,55	0,62	0,27	0,0000	GRID2IP
ENST00000259371	0,34	0,31	0,36	0,34	0,0000	DAB2IP
ENST00000349496	3,38	1,17	0,93	3,17	0,0000	CTNNB1*
ENST00000362061	0,47	0,27	0,45	0,46	0,0001	CACNA1S
ENST00000409345	0,30	0,00	NA	0,28	0,0001	ADRA2B
ENST00000366684	0,30	0,20	0,74	0,31	0,0001	ACTA1
ENST00000519560	0,44	0,58	0,51	0,45	0,0001	SLIT3
ENST00000438926	0,38	0,00	0,11	0,35	0,0001	SLC12A3
ENST00000313077	0,36	0,48	0,42	0,37	0,0002	CYFIP1
ENST00000301030	0,48	0,98	0,35	0,51	0,0002	ANKRD11
ENST00000525144	0,32	0,56	0,30	0,33	0,0002	KIRREL3
ENST00000349830	0,38	0,25	0,60	0,38	0,0002	ARHGEF10
ENST00000322038	0,40	0,24	NA	0,39	0,0002	TSHZ1
ENST00000263270	0,00	0,00	0,81	0,06	0,0002	AP2S1
ENST00000340806	0,31	0,60	0,00	0,32	0,0002	LRRIQ4
ENST00000326277	0,25	0,66	NA	0,27	0,0002	SNX18
ENST00000360215	0,26	0,52	0,00	0,26	0,0002	LHFPL5
ENST00000229829	0,17	0,49	0,00	0,17	0,0003	HLA-DOA
ENST00000312962	0,29	0,15	0,30	0,28	0,0003	KRI1
ENST00000342988	6,22	24,39	2,38	7,03	0,0003	SMAD4*
ENST00000269881	0,25	0,19	0,00	0,24	0,0003	CALR3
ENST00000345378	0,28	0,00	0,23	0,27	0,0004	GCK
ENST00000358024	0,32	0,24	0,00	0,32	0,0004	TMCC2

ENST00000243903	0,33	0,18	0,00	0,31	0,0005	ACTR5
ENST00000372348	0,38	0,69	0,57	0,40	0,0005	ABL1*
ENST00000268154	0,30	0,00	0,00	0,28	0,0006	ZNF710
ENST00000284476	0,45	0,55	0,00	0,45	0,0006	DISP1
ENST00000405093	0,49	0,07	1,08	0,49	0,0006	MYH6
ENST00000268035	0,40	0,69	0,47	0,42	0,0007	IGF1R*
ENST00000297056	0,29	0,77	0,00	0,30	0,0008	DAGLB
ENST00000361246	0,43	0,44	0,32	0,43	0,0008	CDC42BPB
ENST00000343348	0,31	0,22	0,24	0,31	0,0008	SEMA6A
ENST00000257901	0,36	0,32	0,59	0,37	0,0008	KRT85
ENST00000595753	0,09	0,00	0,00	0,09	0,0008	SLC35E1
ENST00000397850	0,36	0,72	0,85	0,40	0,0009	ITGB2
ENST00000439701	0,37	0,34	0,29	0,37	0,0009	IKZF1*
ENST00000265056	0,31	1,10	0,42	0,35	0,0011	MCM2
ENST00000588982	0,26	0,94	0,00	0,28	0,0012	ZBTB7C
ENST00000265969	0,41	0,17	0,55	0,40	0,0015	KCNC1
ENST00000359971	0,24	0,32	0,93	0,25	0,0016	ZNF517
ENST00000413689	0,55	0,69	0,49	0,55	0,0016	SCN5A
ENST00000358649	0,45	0,63	0,00	0,44	0,0016	PASK
ENST00000300131	0,20	0,00	0,00	0,18	0,0016	NAB2*
ENST00000274721	0,26	0,00	0,00	0,24	0,0016	GFRA3
ENST00000315082	0,10	0,00	0,00	0,09	0,0017	RHOT2
ENST00000280333	0,45	0,42	0,33	0,44	0,0017	DOCK1
ENST00000389840	0,58	0,65	0,67	0,58	0,0019	DNAH17
ENST00000375108	0,17	0,56	0,33	0,20	0,0019	PLA2G5
ENST00000369850	0,58	0,68	0,47	0,58	0,0019	FLNA
ENST00000378910	0,47	0,28	0,48	0,47	0,0019	NPHS1
ENST00000443236	0,50	0,38	0,10	0,48	0,0019	CPAMD8
ENST00000278616	1,87	6,04	1,64	2,13	0,0022	ATM*
ENST00000314933	0,29	0,00	0,00	0,27	0,0023	C1QB
ENST00000319914	0,29	0,00	0,00	0,27	0,0023	ST3GAL1
ENST00000238788	0,27	0,00	0,47	0,27	0,0026	TMEM214
ENST00000453981	0,54	0,10	0,17	0,49	0,0030	DOCK8
ENST00000445369	0,08	0,00	NA	0,07	0,0033	CLDN9
ENST00000396671	0,28	0,00	0,36	0,27	0,0033	THRB
ENST00000261381	0,47	0,40	0,40	0,47	0,0034	XYLT1
ENST00000301067	1,30	9,97	1,18	1,71	0,0035	KMT2D*
ENST00000393331	8,76	6,32	3,23	8,33	0,0035	SPOP*
ENST00000525985	0,55	0,41	NA	0,54	0,0035	EPPK1*
ENST00000357195	0,39	0,87	0,00	0,41	0,0035	BCL11B*
ENST00000525115	0,33	0,64	0,30	0,35	0,0035	PGBD5
ENST00000250160	0,35	0,20	1,03	0,36	0,0035	WISP1
ENST00000284523	0,31	0,44	0,53	0,32	0,0035	WNT3A
ENST00000528793	0,20	0,00	0,31	0,20	0,0035	GRK6
ENST00000361069	0,48	0,78	0,58	0,50	0,0037	LAMC3
ENST00000448023	0,44	0,21	0,51	0,43	0,0037	NLRC3
ENST00000399410	0,46	0,39	0,59	0,47	0,0038	ABCC1
ENST00000310581	0,38	0,27	0,29	0,37	0,0038	TERT*
ENST00000262450	0,52	0,77	0,73	0,54	0,0040	CHD5
ENST00000525621	0,42	0,90	0,68	0,45	0,0042	TYK2

ENST00000322957	0,00	0,00	0,00	0,00	0,0045	FOXJ1
ENST00000455140	0,32	0,00	1,10	0,35	0,0046	EPS15L1
ENST00000155926	0,31	0,50	0,00	0,31	0,0046	TRIB2
ENST00000262738	0,55	0,51	1,23	0,57	0,0046	CELSR1
ENST00000397928	0,50	0,37	0,41	0,49	0,0048	TRPM2
ENST00000395842	0,29	0,76	0,00	0,31	0,0049	PANX2
ENST00000219301	0,24	0,35	0,00	0,24	0,0052	PRSS54
ENST00000222345	0,52	0,13	0,57	0,51	0,0052	SIPA1L3
ENST00000444914	0,40	0,27	0,41	0,40	0,0053	SLC13A2
ENST00000598975	0,08	0,00	0,00	0,08	0,0057	no _symbol
ENST00000399899	0,17	0,00	NA	0,16	0,0057	CLDN8
ENST00000261574	0,39	0,32	0,30	0,38	0,0058	IPO5
ENST00000418971	0,32	0,29	0,24	0,31	0,0062	COLEC11
ENST00000406427	0,39	0,84	0,62	0,42	0,0063	PNPLA7
ENST00000262483	0,44	0,41	0,30	0,43	0,0064	PITPNM3
ENST00000304568	0,20	0,00	0,00	0,19	0,0065	TM4SF20
ENST00000355585	0,48	0,42	0,29	0,47	0,0066	SYNJ2
ENST00000404767	0,53	0,15	0,30	0,50	0,0066	INTS1
ENST00000370597	0,42	0,46	0,31	0,41	0,0066	CRTAC1
ENST00000354638	0,53	0,77	0,15	0,52	0,0066	OCA2
ENST00000390241	0,09	0,00	NA	0,08	0,0066	no _symbol
ENST00000428314	0,54	0,71	0,17	0,53	0,0066	MYO7B
ENST00000498124	3,74	58,32	8,73	5,41	0,0066	CDKN2A*
ENST00000291900	0,29	0,35	0,31	0,29	0,0069	ZER1
ENST00000380172	0,17	0,47	0,00	0,17	0,0071	MTAP
ENST00000450736	0,32	0,37	0,00	0,32	0,0071	SLC16A5
ENST00000366794	0,42	0,30	0,44	0,42	0,0072	PARP1*
ENST00000296871	0,00	1,02	0,00	0,05	0,0073	CSF2
ENST00000307149	0,26	0,31	0,83	0,29	0,0076	COG7
ENST00000355537	0,65	0,95	0,33	0,67	0,0076	ZNF536
ENST00000388887	0,56	0,75	0,58	0,57	0,0076	STAB2
ENST00000396352	0,33	1,08	0,00	0,36	0,0076	HOXA3
ENST00000355527	0,35	0,00	0,39	0,33	0,0079	DHCR7
ENST00000360273	0,55	0,12	0,43	0,53	0,0079	ATP2B2
ENST00000361488	0,39	0,44	NA	0,39	0,0084	FAM110B
ENST00000245663	0,37	0,67	0,00	0,38	0,0085	ZBTB46
ENST00000261655	0,53	1,02	0,45	0,55	0,0086	RIMBP2
ENST00000303635	0,57	0,66	0,35	0,57	0,0088	CAMTA1*
ENST00000368184	0,43	0,52	0,00	0,42	0,0088	FCRL3
ENST00000361900	0,25	0,00	0,00	0,22	0,0088	SCAMP5
ENST00000336643	0,22	0,00	0,00	0,20	0,0090	SLC39A9
ENST00000264276	0,45	0,74	0,15	0,45	0,0090	ALS2
ENST00000394967	0,00	0,00	8,93	0,05	0,0091	no _symbol
ENST00000307741	0,38	0,24	0,00	0,36	0,0093	THOP1
ENST00000339824	0,53	0,37	0,22	0,51	0,0098	KSR2
ENST00000388812	0,35	0,63	0,27	0,36	0,0098	GPR56
ENST00000420323	0,61	0,41	0,64	0,60	0,0098	DNAH1
ENST00000582970	0,63	0,64	0,18	0,61	0,0098	RNF213
ENST00000507866	0,42	0,62	0,45	0,43	0,0098	SORCS2
ENST00000467148	0,52	0,41	0,63	0,52	0,0099	HELZ2

ENST00000319234	0,24	0,81	0,00	0,27	0,0101	SHISA3
ENST00000388948	0,50	0,49	0,57	0,50	0,0102	LRRK1
ENST00000249389	0,24	0,00	0,00	0,22	0,0102	OPN1SW
ENST00000429989	0,26	0,00	0,00	0,23	0,0102	TSPAN14
ENST00000428368	0,56	0,92	0,51	0,58	0,0103	MYT1L
ENST00000311601	0,41	0,36	0,80	0,42	0,0103	SH3PXD2B
ENST00000413366	0,37	0,40	0,50	0,38	0,0103	PRKCA
ENST00000380393	0,42	0,65	0,39	0,43	0,0103	PTPRA
ENST00000374115	0,38	0,34	NA	0,38	0,0103	SLC18A3
ENST00000311519	0,40	0,50	0,00	0,39	0,0108	GPR113
ENST00000368498	0,21	0,42	0,00	0,21	0,0116	GOPC*
ENST00000450295	0,43	0,55	0,00	0,42	0,0117	SEMA4D
ENST00000259324	0,43	0,20	2,86	0,42	0,0120	LRRC8A
ENST00000329321	0,38	0,26	0,00	0,37	0,0120	GPR39
ENST00000399583	0,45	1,00	0,25	0,47	0,0125	RADIL
ENST00000394319	0,24	0,40	0,62	0,27	0,0129	LHX6
ENST00000374695	0,60	1,08	0,53	0,62	0,0129	HSPG2
ENST00000024061	0,44	0,59	0,71	0,45	0,0129	SLC45A4
ENST00000262305	0,29	0,39	0,37	0,30	0,0129	RAB11FIP3
ENST00000389232	0,71	0,97	0,69	0,72	0,0130	RYR3
ENST00000357701	0,50	0,58	0,00	0,48	0,0132	MYBPC2
ENST00000263666	0,57	0,21	0,44	0,55	0,0134	PDZRN3
ENST00000263519	0,53	0,90	0,50	0,54	0,0134	ATP2B3*
ENST00000378711	0,00	0,00	0,00	0,00	0,0134	C2orf91
ENST00000262139	0,28	0,38	0,00	0,26	0,0134	WIPI1
ENST00000285039	0,50	0,35	0,48	0,49	0,0135	MYO5B
ENST00000264938	0,37	0,00	0,50	0,37	0,0136	SLC9A3
ENST00000240587	0,60	1,00	0,00	0,62	0,0137	TSHZ3*
ENST00000419573	0,50	0,89	0,67	0,53	0,0138	RASGRF1
ENST00000243077	0,68	0,12	0,38	0,64	0,0141	LRP1
ENST00000322054	0,40	0,69	0,41	0,41	0,0141	EHD3
ENST00000262113	0,56	1,02	0,61	0,59	0,0141	MYOM2
ENST00000389048	0,55	0,90	0,42	0,56	0,0141	ALK*
ENST00000359246	0,44	0,67	0,34	0,44	0,0141	PHF2
ENST00000344420	0,25	0,00	0,60	0,25	0,0141	SLC35F6
ENST00000393229	0,43	0,40	0,00	0,41	0,0141	NTNG2
ENST00000545797	0,27	0,00	0,00	0,25	0,0143	DAPK3
ENST00000331427	0,33	0,38	0,61	0,34	0,0144	OTOP2
ENST00000380525	0,36	0,40	0,65	0,38	0,0146	CARS*
ENST00000262545	0,45	0,71	0,18	0,45	0,0154	PCSK2
ENST00000372648	0,33	0,59	0,00	0,32	0,0154	TBC1D13
ENST00000267101	2,82	1,28	3,20	2,76	0,0158	ERBB3*
ENST00000377577	0,32	0,82	0,20	0,34	0,0158	DNAJC11
ENST00000219476	0,47	1,22	0,13	0,48	0,0158	TSC2*
ENST00000002596	0,34	0,31	NA	0,34	0,0158	HS3ST1
ENST00000298472	0,38	0,64	0,17	0,37	0,0158	SLC18A2
ENST00000379919	0,43	0,17	NA	0,42	0,0158	MAB21L1
ENST00000374632	0,47	0,37	0,20	0,46	0,0158	EPHB2*
ENST00000338492	0,45	0,26	0,16	0,43	0,0159	DPYSL4
ENST00000222982	0,34	0,00	0,00	0,31	0,0159	CYP3A5*

ENST00000301732	0,50	1,16	0,40	0,52	0,0161	ABCA3
ENST00000504120	0,60	0,53	0,00	0,60	0,0161	PCDHA1
ENST00000267163	1,15	24,74	1,38	2,58	0,0162	RB1*
ENST00000371757	0,49	0,76	0,68	0,51	0,0163	KCNT1
ENST00000254037	0,22	0,00	0,61	0,22	0,0163	ZCCHC9
ENST00000338316	0,58	0,47	0,60	0,57	0,0163	ADCY2
ENST00000159111	0,50	0,31	0,16	0,48	0,0163	KDM4B
ENST00000219070	0,47	0,76	0,31	0,48	0,0164	MMP2
ENST00000390319	0,42	0,33	0,86	0,43	0,0164	no_symbol
ENST00000499869	0,29	0,00	0,59	0,29	0,0170	WDR1
ENST00000229771	0,33	0,36	0,23	0,33	0,0170	TULP1
ENST00000254958	0,41	1,03	0,31	0,44	0,0174	JAG1
ENST00000316902	0,34	0,00	0,00	0,31	0,0177	SLC7A8
ENST00000220764	0,27	0,83	0,26	0,29	0,0177	DECR1
ENST00000262519	0,53	0,42	0,00	0,51	0,0177	SETD1A
ENST00000367602	0,42	0,00	0,30	0,40	0,0178	QSOX1
ENST00000369208	0,55	0,31	0,64	0,54	0,0181	SIM1
ENST00000543663	0,40	0,28	0,00	0,37	0,0181	TUBGCP2
ENST00000379999	0,42	0,86	0,53	0,44	0,0183	FBXO18
ENST00000264031	0,00	0,00	0,00	0,00	0,0183	UPK2
ENST00000361725	0,18	0,00	0,00	0,16	0,0186	DLX1
ENST00000315087	0,37	0,64	0,54	0,39	0,0186	ST8SIA5
ENST00000304883	0,42	1,26	0,00	0,45	0,0188	TACR3
ENST00000511217	0,37	0,91	0,00	0,38	0,0188	SH3RF2
ENST00000377939	0,23	0,49	0,33	0,25	0,0188	RNF207
ENST00000379938	0,52	1,18	0,00	0,54	0,0189	RREB1
ENST00000405846	0,41	0,51	NA	0,41	0,0190	GPR12
ENST00000314890	0,00	0,00	NA	0,00	0,0195	ANXA2R
ENST00000225512	0,31	0,67	0,00	0,32	0,0197	WNT3
ENST00000282146	0,38	1,25	0,00	0,41	0,0198	KCNK13
ENST00000296127	0,22	1,20	0,00	0,26	0,0198	ZDHHC3
ENST00000218348	0,41	0,20	1,06	0,43	0,0198	USP11
ENST00000263707	0,28	1,40	0,55	0,35	0,0200	TFCP2L1
ENST00000357166	0,33	0,35	0,28	0,33	0,0202	ZDHHC9
ENST00000264977	0,45	0,28	0,63	0,45	0,0207	PPP2R3A
ENST00000394830	1,99	17,16	1,91	2,86	0,0207	PBRM1*
ENST00000269346	0,36	0,33	0,62	0,38	0,0212	TTYH2
ENST00000309311	0,38	0,33	0,00	0,36	0,0212	EEF2
ENST00000377981	0,31	0,00	NA	0,30	0,0220	OR13J1
ENST00000335793	0,30	1,28	NA	0,34	0,0220	CDC42EP4
ENST00000338101	0,56	1,53	0,25	0,59	0,0220	NOS1
ENST00000297440	0,41	0,34	0,33	0,41	0,0220	HEATR2
ENST00000540677	0,34	0,34	0,60	0,35	0,0222	SYT7
ENST00000247956	0,37	0,43	0,56	0,38	0,0223	ZNF317
ENST00000546292	0,63	0,35	0,28	0,60	0,0223	ZAN
ENST00000311956	0,35	0,00	0,00	0,31	0,0224	ARHGAP1
ENST00000398165	0,37	0,67	0,20	0,37	0,0225	CBS
ENST00000374316	0,61	0,42	0,24	0,58	0,0225	ITPR3
ENST00000374399	0,33	0,00	0,90	0,35	0,0228	NIPAL3
ENST00000291527	0,06	0,00	0,00	0,06	0,0228	TFF1

ENST00000320848	0,18	0,00	NA	0,17	0,0239	MRFAP1L1
ENST00000332710	0,27	0,00	0,00	0,25	0,0239	TBX1
ENST00000253055	0,39	0,00	0,00	0,37	0,0239	MAP3K10
ENST00000335867	0,50	0,98	0,00	0,52	0,0240	DACT1
ENST00000359526	0,54	0,13	0,32	0,51	0,0240	DNMT1*
ENST00000379597	0,33	0,00	1,28	0,32	0,0241	GCNT2
ENST00000521861	0,29	0,35	0,00	0,27	0,0244	EIF3H
ENST00000302754	0,06	0,00	NA	0,06	0,0248	JUNB
ENST00000254901	0,25	0,00	0,34	0,24	0,0253	REEP2
ENST00000303375	0,43	0,69	0,53	0,45	0,0255	MRC2
ENST00000371989	0,23	0,56	0,00	0,23	0,0257	SURF4
ENST00000435030	0,45	0,88	0,38	0,47	0,0257	KIF5C
ENST00000265709	0,62	0,86	0,27	0,61	0,0262	ANK1
ENST00000317552	0,44	0,76	0,00	0,44	0,0263	MID1
ENST00000356792	0,44	0,48	0,20	0,43	0,0263	EDRF1
ENST00000216487	0,47	0,62	0,36	0,47	0,0265	RIN3
ENST00000260264	0,33	0,00	0,48	0,32	0,0267	POU2F3
ENST00000418710	0,22	0,00	5,07	0,24	0,0268	SP8
ENST00000331664	0,54	0,60	0,00	0,54	0,0268	C2orf71
ENST00000396065	0,27	1,12	NA	0,32	0,0269	GCNT3
ENST00000166244	0,52	0,93	0,32	0,53	0,0269	EPHA8
ENST00000369769	0,50	0,57	NA	0,50	0,0269	KCNA3
ENST00000426263	0,36	0,90	0,00	0,36	0,0272	SLC2A1
ENST00000262803	0,52	0,35	0,00	0,48	0,0274	UPF1
ENST00000396573	0,64	0,45	0,54	0,63	0,0282	GRIN2A*
ENST00000370096	1,50	2,85	1,54	1,57	0,0285	COL11A1
ENST00000306749	0,59	0,13	0,60	0,58	0,0286	FASN
ENST00000261497	0,39	0,27	0,00	0,36	0,0286	USP22
ENST00000270162	0,40	0,00	0,00	0,37	0,0286	SIK1
ENST00000283415	0,38	0,27	0,93	0,41	0,0290	LPCAT1
ENST00000328557	0,48	0,68	0,00	0,49	0,0290	NRROS
ENST00000372476	0,56	0,67	0,24	0,55	0,0290	TIE1
ENST00000233154	0,41	0,24	0,00	0,40	0,0293	NCK2
ENST00000315396	0,43	0,00	0,26	0,40	0,0295	CCDC114
ENST00000313766	0,17	0,00	0,00	0,15	0,0295	FAM20C
ENST00000263253	1,96	17,37	6,37	2,88	0,0298	EP300*
ENST00000552695	0,28	0,47	0,00	0,28	0,0303	FICD
ENST00000301656	0,38	0,54	0,00	0,38	0,0304	KRT27
ENST00000315323	0,46	0,43	NA	0,46	0,0304	FZD2
ENST00000372027	0,38	0,33	0,00	0,37	0,0304	MMRN2
ENST00000322753	0,00	0,00	0,00	0,00	0,0305	MINOS1
ENST00000338821	0,50	0,99	0,41	0,52	0,0317	ATP9A
ENST00000356805	0,57	1,29	0,23	0,59	0,0317	SPTBN1
ENST00000315757	0,49	0,55	0,60	0,50	0,0317	ITGA11
ENST00000338820	0,27	0,51	0,00	0,27	0,0319	DNAJB12
ENST00000270560	0,20	0,00	0,00	0,18	0,0319	TM4SF5
ENST00000310942	0,34	0,46	0,00	0,34	0,0320	CBX2
ENST00000269228	0,47	0,39	0,38	0,46	0,0333	NPC1
ENST00000314128	0,32	0,00	0,30	0,30	0,0333	STAT2
ENST00000343216	0,37	0,41	0,00	0,37	0,0340	CXXC11

ENST00000260402	0,44	0,23	0,33	0,42	0,0340	PLCB2
ENST00000352766	0,29	0,00	0,00	0,27	0,0341	PRR5-ARHGAP8
ENST00000333834	0,17	0,00	NA	0,16	0,0345	LENG9
ENST00000564996	0,39	0,54	0,00	0,39	0,0348	KLHL36
ENST00000335757	0,41	0,00	0,35	0,39	0,0358	SLC44A2
ENST00000418404	0,53	0,96	0,79	0,57	0,0364	MYH13
ENST00000447648	0,42	0,59	0,80	0,45	0,0364	TECPR1
ENST00000529694	0,42	0,51	0,00	0,42	0,0364	KCNJ5*
ENST00000394456	0,33	0,00	0,00	0,29	0,0364	GTF2F1
ENST00000295702	0,21	0,00	0,48	0,22	0,0371	SSR2
ENST00000394128	0,45	0,29	0,00	0,41	0,0375	MED24
ENST00000546361	0,34	0,63	0,66	0,37	0,0376	CHERP
ENST00000586748	0,13	0,00	0,69	0,16	0,0376	YIPF2
ENST00000390649	0,60	0,77	0,44	0,60	0,0376	NLRP5
ENST00000309035	0,45	0,00	1,50	0,45	0,0376	CTBP2
ENST00000599957	0,45	0,34	0,00	0,45	0,0376	LRRC4B
ENST00000328041	0,45	0,57	0,17	0,43	0,0376	SLC24A3
ENST00000252804	0,65	0,62	0,29	0,63	0,0376	PXDN
ENST00000357164	0,22	0,00	0,00	0,20	0,0376	GM2A
ENST00000268676	0,35	0,37	0,00	0,33	0,0376	DEF8
ENST00000295057	0,44	0,69	NA	0,46	0,0381	LRRTM1
ENST00000366618	0,45	0,62	0,28	0,45	0,0381	SLC35F3
ENST00000258168	0,37	0,78	0,29	0,39	0,0387	BCMO1
ENST00000296755	0,61	0,89	0,60	0,63	0,0396	MAP1B
ENST00000561648	0,63	0,47	0,55	0,62	0,0396	ZNF423
ENST00000355072	0,59	0,54	0,88	0,61	0,0400	HTT
ENST00000264028	0,36	0,00	0,00	0,32	0,0401	ARCN1
ENST00000546057	0,27	0,37	0,79	0,30	0,0404	P2RX7
ENST00000312358	0,60	0,62	0,46	0,60	0,0404	SPEG
ENST00000305632	0,39	0,35	0,00	0,38	0,0404	TBL2
ENST00000397256	0,19	1,08	1,10	0,25	0,0408	ARPC4-TTLL3
ENST00000322773	1,93	4,05	1,23	1,96	0,0414	COL19A1
ENST00000392678	0,15	0,00	0,00	0,14	0,0416	TIRAP
ENST00000521400	0,39	0,00	0,36	0,37	0,0416	EPHX2
ENST00000393504	0,53	0,27	1,24	0,53	0,0416	CNGA3
ENST00000287934	0,47	0,23	NA	0,46	0,0418	FZD1*
ENST00000396720	0,32	1,34	0,00	0,35	0,0419	GLTSCR1
ENST00000368124	0,33	1,36	0,29	0,37	0,0422	CADM3
ENST00000404795	0,20	0,00	0,00	0,19	0,0422	DMRTA2
ENST00000524558	0,39	0,35	0,00	0,38	0,0424	SERPINH1
ENST00000570689	0,43	1,16	0,75	0,48	0,0425	UMOD
ENST00000359827	0,70	0,25	0,43	0,67	0,0427	PLXNA4
ENST00000373600	0,30	0,46	0,00	0,28	0,0428	NEK6
ENST00000367474	0,15	0,00	0,00	0,14	0,0429	SAMD5
ENST00000217964	0,44	0,49	0,19	0,43	0,0429	TBL1X
ENST00000311457	0,30	0,53	0,00	0,28	0,0429	REC8
ENST00000570156	0,72	1,05	0,70	0,73	0,0429	OBSCN
ENST00000373344	1,29	6,72	2,78	1,71	0,0429	ATRX*
ENST00000479441	0,35	0,00	0,99	0,39	0,0430	CACNA2D2
ENST00000319380	0,48	0,43	0,00	0,46	0,0431	UBE2O

ENST00000311085	1,96	4,08	0,64	2,04	0,0431	DMXL1
ENST00000454309	0,35	1,61	0,00	0,41	0,0432	FGF12*
ENST00000263246	0,35	1,13	0,28	0,39	0,0432	PACSIN2
ENST00000394975	0,07	0,00	0,00	0,06	0,0432	VKORC1
ENST00000247986	0,53	0,34	0,22	0,51	0,0432	KIF17
ENST00000302538	0,47	0,19	0,70	0,47	0,0437	ABR
ENST00000286794	0,33	0,59	NA	0,35	0,0442	NAA11
ENST00000222673	0,55	0,32	0,27	0,53	0,0448	OGDH
ENST00000323274	0,13	0,79	0,00	0,16	0,0448	TYMS*
ENST00000272190	0,32	0,38	0,63	0,34	0,0448	REN
ENST00000421627	0,56	0,67	0,23	0,55	0,0448	DLGAP2
ENST00000593496	0,00	0,00	NA	0,00	0,0456	no _symbol
ENST00000394562	0,29	0,99	0,00	0,34	0,0457	SLFN12
ENST00000314045	0,33	0,41	0,65	0,35	0,0462	DDX54
ENST00000575354	2,07	4,99	0,00	2,11	0,0462	CIC*
ENST00000334920	0,39	0,00	NA	0,38	0,0462	OR10H1
ENST00000222256	0,22	0,00	0,00	0,20	0,0462	RAB3A
ENST00000396757	0,25	0,00	0,00	0,22	0,0462	CD72
ENST00000355497	0,33	0,40	0,00	0,32	0,0463	LMX1B
ENST00000324134	0,62	0,46	0,62	0,61	0,0474	NLRP12
ENST00000261517	1,64	5,22	2,10	1,87	0,0476	VPS13C
ENST00000527786	0,46	0,59	0,00	0,44	0,0479	FLI1*
ENST00000377918	0,67	0,95	0,00	0,67	0,0479	PCDH17
ENST00000261383	0,69	1,27	0,69	0,72	0,0481	DNAH3
ENST00000304623	0,62	1,17	0,57	0,65	0,0481	CTNND2
ENST00000295373	0,47	0,50	0,24	0,46	0,0481	DHX57
ENST00000504087	2,32	3,68	0,00	2,40	0,0481	ANKRD50
ENST00000264230	0,31	0,00	0,00	0,29	0,0481	NOA1
ENST00000389418	0,59	0,42	0,31	0,57	0,0481	PTPRN2
ENST00000380620	0,15	0,00	0,00	0,14	0,0481	B3GALT5
ENST00000238738	0,15	0,00	0,00	0,14	0,0482	RHOQ
ENST00000278317	0,39	0,23	0,31	0,37	0,0482	TNNT3
ENST00000261439	0,47	1,22	0,41	0,50	0,0486	TBC1D1
ENST00000427624	0,07	0,00	NA	0,07	0,0486	ZNF837
ENST00000400454	0,69	0,99	0,14	0,68	0,0486	DSCAM
ENST00000471889	0,40	0,69	1,05	0,43	0,0486	SLC45A1
ENST00000254466	0,38	0,94	0,00	0,40	0,0486	GAS2L2
ENST00000265846	0,23	0,00	0,00	0,20	0,0487	ADAP1
ENST00000299853	0,40	0,96	0,50	0,44	0,0495	POLR3E
ENST00000396324	0,54	0,81	0,76	0,57	0,0501	MYH11*
ENST00000257934	0,54	1,22	0,18	0,56	0,0501	ESPL1
ENST00000220592	0,49	0,54	0,62	0,50	0,0501	AGO2
ENST00000373115	0,39	0,32	0,00	0,39	0,0501	CHST3
ENST00000377574	0,49	0,31	0,47	0,48	0,0502	SLC22A12
ENST00000538904	0,38	0,00	0,56	0,37	0,0502	EVI5L
ENST00000378061	0,54	0,12	0,00	0,51	0,0502	ZNF425
ENST00000307169	0,22	0,00	NA	0,22	0,0505	INSM2
ENST00000395925	0,65	1,13	0,16	0,66	0,0507	GLI3
ENST00000330753	0,58	1,00	NA	0,60	0,0508	FLRT2
ENST00000404857	0,53	0,86	0,47	0,54	0,0508	KCNMA1

ENST00000534961	0,19	0,62	0,00	0,20	0,0508	RNF170
ENST00000252699	0,45	0,23	0,57	0,45	0,0508	ACTN4
ENST00000300091	0,33	0,33	0,00	0,32	0,0511	C18orf54
ENST00000389022	0,21	0,00	0,00	0,19	0,0511	NT5M
ENST00000355841	0,27	0,00	0,77	0,29	0,0519	PDLIM7
ENST00000383829	0,50	0,19	0,00	0,47	0,0519	BRPF1
ENST00000391736	0,39	0,78	0,78	0,43	0,0524	LILRB4
ENST00000454376	0,17	0,00	0,54	0,18	0,0524	PRMT1
ENST00000224600	0,52	0,00	0,00	0,50	0,0524	RBP3
ENST00000582783	0,39	0,00	0,00	0,35	0,0524	C19orf47
ENST00000538872	0,37	0,00	0,00	0,34	0,0525	IQSEC3
ENST00000012443	0,32	0,00	0,00	0,28	0,0534	PPP5C
ENST00000296327	0,22	0,00	1,59	0,30	0,0544	SLC51A
ENST00000431231	0,28	1,25	0,00	0,31	0,0545	ARHGAP23
ENST00000303694	0,42	0,47	0,00	0,41	0,0545	CHST11
ENST00000357183	0,12	0,00	0,00	0,11	0,0545	RAET1E
ENST00000264852	0,50	0,23	0,14	0,47	0,0545	SIDT1
ENST00000297405	1,23	2,29	1,08	1,28	0,0545	CSMD3
ENST00000261769	2,53	20,43	3,83	3,33	0,0545	CDH1*
ENST00000337714	0,48	0,21	0,81	0,47	0,0546	AKAP1
ENST00000354042	0,43	0,33	0,42	0,43	0,0551	SLC13A4
ENST00000282007	1,91	8,93	0,74	2,35	0,0551	ZC3H13
ENST00000262776	0,36	0,34	0,95	0,37	0,0552	LGALS3BP
ENST00000373187	0,66	1,13	0,67	0,68	0,0557	PTPRT
ENST00000381309	0,50	0,74	0,47	0,52	0,0557	KDM4C
ENST00000320634	0,25	0,65	0,30	0,27	0,0557	FAIM2
ENST00000529308	0,47	0,00	0,97	0,46	0,0561	USP35
ENST00000406875	0,52	0,71	0,27	0,52	0,0569	HIPK2
ENST00000325089	0,63	0,45	NA	0,62	0,0571	SLITRK5
ENST00000399151	0,58	0,89	0,64	0,60	0,0572	DOPEY2
ENST00000270642	0,31	0,00	0,00	0,28	0,0572	IGLON5
ENST00000370225	0,61	1,32	0,51	0,63	0,0575	ABCA4
ENST00000306318	0,18	1,47	3,85	0,26	0,0575	GBX2
ENST00000256759	0,31	0,79	0,90	0,36	0,0575	FST
ENST00000419421	0,16	0,00	0,00	0,15	0,0575	PRR22
ENST00000312932	0,27	0,00	0,00	0,24	0,0577	SEC14L2
ENST00000409230	0,00	0,00	0,00	0,00	0,0581	C2orf82
ENST00000253754	0,26	0,68	0,55	0,29	0,0584	PDLIM4
ENST00000392620	0,28	0,43	NA	0,29	0,0584	C17orf77
ENST00000422970	0,34	0,31	0,36	0,34	0,0587	IPCEF1
ENST00000355480	0,52	0,90	0,67	0,54	0,0594	COL18A1
ENST00000371571	0,38	0,00	0,00	0,36	0,0615	KCNG1
ENST00000316218	0,49	0,20	0,14	0,44	0,0615	PDIA5
ENST00000355703	0,50	0,69	0,61	0,51	0,0616	PCNXL3
ENST00000361727	0,69	0,71	0,62	0,69	0,0616	CNTNAP2
ENST00000390444	0,12	0,00	0,00	0,11	0,0619	no _symbol
ENST00000426508	0,49	0,42	0,39	0,48	0,0619	IFT140
ENST00000268603	0,58	0,63	1,21	0,61	0,0620	CDH11*
ENST00000263925	0,39	0,00	0,53	0,37	0,0620	LNX1
ENST00000247706	0,15	1,87	0,00	0,22	0,0623	ABHD8

ENST00000409792	1,68	6,49	1,15	1,97	0,0627	SETD2*
ENST00000302345	0,36	0,00	0,00	0,34	0,0629	CANT1*
ENST00000513610	0,55	1,48	0,67	0,61	0,0631	MYO10
ENST00000318407	0,13	0,00	0,00	0,12	0,0632	BOK
ENST00000324589	0,49	1,09	0,13	0,49	0,0632	GALNT14
ENST00000253796	0,00	0,00	0,00	0,00	0,0632	RAMP2
ENST00000221515	0,00	0,00	0,00	0,00	0,0632	RETN
ENST00000397298	0,13	0,00	0,00	0,11	0,0632	MRPL23
ENST00000298351	0,48	0,00	0,34	0,45	0,0636	TMEM63C
ENST00000331238	0,38	0,47	0,00	0,38	0,0636	RTN4RL1
ENST00000334414	0,60	1,16	0,75	0,63	0,0636	PTPRB*
ENST00000370193	0,22	1,35	0,00	0,27	0,0636	LBX1
ENST00000539243	0,45	0,75	0,66	0,47	0,0636	HMHA1
ENST00000251588	0,30	0,97	0,00	0,32	0,0636	NARFL
ENST00000556143	0,45	0,49	0,00	0,44	0,0638	ZFYVE1
ENST00000304920	0,08	0,00	0,00	0,07	0,0641	IL17D
ENST00000354646	1,87	3,58	2,33	1,99	0,0642	WNK3
ENST00000332235	0,16	0,00	NA	0,15	0,0643	C2CD4C
ENST00000263610	0,36	0,00	0,00	0,33	0,0646	BARHL1
ENST00000255977	0,30	0,38	0,00	0,29	0,0647	MKRN1
ENST00000247207	0,40	0,37	NA	0,40	0,0649	HSPA2
ENST00000354484	0,37	0,72	0,00	0,37	0,0649	ADAMTSL2
ENST00000404914	0,30	0,00	0,31	0,29	0,0649	ATG4B
ENST00000518868	0,42	0,68	0,00	0,41	0,0659	C14orf159
ENST00000434354	0,39	0,36	0,00	0,36	0,0659	PEX5
ENST00000019317	0,14	0,67	0,00	0,17	0,0669	RALBP1
ENST00000550722	0,63	1,11	0,86	0,67	0,0670	HECTD4
ENST00000397899	0,49	0,00	1,03	0,48	0,0670	KIAA1211L
ENST00000202625	0,53	0,88	0,18	0,53	0,0673	TGM6
ENST00000447017	0,44	0,30	0,00	0,40	0,0673	ABLIM2
ENST00000258888	0,58	0,88	0,66	0,59	0,0673	ALPK3
ENST00000230321	0,19	0,00	0,00	0,18	0,0673	MDFI
ENST00000524993	0,36	0,00	0,37	0,35	0,0673	TMPRSS13
ENST00000295108	0,36	0,00	NA	0,34	0,0673	NEUROD1
ENST00000461988	0,26	0,51	0,89	0,31	0,0675	POR
ENST00000397062	4,63	1,20	0,00	4,37	0,0675	NFE2L2*
ENST00000313546	0,30	0,38	0,40	0,31	0,0675	POC1B
ENST00000261726	0,58	0,81	0,18	0,57	0,0675	CUX2
ENST00000422318	0,34	0,00	0,00	0,31	0,0675	NT5DC2
ENST00000311745	0,61	0,27	0,41	0,58	0,0675	RBFOX1
ENST00000184266	0,17	0,00	0,00	0,15	0,0676	NDUFB4
ENST00000454036	0,59	0,42	0,77	0,60	0,0680	SLC12A5
ENST00000374490	0,23	0,00	0,53	0,24	0,0680	HMGCL
ENST00000537592	0,60	0,50	0,00	0,60	0,0680	SALL3
ENST00000322357	0,37	0,00	0,00	0,36	0,0680	ZBTB7A
ENST00000268482	0,53	0,20	0,00	0,49	0,0680	DHX38
ENST00000272252	0,26	0,00	0,65	0,26	0,0681	GALM
ENST00000240100	0,31	0,00	0,00	0,29	0,0684	DUSP4
ENST00000252826	0,54	0,60	0,61	0,54	0,0684	TRPM4
ENST00000393096	0,44	0,73	0,00	0,45	0,0685	SERPINA10

ENST00000467963	0,41	0,00	0,26	0,38	0,0685	BRD9
ENST00000264972	0,52	0,54	0,20	0,50	0,0688	ZAP70
ENST00000235345	0,31	0,85	0,49	0,35	0,0689	SLC35D1
ENST00000262971	0,35	0,00	0,42	0,34	0,0689	PIAS4
ENST00000332582	0,40	0,00	NA	0,39	0,0691	GPR173
ENST00000323084	0,49	0,99	0,00	0,49	0,0691	TSPEAR
ENST00000487903	0,54	0,61	0,37	0,53	0,0695	ATP11A
ENST00000324093	0,53	0,72	1,03	0,56	0,0699	PLXND1
ENST00000393409	0,63	0,65	0,67	0,63	0,0699	PLXNA1
ENST00000310248	0,28	0,00	NA	0,27	0,0699	OR10AD1
ENST00000304381	0,47	0,00	0,45	0,45	0,0699	TMC7
ENST00000544455	1,59	3,46	2,33	1,73	0,0706	BRCA2*
ENST00000289749	0,12	0,00	0,00	0,12	0,0709	NBL1
ENST00000437212	0,37	0,00	0,28	0,35	0,0711	TMPRSS4
ENST00000413817	0,21	0,00	0,44	0,22	0,0722	DENND6B
ENST00000372874	0,26	0,48	0,88	0,32	0,0722	ADA
ENST00000358526	0,56	0,67	0,00	0,56	0,0722	AKAP4
ENST00000378536	0,39	0,71	0,00	0,40	0,0730	SKI
ENST00000327979	0,52	0,99	0,00	0,51	0,0730	FAM65C
ENST00000226284	0,47	0,45	0,27	0,46	0,0730	IBSP
ENST00000296674	0,08	0,00	0,00	0,07	0,0730	RPS23
ENST00000393980	0,30	0,00	0,35	0,29	0,0744	FABP6
ENST00000369096	0,46	0,47	0,63	0,46	0,0746	PRDM1*
ENST00000310373	0,25	0,61	0,00	0,26	0,0747	GP6
ENST00000397820	0,17	0,00	0,00	0,15	0,0747	C19orf38
ENST00000310226	0,34	0,00	0,60	0,34	0,0749	FADS6
ENST00000343846	0,40	1,26	0,35	0,44	0,0751	SUSD4
ENST00000200676	0,40	0,39	0,57	0,41	0,0753	CETP
ENST00000356404	0,50	0,00	0,48	0,48	0,0753	PITPNM1
ENST00000221462	0,00	0,00	0,00	0,00	0,0757	PPP1R37
ENST00000376561	0,46	0,78	0,17	0,46	0,0758	CTTN
ENST00000299106	0,35	0,42	0,31	0,35	0,0758	JAM3
ENST00000400286	2,12	2,74	NA	2,16	0,0759	SLITRK6
ENST00000262065	0,20	0,53	0,92	0,26	0,0763	MMD
ENST00000378191	0,48	0,32	0,46	0,48	0,0763	AJAP1
ENST00000392550	0,48	0,80	0,52	0,49	0,0764	LLGL2
ENST00000370060	0,62	0,54	0,54	0,62	0,0766	L1CAM
ENST00000411851	0,38	0,00	0,00	0,37	0,0766	TBXA2R
ENST00000305817	0,35	0,35	NA	0,35	0,0766	PRND
ENST00000373764	0,31	0,36	0,00	0,30	0,0766	MORN5
ENST00000344157	3,12	7,01	3,10	3,38	0,0766	YTHDC1
ENST00000390560	0,27	0,00	NA	0,23	0,0766	no _symbol
ENST00000409134	0,32	0,00	0,00	0,28	0,0766	ALDH7A1
ENST00000399503	1,85	10,56	3,47	2,39	0,0766	MAP3K1*
ENST00000217185	0,23	0,56	1,33	0,29	0,0767	PTK6
ENST00000358625	0,24	1,11	0,00	0,26	0,0767	WDR5
ENST00000483722	0,31	0,00	0,00	0,29	0,0767	TREML2
ENST00000441003	0,65	0,24	0,55	0,63	0,0767	MUC2
ENST00000412232	0,51	0,32	0,00	0,49	0,0769	GPR124*
ENST00000356530	0,16	0,00	NA	0,15	0,0771	HIST1H2BE

ENST00000340726	0,54	1,03	0,00	0,53	0,0778	EYA1
ENST00000318602	0,57	1,04	0,70	0,61	0,0780	A2M
ENST00000370388	0,22	0,00	0,00	0,20	0,0785	KHDC1L
ENST00000304613	0,55	1,38	0,93	0,60	0,0786	KNDC1
ENST00000331780	0,31	0,90	0,00	0,33	0,0786	SPATA32
ENST00000338352	0,43	0,51	NA	0,44	0,0786	OTUD6A
ENST00000239151	0,24	0,00	0,00	0,22	0,0786	HOXB5
ENST00000402395	0,00	0,00	0,00	0,00	0,0786	no_symbol
ENST00000319562	0,48	1,04	0,32	0,50	0,0786	FARP1
ENST00000223114	0,37	0,00	0,82	0,38	0,0787	MOGAT3
ENST00000356896	0,30	0,00	0,00	0,27	0,0787	IFFO1
ENST00000359106	0,63	0,39	0,85	0,63	0,0794	CACNA1G
ENST00000307161	0,35	0,00	0,00	0,33	0,0797	MTNR1A
ENST00000260324	0,29	1,34	0,00	0,34	0,0801	SQRDL
ENST00000265922	0,60	0,88	0,50	0,61	0,0801	BRINP1
ENST00000600128	0,66	0,77	0,68	0,67	0,0801	FBN3
ENST00000226193	0,36	0,36	0,00	0,35	0,0801	RCVRN
ENST00000375559	0,52	0,00	0,00	0,48	0,0801	F10
ENST00000571088	0,29	0,52	0,92	0,32	0,0809	no_symbol
ENST00000315190	0,24	0,60	NA	0,26	0,0809	SERTM1
ENST00000379959	0,28	0,45	0,39	0,30	0,0809	IL2RA
ENST00000301335	0,37	0,98	0,00	0,38	0,0809	SLC43A2
ENST00000296603	4,72	9,47	4,41	5,00	0,0809	LMBRD2
ENST00000224605	0,49	0,00	0,00	0,47	0,0809	GDF10
ENST00000274376	2,24	17,94	2,04	3,10	0,0809	RASA1*
ENST00000237853	4,66	12,35	2,08	5,00	0,0814	ELL2
ENST00000219406	0,17	0,00	0,00	0,16	0,0814	PDIA2
ENST00000381269	0,55	0,95	1,82	0,59	0,0826	SLC8A3
ENST00000262623	0,54	0,38	1,01	0,55	0,0828	ATP4A
ENST00000337672	0,13	0,00	0,00	0,12	0,0829	MED19
ENST00000512701	0,43	0,74	0,00	0,41	0,0829	TTC39B
ENST00000366953	0,48	0,43	0,90	0,49	0,0831	SLC22A2*
ENST00000527372	0,59	0,51	0,62	0,59	0,0831	MYO18A
ENST00000372054	0,08	0,00	NA	0,08	0,0831	GNG5P2
ENST00000341083	0,43	0,52	0,00	0,42	0,0831	BEND7
ENST00000398721	0,00	0,00	0,00	0,00	0,0831	DEFB133
ENST00000216268	0,52	0,71	NA	0,53	0,0831	ZBED4
ENST00000262366	0,18	0,00	1,66	0,21	0,0836	GLIS2
ENST00000371225	0,22	0,00	NA	0,21	0,0836	TACSTD2
ENST00000217426	0,34	0,00	0,00	0,31	0,0836	AHCY
ENST00000416569	0,45	1,37	0,00	0,49	0,0837	XKR6
ENST00000372343	0,45	0,81	0,48	0,47	0,0837	IPO13
ENST00000250003	0,39	0,54	0,00	0,39	0,0837	MYOD1
ENST00000593685	0,42	0,81	0,00	0,42	0,0837	DYRK1B
ENST00000266718	4,81	6,72	12,41	4,97	0,0837	LUM
ENST00000239243	0,00	0,00	0,00	0,00	0,0837	MSX2
ENST00000436066	0,29	0,00	NA	0,28	0,0837	PLEKHF1
ENST00000381569	0,52	0,56	0,00	0,52	0,0839	LZTS1
ENST00000373493	0,38	0,00	0,00	0,34	0,0840	RBBP4
ENST00000352732	0,45	0,00	0,00	0,40	0,0846	ABHD2

ENST00000267291	0,35	0,00	0,00	0,33	0,0847	RNF113B
ENST00000380067	0,17	0,00	0,00	0,16	0,0847	SLC25A35
ENST00000310823	0,41	0,53	0,27	0,41	0,0847	ADAM17
ENST00000233840	0,42	0,90	0,00	0,44	0,0853	NEU2
ENST00000284770	0,34	1,24	0,37	0,38	0,0858	PDLIM3
ENST00000244314	0,49	0,28	NA	0,48	0,0858	IRGC
ENST00000406855	0,08	0,00	0,00	0,08	0,0858	DERL3
ENST00000372190	0,56	0,35	0,31	0,54	0,0858	RAPGEF1
ENST00000373921	0,41	0,68	0,77	0,43	0,0859	THEMIS2
ENST00000373048	0,53	0,61	0,81	0,55	0,0859	EPHA10
ENST00000254337	0,30	0,53	0,88	0,34	0,0861	DCAF15
ENST00000308020	0,32	0,00	0,44	0,31	0,0862	PTDSS2
ENST00000494426	0,07	0,00	1,12	0,12	0,0865	CLIC3
ENST00000373451	0,41	0,61	0,63	0,43	0,0871	CMTR1
ENST00000241125	0,34	0,54	NA	0,35	0,0871	GJA3
ENST00000343090	0,57	0,84	0,38	0,57	0,0871	UBE4B
ENST00000251582	0,61	0,52	0,51	0,61	0,0871	ADAMTS2
ENST00000522917	0,53	0,90	1,00	0,57	0,0873	FER1L6
ENST00000272224	0,33	0,00	0,00	0,32	0,0873	GDF7
ENST00000319098	0,32	0,00	NA	0,30	0,0876	PSAPL1
ENST00000406477	0,32	0,97	0,48	0,36	0,0883	PARVB
ENST00000334379	0,34	0,00	0,35	0,32	0,0883	TRMT11
ENST00000205890	0,69	0,62	0,33	0,67	0,0883	MYO15A
ENST00000392278	0,30	0,00	0,00	0,28	0,0885	C19orf12
ENST00000565488	0,33	0,38	0,47	0,34	0,0899	SLC25A24
ENST00000222800	0,08	0,00	0,00	0,08	0,0899	ABHD11
ENST00000450189	0,43	0,33	0,50	0,43	0,0903	CELF2
ENST00000245105	0,54	1,37	0,48	0,57	0,0908	SH3TC1
ENST00000256339	0,69	1,05	0,54	0,70	0,0908	UNC79
ENST00000350881	0,46	0,59	0,23	0,45	0,0908	THBS4
ENST00000449022	0,51	0,76	0,82	0,54	0,0909	CADPS2
ENST00000371372	0,66	0,31	1,01	0,66	0,0911	CACNA1B
ENST00000258106	0,29	0,00	0,00	0,27	0,0911	EMX1
ENST00000543196	0,45	0,00	0,76	0,44	0,0913	GALNT6
ENST00000262189	1,12	8,03	0,93	1,51	0,0913	KMT2C*
ENST00000252242	0,49	1,04	0,83	0,52	0,0919	KRT5
ENST00000297954	0,60	0,38	0,22	0,58	0,0919	WNK2
ENST00000293805	0,35	0,44	1,47	0,40	0,0920	BCL6B
ENST00000248901	0,37	0,65	0,78	0,42	0,0920	CYTH4
ENST00000290399	0,38	1,15	0,00	0,40	0,0920	SIM2
ENST00000265441	0,34	0,46	0,00	0,34	0,0920	WNT2
ENST00000399599	0,09	0,00	0,00	0,08	0,0920	CDIP1
ENST00000216180	0,38	0,73	1,10	0,43	0,0928	PNPLA3
ENST00000322810	0,74	0,85	0,41	0,74	0,0928	PLEC
ENST00000280155	0,42	0,00	NA	0,40	0,0928	ADRA2A
ENST00000393946	0,44	0,00	0,00	0,40	0,0928	SYT12
ENST00000316754	0,38	0,53	0,98	0,41	0,0929	RHOJ
ENST00000390609	0,45	0,00	0,00	0,41	0,0929	no _symbol
ENST00000280057	0,44	0,70	0,00	0,44	0,0930	FAM124A
ENST00000264605	0,44	0,33	0,00	0,41	0,0930	MLPH

ENST00000287878	0,40	0,40	0,26	0,39	0,0930	PRKAG2
ENST00000264234	0,21	0,65	0,59	0,26	0,0932	UPK1B
ENST00000171111	2,14	3,68	1,18	2,20	0,0934	KEAP1*
ENST00000215980	0,14	0,00	0,00	0,13	0,0934	CENPM
ENST00000361813	0,50	0,89	0,67	0,53	0,0934	SMG5
ENST00000394729	0,46	1,03	0,00	0,46	0,0937	PRKCD
ENST00000354475	0,47	0,29	0,00	0,44	0,0937	WDR91
ENST00000398389	0,45	0,00	0,22	0,41	0,0939	MPP3
ENST00000298047	0,77	1,25	0,68	0,79	0,0951	FAT3*
ENST00000252999	0,62	0,70	0,69	0,62	0,0951	LAMA5
ENST00000378247	0,44	0,00	0,87	0,44	0,0951	SPIRE2
ENST00000335290	0,35	0,49	0,00	0,35	0,0951	WDR25
ENST00000278742	0,48	0,26	0,65	0,48	0,0955	ST14
ENST00000292672	0,40	0,53	0,00	0,38	0,0955	CELF5
ENST00000419348	0,18	1,36	0,71	0,27	0,0958	HTATIP2
ENST00000327111	0,44	0,30	2,25	0,46	0,0958	NR2F1
ENST00000561311	0,63	0,70	0,83	0,64	0,0962	TLN2
ENST00000360184	0,69	1,00	0,38	0,69	0,0963	DYNC1H1
ENST00000285398	0,52	0,55	0,22	0,51	0,0963	ERCC3*
ENST00000372262	0,51	0,52	1,68	0,55	0,0963	CDH22
ENST00000392476	0,44	0,69	0,89	0,47	0,0963	SEC14L1
ENST00000389740	0,55	0,44	2,16	0,57	0,0963	GPR149
ENST00000368130	0,42	0,83	0,00	0,43	0,0963	AIM2
ENST00000272238	0,31	0,54	0,00	0,30	0,0963	ATP6V1C2
ENST00000378681	0,36	0,37	0,00	0,33	0,0963	UCMA
ENST00000255039	0,28	0,00	0,00	0,26	0,0963	HAPLN2
ENST00000246080	0,36	0,00	0,00	0,34	0,0964	TCF15
ENST00000265825	0,39	0,34	1,20	0,41	0,0969	FSCN3
ENST00000315033	0,21	0,00	NA	0,20	0,0969	GPR88
ENST00000359271	0,49	0,00	0,70	0,48	0,0969	SLC2A10
ENST00000377617	0,51	0,63	0,61	0,52	0,0970	ATXN2
ENST00000380752	0,48	0,55	0,57	0,49	0,0974	SLC7A1
ENST00000343677	0,50	0,00	NA	0,48	0,0974	HIST1H1C*
ENST00000375377	0,61	0,46	1,61	0,60	0,0977	KIAA1462
ENST00000258930	0,18	1,05	0,00	0,20	0,0978	CIB2
ENST00000360001	0,31	0,00	0,74	0,31	0,0978	SDF4
ENST00000246186	0,25	0,49	2,89	0,35	0,0980	MMP24
ENST00000347598	0,63	0,86	0,89	0,66	0,0980	CACNA1C
ENST00000377962	0,40	0,64	0,00	0,39	0,0980	LECT1
ENST00000376503	0,51	0,71	0,26	0,50	0,0980	SLC15A1
ENST00000440886	0,42	0,00	0,00	0,40	0,0980	LPAR3
ENST00000532870	0,17	0,00	0,00	0,16	0,0980	TAGLN
ENST00000292823	0,34	0,42	0,45	0,35	0,0985	PCYT1A
ENST00000333407	0,41	0,43	0,00	0,40	0,0985	FAM83F
ENST00000430479	0,50	0,78	0,00	0,48	0,0989	ETV1*
ENST00000260356	0,45	0,55	0,58	0,46	0,0995	THBS1

Table 15.1: List of significant genes under selection. Asterisk indicates that this gene is part of Cancer Gene Census or a published list of driver genes [Xie et al., 2014].

Part V

Global Discussion

> ”Nothing in Biology Makes Sense Except in the Light of Evolution”
>
> *Theodosius Dobzhansky*

The three parts of this thesis encompas different aspects of cancer evolution, a vision first proposed in the 1970s by Peter Nowell. In his seminal work, he described the acquisition of mutations as a step-wise process where advantageous genotypes lead to increased proliferation, clonal expansion, and ultimately, cancer development [Nowell, 1976]. In our first project (part II), we studied a chronic lymphocytic leukemia patient over an 11-year period, where he displayed different cancer clones coexisting and evolving together. In the second project (part III), we identified cancer driver genes based on the signatures of positive selection imprinted on cancer genomes. Lastly (part IV), we have presented evidence of purifying selection on the evolution of tumors, as well as how we can exploit such a process to identify novel therapeutic targets for cancer treatment.

The results presented in part II show the importance of longitudinal genomic analysis, where multiple time-points of the same individual are sequenced. This analysis revealed the dynamics of tumor evolution where the initial cancer clone had to temporarily coexist with a ”daugther” clone. The daughter clone harbored a catastrophic event thought to be related to tumor initiation, chromothripsis[Campbell et al., 2010]. Nonetheless, chromothripsis was present only after diagnosis, since in the time-point collected after treatment it was completely absent. In this work, we developed a mathematical formula to calculate the clonal fraction (CCF) of somatic single nucleotide variants (SNVs) and copy-number altered (CNA) regions in a single cancer population. We estimate CCF using allele frequencies of SNVs

observed from whole-genome, exome and targeted re-sequencing data. Thus, we show that 40% of chromothripsis-cancer cells were present in the 2002 sample, implying that chromothripsis was not the cancer driver event. In addition, we show that it was the founder clone that was resistant to the treatment, provoking the relapse of the disease [Bassaganyas et al., 2013].

The dynamics of intratumor heterogeneity uncovered through sequencing revealed a strong effect of selection acting at the cellular level. While this selection level was proposed decades ago [Lewontin, 1970], only the molecular and individual levels of selection have been exploited to detect driver genes. Nuria Lopez's group has used molecular signals of driver mutations [Tamborero et al., 2013c], such as functional impact [Gonzalez-Perez and Lopez-Bigas, 2012] and position-based clustering [Tamborero et al., 2013d], to uncover mutational driver genes. Other algorithms exploiting the recurrence of somatic mutations have been developed [Lawrence et al., 2013][Dees et al., 2012][Gonzalez-Perez et al., 2013], nevertheless all of these non-probabilistic models disregard cancer cell fraction as a signature to identify driver genes. To fill this conceptual gap, we implement cDriver, a Bayesian approach that integrates multiple signatures of positive selection to detect cancer driver genes.

Studying cancer from an evolutionary perspective brings to mind the discussion introduced by Richard Dawkins in his popular book, "The selfish gene" [Dawkins et al., 2016]. He defined genes as the replicators where selection is acting upon, placing the organism as a mere vehicle of transmission. In cancer, the relationship between the gene and the organism has to be extended to include the cell [Merlo et al., 2006]. Unlike Dawkins, we believe that multiple levels of selection should be considered when studying the evolutionary strategies that cancer cells adopt to proliferate. Selection at the cellular level is evident when a cell undergoes a clonal expansion given by a selective advantage. Therefore, somatic mutations found in driver genes should be at higher cancer cell fraction than in passenger genes, an observation also noted by other groups [McGranahan et al., 2015].

In the first section of part II, we demonstrated that CCF of somatic mutations is a key signature for prediction of cancer driver genes. cDriver integrates CCF, population recurrence, and functional impact of somatic mutations, allowing us to increase the precision on the identification of gold standard genes. We benchmarked cDriver against four canonical methods for driver gene prediction used by the TCGA consortium. We demonstrated that cDriver outperformed all competing methods in terms of precision, recall and F-score, therefore increasing the chance of finding prognostic markers in a patient.

The second section of the project addressed the problem of driver genes mutated in a low fraction of cancer patients, in other words, inter-patient heterogeneity. Patient-specific genotypes may change the outcome of somatic mutations and therefore the options for tumor initiation [Yates and Campbell, 2012]. Thus, it is not surprising that heterogeneity is so common. For some tumor types, such as chronic lymphocytic leukemia, all currently identified driver genes are found mutated in less than 40% of patients. Recurrence-based models have failed to identify driver genes present only in 1% to 5% of cancer patients [Vogelstein et al., 2013]. Moreover, known driver genes frequently mutated in one tumor type are often neglected in other cancers. To tackle such problems, we applied our algorithm to 6,870 exomes spanning 21 tumor types to uncover a "tumor type-driver gene" (TTDG) landscape. We identified 670 TTDG high confidence connections harboring 234 driver genes. Next, by text mining PubMed, we showed that chromatin modifiers have been largely underrepresented in current studies. Furthermore, we showed that for every tumor type, at least one chromatin modifying protein had prognostic relevance.

Next, we reasoned that if positive selection is acting on cancer genomes at multiple levels, then negative selection should also be present. We observed a very interesting case of purifying selection in the first story of this thesis. The CLL patient had a chromothripsis-harboring clone that disappeared after the patient underwent chemotherapy in 2006 [Bassaganyas et al., 2013]. The treatment acted as a negative selector for this clone, likely due to a high susceptibility given by the intrinsic chromosome instability of chromothripsis. We think that identifying genes and functions conferring clonal susceptibility/resistance to treatment will be key to improve a rational design of drug combination therapies [Zhao et al., 2016].

Previous studies have claimed that negative selection is relaxed in cancer genomes [Ostrow et al., 2014][Woo and Li, 2012] [Beckman and Loeb, 2005]. This conclusion has largely come from the comparison of selective values of germline versus somatic variation. We believe that conclusions made from comparing germline K_n/K_s to somatic K_n/K_s should be taken cautiously. A reason for the latter claim is that the time scale for selection and fixation to act upon cancer genomes is different from inter-species evolution [Rocha et al., 2006], forcing us to adapt old models. Furthermore, most previous studies have neglected subclonal variation, so estimates of K_n/K_s are based on clonal mutations, most likely drivers, which in turn are biased towards positive selection.

Nonetheless, other studies have demonstrated that cancer related genes are under stronger purifying selection than non-cancer genes [Thomas et al., 2003] [Ovens and Naugler, 2012] [Pyatnitskiy et al., 2015], and that substitution rates of cancer

genes are different from other genes [Gojobori and Yokoyama, 1987]. It is important to note that these studies have been performed in small datasets and only in one tumor type. Current next generation sequencing technologies have provided us with a large amount of data from thousands of cancer patients to successfully test the extent of negative selection. We found that negative selection is 10 times more prevalent than positive selection when comparing the number of genes affected. To conclusively test the extent of cancer specific negative selection, we still need to create a catalog of somatic K_n/K_s values by sequencing healthy tissues of many individuals.

Our last project has uncovered a plethora of genes under strong purifying selection across multiple tumors. Among these, two of them have been extensively studied in cancer, *ABL1* and *TERT*. Both are linked to tumorigenesis due to positive selection of non-coding activating mutations, revealing that the interplay between negative and positive selection forces is fundamental for the cell to evolve into a malignant entity. We also identified a promising signaling complex, *P2X7*, with prognostic value across multiple tumor types. Currently, many other genes remain to be explored in the light of negative selection.

To achieve a model of evolution of a cancer cell, we need to account for multilevel selection, and expand our notion of cancer-related genes. A full catalog of cancer-related genes requires the systematic identification of negative, positive, and neutral selection forces. In other words, we must embrace evolution to understand the processes underlying cancer etiology, development, and metastasis, thus enabling us to deploy personalized treatment strategies, and eventually eradicate cancer.

Future perspectives

In the future, modeling the direction and the levels of selection in cancer genomes will allow us to determine cancer essential genes [Merlo et al., 2006]. Figure 15.1 shows six molecular evolutionary signatures that can be detected in cancer genomic data. We propose that natural selection is acting on: (a) activating non-silent mutations/translocations in oncogenes (e.g. *ABL1*); (b) Inactivating non-silent mutations/alterations in tumor suppressor genes (e.g. *TP53*); (c) Increased expression given by enhanced transcription (e.g. *TERT* [Smith et al., 2015]) or dosage increase (e.g. *MYC*, [Meyer and Penn, 2008]); (d) Reduced expression given by regulatory variants [Heyn, 2016] or deletions; (e) Functional silent mutations such as splicing modifiers [Supek et al., 2014]; and (f) non-silent mutations related to cancer essential functions. [Davoli et al., 2013] have explored the differences of signatures (c) and (d), defining them as haploinsuficiency of tumor

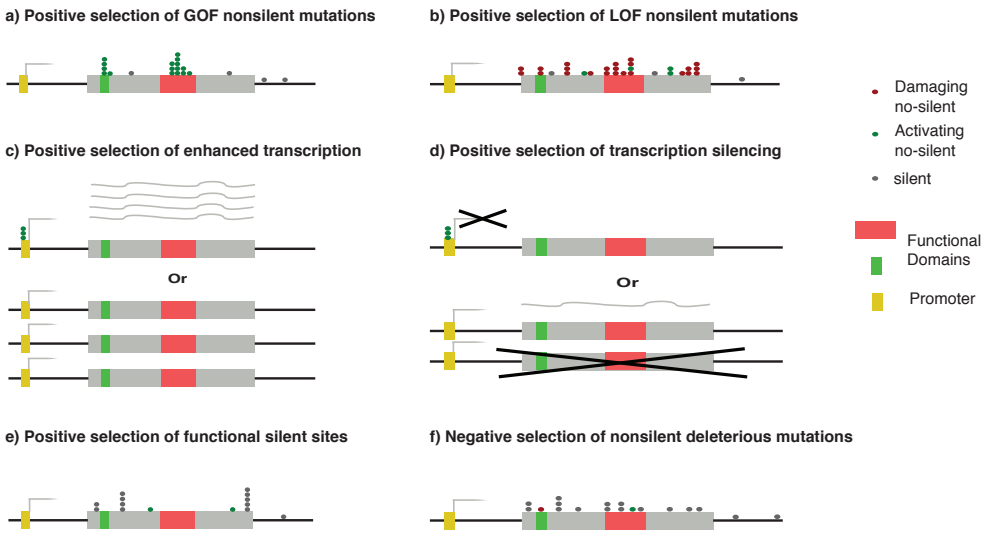

Figure 15.1: **Models of evolutionary forces acting upon cancer genes.** (a) Positive selection of Gain of function non-silent mutations. (b) Positive selection of Loss of function non-silent mutations. (c) Positive selection of enhanced transcription given by copy number alterations or regulatory modifications. (d) Positive selection of transcription silencing. (e) Positive selection of functional silent sites and (f) Negative or purifying selection of deleterious mutations.

suppressor genes and triple sensitiveness of oncogenes, respectively.

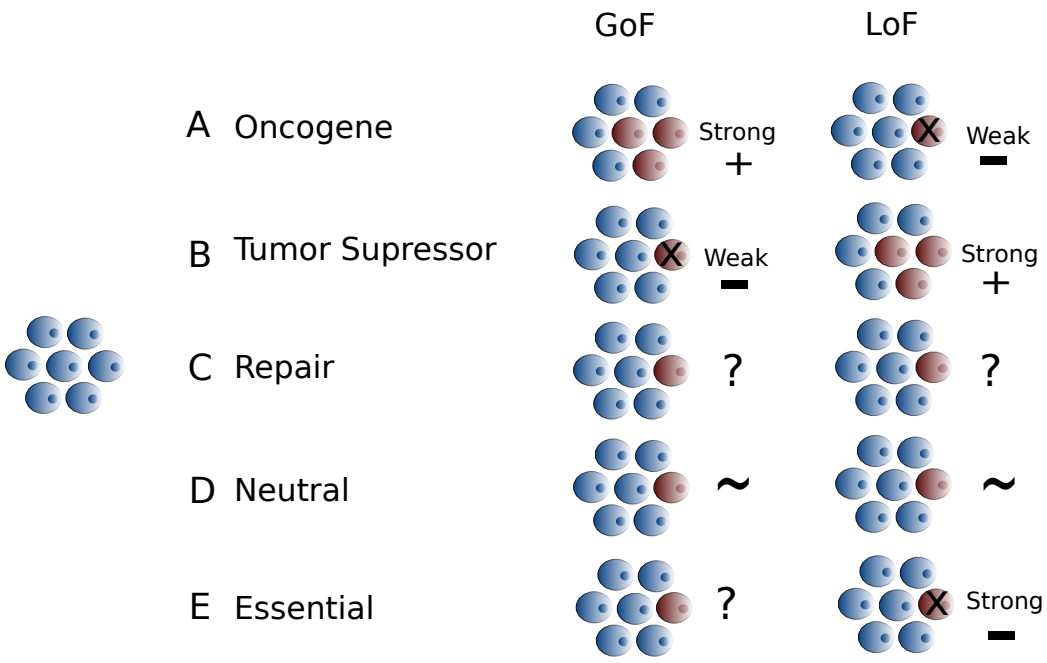

Figure 15.2: **Somatic mutations and selection.** A normal cell acquires a somatic mutation in one of five gene types. Selection strength and direction vary depending on the gene type and mutation type (GoF, gain of function or LoF, Loss of function).

In addition, we extend the previous model to include more underlying mutation types on five gene locus categories (Fig 15.2): (i) Oncogenes, whose function is to regulate the birth rate of cells, (ii) Tumor suppressor genes that regulate the death rate of cells, (iii) Repair genes that regulate the mutation rate of cells, (iv) Neutral genes that are not relevant for the cell at a particular time, but may confer a selective advantage/disadvantage later, (v) Cancer essential genes that regulate metabolism, signaling, translation, transcription (i.e. essential processes for the proper functioning of the cell). We can model the mutational trajectories for tumor initiation based on these five genetic entities interacting with activating and inactivating (GOF and LOF) mutations under the selection umbrella.

For example, the genotype $A^{GoF}B^{LoF}C^{LoF}DE$ represents the typical case of *KRAS* activated, *TP53* inactivated, and a faulty mismatch repair gene. In locus C, LoF mutations are not initially under direct selection because they only increase mutation rate, allowing faster exploration of the fitness landscape. Once

the malignancy has been established (i.e. the clonal population has reached a fitness peak), the accumulation of LoF mutations in repair genes is most likely under weak negative selection. Interestingly, our initial aim was to identify signatures of E^{LoF}, but we observed that, for instance, acquiring genotype A^{GoF} from a translocation increases the strength of selection for A^{LoF} mutations (gene *ABL1*). Therefore, by understanding the temporal order of acquisition of mutations using the aforementioned 5-loci model, we aim to improve the explanation behind the causes of cancer initiation and development.

Finally, we would like to propose that an effective treatment cannot be designed by only targeting driver genes, since at least one cancer cell will carry a resistant mutation. Ergo, the quest for a successful universal treatment strictly relies on targeting cancer essential functions, where, in principle, no malignant cell is able to survive the treatment. We propose that such functions can be uncovered by exploring the negatively selected genes identified in this thesis. Furthermore, to specifically target cancer cells, we can take advantage of the somatically acquired mutations in the common ancestor of the tumor clones. Once we can selectively act against the tumor, we can successfully knockdown the essential functions of the cell. Essentially, we want to recapitulate evolution in a way where we impose negative selection as the main force directing the evolution of the cancer genome.

Barcelona,
July 2016.

Part VI

Appendix

Chapter 16

List of publications during PhD studies

Genomics

1. Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J.R., Rawat, V., Rowan, B.A., de Jong, D.F., de Snoo, C.B., **Zapata, L.**, Huettel, B., de Jong, H., Ossowski, S., Weigel, D., Koornneef, M., Keurentjes, J.J. & Schneeberger, K., 2013, The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana, eLife, 2, p. e01426

2. Willmann, M., Bezdan, D., **Zapata, L.**, Susak, H., Vogel, W., Schrppel, K., Liese, J., Weidenmaier, C., Autenrieth, I.B. & Ossowski, S., 2015, Analysis of a long-term outbreak of XDR Pseudomonas aeruginosa: a molecular epidemiological study, Journal of Antimicrobial Chemotherapy, p. dku546

3. Henaff, E., **Zapata, L.**, Casacuberta, J.M. & Ossowski, S., 2015, Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution, BMC genomics, 16(1)

4. **Zapata, L.**, Ding, J., Willing, E.M., Hartwig, B., Bezdan, D., Jiao, W.B., Pa-

tel, V., Velikkakam James, G., Koornneef, M., Ossowski, S. & Schneeberger, K., 2016, Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms, Proceedings of the National Academy of Sciences of the United States of America. (Original copy attached)

Cancer

5. Bassaganyas, L., Be, S., Escaramis, G., Tornador, C., Salaverria, I., **Zapata, L.**, Drechsel, O., Ferreira, P.G., Rodriguez-Santiago, B., Tubio, J.M., Navarro, A., Martin-Garcia, D., Lopez, C., Martinez-Trillos, A., Lopez-Guillermo, A., Gut, M., Ossowski, S., Lopez-Otin, C., Campo, E. & Estivill, X., 2013, Sporadic and reversible chromothripsis in chronic lymphocytic leukemia revealed by longitudinal genomic analysis, Leukemia

6. Prasad, A., Rabionet, R., Espinet, B., **Zapata, L.**, Puiggros, A., Melero, C., Puig, A., Sarria-Trujillo, Y., Ossowski, S. & Garcia-Muret, M.P., 2016, Identification of Gene Mutations and Fusion Genes in Patients with Sezary Syndrome, Journal of Investigative Dermatology

Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, Patel V, Velikkakam James G, Koornneef M, Ossowski S, Schneeberger K. Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci U S A. 2016 Jul 12;113(28):E4052-60. doi: 10.1073/pnas.1607532113.

Bibliography

[Adinolfi et al., 2002] Adinolfi, E., Melchiorri, L., Falzoni, S., Chiozzi, P., Morelli, A., Tieghi, A., Cuneo, A., Castoldi, G., Di Virgilio, F., and Baricordi, O. R. (2002). P2x7 receptor expression in evolutive and indolent forms of chronic b lymphocytic leukemia. *Blood*, 99(2):706–708.

[Akavia et al., 2010] Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., and Pe'er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–17.

[Alexandrov et al., 2013] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Borresen-Dale, A.-L. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., and Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–21.

[Almendro et al., 2014] Almendro, V., Cheng, Y.-K. K., Randles, A., Itzkovitz, S., Marusyk, A., Ametller, E., Gonzalez-Farre, X., Muñoz, M., Russnes, H. G., Helland, A., Rye, I. H., Borresen-Dale, A.-L. L., Maruyama, R., van Oudenaarden, A., Dowsett, M., Jones, R. L., Reis-Filho, J., Gascon, P., Gönen, M., Michor, F., and Polyak, K. (2014). Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Rep*, 6(3):514–27.

[Altrock et al., 2015] Altrock, P. M., Liu, L. L., and Michor, F. (2015). The mathematics of cancer: integrating quantitative models. *Nat Rev Cancer*, 15(12):730–45.

[Álvarez Silva et al., 2015] Álvarez Silva, M. C., Yepes, S., Torres, M. M., and Barrios, A. F. G. (2015). Proteins interaction network and modeling of igvh mutational status in chronic lymphocytic leukemia. *Theor Biol Med Model*, 12:12.

[Andor et al., 2013] Andor, N., Harness, J., Mewes, H. W., and Petritsch, C. (2013). Expands: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, page btt622.

[Armitage and Doll, 1954] Armitage, P. and Doll, R. (1954). The age distribution of cancer and a multi-stage theory of carcinogenesis. *British journal of cancer*, 8(1):1.

[Ashley, 1969] Ashley, D. J. (1969). The two" hit" and multiple" hit" theories of carcinogenesis. *British journal of cancer*, 23(2):313.

[Babenko et al., 2006] Babenko, V. N., Basu, M. K., Kondrashov, F. A., Rogozin, I. B., and Koonin, E. V. (2006). Signs of positive selection of somatic mutations in human cancers detected by est sequence analysis. *BMC Cancer*, 6:36.

[Bailey et al., 2016] Bailey, P., Chang, D. K., Nones, K., Johns, A. L., Patch, A.-M. M., Gingras, M.-C. C., Miller, D. K., Christ, A. N., Bruxner, T. J. C., Quinn, M. C., Nourse, C., Murtaugh, L. C., Harliwong, I., Idrisoglu, S., Manning, S., Nourbakhsh, E., Wani, S., Fink, L., Holmes, O., Chin, V., Anderson, M. J., Kazakoff, S., Leonard, C., Newell, F., Waddell, N., Wood, S., Xu, Q., Wilson, P. J., Cloonan, N., Kassahn, K. S., Taylor, D., Quek, K., Robertson, A., Pantano, L., Mincarelli, L., Sanchez, L. N., Evers, L., Wu, J., Pinese, M., Cowley, M. J., Jones, M. D., Colvin, E. K., Nagrial, A. M., Humphrey, E. S., Chantrill, L. A., Mawson, A., Humphris, J., Chou, A., Pajic, M., Scarlett, C. J., Pinho, A. V., Giry-Laterriere, M., Rooman, I., Samra, J. S., Kench, J. G., Lovell, J. A., Merrett, N. D., Toon, C. W., Epari, K., Nguyen, N. Q., Barbour, A., Zeps, N., Moran-Jones, K., Jamieson, N. B., Graham, J. S., Duthie, F., Oien, K., Hair, J., Grützmann, R., Maitra, A., Iacobuzio-Donahue, C. A., Wolfgang, C. L., Morgan, R. A., Lawlor, R. T., Corbo, V., Bassi, C., Rusev, B., Capelli, P., Salvia, R., Tortora, G., Mukhopadhyay, D., Petersen, G. M., Munzy, D. M., Fisher, W. E., Karim, S. A., Eshleman, J. R., Hruban, R. H., Pilarsky, C., Morton, J. P., Sansom, O. J., Scarpa, A., Musgrove, E. A., Bailey, U.-M. H. M., Hofmann, O., Sutherland, R. L., Wheeler, D. A., Gill, A. J., Gibbs, R. A., Pearson, J. V., Waddell, N., Biankin, A. V., and Grimmond, S. M. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52.

[Balmain, 2001] Balmain, A. (2001). Cancer genetics: from boveri and mendel to microarrays. *Nature Reviews Cancer*, 1(1):77–82.

[Bassaganyas et al., 2013] Bassaganyas, L., Beà, S., Escaramís, G., Tornador, C., Salaverria, I., Zapata, L., Drechsel, O., Ferreira, P. G., Rodriguez-Santiago, B., Tubio, J. M. C., Navarro, A., Martín-García, D., López, C., Martínez-Trillos, A., López-Guillermo, A., Gut, M., Ossowski, S., López-Otín, C., Campo, E., and Estivill, X. (2013). Sporadic and reversible chromothripsis in chronic lymphocytic leukemia revealed by longitudinal genomic analysis. *Leukemia*.

[Beckman and Loeb, 2005] Beckman, R. A. and Loeb, L. A. (2005). Negative clonal selection in tumor evolution. *Genetics*, 171(4):2123–31.

[Bentley et al., 2008] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R.,

Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9.

[Bertrand et al., 2015] Bertrand, D., Chng, K. R., Sherbaf, F. G., Kiesel, A., Chia, B. K. H., Sia, Y. Y., Huang, S. K., Hoon, D. S. B., Liu, E. T., Hillmer, A., and Nagarajan, N. (2015). Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res*.

[Biegel et al., 2014] Biegel, J. A., Busse, T. M., and Weissman, B. E. (2014). Swi/snf chromatin remodeling complexes and cancer. *Am J Med Genet C Semin Med Genet*, 166C(3):350–66.

[Billaud and Santoro, 2011] Billaud, M. and Santoro, M. (2011). Is co-option a prevailing mechanism during cancer progression? *Cancer Res*, 71(21):6572–5.

[Bissell and Hines, 2011] Bissell, M. J. and Hines, W. C. (2011). Why don't we get more cancer? a proposed role of the microenvironment in restraining cancer progression. *Nature medicine*, 17(3):320–329.

[Bolli et al., 2014] Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., Dawson, K. J., Iorio, F., Nik-Zainal, S., Bignell, G. R., Hinton, J. W., Li, Y., Tubio, J. M. C., McLaren, S., O' Meara, S., Butler, A. P., Teague, J. W., Mudie, L., Anderson, E., Rashid, N., Tai, Y.-T. T., Shammas, M. A., Sperling, A. S., Fulciniti, M., Richardson, P. G., Parmigiani, G., Magrangeas, F., Minvielle, S., Moreau, P., Attal, M., Facon, T., Futreal, P. A., Anderson, K. C., Campbell, P. J., and Munshi, N. C. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun*, 5:2997.

[Boveri, 1914] Boveri, T. (1914). *Zur frage der entstehung maligner tumoren*. Gustav Fischer.

[Braggio et al., 2012] Braggio, E., Kay, N. E., Vanwier, S., Tschumper, R. C., Smoley, S., Eckel-Passow, J. E., Sassoon, T., Barrett, M., Van Dyke, D. L., and Byrd, J. C. (2012). Longitudinal genome-wide analysis of patients with

chronic lymphocytic leukemia reveals complex evolution of clonal architecture at disease progression and at the time of relapse. *Leukemia*, 26(7):1698.

[Cahill et al., 1999] Cahill, D. P., Kinzler, K. W., Vogelstein, B., and Lengauer, C. (1999). Genetic instability and darwinian selection in tumours. *Trends in cell biology*, 9(12):M57–M60.

[Cai et al., 2014] Cai, Y., Geutjes, E. J., De Lint, K., Roepman, P., Bruurs, L., Yu, L. R., Wang, W., van Blijswijk, J., Mohammad, H., and de Rink, I. (2014). The nurd complex cooperates with dnmts to maintain silencing of key colorectal tumor suppressor genes. *Oncogene*, 33(17):2157–2168.

[Campbell et al., 2010] Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., McLaren, S., Lin, M.-L. L., McBride, D. J., Varela, I., Nik-Zainal, S. A., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Griffin, C. A., Burton, J., Swerdlow, H., Quail, M. A., Stratton, M. R., Iacobuzio-Donahue, C., and Futreal, P. A. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–13.

[Campisi and d'Adda di Fagagna, 2007] Campisi, J. and d'Adda di Fagagna, F. (2007). Cellular senescence: when bad things happen to good cells. *Nature Reviews Molecular Cell Biology*, 8(9):729–740.

[Carter et al., 2012] Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., and Getz, G. (2012). Absolute quantification of somatic dna alterations in human cancer. *Nat Biotechnol*, 30(5):413–21.

[Castro-Giner et al., 2015] Castro-Giner, F., Ratcliffe, P., and Tomlinson, I. (2015). The mini-driver model of polygenic cancer evolution. *Nat Rev Cancer*, 15(11):680–5.

[Chabner and Roberts, 2005] Chabner, B. A. and Roberts, T. G. (2005). Timeline: Chemotherapy and the war on cancer. *Nat Rev Cancer*, 5(1):65–72.

[Charlesworth, 2009] Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, 10(3):195–205.

[Cheung et al., 2011] Cheung, H. W., Cowley, G. S., Weir, B. A., Boehm, J. S., Rusin, S., Scott, J. A., East, A., Ali, L. D., Lizotte, P. H., and Wong, T. C. (2011). Systematic investigation of genetic vulnerabilities across cancer cell

lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences*, 108(30):12372–12377.

[Ch'ng and Kumanogoh, 2010] Ch'ng, E. S. and Kumanogoh, A. (2010). Roles of sema4d and plexin-b1 in tumor progression. *Molecular Cancer*, 9(1):1.

[Chudnovsky et al., 2014] Chudnovsky, Y., Kim, D., Zheng, S., Whyte, W. A., Bansal, M., Bray, M.-A. A., Gopal, S., Theisen, M. A., Bilodeau, S., Thiru, P., Muffat, J., Yilmaz, O. H., Mitalipova, M., Woolard, K., Lee, J., Nishimura, R., Sakata, N., Fine, H. A., Carpenter, A. E., Silver, S. J., Verhaak, R. G. W., Califano, A., Young, R. A., Ligon, K. L., Mellinghoff, I. K., Root, D. E., Sabatini, D. M., Hahn, W. C., and Chheda, M. G. (2014). Zfhx4 interacts with the nurd core member chd4 and regulates the glioblastoma tumor-initiating cell state. *Cell Rep*, 6(2):313–24.

[Cibulskis et al., 2013] Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*.

[Cooper et al., 2015] Cooper, C. S., Eeles, R., Wedge, D. C., Van Loo, P., Gundem, G., Alexandrov, L. B., Kremeyer, B., Butler, A., Lynch, A. G., Camacho, N., Massie, C. E., Kay, J., Luxton, H. J., Edwards, S., Kote-Jarai, Z., Dennis, N., Merson, S., Leongamornlert, D., Zamora, J., Corbishley, C., Thomas, S., Nik-Zainal, S., Ramakrishna, M., O'Meara, S., Matthews, L., Clark, J., Hurst, R., Mithen, R., Bristow, R. G., Boutros, P. C., Fraser, M., Cooke, S., Raine, K., Jones, D., Menzies, A., Stebbings, L., Hinton, J., Teague, J., McLaren, S., Mudie, L., Hardy, C., Anderson, E., Joseph, O., Goody, V., Robinson, B., Maddison, M., Gamble, S., Greenman, C., Berney, D., Hazell, S., Livni, N., Fisher, C., Ogden, C., Kumar, P., Thompson, A., Woodhouse, C., Nicol, D., Mayer, E., Dudderidge, T., Shah, N. C., Gnanapragasam, V., Voet, T., Campbell, P., Futreal, A., Easton, D., Warren, A. Y., Foster, C. S., Stratton, M. R., Whitaker, H. C., McDermott, U., Brewer, D. S., and Neal, D. E. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet*, 47(4):367–72.

[Cooper Geoffrey, 2000] Cooper Geoffrey, M. (2000). The cell a molecular approach.

[Darwin, 1872] Darwin, C. (1872). *The origin of species*. Lulu. com.

[Davoli et al., 2013] Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., and Elledge, S. J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962.

[Dawkins et al., 2016] Dawkins, R. et al. (2016). *The selfish gene*. Oxford university press.

[Dees et al., 2012] Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K., and Ding, L. (2012). Music: identifying mutational significance in cancer genomes. *Genome Res*, 22(8):1589–98.

[Di Virgilio and Adinolfi, 2016] Di Virgilio, F. and Adinolfi, E. (2016). Extracellular purines, purinergic receptors and tumor growth. *Oncogene*.

[Dieci et al., 2016] Dieci, M. V., Smutná, V., Scott, V., Yin, G., Xu, R., Vielh, P., Mathieu, M.-C. C., Vicier, C., Laporte, M., Drusch, F., Guarneri, V., Conte, P., Delaloge, S., Lacroix, L., Fromigué, O., André, F., and Lefebvre, C. (2016). Whole exome sequencing of rare aggressive breast cancer histologies. *Breast Cancer Res Treat*, 156(1):21–32.

[Ding et al., 2010] Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L., Abbott, R. M., Hoog, J., Dooling, D. J., Koboldt, D. C., Schmidt, H., Kalicki, J., Zhang, Q., Chen, L., Lin, L., Wendl, M. C., McMichael, J. F., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Appelbaum, E., Deschryver, K., Davies, S., Guintoli, T., Lin, L., Crowder, R., Tao, Y., Snider, J. E., Smith, S. M., Dukes, A. F., Sanderson, G. E., Pohl, C. S., Delehaunty, K. D., Fronick, C. C., Pape, K. A., Reed, J. S., Robinson, J. S., Hodges, J. S., Schierding, W., Dees, N. D., Shen, D., Locke, D. P., Wiechert, M. E., Eldred, J. M., Peck, J. B., Oberkfell, B. J., Lolofie, J. T., Du, F., Hawkins, A. E., O'Laughlin, M. D., Bernard, K. E., Cunningham, M., Elliott, G., Mason, M. D., Thompson, D. M., Ivanovich, J. L., Goodfellow, P. J., Perou, C. M., Weinstock, G. M., Aft, R., Watson, M., Ley, T. J., Wilson, R. K., and Mardis, E. R. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464(7291):999–1005.

[Ding et al., 2014a] Ding, L., Kim, M., Kanchi, K. L., Dees, N. D., Lu, C., Griffith, M., Fenstermacher, D., Sung, H., Miller, C. A., Goetz, B., Wendl, M. C., Griffith, O., Cornelius, L. A., Linette, G. P., McMichael, J. F., Sondak, V. K., Fields, R. C., Ley, T. J., Mulé, J. J., Wilson, R. K., and Weber, J. S. (2014a).

Clonal architectures and driver mutations in metastatic melanomas. *PLoS One*, 9(11):e111153.

[Ding et al., 2012] Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., and McLellan, M. D. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510.

[Ding et al., 2014b] Ding, L., Wendl, M. C., McMichael, J. F., and Raphael, B. J. (2014b). Expanding the computational toolbox for mining cancer genomes. *Nat Rev Genet*, 15(8):556–70.

[Fabbri and Dalla-Favera, 2016] Fabbri, G. and Dalla-Favera, R. (2016). The molecular pathogenesis of chronic lymphocytic leukaemia. *Nat Rev Cancer*, 16(3):145–62.

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

[Fischer et al., 2014] Fischer, A., Vázquez-García, I., Illingworth, C. J. R., and Mustonen, V. (2014). High-definition reconstruction of clonal composition in cancer. *Cell Rep*.

[Friend et al., 1986] Friend, S. H., Bernards, R. A., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M., and Dryja, T. P. (1986). A human dna segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*, 323(6089):643–646.

[Futreal et al., 2004] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer*, 4(3):177–83.

[Gerlinger et al., 2012] Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., and Tarpey, P. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892.

[Gerstung et al., 2012] Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun*, 3:811.

[Giannuzzo et al., 2015] Giannuzzo, A., Pedersen, S. F., and Novak, I. (2015). The p2x7 receptor regulates cell survival, migration and invasion of pancreatic ductal adenocarcinoma cells. *Molecular cancer*, 14(1):1.

[Gojobori and Yokoyama, 1987] Gojobori, T. and Yokoyama, S. (1987). Molecular evolutionary rates of oncogenes. *Journal of molecular evolution*, 26(1-2):148–156.

[Gonzalez-Perez and Lopez-Bigas, 2012] Gonzalez-Perez, A. and Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Res*, 40(21):e169.

[Gonzalez-Perez et al., 2013] Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R. S., Creixell, P., Karchin, R., Vazquez, M., Fink, J. L., Kassahn, K. S., Pearson, J. V., Bader, G. D., Boutros, P. C., Muthuswamy, L., Ouellette, B. F. F., Reimand, J., Linding, R., Shibata, T., Valencia, A., Butler, A., Dronov, S., Flicek, P., Shannon, N. B., Carter, H., Ding, L., Sander, C., Stuart, J. M., Stein, L. D., and Lopez-Bigas, N. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods*, 10(8):723–9.

[Greenman et al., 2007] Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y.-E. E., DeFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M.-H. H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., and Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–8.

[Greenman et al., 2006] Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R., and Easton, D. F. (2006). Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 173(4):2187–98.

[Greuber et al., 2013] Greuber, E. K., Smith-Pearson, P., Wang, J., and Pendergast, A. M. (2013). Role of abl family kinases in cancer: from leukaemia to solid tumours. *Nat Rev Cancer*, 13(8):559–71.

[Griffith et al., 2015] Griffith, M., Miller, C. A., Griffith, O. L., Krysiak, K., Skidmore, Z. L., Ramu, A., Walker, J. R., Dang, H. X., Trani, L., Larson, D. E., Demeter, R. T., Wendl, M. C., McMichael, J. F., Austin, R. E., Magrini, V., McGrath, S. D., Ly, A., Kulkarni, S., Cordes, M. G., Fronick, C. C., Fulton, R. S., Maher, C. A., Ding, L., Klco, J. M., Mardis, E. R., Ley, T. J., and Wilson, R. K. (2015). Optimizing cancer genome sequencing and analysis. *Cell Syst*, 1(3):210–223.

[Hamblin et al., 1999] Hamblin, T. J., Davis, Z., Gardiner, A., Oscier, D. G., and Stevenson, F. K. (1999). Unmutated ig vh genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood*, 94(6):1848–1854.

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74.

[Hao et al., 2016] Hao, D., Wang, L., and Di, L.-j. J. (2016). Distinct mutation accumulation rates among tissues determine the variation in cancer risk. *Sci Rep*, 6:19458.

[Hayashi et al., 2006] Hayashi, Y., Aita, T., Toyota, H., Husimi, Y., Urabe, I., and Yomo, T. (2006). Experimental rugged fitness landscape in protein sequence space. *PLoS One*, 1:e96.

[Hayden, 2014] Hayden, E. C. (2014). Is the $1,000 genome for real? *Nature News*.

[Heyn, 2016] Heyn, H. (2016). Quantitative trait loci identify functional noncoding variation in cancer. *PLoS Genet*, 12(3):e1005826.

[Horn et al., 2013] Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., and Hemminki, K. (2013). Tert promoter mutations in familial and sporadic melanoma. *Science*, 339(6122):959–961.

[Howlader et al., 2015] Howlader, N., Noone, A. M., Krapcho, M., Garshell, J., Miller, D., Altekruse, S. F., Kosary, C. L., Yu, M., Ruhl, J., and Tatalovich, Z. (2015). Seer cancer statistics review, 1975–2011. national cancer institute; bethesda, md: 2014.

[Hua et al., 2013] Hua, X., Xu, H., Yang, Y., Zhu, J., Liu, P., and Lu, Y. (2013). Drgap: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *Am J Hum Genet*, 93(3):439–51.

[Kandoth et al., 2013] Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., Robertson, A. G., Pashtan, I., Shen, R., Benz, C. C., Yau, C., Laird, P. W., Ding, L., Zhang, W., Mills, G. B., Kucherlapati, R., Mardis, E. R., and Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73.

[Kim et al., 2001] Kim, M., Jiang, L.-H. . H., Wilson, H. L., North, R. A., and Surprenant, A. (2001). Proteomic and functional evidence for a p2x7 receptor signalling complex. *The EMBO Journal*, 20(22):6347–6358.

[Kircher et al., 2014] Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310–5.

[Knight et al., 2012] Knight, S. J. L., Yau, C., Clifford, R., Timbs, A. T., Sadighi Akha, E., Dréau, H. M., Burns, A., Ciria, C., Oscier, D. G., Pettitt, A. R., Dutton, S., Holmes, C. C., Taylor, J., Cazier, J.-B. B., and Schuh, A. (2012). Quantification of subclonal distributions of recurrent genomic aberrations in paired pre-treatment and relapse samples from patients with b-cell chronic lymphocytic leukemia. *Leukemia*, 26(7):1564–75.

[Knudson, 1971] Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823.

[Knudson, 2001] Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. *Nature Reviews Cancer*, 1(2):157–162.

[Konopka et al., 1985] Konopka, J. B., Watanabe, S. M., Singer, J. W., Collins, S. J., and Witte, O. N. (1985). Cell lines and clinical isolates derived from ph1-positive chronic myelogenous leukemia patients express c-abl proteins with a common structural alteration. *Proceedings of the National Academy of Sciences*, 82(6):1810–1814.

[Lai and Wade, 2011] Lai, A. Y. and Wade, P. A. (2011). Cancer biology and nurd: a multifaceted chromatin remodelling complex. *Nat Rev Cancer*, 11(8):588–96.

[Landau et al., 2013] Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., Wan, Y., Zhang, W., Shukla, S. A., Vartanov, A., Fernandes, S. M., Saksena, G., Cibulskis, K., Tesar, B., Gabriel, S., Hacohen, N., Meyerson, M., Lander, E. S., Neuberg, D., Brown, J. R., Getz, G., and Wu, C. J. (2013).

Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 152(4):714–26.

[Landau et al., 2015] Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Böttcher, S., Carter, S. L., Cibulskis, K., Mertens, D., Sougnez, C. L., Rosenberg, M., Hess, J. M., Edelmann, J., Kless, S., Kneba, M., Ritgen, M., Fink, A., Fischer, K., Gabriel, S., Lander, E. S., Nowak, M. A., Döhner, H., Hallek, M., Neuberg, D., Getz, G., Stilgenbauer, S., and Wu, C. J. (2015). Mutations driving cll and their evolution in progression and relapse. *Nature*, 526(7574):525–30.

[Landau and Wu, 2013] Landau, D. A. and Wu, C. J. (2013). Chronic lymphocytic leukemia: molecular heterogeneity revealed by high-throughput genomics. *Genome Med*, 5(5):47.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A.,

Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Lawrence et al., 2014] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501.

[Lawrence et al., 2013] Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., Dicara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A. A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*.

[Le Gallo et al., 2012] Le Gallo, M., O'Hara, A. J., Rudd, M. L., Urick, M. E., Hansen, N. F., O'Neil, N. J., Price, J. C., Zhang, S., England, B. M., Godwin, A. K., Sgroi, D. C., Hieter, P., Mullikin, J. C., Merino, M. J., and Bell,

D. W. (2012). Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nat Genet*, 44(12):1310–5.

[Lee et al., 2014] Lee, J., Mueller, P., Sengupta, S., Gulukota, K., and Ji, Y. (2014). Bayesian inference for tumor subclones accounting for sequencing and structural variants. *arXiv preprint arXiv:1409.7158*.

[Lee et al., 2015] Lee, J.-Y. . Y., Yoon, J.-K. . K., Kim, B., Kim, S., Kim, M. A., Lim, H., Bang, D., and Song, Y.-S. . S. (2015). Tumor evolution and intratumor heterogeneity of an epithelial ovarian cancer investigated using next-generation sequencing. *BMC Cancer*, 15(1).

[Leiserson et al., 2013] Leiserson, M. D., Blokh, D., Sharan, R., and Raphael, B. J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS computational biology*, 9(5):e1003054.

[Lewontin, 1970] Lewontin, R. C. (1970). The units of selection. *Annual Review of Ecology and Systematics*, 1:1–18.

[Li and Li, 2014] Li, B. and Li, J. Z. (2014). A general framework for analyzing tumor subclonality using snp array and dna sequencing data. *Genome Biol*, 15(9):473.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li et al., 2009a] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009a). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9.

[Li et al., 2014] Li, S. C., Tachiki, L. M. L., Kabeer, M. H., Dethlefs, B. A., Anthony, M. J., and Loudon, W. G. (2014). Cancer genomic research at the crossroads: realizing the changing genetic landscape as intratumoral spatial and temporal heterogeneity becomes a confounding factor. *Cancer Cell Int*, 14(1):115.

[Li et al., 2009b] Li, X., Cassidy, J. J., Reinke, C. A., Fischboeck, S., and Carthew, R. W. (2009b). A microrna imparts robustness against environmental fluctuation during development. *Cell*, 137(2):273–82.

[Ling et al., 2015] Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., Hao, L., Chen, Q., Gong, Q., Wu, D., Li, W., Zhao, W., Tian, X., Hao, C., Hungate, E. A., Catenacci, D. V. T., Hudson, R. R., Li, W.-H. H., Lu, X., and Wu, C.-I. I. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-darwinian cell evolution. *Proc Natl Acad Sci U S A*, 112(47):E6496–505.

[Lönnstedt et al., 2014] Lönnstedt, I. M., Caramia, F., Li, J., Fumagalli, D., Salgado, R., Rowan, A., Salm, M., Kanu, N., Savas, P., and Horswell, S. (2014). Deciphering clonality in aneuploid tumors using snp array and sequencing data. *Genome biology*, 15(9):470–470.

[Lynch, 2016] Lynch, M. (2016). Mutation and human exceptionalism: Our future genetic load. *Genetics*, 202(3):869–75.

[Malumbres and Barbacid, 2001] Malumbres, M. and Barbacid, M. (2001). Milestones in cell division: to cycle or not to cycle: a critical decision in cancer. *Nature Reviews Cancer*, 1(3):222–231.

[Marco-Sola et al., 2012] Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The gem mapper: fast, accurate and versatile alignment by filtration. *Nature Methods*, 9(12):1185–1188.

[Mardis, 2008] Mardis, E. R. (2008). Next-generation dna sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402.

[Margulies et al., 2005] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J. . J., and Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

[Martincorena et al., 2015a] Martincorena, I. n., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., and Tubio, J. M. (2015a). High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886.

[Martincorena et al., 2015b] Martincorena, I. n., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., and Campbell, P. J. (2015b). Tumor evolution. high burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–6.

[Marusyk et al., 2012] Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer*, 12(5):323–34.

[Marx, 2014] Marx, V. (2014). Cancer genomes: discerning drivers from passengers. *Nature methods*, 11(4):375–379.

[Mathon and Lloyd, 2001] Mathon, N. F. and Lloyd, A. C. (2001). Milestones in cell division: Cell senescence and cancer. *Nature Reviews Cancer*, 1(3):203–213.

[McFarland et al., 2013] McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R., and Mirny, L. A. (2013). Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci U S A*, 110(8):2910–5.

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

[McKenna et al., 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303.

[McLendon et al., 2008] McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., and Aldape, K. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068.

[Merlo et al., 2006] Merlo, L. M. F., Pepper, J. W., Reid, B. J., and Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nat Rev Cancer*, 6(12):924–35.

[Meyer and Penn, 2008] Meyer, N. and Penn, L. Z. (2008). Reflecting on 25 years with myc. *Nature Reviews Cancer*, 8(12):976–990.

[Michor et al., 2003] Michor, F., Frank, S. A., May, R. M., Iwasa, Y., and Nowak, M. A. (2003). Somatic selection for and against cancer. *Journal of Theoretical Biology*, 225(3):377–382.

[Michor and Polyak, 2010] Michor, F. and Polyak, K. (2010). The origins and implications of intratumor heterogeneity. *Cancer prevention research*, 3(11):1361–1364.

[Miller et al., 2014] Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., Ellis, M. J., Schierding, W., DiPersio, J. F., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2014). Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8):e1003665.

[Morales et al., 2005] Morales, C., Ribas, M., Aiza, G., and Peinado, M. A. (2005). Genetic determinants of methotrexate responsiveness and resistance in colon cancer cells. *Oncogene*, 24(45):6842–7.

[Mwenifumbo and Marra, 2013] Mwenifumbo, J. C. and Marra, M. A. (2013). Cancer genome-sequencing study design. *Nat Rev Genet*, 14(5):321–32.

[Navin et al., 2011] Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., and Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–4.

[Navin et al., 2010] Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Maanér, S., Zetterberg, A., Hicks, J., and Wigler, M. (2010). Inferring tumor progression from genomic heterogeneity. *Genome Res*, 20(1):68–80.

[Nielsen, 2005] Nielsen, R. (2005). Molecular signatures of natural selection. *Annu Rev Genet*, 39:197–218.

[Nik-Zainal et al., 2016] Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjaerde, O. C., Langerod, A., Ringnér, M., Ahn, S.-M. M., Boyault, S., Brock, J. E., Broeks, A., Butler, A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J. A., Gerstung, M., Hooijer, G. K. J., Jang, S. J., Jones, D. R., Kim, H.-Y. Y., King, T. A., Krishnamurthy, S., Lee, H. J., Lee, J.-Y. Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporté, I., Pivot, X., Purdie, C. A., Raine, K., Ramakrishnan, K., Rodríguez-González, F. G., Romieu, G., Sieuwerts, A. M., Simpson, P. T., Shepherd, R., Stebbings, L., Stefansson, O. A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G. G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., Veer, L. V., Tutt, A., Knappskog, S., Tan, B. K. T., Jonkers, J., Borg, A., Ueno, N. T., Sotiriou, C., Viari, A., Futreal, P. A., Campbell, P. J., Span, P. N., Van Laere, S., Lakhani, S. R., Eyfjord, J. E., Thompson, A. M., Birney,

E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W. M., Borresen-Dale, A.-L. L., Richardson, A. L., Kong, G., Thomas, G., and Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*.

[Nordling, 1953] Nordling, C. O. (1953). A new theory on the cancer-inducing mechanism. *British journal of cancer*, 7(1):68.

[Nowak et al., 2002] Nowak, M. A., Komarova, N. L., Sengupta, A., Jallepalli, P. V., Shih, I.-M. . M., Vogelstein, B., and Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. *Proceedings of the National Academy of Sciences*, 99(25):16226–16231.

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

[Nowell and Hungerford, 1960] Nowell, P. C. and Hungerford, D. A. (1960). Chromosome studies on normal and leukemic human leukocytes. *Journal of the National Cancer Institute*, 25(1):85–109.

[Oesper et al., 2013] Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). Theta: Inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biology*, 14(7):R80.

[Orvis et al., 2014] Orvis, T., Hepperla, A., Walter, V., Song, S., Simon, J., Parker, J., Wilkerson, M. D., Desai, N., Major, M. B., Hayes, D. N., Davis, I. J., and Weissman, B. (2014). Brg1/smarca4 inactivation promotes non-small cell lung cancer aggressiveness by altering chromatin organization. *Cancer Res*, 74(22):6486–98.

[Ostrow et al., 2014] Ostrow, S. L., Barshir, R., DeGregori, J., Yeger-Lotem, E., and Hershberg, R. (2014). Cancer evolution is associated with pervasive positive selection on globally expressed genes. *PLoS Genet*, 10(3):e1004239.

[Ovens and Naugler, 2012] Ovens, K. and Naugler, C. (2012). Preliminary evidence of different selection pressures on cancer cells as compared to normal tissues. *Theor Biol Med Model*, 9:44.

[Parker et al., 2016] Parker, N. R., Hudson, A. L., Khong, P., Parkinson, J. F., Dwight, T., Ikin, R. J., Zhu, Y., Cheng, Z. J., Vafaee, F., Chen, J., Wheeler, H. R., and Howell, V. M. (2016). Intratumoral heterogeneity identified at the epigenetic, genetic and transcriptional level in glioblastoma. *Sci Rep*, 6:22477.

[Podlaha et al., 2012] Podlaha, O., Riester, M., De, S., and Michor, F. (2012). Evolution of the cancer genome. *Trends in Genetics*, 28(4):155–163.

[Proctor, 2001] Proctor, R. N. (2001). Tobacco and the global lung cancer epidemic. *Nature Reviews Cancer*, 1(1):82–86.

[Puente et al., 2015] Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., Munar, M., Rubio-Pérez, C., Jares, P., Aymerich, M., Baumann, T., Beekman, R., Belver, L., Carrio, A., Castellano, G., Clot, G., Colado, E., Colomer, D., Costa, D., Delgado, J., Enjuanes, A., Estivill, X., Ferrando, A. A., Gelpí, J. L., González, B., González, S., González, M., Gut, M., Hernández-Rivas, J. M., López-Guerra, M., Martín-García, D., Navarro, A., Nicolás, P., Orozco, M., Payer, A. R., Pinyol, M., Pisano, D. G., Puente, D. A., Queirós, A. C., Quesada, V., Romeo-Casabona, C. M., Royo, C., Royo, R., Rozman, M., Russiñol, N., Salaverría, I., Stamatopoulos, K., Stunnenberg, H. G., Tamborero, D., Terol, M. J., Valencia, A., López-Bigas, N., Torrents, D., Gut, I., López-Guillermo, A., López-Otín, C., and Campo, E. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 526(7574):519–24.

[Puente et al., 2011] Puente, X. S., Pinyol, M., Quesada, V., Conde, L., Ordóñez, G. R., Villamor, N., Escaramis, G., Jares, P., Beà, S., González-Díaz, M., Bassaganyas, L., Baumann, T., Juan, M., López-Guerra, M., Colomer, D., Tubío, J. M. C., López, C., Navarro, A., Tornador, C., Aymerich, M., Rozman, M., Hernández, J. M., Puente, D. A., Freije, J. M. P., Velasco, G., Gutiérrez-Fernández, A., Costa, D., Carrió, A., Guijarro, S., Enjuanes, A., Hernández, L., Yagüe, J., Nicolás, P., Romeo-Casabona, C. M., Himmelbauer, H., Castillo, E., Dohm, J. C., de Sanjosé, S., Piris, M. A., de Alava, E., San Miguel, J., Royo, R., Gelpí, J. L., Torrents, D., Orozco, M., Pisano, D. G., Valencia, A., Guigó, R., Bayés, M., Heath, S., Gut, M., Klatt, P., Marshall, J., Raine, K., Stebbings, L. A., Futreal, P. A., Stratton, M. R., Campbell, P. J., Gut, I., López-Guillermo, A., Estivill, X., Montserrat, E., López-Otín, C., and Campo, E. (2011). Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 475(7354):101–5.

[Pyatnitskiy et al., 2015] Pyatnitskiy, M., Karpov, D., Poverennaya, E., Lisitsa, A., and Moshkovskii, S. (2015). Bringing down cancer aircraft: Searching for essential hypomutated proteins in skin melanoma. *PLoS One*, 10(11):e0142819.

[Quesada et al., 2012] Quesada, V., Conde, L., Villamor, N., Ordóñez, G. R., Jares, P., Bassaganyas, L., Ramsay, A. J., Beà, S., Pinyol, M., Martínez-Trillos,

A., López-Guerra, M., Colomer, D., Navarro, A., Baumann, T., Aymerich, M., Rozman, M., Delgado, J., Giné, E., Hernández, J. M., González-Díaz, M., Puente, D. A., Velasco, G., Freije, J. M. P., Tubío, J. M. C., Royo, R., Gelpí, J. L., Orozco, M., Pisano, D. G., Zamora, J., Vázquez, M., Valencia, A., Himmelbauer, H., Bayés, M., Heath, S., Gut, M., Gut, I., Estivill, X., López-Guillermo, A., Puente, X. S., Campo, E., and López-Otín, C. (2012). Exome sequencing identifies recurrent mutations of the splicing factor sf3b1 gene in chronic lymphocytic leukemia. *Nat Genet*, 44(1):47–52.

[Reimand et al., 2016] Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g: Profilera web server for functional interpretation of gene lists (2016 update). *Nucleic acids research*, page gkw199.

[Reimand and Bader, 2013] Reimand, J. and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol*, 9:637.

[Ren, 2005] Ren, R. (2005). Mechanisms of bcr–abl in the pathogenesis of chronic myelogenous leukaemia. *Nature Reviews Cancer*, 5(3):172–183.

[Reuter et al., 2015] Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Mol Cell*, 58(4):586–97.

[Roberts and Orkin, 2004] Roberts, C. W. M. and Orkin, S. H. (2004). The swi/snf complex–chromatin and cancer. *Nat Rev Cancer*, 4(2):133–42.

[Rocha et al., 2006] Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T. G., Cooper, J. E., Smith, N. H., and Feil, E. J. (2006). Comparisons of dn/ds are time dependent for closely related bacterial genomes. *J Theor Biol*, 239(2):226–35.

[Roger and Pelegrin, 2011] Roger, S. and Pelegrin, P. (2011). P2x7 receptor antagonism in the treatment of cancers. *Expert opinion on investigational drugs*, 20(7):875–880.

[Rosowsky et al., 1985] Rosowsky, A., Wright, J. E., Cucchi, C. A., Lippke, J. A., Tantravahi, R., Ervin, T. J., and Frei, E. (1985). Phenotypic heterogeneity in cultured human head and neck squamous cell carcinoma lines with low-level methotrexate resistance. *Cancer research*, 45(12 Part 1):6205–6212.

[Roth et al., 2014] Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014). Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*.

[Sakoparnig et al., 2015] Sakoparnig, T., Fried, P., and Beerenwinkel, N. (2015). Identification of constrained cancer driver genes based on mutation timing. *PLoS Comput Biol*, 11(1):e1004027.

[Salaro et al., 2016] Salaro, E., Rambaldi, A., Falzoni, S., Amoroso, F. S., Franceschini, A., Sarti, A. C., Bonora, M., Cavazzini, F., Rigolin, G. M., Ciccone, M., Audrito, V., Deaglio, S., Pelegrin, P., Pinton, P., Cuneo, A., and Di Virgilio, F. (2016). Involvement of the p2x7-nlrp3 axis in leukemic cell proliferation and death. *Sci Rep*, 6:26280.

[Sarkisyan et al., 2016] Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., Vlasov, P. K., Egorov, E. S., Logacheva, M. D., Kondrashov, A. S., Chudakov, D. M., Putintseva, E. V., Mamedov, I. Z., Tawfik, D. S., Lukyanov, K. A., and Kondrashov, F. A. (2016). Local fitness landscape of the green fluorescent protein. *Nature*.

[Schroeder et al., 2014] Schroeder, M. P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2014). Oncodriverole classifies cancer driver genes in loss of function and activating mode of action. *Bioinformatics*, 30(17):i549–55.

[Schuh et al., 2012] Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S. M., Grocock, R., Henderson, S., Khrebtukova, I., Kingsbury, Z., Luo, S., McBride, D., Murray, L., Menju, T., Timbs, A., Ross, M., Taylor, J., and Bentley, D. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20):4191–6.

[Simpson, 1944] Simpson, G. G. (1944). *Tempo and mode in evolution*. Columbia University Press.

[Skorski et al., 1997] Skorski, T., Bellacosa, A., Nieborowska-Skorska, M., Majewski, M., Martinez, R., Choi, J. K., Trotta, R., Wlodarski, P., Perrotti, D., and Chan, T. O. (1997). Transformation of hematopoietic cells by bcr/abl requires activation of a pi-3k/akt-dependent pathway. *The EMBO journal*, 16(20):6151–6161.

[Smith and Haigh, 1974] Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical research*, 23(01):23–35.

[Smith et al., 2015] Smith, K. S., Yadav, V. K., Pedersen, B. S., Shaknovich, R., Geraci, M. W., Pollard, K. S., and De, S. (2015). Signatures of accelerated

somatic evolution in gene promoters in multiple cancer types. *Nucleic Acids Res*, 43(11):5307–17.

[Sottoriva et al., 2015] Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., and Curtis, C. (2015). A big bang model of human colorectal tumor growth. *Nat Genet*, 47(3):209–16.

[Sottoriva et al., 2013] Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C., and Tavaré, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014.

[Stephens et al., 2011] Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., McLaren, S., Lin, M.-L. L., McBride, D. J., Varela, I., Nik-Zainal, S., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Quail, M. A., Burton, J., Swerdlow, H., Carter, N. P., Morsberger, L. A., Iacobuzio-Donahue, C., Follows, G. A., Green, A. R., Flanagan, A. M., Stratton, M. R., Futreal, P. A., and Campbell, P. J. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40.

[Stransky et al., 2011] Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., Kryukov, G. V., Lawrence, M. S., Sougnez, C., and McKenna, A. (2011). The mutational landscape of head and neck squamous cell carcinoma. *Science*, 333(6046):1157–1160.

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

[Su et al., 2012] Su, X., Zhang, L., Zhang, J., Meric-Bernstam, F., and Weinstein, J. N. (2012). Purityest: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, 28(17):2265–6.

[Supek et al., 2014] Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–35.

[Surcel et al., 2015] Surcel, A., Ng, W. P., West-Foyle, H., Zhu, Q., Ren, Y., Avery, L. B., Krenc, A. K., Meyers, D. J., Rock, R. S., Anders, R. A., Freel Meyers, C. L., and Robinson, D. N. (2015). Pharmacological activation of myosin ii paralogs to correct cell mechanics defects. *Proc Natl Acad Sci U S A*, 112(5):1428–33.

[Szczurek and Beerenwinkel, 2014] Szczurek, E. and Beerenwinkel, N. (2014). Modeling mutual exclusivity of cancer mutations. *PLoS Comput Biol*, 10(3):e1003503.

[Tabassum and Polyak, 2015] Tabassum, D. P. and Polyak, K. (2015). Tumorigenesis: it takes a village. *Nat Rev Cancer*, 15(8):473–83.

[Tamborero et al., 2013a] Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013a). Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–44.

[Tamborero et al., 2013b] Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M. S., Getz, G., Bader, G. D., and Ding, L. (2013b). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3.

[Tamborero et al., 2013c] Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L., and Lopez-Bigas, N. (2013c). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep*, 3:2650.

[Tamborero et al., 2013d] Tamborero, D., Lopez-Bigas, N., and Gonzalez-Perez, A. (2013d). Oncodrive-cis: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PLoS One*, 8(2):e55489.

[Tao et al., 2015] Tao, Y., Hu, Z., Ling, S., Yeh, S.-H. . H., Zhai, W., Chen, K., Li, C., Wang, Y., Wang, K., and Wang, H.-Y. . Y. (2015). Further genetic diversification in multiple tumors and an evolutionary perspective on therapeutics. *bioRxiv*, page 025429.

[TCGANetwork, 2008] TCGANetwork (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8.

[TCGANetwork, 2011] TCGANetwork (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

[TCGANetwork, 2012a] TCGANetwork (2012a). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–25.

[TCGANetwork, 2012b] TCGANetwork (2012b). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–7.

[TCGANetwork, 2012c] TCGANetwork (2012c). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

[Thomas et al., 2003] Thomas, M. A., Weston, B., Joseph, M., Wu, W., Nekrutenko, A., and Tonellato, P. J. (2003). Evolutionary dynamics of oncogenes and tumor suppressor genes: higher intensities of purifying selection than other genes. *Molecular biology and evolution*, 20(6):964–968.

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

[Torres et al., 2007] Torres, L., Ribeiro, F. R., Pandis, N., Andersen, J. A., Heim, S., and Teixeira, M. R. (2007). Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast cancer research and treatment*, 102(2):143–155.

[Trusolino and Comoglio, 2002] Trusolino, L. and Comoglio, P. M. (2002). Scatter-factor and semaphorin receptors: cell signalling for invasive growth. *Nat Rev Cancer*, 2(4):289–300.

[Van Loo et al., 2010] Van Loo, P., Nordgard, S. H., Lingjaerde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Borresen-Dale, A.-L. L., and Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*, 107(39):16910–5.

[Vogelstein et al., 2013] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *science*, 339(6127):1546–1558.

[Vohra and Biggin, 2013] Vohra, S. and Biggin, P. C. (2013). Mutationmapper: a tool to aid the mapping of protein mutation data. *PLoS One*, 8(8):e71711.

[Vorontsov et al., 2015] Vorontsov, I. E., Kulakovskiy, I. V., Khimulya, G., Lukianova, E. N., Nikolaeva, D. D., Eliseeva, I. A., and Makeev, V. J. (2015). Negative selection maintains transcription factor binding motifs in human cancer. *arXiv preprint arXiv:1511.02842*.

[Wang et al., 2014a] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014a). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*.

[Wang et al., 2015] Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., and Sabatini, D. M. (2015). Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101.

[Wang et al., 2014b] Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., Zhang, H., Zhao, R., Michor, F., Meric-Bernstam, F., and Navin, N. E. (2014b). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–60.

[Warburg et al., 1924] Warburg, O., Negelein, E., and Posener, K. (1924). Versuche an überlebendem carcinomgewebe. *Journal of Molecular Medicine*, 3(24):1062–1064.

[Weinberg, 1996] Weinberg, R. A. (1996). How cancer arises. *Scientific American*, 275(3):62–71.

[Weir et al., 2004] Weir, B., Zhao, X., and Meyerson, M. (2004). Somatic alterations in the human cancer genome. *Cancer cell*, 6(5):433–438.

[Williams et al., 2016] Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nature genetics*.

[Wolman, 1986] Wolman, S. R. (1986). Cytogenetic heterogeneity: its role in tumor evolution. *Cancer genetics and cytogenetics*, 19(1):129–140.

[Woo and Li, 2012] Woo, Y. H. and Li, W.-H. H. (2012). Dna replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat Commun*, 3:1004.

[Wright, 1932] Wright, S. (1932). *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*, volume 1. na.

[Wu et al., 2012] Wu, C. C., Ye, R., Jasinovica, S., Wagner, M., Godiska, R., Tong, A. H. Y., Lok, S., Krerowicz, A., Knox, C., and Mead, D. (2012). Long-span, mate-pair scaffolding and other methods for faster next-generation sequencing library creation. *Nature Methods*, 1.

[Wu, 2012] Wu, C. J. (2012). Cll clonal heterogeneity: an ecology of competing subpopulations. *Blood*, 120(20):4117–8.

[Wu et al., 2016] Wu, S., Powers, S., Zhu, W., and Hannun, Y. A. (2016). Substantial contribution of extrinsic risk factors to cancer development. *Nature*, 529(7584):43–47.

[Xie et al., 2014] Xie, M., Lu, C., Wang, J., McLellan, M. D., Johnson, K. J., Wendl, M. C., McMichael, J. F., Schmidt, H. K., Yellapantula, V., Miller, C. A.,

Ozenberger, B. A., Welch, J. S., Link, D. C., Walter, M. J., Mardis, E. R., Dipersio, J. F., Chen, F., Wilson, R. K., Ley, T. J., and Ding, L. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med*, 20(12):1472–8.

[Yachida et al., 2010] Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., and Nowak, M. A. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature*, 467(7319):1114–1117.

[Yates and Campbell, 2012] Yates, L. R. and Campbell, P. J. (2012). Evolution of the cancer genome. *Nat Rev Genet*, 13(11):795–806.

[Yates et al., 2015] Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P.-Y. Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lonning, P. E., Wedge, D. C., and Campbell, P. J. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med*, 21(7):751–9.

[Youn and Simon, 2011] Youn, A. and Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, 27(2):175–181.

[Zhang et al., 2014] Zhang, J., Fujimoto, J., Zhang, J., Wedge, D. C., Song, X., Zhang, J., Seth, S., Chow, C.-W. W., Cao, Y., Gumbs, C., Gold, K. A., Kalhor, N., Little, L., Mahadeshwar, H., Moran, C., Protopopov, A., Sun, H., Tang, J., Wu, X., Ye, Y., William, W. N., Lee, J. J., Heymach, J. V., Hong, W. K., Swisher, S., Wistuba, I. I., and Futreal, P. A. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science*, 346(6206):256–9.

[Zhang et al., 1999] Zhang, X., Mar, V., Zhou, W., Harrington, L., and Robinson, M. O. (1999). Telomere shortening and apoptosis in telomerase-inhibited human tumor cells. *Genes & development*, 13(18):2388–2399.

[Zhao et al., 2014a] Zhao, B., Hemann, M. T., and Lauffenburger, D. A. (2014a). Intratumor heterogeneity alters most effective drugs in designed combinations. *Proc Natl Acad Sci U S A*, 111(29):10773–8.

[Zhao et al., 2016] Zhao, B., Hemann, M. T., and Lauffenburger, D. A. (2016). Modeling tumor clonal evolution for drug combinations design. *Trends in Cancer*, 2(3):144–158.

[Zhao et al., 2014b] Zhao, J., Xu, H., He, M., Wang, Z., and Wu, Y. (2014b). Rho gtpase-activating protein 35 rs1052667 polymorphism and osteosarcoma risk and prognosis. *Biomed Res Int*, 2014:396947.