Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

Vasilisa Rudneva, M.Sc.

born in Moscow, Russia

Oral examination: 20th July 2016

Computational approaches for identifying somatic intergenic mutations of relevance in cancer

Referees:

Dr. Wolfgang Huber

Prof. Dr. Michael Boutros

Abstract

Cancer is a complex genomic disease characterized by accumulation of somatic mutations over the lifetime of a patient. Identification of somatic driver mutations that contribute to tumorigenesis is a major goal of cancer genomics. With the recent advances in the sequencing technologies it became possible to study somatic mutations on the whole-genome scale in multiple cancers. While most of the cancer genomics studies were previously focused on identification of driver mutations affecting exons, several examples of driver events within the non-protein-coding regions of the genome were identified, including the recurrent *TERT* promoter mutations. Such findings have spurred searches for similar examples of recurrent non-coding mutations using computational cancer genomics. In my PhD thesis, I present several computational approaches aimed to identify somatic driver mutations with a specific focus on intergenic regions of the genome.

The first part of this thesis focuses on the somatic mutational patterns along the cancer genome and addresses a fundamental problem of computational identification of recurrently mutated regions – regional mutational heterogeneity. Here I studied the correlation of specific genomic features with background somatic mutation rates and devised a background model that accounts for regional mutational heterogeneity.

The second part of this thesis describes three different computational approaches designed to identify somatic driver events of functional relevance in cancer. The first approach integrates somatic mutation calls with gene expression data to identify variants associated with altered mRNA levels. The second approach is designed to predict changes in transcription factor binding sites in presence of recurrent somatic mutations. The third approach uses cross-validation scheme to enable parameter tuning in screens for recurrently somatically mutated regions in cancer genomes in an unbiased genome-wide manner. Using this approach, we identify several known cancer-relevant targets, both exonic (e.g., the *TP53*, *MYC*, and *SMARCA4* genes) as well as non-coding regulatory regions (e.g., the *TERT* promoter) and uncover novel candidate regulatory driver regions. Among those, a cluster of recurrent intergenic mutations, occurring in an enhancer element near the *FADS2* gene, which encodes a critical enzyme in the biosynthesis of long chain polyunsaturated fatty acids and has been previously implicated in cancer.

Collectively, the computational approaches presented here helped in uncovering novel somatic candidate events of relevance in cancer and can be further used for various applications in cancer genomics.

Zusammenfassung

Krebs ist eine komplexe genomische Erkrankung, die v.a. durch somatische Mutationen charakterisiert ist, die sich im Laufe des Lebens im Patienten anhäufen. Einer der wesentlichen Ziele in der Krebsforschung/genomik ist somit die Identifikation der somatischen Mutationen, die zur Tumorbildung beitragen- sogenannte „somatic driver mutations“. Neueste Fortschritte in Sequencing-Technologien ermöglichen die Suche nach diesen Mutationen im gesamten Genom in verschiedenen Krebstypen. Während die meisten Genomstudien in der Krebsforschung sich hauptsächlich auf „driver“ Mutationen, die Exone betreffen, fokussieren, gibt es bereits einige Evidenzen für *driver mutations*, die in nicht-kodierenden Regionen des Genoms auftreten, wie z.B. wiederkehrende *TERT* Promoter Mutationen. Solcherlei Beobachtungen führten dementsprechend zu Studien von weiteren Mutationen unter Anwendungen von bioinformatischen Methoden. In meiner hier vorgelegten PhD-Arbeit präsentiere ich mehrere bioinformatische Angehensweisen zur Identifizierung von somatischen *driver mutations* mit Fokus auf jene, die außerhalb von protein-kodierenden Bereichen des Genoms liegen.

Der erste Teil dieser Arbeit behandelt das Auftreten von somatischen Mutationen im Krebsgenom und adressiert das fundamentale Probleme bei der Identifizierung von wiederkehrenden Mutationen, nämlich regionale Mutationsheterogenität. Ich untersuche in diesem Zusammenhang die Korrelation von spezifischen genomischen Merkmalen mit im Hintergrund auftretenden Mutationsraten und entwickle eine Modell, das Nulleffekte sowie Heterogenität von Mutationen in spezifischen Regionen, berücksichtigt.

Der zweite Teil meiner Arbeit beschreibt drei verschiedene bioinformatische Ansätze zur Identifikation von somatischen Mutationsereignissen mit funktioneller Relevanz zur Krebsentwicklung. Der erste Ansatz verbindet signifikante Mutationsereignisse mit Genexpression um eine genomische Variante zu detektieren, die zu veränderten mRNA-Levels führt. Der zweite Ansatz, hingegen, versucht Veränderungen an Bindungstellen von Transkriptionsfaktoren aufzuzeigen, die aufgrund von wiederkehrenden somatischen Mutationen auftreten können. Schließlich stelle ich das dritte Konzept vor, das eine Kreuzvalidierung umfasst, die einen Parameterabgleich in der Suche nach relevanten Mutationen im Krebsgenom auf eine objektive Art und Weise ermöglicht. Unter Anwendung dieser Methode identifizieren wir etliche krebs-relevante Targets, ob in Exon angesiedelte Gene (wie z.B. *TP53, MYC*, und *SMARCA4*), oder in nicht-kodierenden Bereichen (wie z.B. der *TERT* promoter). Allerdings detektieren wir auch neue Kandidat-Regionen, die als *driver* Mutations in Frage kommen, u.a. ein Cluster an Mutationen im Enhancer-Bereich in der Nähe des *FADS2*-Gens. Interessanterweise ist dieses Gen, das ein kritisches Enzym für die Biosynthese von langkettigen nicht-saturierten Fettsäuren kodiert, bereits in anderen Studien mit Krebs in Verbindung gebracht worden.

Zusammenfassend sind die vorgestellten bioinformatischen Methoden- die sicherlich auch Anwendung in anderen Bereichen der Krebsforschung finden könnten- ein Beitrag zur Identifikation von neuen somatischen Mutationsereignissen, die für die Krebsentwicklung von Relevanz sind.

Acknowledgements

First and foremost, I want to thank my supervisor Jan Korbel for giving me the opportunity to work in his lab and for his excellent supervision throughout my PhD. Without his guidance, support and enthusiasm this work would not have been possible.

Further I would like to thank my TAC members Prof. Dr. Michael Boutros, Dr. Wolfgang Huber and Dr. Martin Jechlinger for their insightful feedback and advice on the project that helped me to reflect on my progress and to move forward. I appreciate their different backgrounds and views on my work and I believe, they helped me to develop my work in the best possible way.

I would like to thank my collaborator and coauthor Simon Anders for the constant support and encouragement that he provided me with for the 3 years we spend together at EMBL. It was a pleasure to work with him and shape this project together.

Thanks to thank all former and current members of the Korbel research group for the creating a great working environment and for the useful comments and feedback on this work during the lab meetings. Especially, I want to acknowledge Verena Tischler, Andreas Schlattl, Markus Fritz , Tobias Rausch and Thomas ZIchner for their constant support and patience through the first month of my PhD. Special thanks to Christopher Buccitelli and Balca Mardin.

Last, but by no means least, I would like to acknowledge the people who supported me in my personal life. First, I want to thank my parents and my family for their constant aid and encouragement as well as for always believing in me and the things I do. Furthermore, I want to acknowledge my dear friends, Vilma Jimenez Sabinina and Natalie Romanov, for always being there for me. And finally, I want to thank Marco Faini for his endless support and encouragement throughout most of my PhD time.

Моим родителям

Contents

List of Abbreviations

BMR Background mutation rate

bp Base pair(s)

chr Chromosome

CNV Copy number variant

DNA Deoxyribonucleic acid Doxo Doxorubicin

EMBL European Molecular Biology Laboratory

FDR False discovery rate

HapMap The International HapMap Project

Indel Short insertion/deletion

kb Kilobase(s)

Mb Megabase(s)

lncRNA Long non-coding RNA

NGS Next-generation sequencing

RNA Ribonucleic acid

RNA-Seq RNA sequencing

ROS Reactive oxygen species

SNP Single nucleotide polymorphism

SNV Single nucleotide variant

SV Structural variant

TF Transcription factor

TFBS Transcription factor binding site

UTR Untranslated region

UV Ultraviolet

WGS Whole-genome sequencing

WT Wild type

1. Introduction

Cancer is a complex disease widely spread in the world. According to the Cancer Research UK, 14.1 million of new cancer cases occurred worldwide in 2012 and an estimated 8.2 million people died from the disease that year.

According to the classical model of tumorigenesis, cancer is a genetic disease characterized by accumulation of somatic alterations during the lifetime of an individual. The most common type of somatic alterations known in cancer is single nucleotide variants (SNVs) also referred to as point mutations. It is estimated that most of cancers carry 1,000 to 20,000 somatic mutations and only few to hundreds of other types of somatic genomic alterations (Martincorena and Campbell, 2015). Most somatic mutations in cancer occur in the non-coding regions due to generally weaker purifying selection of these regions when compared to exons and other types of genomic elements (Khurana, Fu, Colonna, Mu, Kang, T. Lappalainen, *et al.*, 2013). Despite that, the absolute majority of the known cancer driver events known to-date occur within the coding genome (Forbes *et al.*, 2015). Various reasons have prevented researches from studying the non-coding mutation in cancer and its potential consequences for tumorigenesis. Among them, financial reasons such as high sequencing costs; technical reasons, for example, the regional mutations heterogeneity; as well various methodological reasons such as the lack of computational approaches specially designed to detect somatic drivers outside of protein-coding regions. However, with the recent decrease in whole-genome sequencing costs the situations in the cancer genomics field is rapidly changing. For example, a driver event occurring the promoter regions of a cancer gene, *TERT*, was identified at first in melanoma (Horn *et al.*, 2013; Huang *et al.*, 2013) and later in other malignancies (Vinagre *et al.*, 2013). This example has motivated researchers to shift their focus to the non-coding genome. Recent attempts on identification of novel intergenic drivers using both computational and experimental approaches have extended the list of known driver events within the non-coding genome as well as broadened our knowledge on how such events might contribute to tumorigenesis. With regards to the methodological advances in the non-coding driver detection using computational approaches, several recent studies aimed to identify somatic drivers that may play a role in cancer development. However, despite the significant overlap between the datasets and generally similar methodology used, the overlap between the findings was relatively limited with the *TERT* promoter being a remarkable exception.

1.1 Motivation and outline of this thesis

The aim of my PhD work was to develop computational approaches to identify novel examples of driver events in cancer with the primary focus on the non-coding genomic regions.

In the remainder of this Chapter, I introduce some background concepts related to the non-coding somatic mutation in cancer. First, I describe the types of genomic alterations that are known in cancer and provide examples of how such alterations, within the intergenic regions, may contribute to cancer. Next, I focus on mutational processes in cancer and I discuss the challenges they provide for computational cancer genomics. And lastly, I review the existing computational approaches to identify somatic driver events in cancer.

In **Chapter 2**, I address the major problem for computational identification of recurrently mutated regions in cancer – the regional mutational heterogeneity. To overcome this problem, I first studied the mutational patterns in cancer and their correlation with various genetic and epigenetic features, and then I identified the best correlate with somatic mutation rates. Finally, I stratified the entire genome into groups of regions with comparable genetic and epigenetic background that were later used to control for the background mutation rates in the following analyses.

Chapter 3 describes three different computational approaches that I designed to identify somatic driver events of functional relevance in cancer, with a specific focus on non-coding genomic regions. In the first approach, I integrated mutation calls with gene expression data to identify somatic mutations associated with altered mRNA levels. The second approach was developed to predict changes in transcription factor binding sites in presence of recurrent somatic mutations. The third approach was mainly focused on identification of recurrently mutated regions by implementing the background model for somatic mutations across multiple cancer types established in Chapter 2. In the end of the Chapter I discuss some statistical aspects of identification of recurrently mutated regions and propose and approach based on cross-validations to identify optimal parameters for such computational genomics studies.

Please note, that for all of the approaches, I here describe my work and clearly indicate other people's contributions in the text when necessary.

Finally, in the last Chapter of this thesis I summarize my main results and conclusions, and give future perspectives on potential development of the computational approaches. Supplementary Information is provided in the Appendix A. All computational methods used in this work are described in Appendix B. A list of publications in which I was involved during my PhD is included in Appendix C.

1.2 Genomic alterations in cancer

According to the classical model, tumorigenesis is driven by somatic alterations, which accumulate during the lifetime of an individual (Cavenee & White, 1995). *Somatic* alterations are the genomic alterations that are acquired by the tumor during its development and progression and are usually defined in the context of cancer sequencing studies as those observed in the tumor samples of an individual but in the matching control sample. The alterations observed in both, the control and the tumor samples, are usually referred as *germline*.

Somatic alterations vary in size and based on this can be divided into two groups. By definition, *large-scale structural variants* (SVs) represent a class of genomic alterations of larger than 50 bp size, that includes duplications, deletions, insertions, inversions and translocations. In contrast, *small-scale variants* are defined as alterations smaller than 50 bp in size and consist of two types of events: *indels* (small insertions or deletions) and *single nucleotide variants* (SNVs). SNVs, commonly referred as point mutations, are single base pair exchanges. Germline SNVs observed in a given populations with a frequency higher than 1%, are usually considered polymorphisms and therefore termed *single nucleotide polymorphisms* (SNPs).

1.2.1 Large-scale variants

Large-scale structural variants (SVs) can involve both microscopic and submicroscopic events, ranging in size from several kilobases up to a few megabases (Baker, 2012; Feuk and Carson, 2006). Some SVs can be balanced in terms of copy number (inversions and translocations), while others are unbalanced (deletions, duplications, and insertions). SVs tend to occur more frequently in some regions of the genome compared to the others creating hotspots of recurrent variation (Mills *et al.*, 2011). Among the factors that contribute to such clustering are sequence context and local genomic architecture (Stankiewicz and Lupski, 2002; Shaw, 2004).

Since SVs affect a larger fraction of the genome in comparison to SNVs, their consequent phenotypic impact is larger than that of SNVs. Unsurprisingly, SVs were associated with both disease and normal traits variation (Onishi-Seebacher and Korbel, 2011; Weischenfeldt *et al.*, 2013).

The generation of some forms of SVs can be physiological. For example, structural variation is a necessary part of maturation at the IG locus of cells of the immune system. However, the same machinery when misregulated might also drive tumorigenesis. For example, RAG proteins are required for structural rearrangements between the IGH locus and the B-cell CLL/lymphoma 2 (BCL2), a driver event in follicular lymphoma; AID protein promotes C-MYC–IGH chromosomal translocations that drive Burkitt's lymphoma. (Helleday *et al.*, 2014)

In cancer, one of the first SVs discovered was a translocation between the chromosomes 9 and 22, known as the Philadelphia chromosome, which is the major driver event in chronic myeloid leukemia (CML) (Nowell and Hungerford, 1960). More recently, SVs have been implicated in the development of different types of tumors (Pleasance *et al.*, 2010; Rausch *et al.*, 2012). For example, structural rearrangements may affect the coding regions of genes, either by removing part of the coding sequence or by creating gene fusions (e.g. the fusion of *TMPRSS2* and ETS transcription factors, *ERG* or *ETV1*, in prostate cancer (Tomlins, 2005)); duplications may lead to the gene dosage changes. High level amplifications are also characteristic for specific cancers and can lead to overexpression of oncogenes, for example, amplification of *TERT* and *MYC* genes in medulloblastoma (Northcott *et al.*, 2012). Furthermore, SVs can also affect the expression of a gene without directly damaging the coding region, e.g. occurring within the non-coding regions. The mechanism they act upon in these cases may involve changes in location of regulatory elements, such as enhancers and isolators. This leads to misregulation of genes that were otherwise not regulated by these elements. An example of such driver event termed as "enhancer hijacking" was recently identified in medulloblastoma. In this scenario, an enhancer element is brought to the proximity of proto-oncogenes *GFI1* and *GFI1B* resulting in their activation (Northcott *et al.*, 2014).

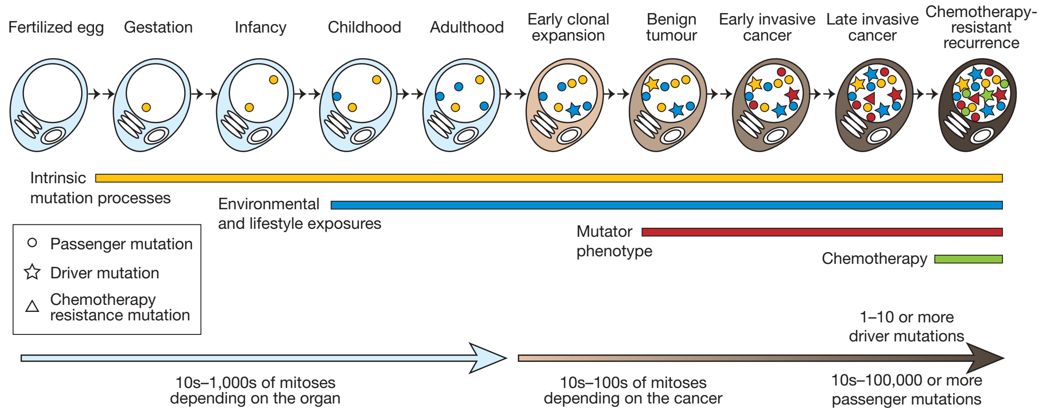
1.2.2 Small-scale variants

Small-scale variants include indels and SNVs. SNVs or mutations can originate from replication errors during DNA synthesis by the DNA polymerase or from DNA damage that is either repaired incorrectly or left unrepaired (Martincorena and Campbell, 2015). DNA damage can be caused by exogenous factors (e.g. UV light), by endogenous factors (e.g. reactive oxygen species (ROS)) or by enzymes involved in DNA repair or genome editing.

The distribution of SNVs along the genome is not homogeneous. Variants in coding regions are less frequent than in non-coding ones, which is explained by the purifying selection acting against mutations with a negative effect on the phenotype (Khurana, Fu, Colonna, Mu, Kang, Tuuli Lappalainen, *et al.*, 2013). Interestingly, many regions of the genome experience patterns of purifying selection that are human-specific, whereas other regions conserved across mammals do not show functional activity and selection in humans (Ward and Kellis, 2012).

SNVs within coding regions are called *synonymous* if they do not change the sequence of the protein encoded by the gene, or *non-synonymous* if they do. Non-synonymous variants occurring within the protein-coding regions of the genes can be further classified into two types based on their effects on amino acids. The *missense* variants lead to a change of amino acid, whereas the *nonsense* variants produce a premature stop codon at the variant site. Depending on the base change caused by an SNV, it can be referred to as a *transition* or a *transversion*. If a purine is replaced by another purine, or if a pyrimidine is replaced by a pyrimidine, such variant will be a transition. In contrast, if a purine is substituted by a pyrimidine, or the vise versa, the variant will be called a transversion. In general, transitions occur more often than transversions (Zhang and Gerstein, 2003). The most frequent type of transition accounting for almost half of human SNVs is C→T, which is caused by spontaneous deamination of 5-methylcytosine (Shen *et al.*, 1994).

Indels are a less abundant form of genomic variations which is observed in humans with an average frequency of one every eight kb per individual (Montgomery *et al.*, 2013). Similarly to SNVs, indels are usually depleted from protein-coding regions with a tendency to cluster within repetitive sequences, forming hotspots of variation with increased indel occurrence compared to the chromosomal average (Mills *et al.*, 2006). One of the possible explanations how these unusual regions of high genetic variation occur, is that DNA segments with a longer evolutionary history have more time to accumulate structural variants. Among other contributing factors, are higher rates of homologous recombination in some genomic regions and a lack of selective pressure on the variants that have little impact on the fitness of the individual (Montgomery *et al.*, 2013). The most common mechanism of indel formation is the slippage of the DNA polymerase during replication, which is especially prevalent in regions of highly repetitive sequences. This mechanism accounts for almost 75% of indels in the human genome including the ones that occur in hotspots (Montgomery *et al.*, 2013). Indels can alter the protein sequence in similar ways to SNVs and with analogous phenotypic consequences. Therefore indels can be associated with increased risks for several diseases, including cancer (Frazer *et al.*, 2009; Yang *et al.*, 2010).



Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-resistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
1–10 or more driver mutations
10s–1,000s of mitoses depending on the organ
10s–100s of mitoses depending on the cancer
10s–100,000 or more passenger mutations

Figure 1. The classical model of tumorigenesis. Accumulation of somatic mutations within the cell lineage during the lifetime of an individual starting from the early development until the relapsed tumor. Figure adapted from Stratton M *et al.*, 2009.

In the context of cancer, somatic mutation plays an important role. For example, it was estimated that around 90% of recurrently mutated cancer genes are altered by somatic mutation and only about 20% of the genes are associated with germline mutations that predispose to cancer (Martincorena and Campbell, 2015). Most cancer genomes harbor between 1,000 to 20,000 somatic point mutations. However, the vast majority of mutations make no contribution to the tumor development and are commonly referred as *passenger* mutations. In contrast, a small fraction of mutations does provide a tumor with a selective advantage; such mutations are usually termed *driver* mutations (Figure 1). Since the driver mutations are advantageous for tumor they are under positive selection during the tumor progression and growth. Therefore, such mutations are observed in the tumor samples with higher frequency compared to passenger mutations. Most of the known examples are affecting the gene-coding regions of known oncogenes. The three most frequently mutated cancer genes across multiple cancer entities are *TP53*, *PIK3CA* and *BRAF* with the corresponding mutational frequencies of 36.1%, 14.3% and 10%, respectively.

The COSMIC database provides a list of all known genes that are recurrently mutated in cancer (Forbes *et al.*, 2015).

According to the most recent estimations, around 98% of the human genome does not encode for protein sequences, but rather contains nearly all regulatory regions, such as promoters, enhancers and insulators. The entire cancer genomics field for many years was focused almost exclusively at the protein-coding regions. In this review I am addressing a question to which extend non-coding somatic mutations may be involved in cancer.

Non-coding somatic mutations in cancer

A number of recent studies have provided multiple lines of evidences suggesting that SNVs within the non-coding genomic regions can also be of relevance in cancer. For example, driver mutations in the promoter region of the *TERT* gene were identified at high frequency in human melanoma (Huang *et al.*, 2013; Horn *et al.*, 2013) and other malignancies (Vinagre *et al.*, 2013). The mutations occur recurrently at two nucleotide positions located only 22 bp away from each other. All of the observed mutations encode for C>T substitutions. According to the proposed mechanism, these mutations create an ETS binding site in the promoter region of the *TERT* gene, which leads to overexpression of the gene.

Two additional recurrent non-coding drivers were recently identified in a study performed on chronic lymphocytic leukaemia (CLL) (Puente *et al.*, 2015). A single mutation in the 3' UTR of *NOTCH1* gene was found in up to 6.7% of CLL cases harboring wild-type IGHV (Puente *et al.*, 2015). This mutation causes a novel splicing event within the last exon of *NOTCH1*, leading to removal of a PEST domain of *NOTCH1* and increased protein stability. The second driver region is located on chromosome 9p13 and contains an active enhancer. Somatic mutations in this enhancer lead to reduced *PAX5* gene expression levels (Puente *et al.*, 2015). Somatic mutations in this region are found also in other types of lymphoma, and account for up to 13% of IGHV-mutant CLLs (Puente *et al.*, 2015). Additional evidence for the relevance of intergenic mutations comes from a recent study reporting mutations recurrently observed in the bidirectional promoter of the *DPH3* and *OXNAD1* genes in multiple skin cancers (Denisova *et al.*, 2015), although the functionality of these mutations has to date remained undetermined.

A different mechanism by which somatic mutations can drive tumorigenesis was discovered in T-cell acute lymphoblastic leukaemia (T-ALL). In this example, heterozygous mutations form a super-enhancer upstream of *TAL1* oncogene that binds MYB transcription factor and therefore leads to mono-allelic overexpression of *TAL1* (Mansour *et al.*, 2014; Poulos *et al.*, 2015).

These recent findings provide multiple lines of evidence of important roles that intergenic somatic mutations may have in cancer. This motivated researchers to shift their focus from exomes to the non-coding regions and encouraged them to search for similar examples of recurrent non-coding mutations also using computational approaches. In the following section of this Chapter some of the challenges for such computational approaches will be discussed.

1.3 Mutational processes in cancer genomes

Cancer arises as a result of accumulation of somatic mutations over the lifetime of a patient. Somatic mutations are the outcomes of multiple *mutational processes* that can be caused by various exogenous and endogenous factors (Martincorena and Campbell, 2015). While each cancer carries only a handful of driver mutations, most of the variants observed in a tumor sample represent passenger mutations that are the product of the mutational processes that occurred through the development of cancer (Helleday *et al.*, 2014).

Some processes generate mutations throughout life at a constant rate and are referred as the clock-like mutational processes, while others act in an episodic manner producing mutations in bursts over short time periods (Alexandrov *et al.*, 2015). The final mutational portrait observed in the tumor is determined by the strength and duration of exposure of each mutational process (Alexandrov *et al.*, 2013). Thus, the passenger mutations can be accumulated at different rates due to various processes and can hence create problems for the computational identification of driver mutations based on the observed recurrence by masking the biological signal. Subclonal populations in cancer can be exposed to different mutational processes leading to an even higher complexity of the final landscape of somatic mutations (Helleday *et al.*, 2014).Therefore it is important to systematically to account for the *mutational heterogeneity* when performing cancer genomics studies.

1.3.1 Somatic mutational heterogeneity

Lawrence *et al.*, 2013 in their study performed on over 3,000 tumour samples across 27 cancer types produced by The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium, described three major types of mutational heterogeneity in cancer: 1) heterogeneity across patients with a given cancer type; 2) regional mutational heterogeneity and 3) heterogeneity in the mutational spectrum of the tumor.

The first type of mutational heterogeneity, heterogeneity across patients with a given cancer type, resulted in more than 1,000-fold differences in the median frequencies of non-synonymous mutations between individual tumors. Such variation in the mutational counts was associated with the tissue type the tumor originated from. For instance, pediatric cancers showed the lowest somatic mutation frequencies while lung cancers and melanoma harbored the highest number of mutations per megabase. Such difference can be explained by the exposure to exogenous factors such as ultraviolet light or tobacco smoke. More surprisingly, there were also remarkable differences in mutational frequencies between patients of the same cancer type, which could be explained by the mutational processes specific to the individuals. In a conclusion, it is important to take into account such effects while performing computational analysis, for example, by normalizing the observed number of SNVs by the average number of mutations in the given cancer type or/and the total number of mutations that the given individual genome harbors.

The second type of the heterogeneity described by Lawrence et al., 2013 is the *regional mutational heterogeneity*, which is the main challenge for computational identification of recurrent somatic variants.

Tumorigenesis is often seen as an evolutionary process, in which the driver mutations are being under positive selection and are therefore observed at higher frequencies in tumor samples. Computational approaches therefore typically aim at identification of individual point mutations or regions of the genome that experience recurrent somatic point mutation. As was already mentioned, somatic mutations in cancer are not evenly distributed across the human genome. In other words, there are regions within the human genome that tend to accumulate higher somatic mutation load than others. For example, a mutational pattern of localized somatic hypermutation called *kataegis*, that was first described in breast cancer (Nik-Zainal *et al.*, 2012) and later observed in many other cancer entities (Alexandrov *et al.*, 2013). The word “kataegis” is derived from an ancient Greek word for “thunder” (καταιγίς) because, when visualized in a plot, kataegis looks like a cluster of multiple somatic mutations that occur within a small genomic window. Such localized hypermutation is commonly observed near regions of somatic genome rearrangements in cancer. Recent studies have identified a link between kataegis and the activity of enzymes of the APOBEC family (Lada *et al.*, 2012; Burns *et al.*, 2013).

Another example of regions with higher mutational load than the averaged rates along the genome is localized hypermutation near the human immunoglobulin heavy-chain locus (IGH) on chromosome 14 in B cells (Richter *et al.*, 2012), which takes place also in healthy individuals as a necessary part of maturation of the immune system.

In contrast, other genomic regions such as closed chromatin appear largely depleted of somatic mutations (Martincorena and Campbell, 2015).

Various genetic and epigenetic features are known to correlate with somatic mutational rates in cancer. For example, Schuster-Boekler *et al.*, 2012 compared correlation between mutational rates and 46 various genetic and epigenetic features at different resolution, ranging between 1 kb and 100 Mb; and identified a feature – H3K9me3 – that most strongly correlated with cancer SNV density (Figure 1a). Additionally, they observed that many of the tested features were correlated and studied this further by principle component analysis. They found that up to 60% of the variance between the features was explained by the first principle component, at 1-Mb resolution.

Lawrence *et al.*, 2013 studied the same phenomenon on a larger cohort of samples and identified two factors that explained mutational heterogeneity at 100-kb-resolution best: gene expression levels and replication timing of a DNA region during the cell cycle (Figure 1b). Other features that correlated with mutational rates were chromatin state estimated from Hi-C data and GC content, at 1-Mb-resolution.

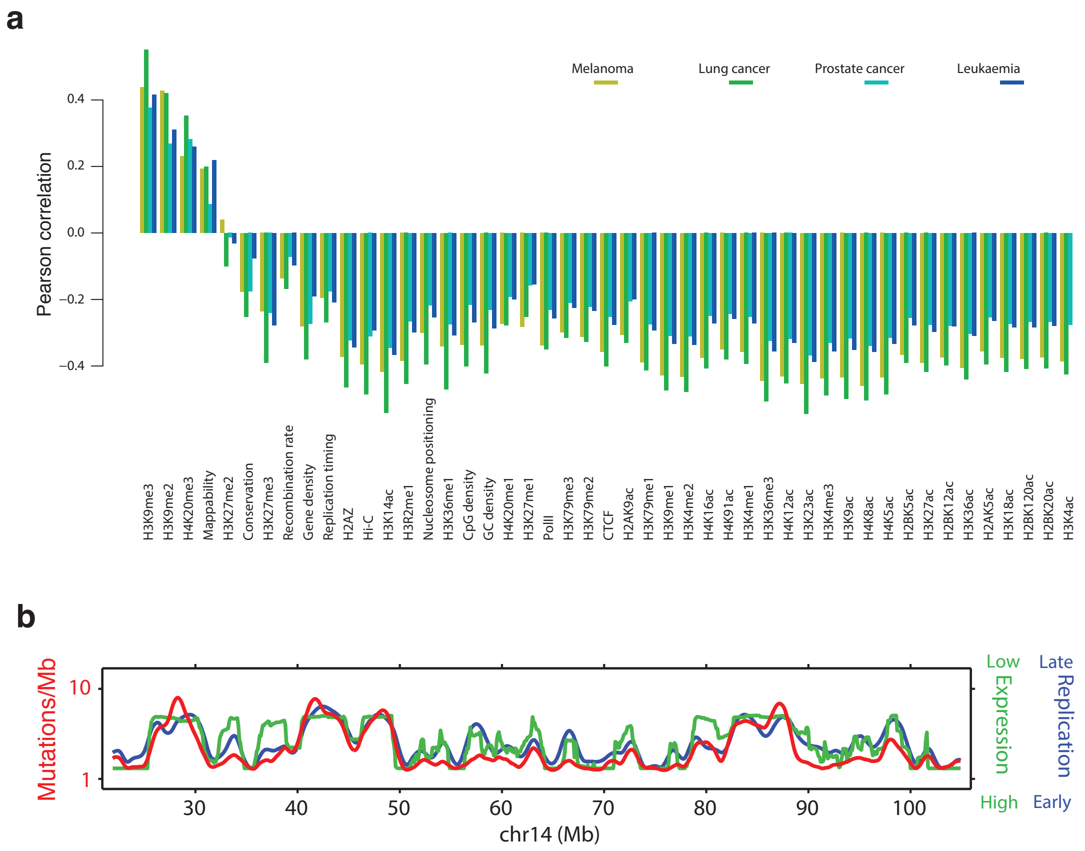


Figure 1. Correlations between SNV density from individual cancers at 1-Mb (a) and 100-kb (b) resolution with diverse genetic and epigenetic features. Figure adapted from Lawrence M et al, 2013.

Based on these results, Lawrence *et al.*, 2013 proposed to use replication timing as a covariate to control for background mutation rates and implemented it in their method MutSigCV. Similarly, other computational studies used DNA replication timing (Weinhold et al., 2014) or a combination of replication timing, base-pair type and transcript region (Melton *et al.*, 2015) for the background mutation rates estimation.

The last type of the mutational heterogeneity addressed by Lawrence *et al.*, 2013 was the heterogeneity in the mutational spectrum of the tumor. Using non-negative matrix factorization method the authors stratified all possible mutations into mutational spectra and represented them in a circular plot to identify clusters of tumor samples characterized by the prevalent mutational processes in the tumors. One their observations was, for example, that lung cancers shared a mutational spectrum of C>A mutations, that are associated with exposure to the polycyclic aromatic hydrocarbons in tobacco. In brief, the observed clustering of the cancer samples was associated with the predominant mutational signatures in the tumors.

1.3.2 Mutational signatures in cancer

Mutational processes leave on each cancer genome their characteristic imprints that are called *mutational signatures*. The number of mutations contributing to each signature is a proxy for the amount of exposure to each mutational process. Some of these processes have happened in the past and are therefore can be termed *historical*. They encode information on the previous exposures and can therefore be important in the context of cancer prevention. In contrast, the *ongoing* processes can be used for prognostic predictions, since they correspond to current factors associated with cancer (Helleday *et al.*, 2014).

Alexandrov *et al.*, 2013 identified 21 different mutational signatures across 30 various cancer types. One of them, the signature 1, correlated with age of diagnosis in some cancer types. Others were associated with known exogenous and endogenous factors. For example, C>T mutations at CpC or TpC dinucleotides induced by UV light were characteristic for signature 7 mainly found in malignant melanoma samples. Additionally, signature 4 showed transcriptional strand bias for C>A mutations associated with tobacco smoking in lung cancer.

Some of the mutational processes cause only one type of somatic mutation. For example, the carcinogen aristolochic acid causes almost exclusively A>T base substitutions. In contrast, the loss of the *BRCA1* or *BRCA2* genes in breast, ovarian, and pancreatic cancers is associated with a combinations of mutations, indels, as well as duplications and deletions occurring at a distinctive pattern (Martincorena and Campbell, 2015).

One of the most common mutational signatures in human cancers is represented by C>T or C>G substitutions at sites preceded by a thymine nucleobase and is caused by off-target modification of DNA by the APOBEC family of proteins (Martincorena and Campbell, 2015)

In conclusion, studying mutational signatures cancer can improve our understanding of the underlying biology of the tumors and some of their predisposing factors. Moreover, it can provide useful biomedical applications.

1.4 Detection of somatic mutations in cancer genomes using next-generation sequencing

Recent technological advances in DNA sequencing allow simultaneous sequencing of millions of short DNA fragments. These high-throughput methods are mainly referred to in the literature as next-generation sequencing (NGS) methods.

NGS is a powerful method to study structural alterations (Mills et al., 2011) and have been successfully applied to cancer research, for example, to prostate cancer (Weischenfeldt et al., 2013) and medulloblastoma (Northcott et al., 2014).

The most relevant application of NGS in context of my research is identification of somatic SNV in cancer. The most common experimental setup used in cancer genomics studies involves sequencing of tumor and matching control material for every patient.

1.4.1 SNV calling

A typical pipeline for somatic SNV calling in cancer from the next-generation sequencing experiments includes the following steps:

1) raw data preprocessing and quality control;

2) alignment of the reads to a reference genome;

3) SNV calling for tumor and control samples;

4) quality-base filtering of resulting SNVs;

5) identification of somatic SNVs as the variants that are present in the tumor but not in the matching control sample.

Various computational tools are available for performing each of the steps. For example, the most commonly used tools for identification of SNVs and small indels based on mapped sequencing data are Samtools mpileup (Li, 2011; Li *et* *al.*, 2009), GATK UnifiedGenotyper (DePristo *et al.*, 2011), and Freebayes (Garrison and Marth, 2012).

Nevertheless, it is important to consider possible sources of errors and biases when performing such analyses. Depending on the sequencing platform used, various biases may occur. For example, the strand bias, when the distribution of forward versus reverse directions in the aligned reads is highly unequal. Additionally, the PCR bias may lead to high duplication of certain reads.

Another important consideration when performing SNV calling is that the sequencing coverage is the crucial parameter (i.e. to have a larger number of reads supporting a variant). It is estimated that an average coverage of 30× is sufficient to call SNVs reliably in about 90% of the genome (Ajay *et al.*, 2011).

1.4.2 Computational methods to identify driver somatic mutations: focus on intergenic mutations

A typical property of somatic driver mutations is that they can be observed recurrently across cancer samples, a characteristic making their identification amendable to cancer genomics. Exome sequencing based cancer genomics studies, for example, have recently provided numerous insights into somatic driver mutations in protein-coding regions in numerous cancer types (Meyerson *et al.*, 2010). With the recent findings supporting important roles of somatic mutations within the non-coding regions of human genome, developing computational that would allow identification of novel candidate drivers became an important task, especially given the rapidly growing amount of large-scale omics data on cancer produced by large consortia such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) project (https://dcc.icgc.org/pcawg) (Stein *et al.*, 2015).

Three recent studies in particular have aimed at identifying functionally relevant intergenic mutations in large-scale cancer genomic datasets consisting of more than 400 whole-genome sequenced tumors (Weinhold *et al.*, 2014; Melton *et al.*, 2015; Fredriksson *et al.*, 2014). The brief summaries of their study design and obtained results are listed below.

Weinhold *et al.*, 2014 in their study on 863 tumors focused on the mutations that are most likely to affect regulatory elements. One of their approaches, for example, focused only on promoter and enhancer regions. To identify recurrently mutated regions the authors compared the observed mutational frequency with the frequencies in the background set of regions, which they constructed from the regions of the same category (for example, promoters) and with similar DNA replication timing. Binomial test was used to compute statistics. Their top identified candidate was the *TERT* promoter region with the mutational recurrence of 56 out of 863 samples. Other top candidates included *PLEKHS1* and *WDR74* regions mutated in 20 and 36 samples respectively.

Fredriksson et al. in their analysis of 505 tumors investigated associations between somatic mutations in regulatory regions and altered mRNA levels. In their analysis, the authors considered only somatic mutations within 0.5-100kb distance from transcription start sites of 445 COSMIC cancer-associated genes. Additionally they required for the mutations to be in DNAse I hypersensitivity sites and either create or disrupt ETS factor consensus binding sequence. They used a windows-based approach with partially overlapping genomic windows of 100 bp size and and identified *TERT* promoter region as a candidate and concluded that it might be an extreme example showing a strong association with the gene expression levels.

Melton *et al.*, 2015 performed their study on 436 cancer genomes, restricting their search space to all regulatory regions according to RegulomeDB (Boyle *et al.*, 2012) annotations. Sample-specific and covariate-corrected background mutation probabilities with Poisson binomial model were used to calculate statistics. For the estimation of the background mutation rates, a combination of replication timing, base-pair type and transcript region were employed. The authors identified the following candidates: *TERT* promoter region, a list of selected coding mutations in know cancer genes (*TP53*, *AKT1*, *PIK3CA*, *PTEN*, *EGFR*, *CDKN2A* and *KRAS*), as well as eight new candidate regions potentially regulating known cancer-associated genes such as *GNAS*, *INPP4B*, *MAP2K2*, *BCL11B*, *NEDD4L*, *ANKRD11*, *TRPM2* and *P2RY8*.

As a common ground, all three studies used window-based approaches to scan the genome and controlled for covariates – mostly replication timing – to estimate the recurrence of events in comparison to the background mutation rates. None of the studies, however, run their analysis in a fully unbiased (e.g. genome-wide) manner, without restricting their search space to a particular class of annotations. Moreover, the limited overlap that we observed in the findings of the studies, especially considering the similarities in the methodology used, raises a question to which extend the choice of parameters (e.g. statistical test or window size) can influence the outcomes of such computational approaches. We designed a study to address these questions and developed a computational approach to identify intergenic regions that are recurrently mutated across multiple cancer samples in an unbiased, genome-wide manner.

.

2. Towards developing a background model for mutational propensities in cancer

2.1 Motivation and background

Somatic mutations in cancer are not distributed evenly across the genome; certain regions are prone to undergo hypermutation while other regions may accumulate very few mutations. This phenomenon makes it difficult for computational biologists to identify genomic regions that are often mutated across multiple cancer patients, that is, recurrently mutated regions.

If we consider a dataset of genome-wide somatic mutations across various cancer types. We can bin the entire human genome into non-overlapping windows of a particular size. For any given window we can calculate the observed number of mutations among all patients. This will give us some estimate of the total mutation load in the given region. Another value that we can calculate from the same data is *mutational recurrence*, which is, how many samples have at least one mutation within the window or, in other words, how many patients have this region mutated. Next, if we want to identify where the driver mutations are, we will likely be more interested in the regions that are more frequently mutated than others, because of the assumption that driver mutations give an evolutionary advantage to the tumor and are therefore under positive selection in cancer. However, it is important to distinguish the regions with high total mutation load and the regions with high mutational recurrence. To better illustrate this, let's discuss how N mutations might be distributed within one genomic window. If all mutations mostly come from only one or few samples, this will likely indicate that these samples are outliers. The possible reasons for this could be both technical (due to batch effects or technical errors in sample preparation, sequencing, alignment or mutation calling) and biological (for example, the high number of somatic mutations in *IGH* locus in lymphomas). Therefore such scenario will not be interesting for us if we aim to identify potential driver mutations. In contrast, if the mutations come mostly from different samples and, especially, if they show a tendency to cluster within a small locus, this may indeed indicate a driver event. Therefore, we should be looking for windows that are mutated in many patients rather than containing many mutations in general. In other words, we should focus on identification of windows with higher mutational recurrence.

The main idea behind the identification of recurrently mutated regions is to compare the observed recurrence of the region to some estimate of the background mutational recurrence for a window of the same size. Since the mutational densities are not evenly distributed across the genome, we will naturally expect to observe high mutational recurrence in some regions and very low in others. Therefore, if our window of interest is among the rarely mutated ones, comparing its recurrence to, say, the average value across the entire genome will result in masking of the real signal. In contrast, the regions that are prone to hypermutation will likely always appear significant in such analysis.

Therefore, it is very important to establish an appropriate background model to correctly estimate expected mutation rates for different types of genomic regions. The main steps in order to do this are 1) to understand the mutational patterns in cancer and 2) to identify features that correlate with mutation rates in cancer.

2.1.1 Contribution

In this part of my PhD project my aim was to devise a mutational background model that accounts for regional mutational heterogeneity in multiple cancer samples and was suitable for the dataset I used. For this I studied correlations between various genetic and epigenetic features with background somatic mutation rates calculated from the dataset of interest.

The results described in this Chapter were included in the submitted manuscript Rudneva V *et al.*, 2016. I contributed all analysis described in this manuscript and I wrote the first and the principal version of this manuscript. Simon Anders and Wolfgang Huber provided additional mentorship and numerous useful suggestions on the design of my analysis approach.

2.2 Regional mutational heterogeneity across multiple cancer types

Dataset of somatic mutations

We obtained a dataset of genome-wide somatic mutations in various cancer types from previously published studies (Table 1). For each patient, a tumor-normal tissue pair of samples was whole-genome sequenced and somatic SNVs were called. The median number of somatic mutations per sample varies greatly between different cancer types which is consistent with the observations made by Alexandrov *et al.*, 2013 (Table 1).

For the analysis described in this Chapter, we considered only those cancer types that were represented by at least 20 patients, which resulted in 698 cancer samples in total, while for the analyses described in the Chapters 3.2.1 and 3.2.2 we used the entire dataset.

Cancer type	**Number of samples**	**Median number of somatic mutations per sample**
ALL	1	
AML	7	
Breast	119[1]	3,579
CLL	28[1]	1,920
Liver	88[1]	9,021
Lung Adenocarcinoma	24[1]	46,634

Lymphoma	76[1]	2,651
Medulloblastoma	209[1,2]	1,064
Pancreas	15	
Prostate	30[1]	2,206
Pilocytic Astrocytoma	101[1]	100

Table 1. Overview of the datasets used. The VCF files with somatic SNV calls were produced in the following studies: 1. (Alexandrov *et al.*, 2013); 2. (Northcott *et al.*, 2014; Kool *et al.*, 2014; Jones *et al.*, 2012; Jager *et al.*, 2013).

Correlation between features

Previous attempts to understand the patterns of mutational heterogeneity in cancer were performed by Schuster-Böckler and Lehner, 2012 and Lawrence *et al.*, 2013. They identified various genetic and epigenetic features that correlated with somatic mutational rates in cancer at 100-kb to 1-Mb resolution. These features included H3K9me3 (Schuster Boeckler et al), gene expression level, DNA replication timing, open vs. closed chromatin states, and GC content (Lawrence et al). They also observed, that some of the features tend to be correlated. For example, as was found using principle component analysis, up to 60% of the variance between the features at the 1-Mb-resolution can be explained by the first principle component, this is believed to reflect chromatin organization (Schuster-Böckler and Lehner, 2012). Since this correlation was not systematically tested at different resolution and for large sets (>400 samples) of cancer genomes, we assessed which of the features above have the strongest correlation with somatic mutation densities in cancer, and whether combined contributions of features correlate with the mutational densities more strongly than the individual features alone.

Feature	**Reference**
Replication Time	Lawrence M. *et al.*, Nature 2013
Averaged gene expression levels (CCLE)	Lawrence M. *et al.*, Nature 2013
GC content	UCSC Genome Browser
H3K9me3	Barski A. *et al.*, Cell 2007
Chromatin state (Hi-C data)	Lieberman-Aiden *et al.*, Science 2009

Table 2. Genetic and epigenetic features used for Principal Component Analysis.

To study the correlation between individual features we performed principal component analysis (PCA) on five features: replication timing, gene expression level, GC content, H3K9me3 and open vs. closed chromatin conformation. In order to obtain results that could be compared to previous studies, we used data from the same sources that were used in work by Schuster-Böckler and Lehner, 2012 and Lawrence *et al.*, 2013 (see Table 2 for the details). We binned the entire human genome (hg 19 build, with no masking of the regions) into 100 kb windows resulting into 28,800 windows in total. For every window we obtained values for each of the features by either taking the values directly from the source or averaging the value for the window if the window corresponded to multiple feature annotations. This way we obtained a 28,800 by 5 matrix, in which rows represented genomic windows and columns represented the five features tested. We performed PCA on this matrix and observed that the first principal component alone explained 60% of variance between the features (Figure 2a), while the second principal component explained only 14% of the variance. This observation is similar with what Schuster-Böckler and Lehner, 2012 have observed in their study assessing a less extensive set of cancer genomes. The fact that we identified one feature explaining most of the variance indicates that there is a strong correlation between several features. In other words, various genetic and epigenetic features reflect a higher-level organization of the genome.

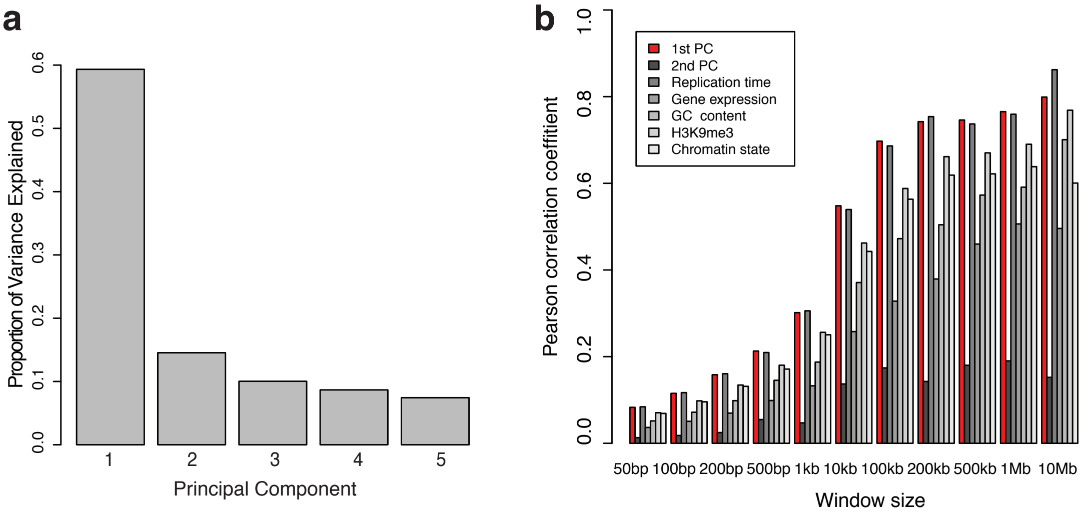


Figure 2. Principal component analysis on five features. (a) Proportion of variance explained by the principal components. (b) Correlation between somatic mutational densities and individual features as well as with 1st and 2nd principal components (PCs) of the five features for different window sizes.

Best correlate with mutation rates

In order to be able to estimate background mutation rates for any given region in the human genome, one should first find a way to stratify the genome into regions with comparable genetic and epigenetic properties. For this, one first needs to identify the best correlate with the mutation rates in the existing data. We therefore addressed the following questions in this part of the analysis: 1) which of the individual features alone correlate best with the mutation rates in cancer? 2) Does the combined contribution of the features captured by the PCA loading vectors show a higher correlation with the mutation rates than the individual features alone?

Since it is expected that the correlations will highly depend on the window size (smaller window sizes will have smaller mutation counts and this will lower the correlations), we studied the above-mentioned effects on eleven window sizes ranging between 50bp and 10Mb.

To perform the analysis we first constructed a matrix for each window size (11 matrices in total). In each matrix:

- rows corresponded to genomic windows;
- columns 1-5 corresponded to the individual features;
- columns 6 & 7 corresponded to the 1st and 2nd PC loading vectors obtained from the PCA as described above;
- column 8 contained the total number of mutations across 698 cancer samples (Table 1).

For each window size, we calculated Pearson correlation coefficient between the mutation rates (8th column) and each of the other 7 columns from the matrix. Since some features are known to correlate positively with the mutational rates (replication timing), while others have a negative correlation with the mutation rates (gene expression level), we considered the modus value of the Pearson correlation coefficient in our analysis.

The results of our analysis are shown in Figure 2b. As was expected, all of the tested correlations improved with increased window sizes, which is likely due to the fact that the higher number of observed mutations per window, when using larger window sizes, allows for more precise estimates of the actual mutation rate. Replication timing had the strongest correlation with mutation rates as an individual feature at all window sizes; this supports the selection of this feature in previous studies to control for local and global background mutation rates. However, when comparing the correlation between mutation rates and replication timing with the 1st PC loading vector we observed a slightly better correlation for the first principal component for window sizes of 50bp-1Mbp. This observation motivated us to propose to use the 1st PC loading vector to estimate the expected regional mutation rate for the window-based approaches where window sizes range between 50bp and 1Mbp.

2.5 Discussion

Based on the results of our analysis we concluded, that the replication timing is the best correlate with somatic mutation rates when compared to individual features. Nevertheless, the combined contribution of the five features captured by the 1st PC loading vector outperforms replication timing at window sizes between 50bp and 1Mbp. We therefore propose to use the 1st PC loading vector to stratify the human genome into groups of regions with comparable genetic and epigenetic properties and thus to estimate the background mutation rate for the any given region of interest. However, it is important to keep in mind, that due to our still limited knowledge about mutational patterns in cancer, even a combined contribution of the five different features is likely to be only a useful approximation of the true background mutation rates and likely does not represent a complete estimate for all of the mutational heterogeneity existing in cancer genomes.

3. Identification of intergenic somatic driver mutations in cancer

3.1 Motivation and background

Examples of somatic driver mutations within the exonic regions of oncogenes such as HER2 and MYC have been known to cancer biologists for a long time. The mechanisms by which they act are mostly well described and studied. Multiple previous studies aimed to identify novel harmful exonic mutations and succeeded in this with multiple computational tools and pipelines now available for this purpose (Wang *et al.*, 2010; Aleman *et al.*, 2014). The recent decrease in whole-genome sequencing costs over the last years has made it possible to search for somatic driver mutations also within the intergenic regions of the genome. There are large datasets on whole-genome sequencing in cancer available to-date, including TCGA and ICGC projects. The computational methods and pipelines for this area are however still in the beginning of development and there is no golden standard approach existing for the identification of somatic driver mutations within intergenic regions.

Driver mutations are by definition advantageous for the tumor evolution. Therefore they are *recurrently* observed in the tumor samples and have a *functional* relevance for the tumor. As described in Chapter 2, the mutational recurrence of an event needs to be compared to the background mutation rate. Several mechanisms by which intergenic mutations play a role in tumorigenesis have been proposed. We used the *TERT* promoter mutations case as an example of functional impact. Driver mutations within the *TERT* promoter region create a transcription factor binding site for an ETS factor that leads to overexpression of the *TERT* oncogene. We took two approaches in our search for novel somatic driver events within intergenic regions in the cancer genome.

1) Identification of intergenic mutations that have functional consequences for the tumor.

- In the first project, we focused on identification of mutations that are associated with changes in gene expression. This study was performed on a dataset that consisted of 23 lymphoma samples for which both somatic mutation and gene expression data were available.

- In the second project, we focused on the somatic mutations within intergenic regions that changed transcription factor binding sites (TFBS). This study was performed on a dataset that consisted of 745 samples from eleven cancer types, for which somatic SNV calls were available.

2) Identification of intergenic somatic driver events based on the mutational recurrence. The study was performed on 698 samples from the dataset mentioned above. Results described in this section were included in the submitted manuscript Rudneva V *et al.*, 2016. All analysis included in this manuscript was performed by me. I also wrote the initial draft of the manuscript. Simon Anders and Wolfgang Huber contributed to this work by providing additional mentorship and useful comments on the study design and the text of the manuscript.

3.2.1 Identification of intergenic somatic mutations associated with gene expression changes in lymphoma

In this approach we aimed to estimate the functional impact of intergenic mutations on the gene expression. For this we aimed to identify SNVs that are located upstream of a gene (ideally, in its promoter region) and are associated with the expression changes of the gene (Figure 3).

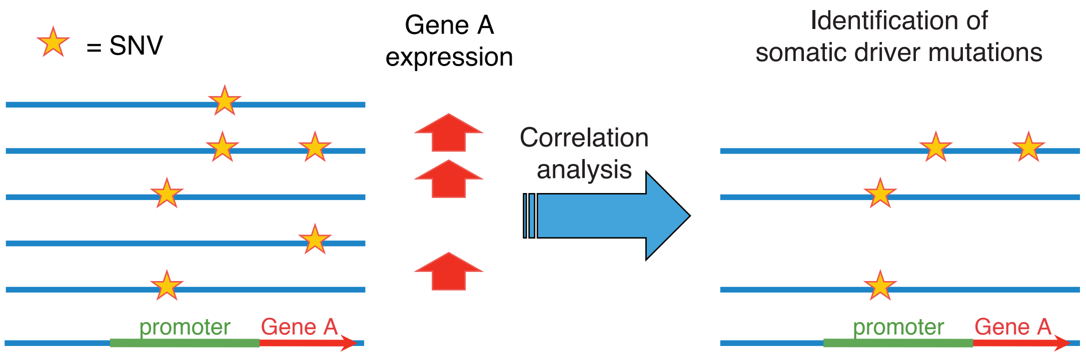


Figure 3. Overview of the approach for identification of intergenic somatic mutations that correlate with expression changes in close by genes

We combined this principle with a straightforward annotation and filtering pipeline (Figure 4) and performed our analysis on 23 lymphoma samples for which both somatic SNVs and gene expression data were available (Richter *et al.*, 2012).

In brief, we obtained a list of somatic mutations that were present at the exact same position in at least two samples; we filtered out mutations in "unreliable" regions and extended the resulting high-confidence mutations to genomic regions of 500bp upstream and downstream of the mutation. We merged the neighboring regions if they overlapped, but kept the window size the same for all windows. We annotated the resulting genomic regions using various sources, such as ENCODE and COSMIC (Forbes *et al.*, 2015). As the last step, we identified genes that had statistically significant changes in the expression values between the samples with and without mutations within the regions of interest.

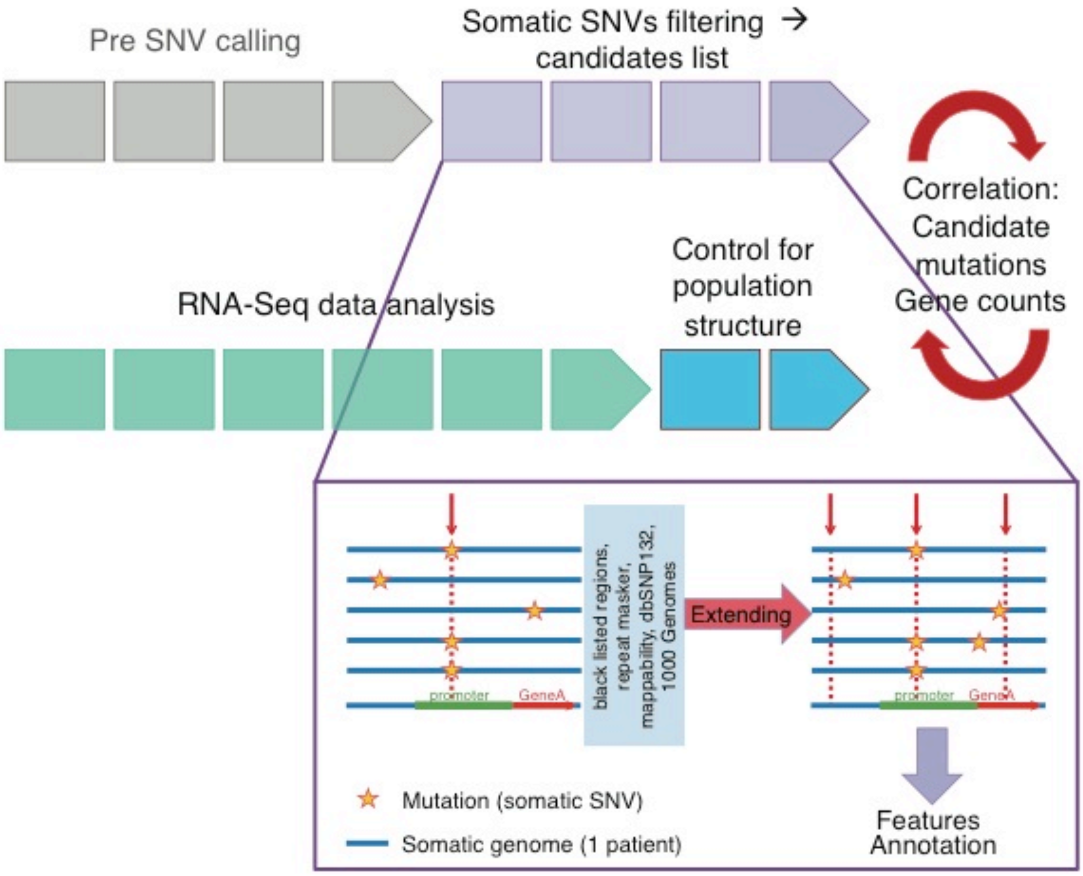

Pre SNV calling
Somatic SNVs filtering → candidates list
Correlation: Candidate mutations Gene counts
RNA-Seq data analysis
Control for population structure
black listed regions, repeat masker, mappability, dbSNP132, 1000 Genomes
Extending
promoter
GeneA
promoter
GeneA
Mutation (somatic SNV)
Somatic genome (1 patient)
Features Annotation

Figure 4. Overview of the SNVs annotation and filtering pipeline combined with transcriptomic data integration.

Somatic SNV calling was performed by Tobias Rausch. By considering only the mutations that were present in at least two individuals, we obtained a list of 121 recurrent somatic mutations. In order to identify only the high confidence somatic variants, we filtered this list based on the following criteria:

1) all mutations occurring within the DAC Blacklisted Regions produced by the ENCODE project were excluded from the analysis. The DAC Blacklisted Regions consist of anomalous, unstructured regions of the genome, as well as the regions with high signal/read counts in next generation sequencing experiments independent of cell line and type of experiment. There regions were initially obtained using 80 open chromatin tracks (DNase and FAIRE datasets) and 20 ChIP-seq input/control tracks spanning ~60 human tissue types/cell lines in total.

2) mutations that overlapped with regions of low mappability were filtered out. Mappability is a metric that represents the regions in the genome that cannot be uniquely mapped given the read length and therefore SNVs that were called in such regions cannot be considered high-confidence.

3) mutations within the highly repetitive regions according to the RepeatMasker (http://www.repeatmasker.org/) were also not considered.

4) mutations that are present in the dbSNP132 database or identified by the 1000 Genomes Project as known polymorphisms. Such mutations are very likely to be germline variants that were mistakenly called as somatic (mainly because of the lack of coverage at that locus in the control sample).

After applying the abovementioned filtering, we identified 42 high confidence SNVs. We extended each of the point mutations to 1000 bp regions (500 bp upstream and 500 bp downstream of the mutation). We merged the overlapping regions and centered the mutations so that each region was of 1000 bp size. This way we ended up with 21 regions of interest (Table 3).

Genomic region	Total number of somatic mutations
chr14: 106326613 - 106327613	117
chr14: 106329616 - 106330616	80

chr14: 106328615-106329615	62
chr2: 89158984-89159984	51
chr14: 106327614-106328614	46
chr2: 89160096-89161096	39
chr18: 60985882-60986882	35
chr18: 60984597-60985597	33
chr3: 187462336-187463336	32
chr8: 128748603-128749603	32
chr22: 23222891-23223891	29
chr14: 106330617-106331617	15
chr3: 187660483-187661483	9
chr14: 106732659-106733659	7
chr18: 60805818-60806818	5
chr3: 4941250949413509	4
chr19: 11133752-11134752	3
chr8: 4042874-4043874	2
chr3: 50293195-50294195	2
chr21: 45381115-45382115	2
chr14: 37874607-37875607	2

Table 3. List of candidate 1 kb genomic regions containing high confidence somatic mutations.

Integration with gene expression data

To identify among the genomic regions of interest those that contain driver events with potential *cis*-effects on the gene expression we applied a simple burden test. For each of the 21 regions of interest we calculated Pearson correlation coefficients between genotypes and expression values of ten closest genes. In our analysis, we called a "genotype" the state of 1kb region: 0 if there were no mutations within the region in the given patient and 1 if the patient had at least one mutation in the 1kb region. This way we constructed a 23 by 21 matrix, in which rows represented individual patients and columns represented the regions of interest; each value corresponded to a genotype (0 or 1).

The main assumption for performing such correlation analysis using a burden test is that somatic SNVs have a strong effect on the expression of the genes. However, some effect on the gene expression levels may come from the genetic variability between the individuals. Indeed, consider two individuals coming from two different populations in which one has a polymorphism in the gene of interest that determines it's higher expression level in comparison to the individual from the other population. Then, assuming that there is an effect of the somatic variant on the gene expression but it is smaller than the effect of a SNP, the fact that the two individuals are coming from two different populations will mask the effect of the somatic SNV. It is therefore very important to control for population structure of the patient samples while performing phenotype-genotype correlation analysis. If patients in the dataset are coming from distinct populations, the common SNPs present in the data (and that could eventually be still present among the somatic mutations) may influence the results of the analysis by driving the correlation. We therefore performed principal component analysis (PCA) on the lymphoma dataset combined with dataset from the HapMap project (the reference on variation between distinct human populations) to make sure that there was no significant population structure between individuals. The full list of HapMap populations used in this analysis and the descriptors are listed in the Table 4.

We observed that most of our lymphoma samples (Figure 5, red dots cluster) were coming from the same population, so there was no need to account for population stratification before performing the correlation analysis.

Descriptor	Population
ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MXL	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscani in Italia
YRI	Yoruba in Ibadan, Nigeria

Table 4. List of HapMap populations used for the population structure analysis.

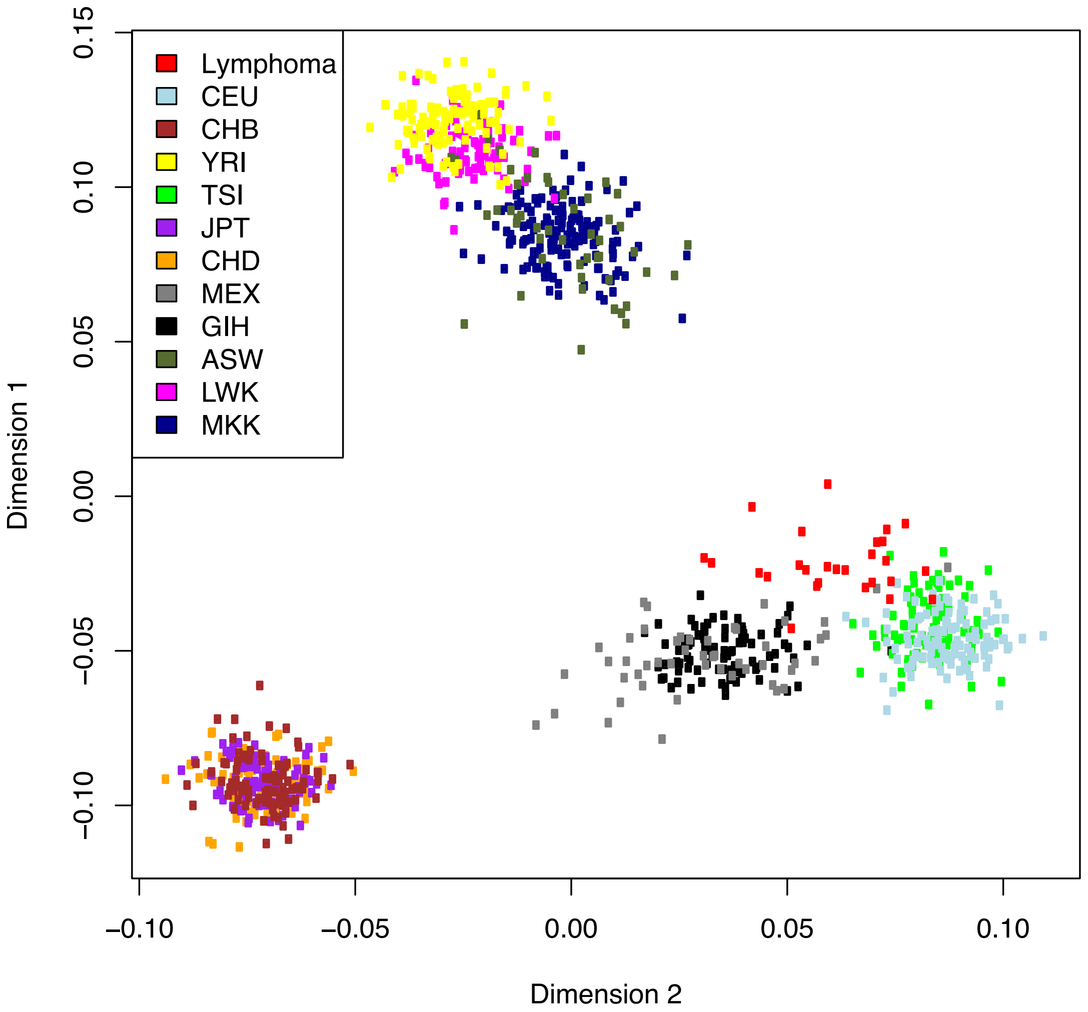


Figure 5. Principle Component Analysis on 23 lymphoma samples and HapMap individuals showing the population structure. The lymphoma samples used in this analysis are shown in red.

The RNA-Seq data preprocessing, quality control assessment from raw reads and alignment were performed using a custom pipeline and various publically available tools (see Methods for more details). As the result of this analysis, gene expression values were represented in reads per kilobase per million (RPKM) values. As we considered only the *cis*-effects of the regions on gene expression in this analysis, we identified ten top closest genes for every candidate genomic region and calculated Pearson correlation coefficients between their expression values and the regions' genotypes. We used FDR for multiple testing correction and a significance threshold of 0.05 for the resulting q-values.

We identified four genomic regions that significantly correlated with expression levels of three genes: *BCL2*, *CBS* and *MIPOL1* (Figure 6).

BCL2 is a gene known to drive follicular lymphoma. The regions of interest that we identified to be associated with its expression changes in the presence of mutations were both located in a close proximity on the chromosome 18. The first region was mutated in eight samples, while the other one was mutated in seven samples. Even though the correlations between the gene expression and the regions genotypes were statistically significant, the genomic locus containing *BCL2* gene is known to undergo structural rearrangements during the lymphoma development and therefore does not provide any novel drivers.

Two more genes, *CBS* and *MIPOL1,* were not previously associated with lymphoma and therefore could serve as interesting candidate mutations for the follow-up studies. Unfortunately, we did not observe that the regions were frequently mutated among our samples: both genomic loci were mutated only in two out of 23 samples each. This is likely the reason why the statistical significance values were borderline for both cases (Figure 6).

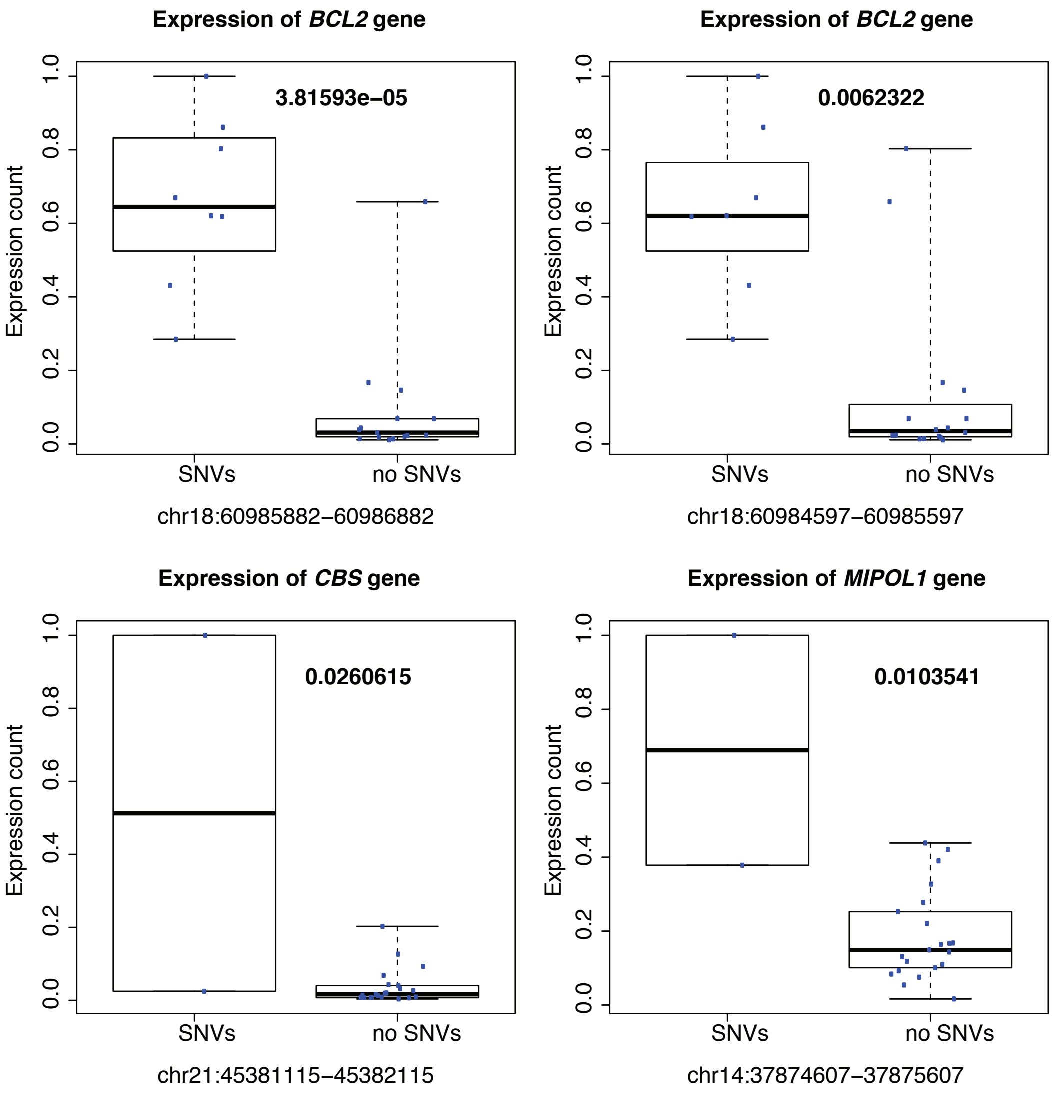


Figure 6. Differences in gene expression between samples harboring mutations within the regions of interest and samples with no mutations within the regions. Only significantly differentially expressed genes are shown (FDR corrected p-value < 0.05, shown on plots).

Discussion

Note that even though association with gene expression is an indication that the region might contain driver mutations, high mutational recurrence of the event (how many samples are mutated) is required to score a potential candidate high. Indeed, if an event is observed in only one or two samples this might be a clear indication that these samples are outliers or the somatic calls are technical artifacts. Moreover, even if the event is indeed present in the samples with such a low frequency, it will be very difficult to experimentally validate and study the mechanisms of action if the event we study is very rare. Lastly, even if the event is a true driver and has indeed a functional role that could be confirmed experimentally in the lab, this might still be a unique case and not represent the common process of tumorigenesis in humans and, therefore, will likely not make its way into medical practice due to the high costs of clinical trials for rare diseases.

Taken into account all of the above points, here we did not propose any new candidate drivers and concluded that we don't have enough statistical power to identify new drivers in this setup. We however proposed three solutions that could help improving the analysis: 1) obtaining a larger dataset, consisting of somatic variant calls and gene expression data matched for all samples; 2) focusing the analysis on other functional consequences of the driver mutations presence (for instance, transcription factor binding changing by somatic mutations); 3) focusing the entire analysis on identification of genomic regions that are recurrently mutated across multiple cancer samples.

Since in our case we could not obtain a significantly larger dataset as described in (1) during the time of this PhD work, we focused on performing the (2) and (3), as will be presented in the following sections of this Chapter.

3.2.2 Identification of intergenic somatic mutations associated with predicted TFBS changes in multiple cancer types

As was mentioned in the previous section, for most of the cancer samples with somatic mutation calls the matching gene expression data were unfortunately not available. We therefore could not estimate the functional impact of somatic mutations by directly comparing gene expression values to the genotypes. However, we could still estimate the effect of the SNVs indirectly, by inferring the changes in transcription factor binding sites (TFBS) sequences introduced by the mutations. This analysis setup was motivated by the *TERT* promoter example, in which two somatic mutations create a TFBS for the ETS transcription factor leading to the overexpression of the *TERT* gene.

Since the *TERT* promoter is not a cancer-type-specific driver but is rather known to be involved in various cancer types (Vinagre *et al.*, 2013), we employed this idea to identify similarly acting driver mutations across different cancer types.

For this purpose we used a dataset of 745 cancer samples of 11 different cancer types (1 ALL, 7 AMLs, 76 lymphomas, 119 breast cancers, 28 CLLs, 88 liver cancers, 24 lung adenocarcinomas, 256 medulloblastomas, 15 pancreas cancers, 30 prostate cancers, 101 pilocytic astrocytomas; see Table 1). Cancer types varied greatly in amounts of somatic mutations observed per sample both in total and for intergenic regions only (Figure 8). Median number of somatic mutations per sample was in a range between 10^2 to 10^5, which is consistent with previous reports (Martincorena and P. J. Campbell, 2015). As expected, the highest mutational load was observed in lung cancers, while the least number of somatic mutations were identified in pediatric samples (e.g. pilocytic astrocytoma). In addition, similarly to Lawrence *et al.*, 2013, we observed variability in mutational densities between the patients with the same cancer type. Our observations are particularly interesting because here we studied the mutational heterogeneity on the largest dataset of whole-genome sequenced tumors available to-date.

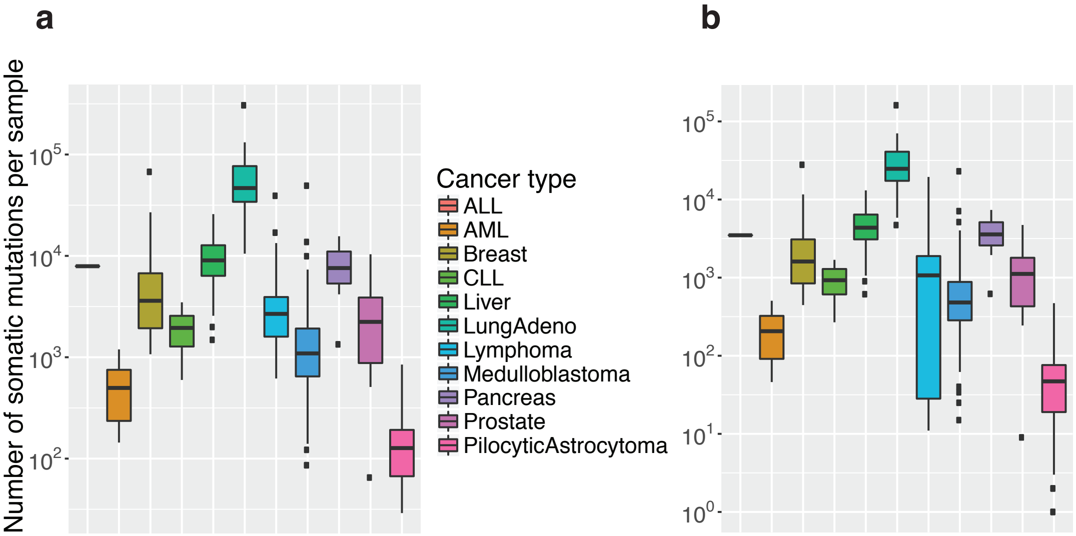


Figure 8. Distribution of somatic mutational counts across different cancer types for (a) all mutations and (b) intergenic mutations only.

Similarly to the analysis performed on lymphoma samples, we first collected a list of all somatic mutations across multiple tumors and then selected SNVs that were present at the exact same nucleotide position in at least two individuals and annotated them as recurrent. We filtered out the resulting list of recurrent SNVs using the following criteria:

1) for each recurrent mutation of interest, there must be another recurrent SNV within 100 bp to make sure that the mutations occur within the same regulatory region of the genome.

2) the two recurrently mutated positions must not be present in the same patient. This comes from an assumption that if there is already one driver mutation that has an affect on the gene expression, there will be no selective pressure for the other position to be mutated in this patient (e.g. mutually exclusive mutations within the *TERT* promoter). We called this an *anti-correlation principle*.

3) each of the mutations must be motif changing. As shown schematically in Figure 5, mutations might create a new TFBS sequence, like in the case of *TERT*, or they might disrupt an already existing one by introducing a mismatch.

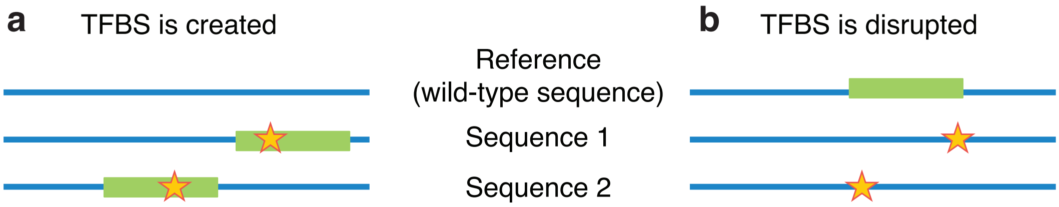


Figure 9. Possible scenarios of motif changing caused by driver somatic mutations. Sequences with somatic mutations (stars) are compared to the reference sequence. (a) a TFBS motif (green rectangle) is being created by introducing a mutation at any of the positions; (b) mutations at any of the nucleotide positions disturb an existing TFBS motif.

Before the filtering steps, we had a list consisting of 159,584 SNVs mutated at the exact same position in at least two out of 745 cancer samples. Following the first filtering step we identified 5,038 genomic loci of 100 bp size in which at least two positions were recurrently mutated. Among them 2,230 loci satisfied our anti-correlation criterion (neighboring recurrent SNVs are never present in the same patient). When considering non-coding loci only, we ended up with 2,211 loci of interest for which we computationally tested if they were motif changing based on the scheme shown in the Figure 9. This resulted in identification of 157 genomic regions of interest.

It is important to note, that the last filtering step is a very strict bottleneck in our analysis pipeline. Because of a large number of candidate loci coming from the previous analysis step, we had to rely on computational tools to predict if the presence of a mutation could create/disrupt a motif. Given the limited knowledge of motif binding patterns in the genome and the still early state of computational motif binding prediction, one should be aware though that this last step may not be highly reliable at the moment. We therefore performed a more detailed analysis of the 157 candidates. Among them, 66 were intergenic. Regions, overlapping with known copy-number variation and commonly repeated regions are a source of possible false positives and we therefore further filtered out such loci where the mutations of interest overlapped. We ended up with nine regions that we could call high confidence. Further investigation showed that among the nine candidate regions, four overlapped with DNase I hypersensitive sites according to the ENCODE project annotations and two overlapped known transcription starts site enhancers according to the FANTOM5 project data.

We continued with the visual investigation of all highly confident regions in the genome browser. We observed that all the individual cases of mutations that were predicted to be motif changing in our computational analysis had very low recurrence. Similarly to conclusions made in the previous section of the Chapter, we could not gain any mechanistic insights into cancer biology for such low recurrent events. We therefore focus our final analysis on identification of recurrently mutated intergenic regions.

3.2.3 Identification of intergenic regions recurrently mutated across multiple cancer types

In this section, I will first present results of a naïve approach to identify recurrently mutated intergenic regions, followed by some statistical issues that we identified to be associated with the approach. Lastly, I will present the results of our refined approach that we developed to overcome those issues. This approach was presented in the submitted manuscript Rudneva V *et al.*, 2016.

Naïve approach for identification of recurrently mutated regions

As discussed in detail in Chapter 2, regional mutational heterogeneity is a confounding factor for computational identification of recurrently mutated regions. Clearly, finding an approach for correcting differences in regional mutational propensities will allow the reduction of mutational biases and increase the statistical power for identifying mutations in intergenic regions of relevance in cancer.

To address this issue we first used a naïve approach that operated as follows (Figure 10). First, we binned the human genome into non-overlapping genomic windows of 100 kb size and employed a list of covariates (both genetic and epigenetic) to select regions with a comparable environment. In brief, we performed principal component analysis on five genetic and epigenetic features that were previously reported to correlate with somatic mutational densities in cancer (see Chapter 2 for the details and Table 2 for the complete list of features). We used the 1st principal component loading vector to stratify the entire genome into 25 groups of genomic regions with comparable genomic background (each group contained 973 or 972 genomic regions). We then ranked genomic regions within each group and identified regions that were mutated more frequently in comparison to other regions with similar background mutational rates (e.g. the regions that were >0.99 quantile outliers within the group). To assess the mutational patterns for each cancer type individually, we selected 100 kb genomic regions that were frequently mutated in the given cancer type but not in other cancers. We assumed those across-all-cancer-types frequently mutated regions do not harbor biologically relevant somatic drivers but rather represent loci of somatic hypermutation. Latter would likely be a consequence of tumorigenesis rather than its cause.

Using the *TERT* promoter as an example, we expected to identify candidate loci of a relatively small size of around 50 bp. We hence searched for regions of 50 bp in size that contained multiple SNVs within the initial windows of 100 kb that we used in the first steps of the analysis. Finally, the resulting regions were manually screened to identify high-confidence candidate loci.

We started the analysis with 100 kb regions for the genome stratification step and only at the latest stage focused on smaller window sizes that were more biologically relevant. We designed our approach this way in order to avoid statistical issues that are known to arise due to low mutation counts in small windows. To be able to select frequently mutated regions in a cancer-type-specific manner (step 3 in the scheme from Figure 10), we required each cancer type to be represented by at least 20 samples (Table 1), which resulted in a dataset of 698 samples.

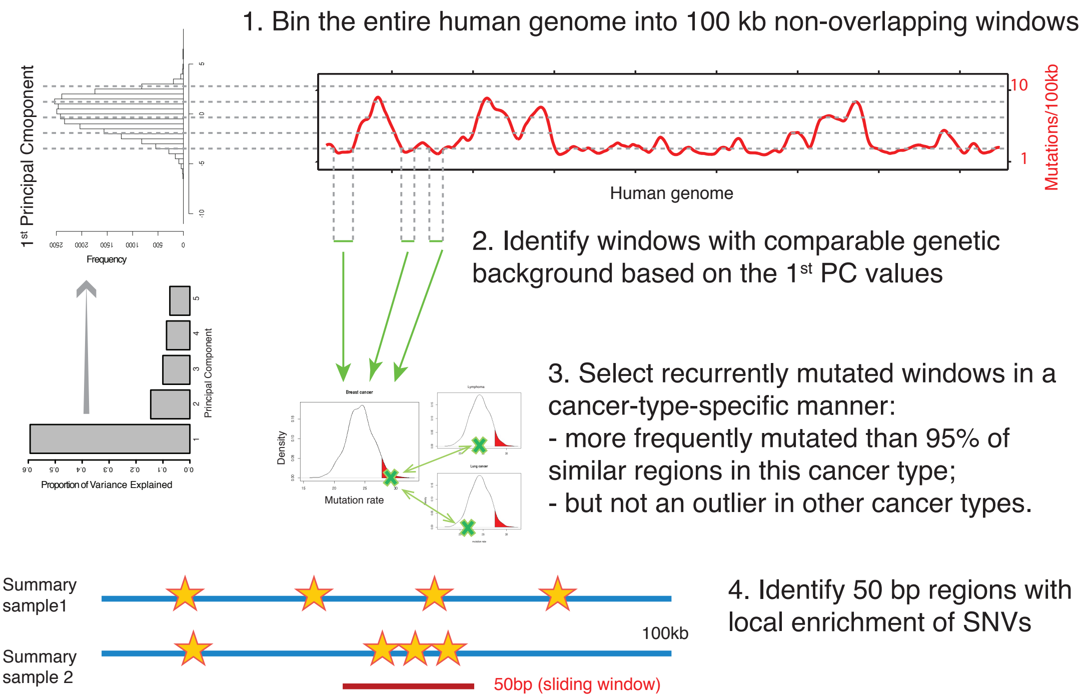

1. Bin the entire human genome into 100 kb non-overlapping windows
1st Principal Cmoponent
Frequency
10
1
Mutations/100kb
Human genome
2. Identify windows with comparable genetic background based on the 1st PC values
Proportion of Variance Explained
Principal Component
3. Select recurrently mutated windows in a cancer-type-specific manner:
- more frequently mutated than 95% of similar regions in this cancer type;
- but not an outlier in other cancer types.
Breast cancer
Density
Mutation rate
Lymphoma
Lung cancer
Summary sample1
Summary sample 2
100kb
50bp (sliding window)
4. Identify 50 bp regions with local enrichment of SNVs

Figure 10. Naïve approach overview. Detailed description is in the main text.

To computationally validate that our pipeline could indeed identify potential candidate somatic drivers, we investigated the functional impact of the mutated regions on expression of the corresponding genes, using the same assumption as in 3.2.1 section of this Thesis. We chose a cancer type that was represented by a large number of samples (Table 1) and for which we had matching expression data available in the lab. Medulloblastoma dataset consisted of 209 samples in total with expression data available for 128 of them. Additionally, our lab had access to the subgroup information, copy-number variants calls and enhancers data for most of the samples.

We ran our pipeline on the dataset of 698 samples from eight cancer types and identified 50 candidate regions of 100 kb size that were frequently mutated in medulloblastoma in comparison to other cancer types. 39 out of the 50 candidate regions overlapped with the list of medulloblastoma-specific enhancers that was provided by Serap Erkek. To further study the candidate regions, we identified loci of 50 bp that contained multiple mutations closely located within the 100 kb candidate regions. For some of 100 kb candidate regions we identified more than one 50 bp locus with clustered mutations. We identified in total 83 such loci and studied them closely.

A detailed study of the candidate loci allowed us to identify frequent exonic mutations in known WNT signaling pathway genes (*CTNNB1*, *SMARCA4*, *PTCH1*). Notably, only the samples with medulloblastoma WNT subtype had mutations in those regions. In addition to these known exonic drivers our pipeline also identified *TERT* promoter mutations as candidates. To confirm that identified mutations affected *TERT* gene expression we compared expression values between samples with and without *TERT* promoter mutations. We used Mann-Whitney U test to assess statistical significance of the difference in gene expression between the two groups of samples. Indeed the mutated samples showed significantly higher expression values (Figure 11a). However, a number of outliers in the non-mutated group had extremely high expression values. *TERT* high-level copy-number amplifications are commonly observed in medulloblastoma and are known to lead to extreme outlier expression of the gene (Northcott *et al.*, 2012). Therefore, it is sensible to not consider samples with high-level TERT copy-number amplification in our gene expression analysis. Being aware of that, we removed two samples with known *TERT* amplifications. Notably, those samples had the highest expression values among all medulloblastoma samples. Excluding those samples from the comparison slightly influenced the expression analysis results (Figure 11, a vs. b; 0.0146 vs 0.0133 Mann-Whitney U test p-values). This let us hypothesize that the remaining outliers could also be due to genomic alterations involving the *TERT* gene region. For example, they may be explained by the fact that *TERT* mutations occur in the context of a highly repetitive promoter sequence, which may have resulted in under-called mutations in the *TERT* promoter. Alternatively, this effect could be caused by distal somatic structural alterations affecting enhancers of *TERT* (Peifer *et al.*, 2015).

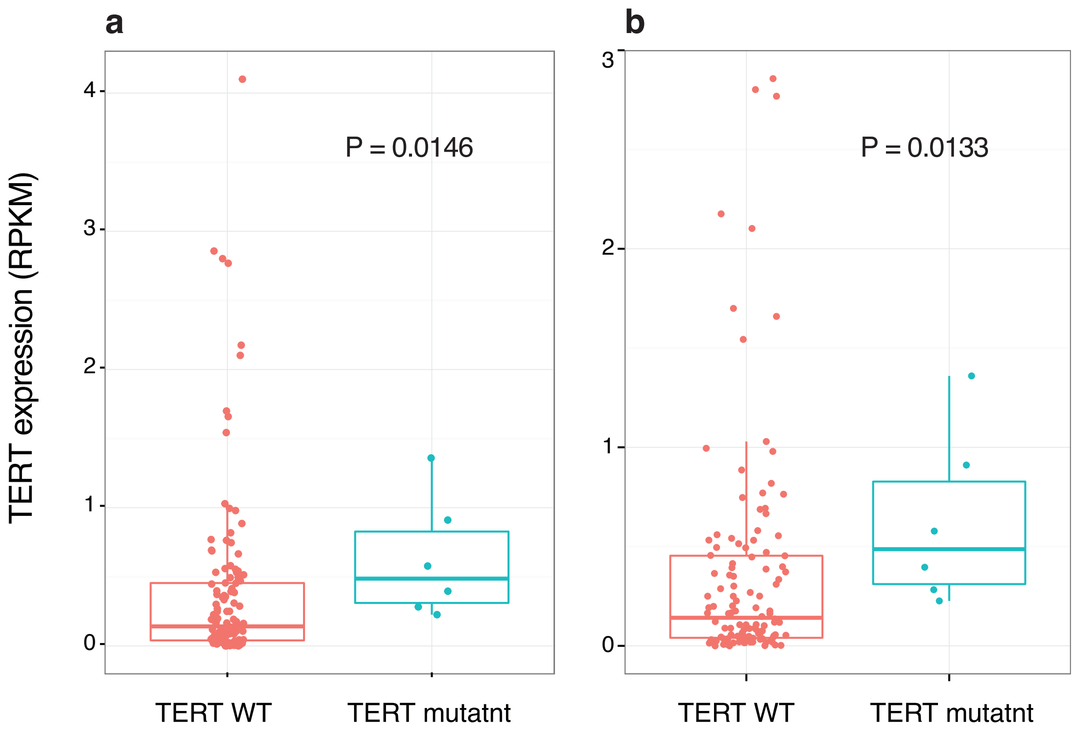


Figure 11. Differences in the *TERT* gene expression in medulloblastoma samples with mutated promoter region and the samples that do not harbor mutations in the region of interest. Gene expression levels are normalized. (**a**) Expression values for all medulloblastoma samples were used: samples with promoter mutations (n = 6; red) vs. samples with wild-type (WT) *TERT* (n = 122; blue). (**b**) Two samples with known TERT amplification status were excluded from analysis.

In summary, using our pipeline we identified a list of known exonic somatic drivers in medulloblastoma and a known intergenic driver event in the promoter region of *TERT* that was associated with expression changes of the gene. The focus of our study was to identify novel intergenic driver events. We hence annotated our list of candidate 50 bp regions by their genomic location (coding or intergenic) and additionally filtered out all the regions that had no overlap with medulloblastoma-specific enhancers. The resulting list of 37 intergenic candidate regions is listed in Table 5.

Genomic region	Mutational recurrence
chr5:1295201-1295251	**11**
chr12:66463226-66463276	9
chr12:66463226-66463276	9
chr12:66463201-66463251	6
chr16:34210301-34210351	4
chr13:25606376-25663301	4
chr16:34210276-34210326	3
chr11:32554601-32554651	2
chr11:39773551-39773601	2
chr12:94499151-94499201	2
chr13:25606376-25606426	2
chr13:25663251-25663301	2
chr13:28526051-28526101	2
chr16:34664551-34664601	2
chr16:34757351-34757401	2
chr17:21987526-21987576	2
chr17:41558651-41558701	2
chr19:15015551-15015601	2

chr19:40281601-40281651	2
chr2:14656651-14656701	2
chr2:14683551-14683601	2
chr2:130330251-130330301	2
chr2:130339076-130339126	2
chr22:22745476-22745526	2
chr22:22799851-22799901	2
chr3:82966026-82966076	2
chr3:82966051-82966101	2
chr3:87353076-87353126	2
chr3:103405101-103405151	2
chr4:122457426-122457476	2
chr4:122487626-122487676	2
chr5:50707701-50707751	2
chr5:101646001-101646051	2
chr6:17013126-17013176	2
chr6:107130801-107130851	2
chr7:36131951-36132001	2
chr7:49810751-49810801	2

Table 5. Candidate 50 bp intergenic regions ranked by mutational recurrence. The genomic region containing *TERT* promoter mutations is shown in bold.

The majority of the candidate regions we identified had a very low recurrence, which made them difficult to validate even using computational approaches. The top most recurrently mutated region was above-mentioned *TERT* promoter. Other regions with high mutational recurrence did not show any significant correlation with the expression changes of the closest genes. Since the regions overlap with medulloblastoma-specific enhancers, we still expect the mutations to be functionally relevant. This, however, might mean that the mutations act as *trans*-regulators rather than *cis*- and this would require additional level of data (e.g. Hi-C) to uncover such mechanisms.

Statistical aspects of recurrently mutated regions identification

The results of the naïve approach described in this section raised a list of questions. First of all, how to prioritize the candidate regions identified using a windows-based approach? As using simple mutational recurrence did not yield many novel candidates, the most intuitive approach would be to perform statistical tests and calculate p-values. The p-values would then be corrected for multiple testing and by choosing a significance threshold one would therefore obtain a list of statistically significant recurrently mutated genomic regions. However, if we consider such small window sizes as 50 bp, as were used here for this naïve approach, without performing any prior filtering of the genome, we will have a very large number of genomic regions ($\sim 10^7$) to start with and consequently, a large number statistical tests performed. This will lead to difficulties in obtaining many (if any) regions that would satisfy the significance threshold after multiple testing correction.

In addition, the data we are dealing with in this analysis is sparse. If we construct a matrix of observed number of somatic mutations, in which rows correspond to genomic windows and columns correspond to individual patients, most of the values of this matrix will be zeros given the estimated rates of somatic mutations in human of 2 to 10 mutations per diploid genome per cell division (Martincorena and Campbell, 2015). One of the obvious consequences will be that errors in somatic mutation calling, especially the false positives, will greatly influence the outcome of the statistical analysis.

The last important point that should be taken into account in the recurrence analysis is the above-mentioned regional mutational heterogeneity. One should consider it when comparing the observed recurrence of a region with the estimated background mutation rates.

Taking these described issues into consideration, we attempt to optimize for a number of statistical parameters, including p-value cut-offs and statistics choice, in the upcoming section of this Chapter. We will also discuss the choice of appropriate window size that will allow detecting correlations and at the same time will not dilute the biological signal.

The results described in the following section were included in the submitted manuscripts Rudneva V. *et al.*, 2016.

Parameter optimization using cross-validations.

A typical workflow to identify recurrently mutated genomic regions using window-based approach operates as follows. Somatic mutational data from multiple whole-genome sequenced cancer samples are obtained through alignment and SNV calling pipelines. Then, the lists of somatic variants are merged and mutational counts for each of the regions of interest, usually between 50 and 200 bp in size, are calculated. For each of the genomic regions, p-values are calculated based on the null hypothesis that the number of observed mutations in the region is not different from the background mutation rates. To estimate background mutation rates either regions with similar DNA replication timing (Weinhold *et al.*, 2014; Lawrence *et al.*, 2013) or a combination of replication timing, base-pair type and transcript region (Melton *et al.*, 2015) are employed. P-values are then computed using Binomial test statistics (Weinhold *et al.*, 2014) or a Poisson binomial model (Melton *et al.*, 2015), and a p-value cut-off is employed to identify regions that are recurrently mutated. Despite the use of fairly similar approaches and largely or partially overlapping datasets, there was no overlap in the intergenic candidate driver identified in previous studies besides the *TERT* promoter region. To address the question to which extent the statistical parameters choice may drive the findings of such studies, we devised the following approach based on cross-validation scheme (Figure 12).

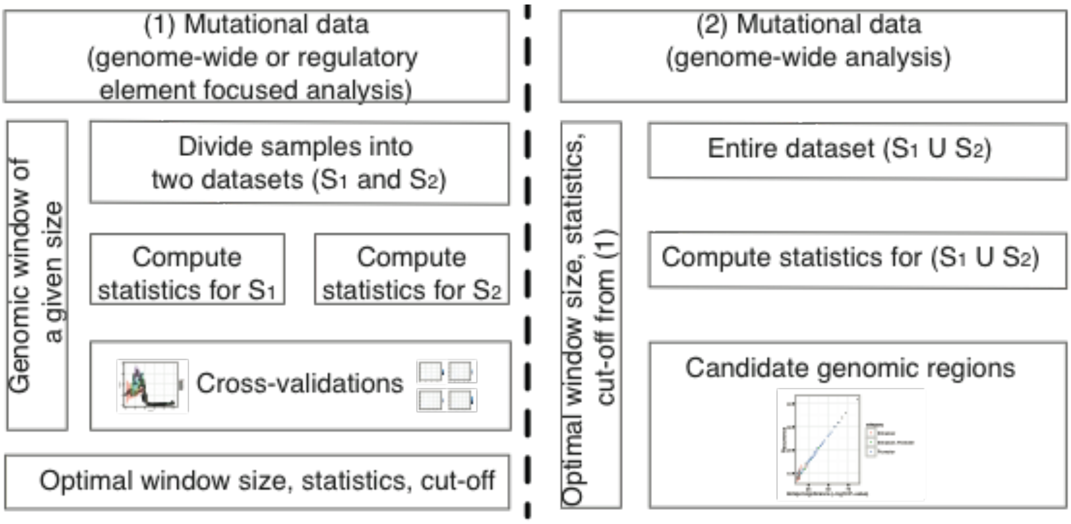

(1) Mutational data
(genome-wide or regulatory element focused analysis)
Genomic window of a given size
Divide samples into two datasets (S1 and S2)
Compute statistics for S1
Compute statistics for S2
Cross-validations
Optimal window size, statistics, cut-off
(2) Mutational data
(genome-wide analysis)
Optimal window size, statistics, cut-off from (1)
Entire dataset (S1 U S2)
Compute statistics for (S1 U S2)
Candidate genomic regions

Figure 12. Overview of the cross-validation based workflow for identification of recurrently mutated genomic regions.

We first binned the entire human genome into non-overlapping windows. We tested different window sizes in our analysis, four in total (50bp, 100bp, 200bp, and 500bp). For every genomic window of a given size we computed its *mutational recurrence* (the number of samples that have at least one somatic mutation in the given window) and compared it to the average mutational recurrence of the genome-wide set of regions with comparable genetic and epigenetic background (*background mutation rate*). To obtain a background set of genomic regions for every region of the genome we used the methodology described in previous sections (see 3.2.3). To study how the choice of window size influences the result and which statistics are most suitable for the data we performed cross-validations. In detail, we randomly divided our dataset into two sets of samples, S_1 and S_2. We constructed these sets in such a manner that each contained a similar number of samples of each cancer type. We performed the analysis independently on sets S_1 and S_2 computing a separate set of test statistics for each (enrichment score, mutational recurrence, Binomial test p-value, Gamma-Poisson test p-value). We calculated an enrichment score as a ratio of observed mutations to the background mutation rate. To compute Binomial test p-values and Gamma-Poisson test p-values we used background mutation rates estimated from the 1st principal component loading vector. We repeated the same analysis for each of the four window sizes: 50bp, 100bp, 200bp, and 500bp. We compared the results obtained in individual runs of the analysis and identified an optimal combination of window size and test statistics that allowed for robust and reproducible results. The optimal combination was defined as the one that gave the highest precision and recall values on a precision-recall curve. To construct precision-recall curves we scanned through cut-off values and calculated the number of cross-validated hits in both sets (intersection or *recall*) as well as the fraction that the recall represent of the total hits surpassing the threshold in at least one set (intersection/union or *precision*). Finally, we employed our approach with the chosen window size, cut-off and test statistics on the entire dataset (S_1 U S_2). As a result, we obtained a list of candidate genomic regions that we subsequently studied in more detail.

Similarly to the previous section, we performed our analysis on a dataset consisting of 698 cancer genomes from eight different cancer types for which each cancer entity was represented by at least 20 samples (Table 1).

Using cross-validations, we compared performance of various statistics and observed that the enrichment score as such does not provide reproducible results for any of the window sizes ranging between 50bp and 1Mbp. This statement can be illustrated by scatter plots, where dots corresponding to individual genomic windows form a sparse cloud indicating there is little agreement between the results obtained on the dataset S_1 and S_2 in individual runs of cross-validation (Figure 13a). Mutational recurrence that was used as statistics in the naïve approach described in the previous section also showed low robustness, especially when used with smaller window sizes such as 50 bp (Figure 13b). The most interesting observations came from genomic regions with mutational recurrence values between two and six out of 698 samples – the exact range of mutational recurrence we previously identified most of our candidates using the naïve approach. For these mutational recurrence values obtained on the dataset S_2, we observed that the same genomic regions could reach all possible values between as low as zero or as high as 14 when the analysis was performed on the dataset S_1. This indicates that the mutational recurrence as such, especially in a combination with 50 bp window sizes is not a robust statistic for such analysis. Moreover, these results illustrate how relying on such statistic could lead to identification of false candidate regions.

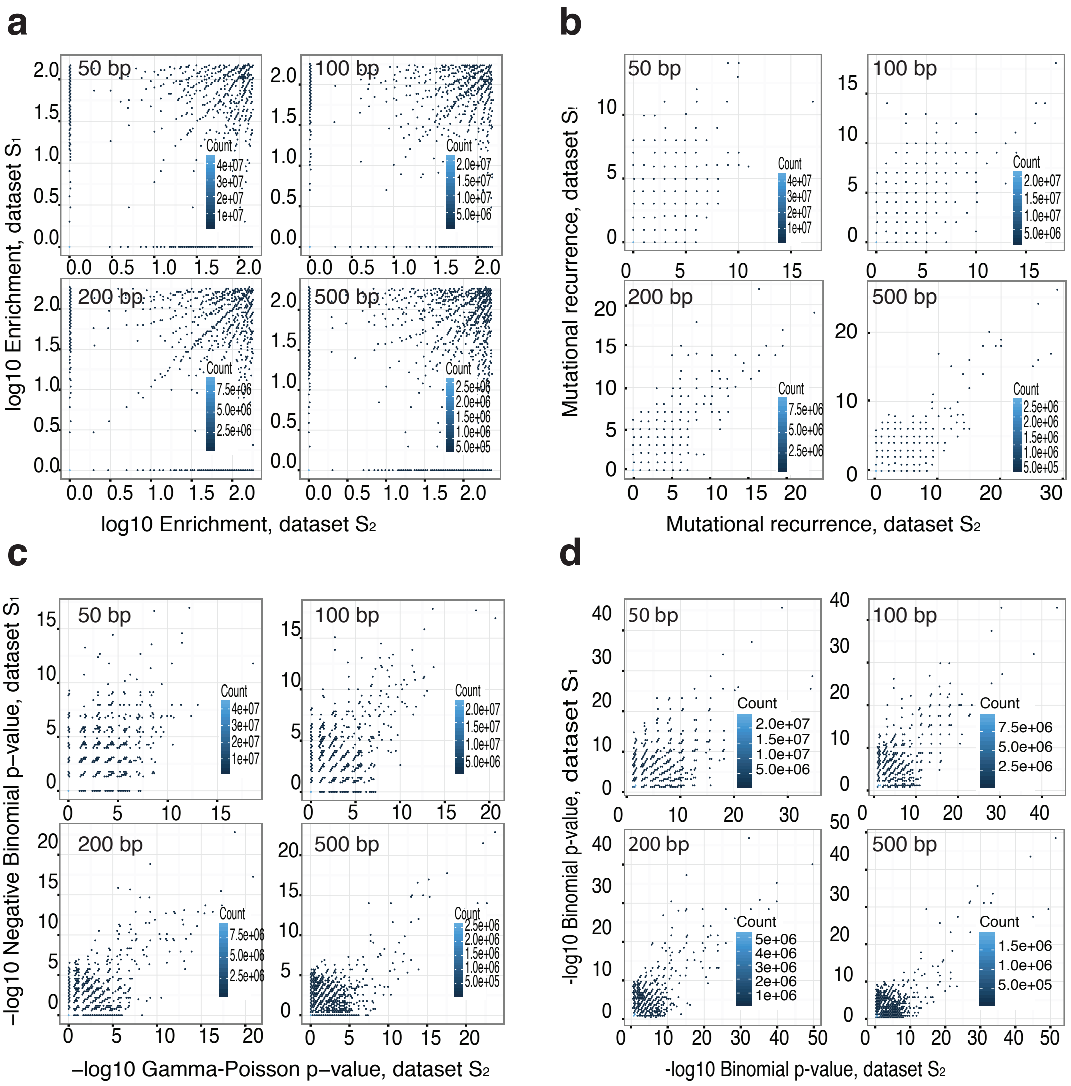


Figure 13. Cross-validation results for various statistics at different window sizes. (**a**) Enrichment score (**b**) Mutational recurrence (**c**) Gamma-Poisson test p-value (**d**) Binomial test p-value.

Using cross-validation scheme we also assessed how robust were two statistical tests more commonly used in cancer genomics studies: Binomial and Gamma-Poisson tests. For every genomic window we compute p-values given the null hypothesis that the expected mutational recurrence of the window is not different from the background mutation rate (BMR). The BMR for a given region was estimated from its background set of genome-wide regions defined using the 1st PC loading vector. It is important to mention, that for the purpose of this study we interpreted both test p-values as test statistics – *i.e.* not as literal p values, but as scores decreasing monotonously with significance. Both, the Binomial test and the Gamma-Poisson test, showed higher robustness in comparison to mutational recurrence and enrichment score (Figure 13). As a general note, the p-values range for Binomial test and Gamma-Poisson test is, as expected, quite different. In this study, the lowest p-values for the Binomial test reached 10^{-50} while for the Gamma-Poisson test p-values they reached 10^{-20}. This means that if one were to use the same standard p-value cut-off for both tests and on the same dataset, choosing the Gamma-Poisson test would generally result in lower number of candidates identified as significant. Since we rather preferred having a larger list of candidates that we could manually study in detail, we choose the Binomial test for our analysis.

By comparing the performance of the Binomial test at different resolution, we observed that it had higher reproducibility at larger window sizes, starting from 100 bp (Figure 13d). We used the precision-recall curves to identify the optimal combination of the Binomial test cut-off and window size for our data.

Typical precision-recall curves for each window sizes are shown in Figure 15. Every dot corresponds to one p-value cut-off; for the given p-value cut-off, precision was calculated as a number of regions (dots on cross-validation plots in Figure 13) that scored higher than the cut-off in *both* attempts of cross-validation (e.g. according to the results obtained on S_1 and S_2 datasets independently). The recall was calculated as a fraction of the recall over the total number of dots above the cut-off in *each* of the datasets. We performed the random splitting of the entire dataset into two halves (S_1 and S_2) ten times in order to compute confidence intervals for precision. The results of this for all window sizes are shown in a compact way in a precision-recall plot in Figure 14a.

We defined the optimal combination of the parameters as the one that allowed reaching at the same time the highest precision and recall values on the precision-recall curve. We observed that the highest recall and precision values were reached when using 200bp window size (Figure 14a). We assumed here that somatic driver mutations when they occur within the intergenic regions should overlap with certain regulatory elements. Given the typical size of *cis*-

DNA regulatory elements of 6-12bp for transcription factor-binding sites and ~500bp for enhancers (Loots, 2008), we considered 200 bp window size as plausible, because it is large enough to encompass whole regulatory elements, and small enough to not contain much more than the regulatory element (*i.e.* as few "uninteresting" bases as possible, as this may dilute the signal).

The highest recall value for 200 bp window size was reached using 10^{-12} p-value cut-off (Figure 15) and lied within the 10^{2} range; at the same time the highest precision value was 0.58 (95% CI: 0.54-0.64). It is important to note here, that given our definition, the precision value we calculated here is a very conservative, downwards-biased estimation of the true precision of the method and is probably underestimated. Indeed, we considered the number of hits found in both attempts of cross-validation as "true positives" (i.e., as numerator) and the hits found in at least one as all discoveries (denominator). We therefore calculated precision as if all the hits found only in one set and not in the other were false positives. Based on this, we concluded that the precision of 0.58 was a sensible result given the definition.

In a conclusion, we choose the following parameters for the further analysis on our data: 200 bp window size and Binomial p-value threshold of 10^{-12}.

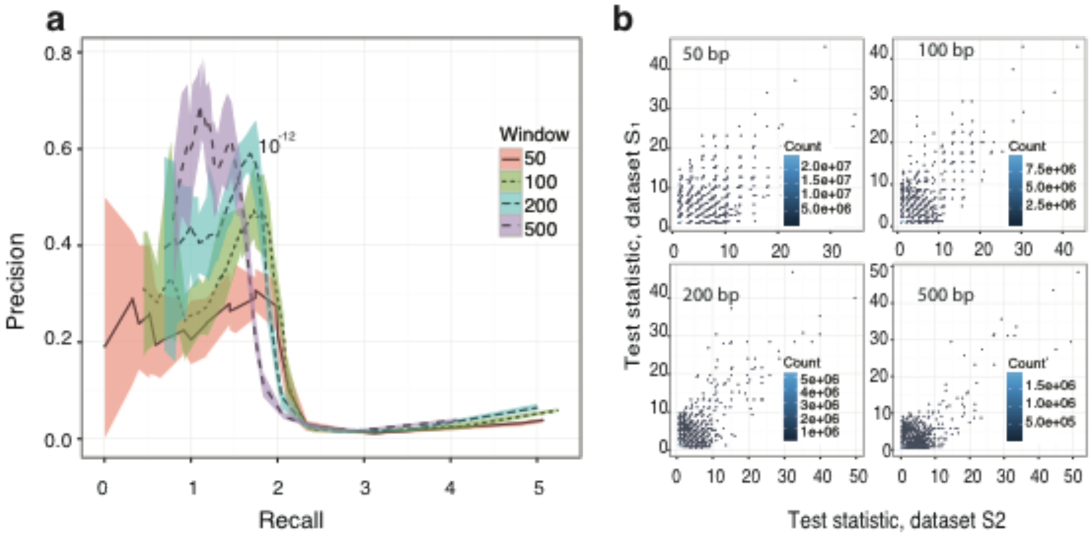


Figure 14. Parameter optimization for the windows-based approach. (**a**) Precision-recall curves for different Binomial p-value cut-offs. Minimal and maximal precision values over 10 random splits of the data set into halves are shown in color, mean precision is shown as dotted line. Colors indicate different

window sizes. (**b**) Cross-validation results for recurrently mutated regions for different window sizes. P-values were calculated using a Binomial test statistic, and scales are logarithmic (-log10). Test statistics were computed independently for the S_1 set (x axis) and S_2 set (y axis).

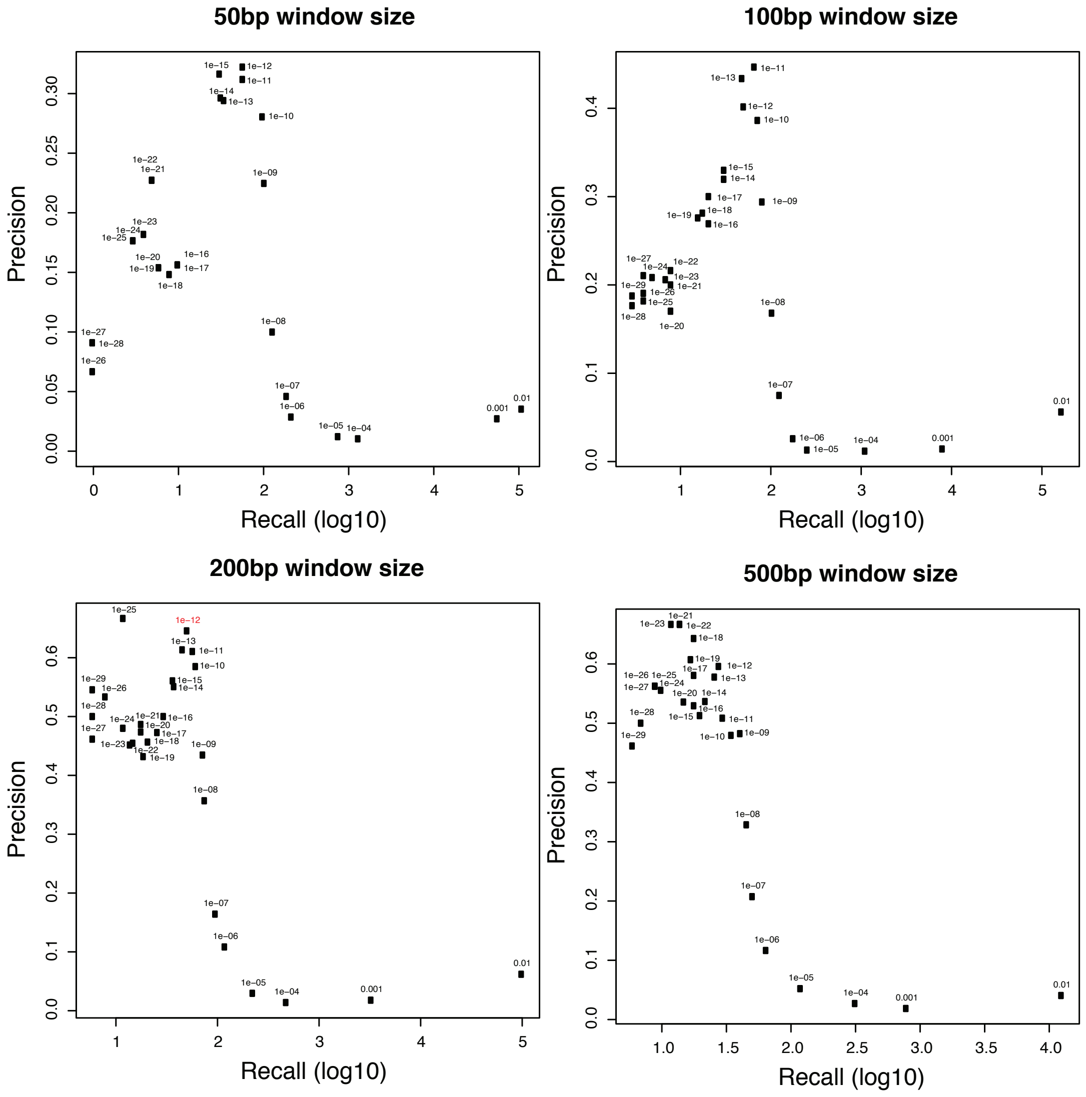


Figure 15. Precision-recall curves for different window sizes. Points correspond to different Binomial p-value cut-offs; the x axis is logarithmic (log10).

Recurrently mutated genomic regions within the regulatory elements of the genome

Recent studies aimed on identification of recurrently mutated somatic mutations focused on subsets of genomic regions, rather than performing their searches in an unbiased genome-wide manner. This strategy offers obvious advantages in terms of statistical power and computational performance because of smaller numbers of statistical tests that are required. Limiting the search space, however, may also result in promising candidates being overlooked. To compare our approach to the previous studies performed on restricted search space (Fredriksson *et al.*, 2014; Melton *et al.*, 2015; Weinhold *et al.*, 2014) we applied our method to a comparable subset of genomic windows. To do so we restricted the complete list of genomic windows that we used in the previous section by overlapping it with:

1) a list of promoters and enhancers that we obtained following the procedure described in Weinhold *et al.*, 2014;

2) a list of regulatory genomic regions according to the RegulomeDB database annotations as described in Melton *et al.*, 2015.

Since in most of the cases, there was no complete overlap between the above-listed regulatory regions and the genome-wide list of windows, we considered two different scenarios for our analysis: 1) genomic windows exhibiting more than one base-pair overlap with the regulatory elements; 2) regulatory elements overlapping with the windows for at least half of the selected window size.

Note that when we applied the second scenario to the list of RegulomeDB regulatory regions, the number of resulting windows was not large enough to perform cross-validations and construct precision-recall curves for window sizes larger than 100 bp. Therefore, we will not be discussing this case here.

Subsequently we performed the same cross-validations analysis as described in the previous section (Figure 12) but this time on the restricted list of windows. To compute test statistics on the restricted list of genomic windows, we used the same restricted search space for estimation of background mutation rates.

The results that we obtained for the restricted analysis setup looked generally similar to the outcomes of the aforementioned whole-genome analysis (Figure 16). Using small window sizes, such as 50 bp, resulted in relatively low precision regardless of the type of regulatory regions used for restricting the search space. This indicated low robustness of the approach with this window size. As observed previously, using larger window sizes improved the precision.

Similarly to the genome-wide analysis setup, the use of 200 bp windows allowed reaching the highest combination of precision and recall values in all cases. Using even larger window sizes led to a decrease in recall while not significantly improving the precision. This made us conclude that 200 bp was an optimal choice of window size for the restricted analysis as well as for the genome-wide analysis setup.

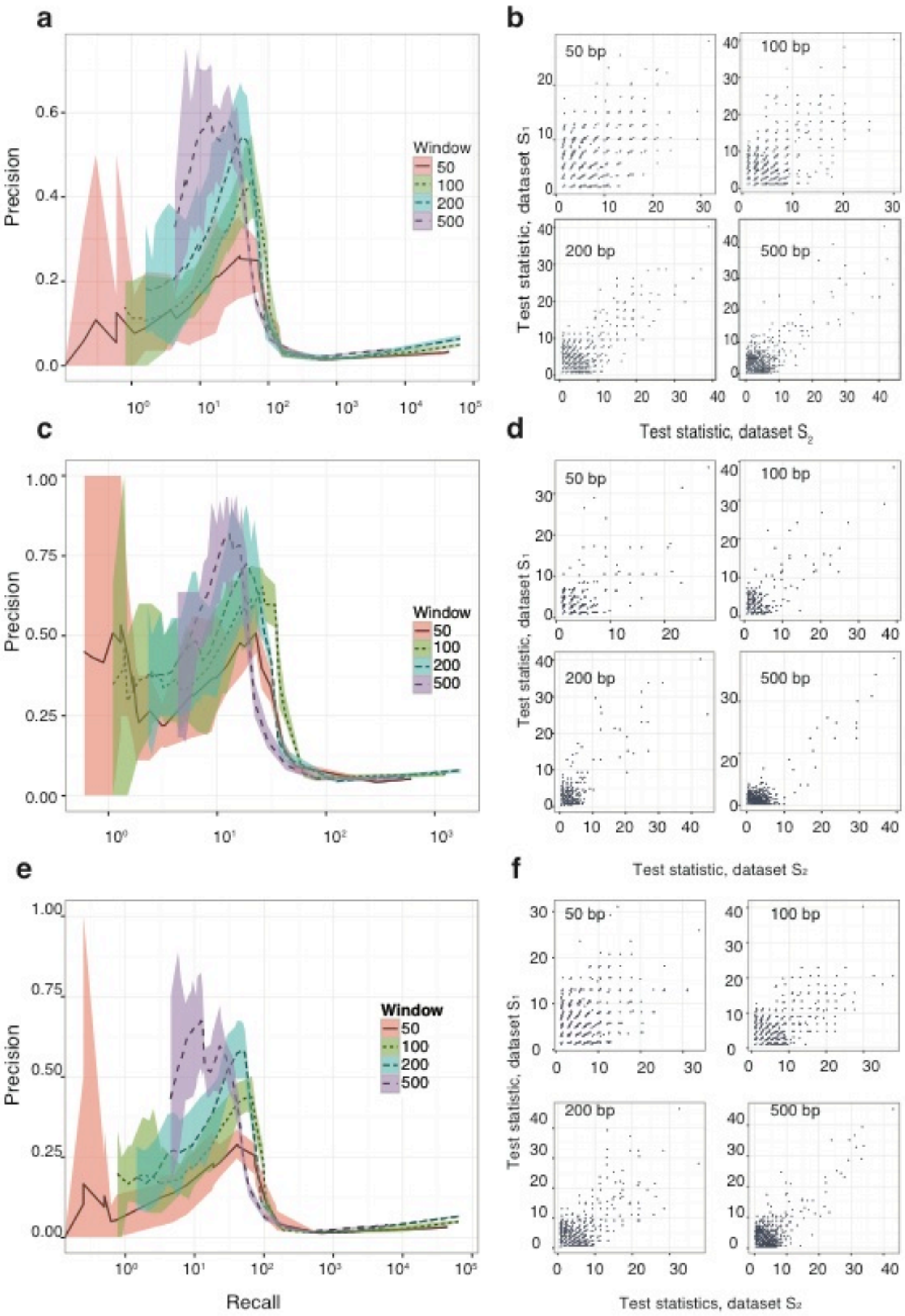


Figure 16. Parameter optimization for the analysis restricted sets of genomic windows. Left panel: Precision-recall curves for different Binomial p-value cut-offs. Minimal and maximal precision values over 10 random splits of the data set

into halves are shown in color, mean precision is shown as dotted line. Colors indicate different window sizes. Right panel: Cross-validations results for recurrently mutated regions for different window sizes. P-values were calculated using a Binomial test statistic, and scales are logarithmic (-log10). Test statistics were computed independently for the S_1 set (x axis) and S_2 set (y axis). (**a**,**b**) Restricted analysis performed on the list of regions overlapping for at least 1bp with enhancers, promoters or UTRs. (**c**,**d**) Restricted analysis performed on the list of regions overlapping for at least 1 bp with regulatory regions according to RegulomeDB annotations. (**e**,**f**) Precision-recall curves and cross-validations results for genomic regions restricted to at least 50% overlap with promoters, enhancers or UTRs.

Overall, we observed that restricting the search space by considering only particular classes of regulatory regions did not bring a marked advantage in terms of robustness when compared to an unbiased genome-wide analysis. Since restricting the search space also potentially limits the findings, we choose to apply the unbiased genome-wide setup to the dataset of 698 cancer samples in order to search for recurrently mutated regions. We ran the final analysis with the parameters that we identified in the previous section: 200bp window size, Binomial p-value test statistic cut-off 10^{-12}.

Prioritizing candidate regions

Using our unbiased genome-wide approach we identified 153 recurrently mutated candidate regions of 200 bp size. Since our approach operated on the genomic windows without any prior filtering, among the candidates we identified were both coding and non-coding regions. We first aimed at filtering out those regions that were unlikely to harbor driver events. After annotating the list of candidates we excluded:
1) 35 regions overlapping with the human immunoglobulin heavy-chain locus (IGH) on chromosome 14. This region is known to be prone to somatic hypermutation in B cells, a process that takes place even in healthy individuals. As a confirmation, we observed that the majority of the samples harboring mutations in these regions corresponded to CLL and lymphoma cases.

2) 27 regions harboring mutations close to the *BCL2* and the *IGLL5* loci in the B-cell lymphoma samples. Similarly to the previous example, these mutations likely arose from somatic hypermutation and are therefore unlikely to represent driver events.

After performing the filtering step, we ended up with 91 candidate regions. The regions varied in mutational recurrence, ranging from seven to twenty-seven out of 698 samples (Supplementary Table 1). It is important to note, that the mutational recurrence of driver events is a potential limitation of the approach, especially in case of cancer-type specific drivers. Indeed, if the tumor type of interest is represented by only a few samples and the driver event is infrequent, while performing cross-validation all of the mutated samples might by chance appear in only one half of the dataset (or in both, but with a very low recurrence) and the driver event will hence not be detected.

As was mentioned previously, cancer driver events, by definition, convey an advantage to the tumor and are therefore expected to overlap with functionally relevant regions of the genome, such as exons, promoters or enhancers. To prioritize our high confidence candidate regions we annotated them using the abovementioned categories (Supplementary Table 1). We observed that the majority of candidate regions (71 out of 91) overlapped with exons of protein-coding genes. A detailed investigation of these genes showed that all were well-known cancer genes such as *TP53*, *BCL2*, *BCL6*, *CTNNB1*, *KRAS*, *MYC* and *SMARCA4* (Supplementary Table 1). This further supported the accuracy of our method to detect genomic regions with somatic cancer drivers.

Since the main goal of our analysis was to identify novel driver regions within the intergenic regions, we further focused our attention on the remaining 20 candidates that were not intersecting exons of protein-coding genes (Table 6). Among these, three overlapped with promoters and eight with enhancers, while one of the regions overlapped with both, a promoter and an enhancer; ten more regions did not overlap with any of the known annotations and were therefore omitted from the remainder of the study.

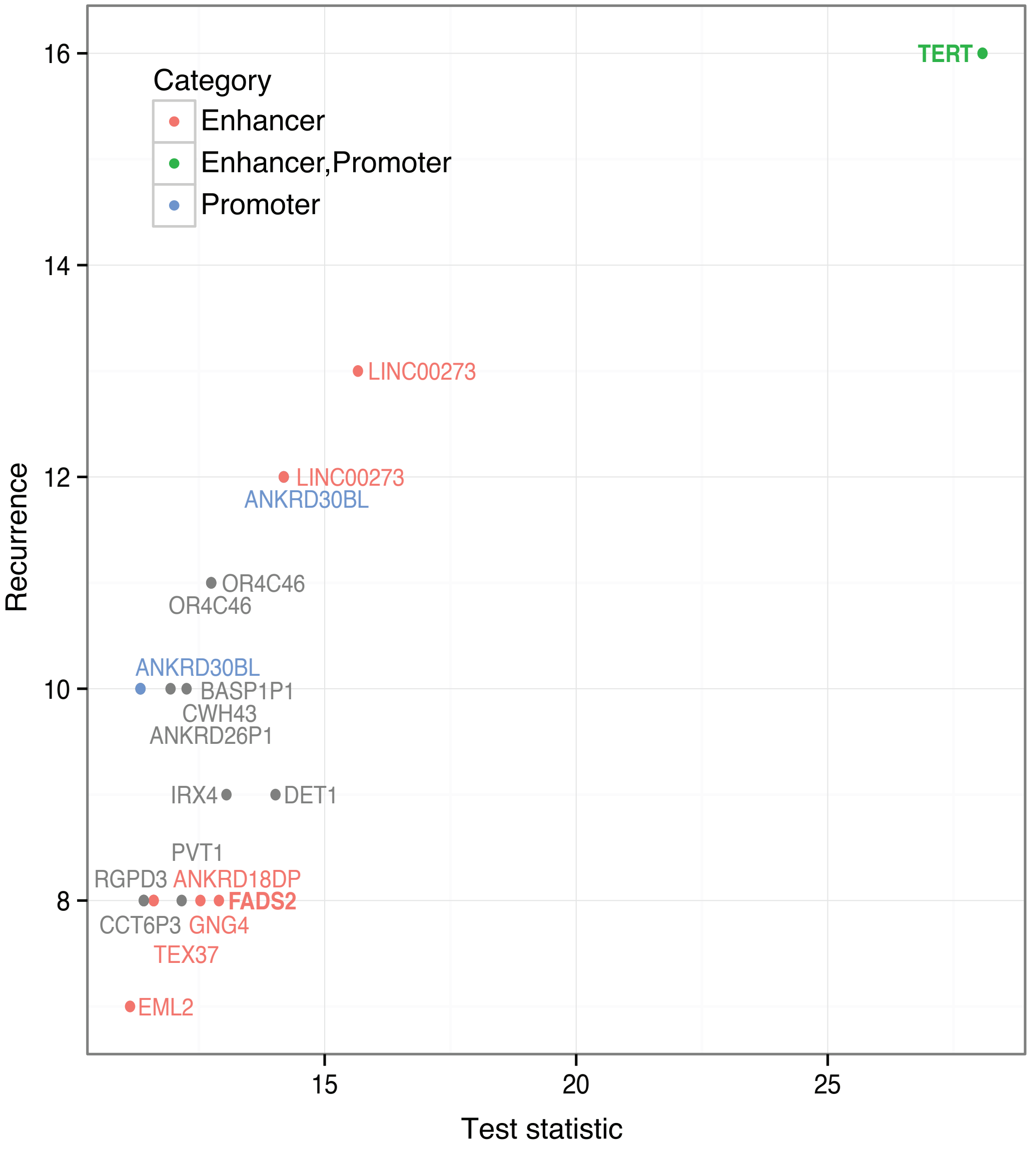


Figure 17. High confidence recurrently mutated intergenic regions identified by the approach. Candidate regions are colored by category of regulatory elements (enhancer, promoter) they overlap with. Candidates that do not overlap any functionally annotated genomic regions are shown in grey. The labels indicate the gene closest to the region.

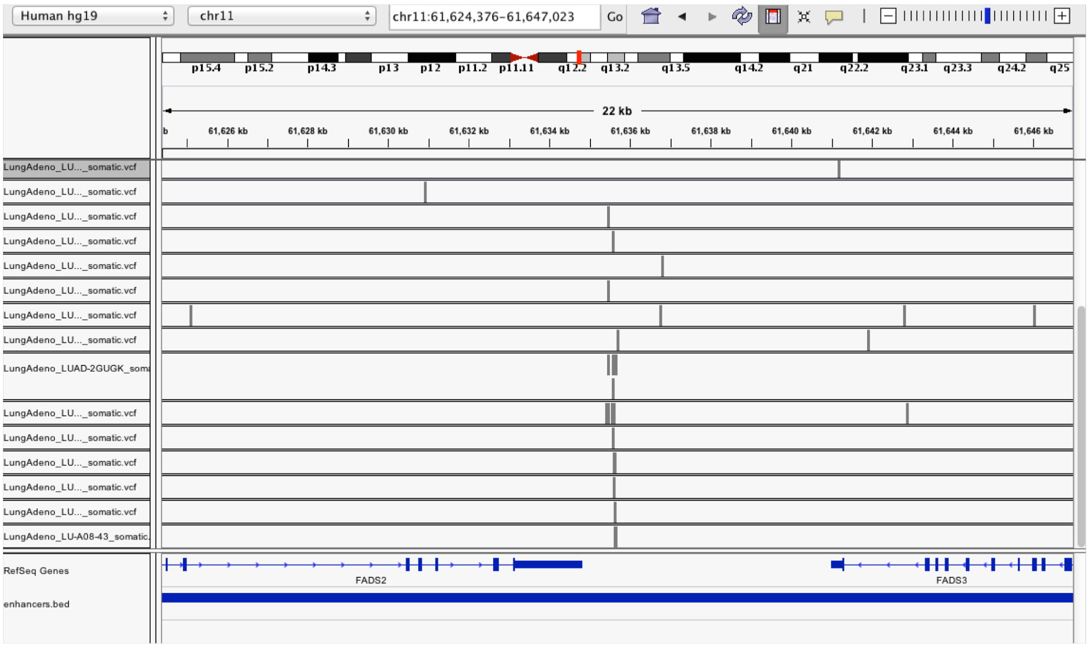
Genomic region	**Mutational Recurrence**	**Closest Gene (Distance in bp)**	**Test Statistic**	**Cancer Types (Number of samples with mutations)**
chr5:1295201-1295400	16	*TERT* (18)	10^{-29}	Liver(5), Medulloblastoma(11)
chr16:33953201-33953400	13	*LINC00273* (7653)	10^{-16}	Prostate(10), Liver(2), Medulloblastoma(1)
chr16:33953001-33953200	12	*LINC00273* (7853)	10^{-15}	Prostate(10), LungAdeno(1), Medulloblastoma(1)

chr2:133018801-133019000	12	*ANKRD30B* (3260)	10^{-15}	Prostate(12)
chr15:89096601-89096800	9	*DET1* (6696)	10^{-15}	Breast(4), LungAdeno(5)
chr5:2145601-2145800	9	*IRX4* (258252)	10^{-14}	LungAdeno(9)
chr11:61635601-61635800	8	*FADS2* (776)	10^{-13}	LungAdeno(8)
chr11:51570001-51570200	11	*OR4C46* (53791)	10^{-13}	Prostate(11)
chr11:51580401-51580600	11	*OR4C46* (64191)	10^{-13}	Prostate(10), Medulloblastoma(1)
chr1:235692001-235692200	8	*GNG4* (18788)	10^{-13}	Liver(1), LungAdeno(7)
chr4:49316401-49316600	10	*CWH43* (252304)	10^{-13}	Prostate(8), LungAdeno(1), Medulloblastoma(1)
chr8:129131001-129131200	8	*PVT1* (17503)	10^{-13}	Breast(1), LungAdeno(7)
chr13:23151201-23151400	10	*BASP1P1* (320081)	10^{-12}	LungAdeno(10)
chr2:88791201-88791400	8	*TEX37* (32770)	10^{-12}	Lymphoma(7), Medulloblastoma(1)
chr3:197825801-197826000	8	*ANKRD18D* (18211)	10^{-12}	Breast(3), LungAdeno(5)
chr7:64574001-64574200	8	*CCT6P3* (38911)	10^{-12}	Prostate(1), Breast(2), LungAdeno(5)
chr2:107016201-107016400	8	*RGPD3* (5047)	10^{-12}	Breast(1), Liver(1), LungAdeno(6)
chr16:46412201-46412400	10	*ANKRD26P1* (90854)	10^{-12}	Prostate(7), LungAdeno(3)
chr2:133019401-133019600	10	*ANKRD30BL* (3860)	10^{-12}	Prostate(9), LungAdeno(1)
chr19:46151601-46151800	7	*EML2* (2715)	10^{-12}	LungAdeno(7)

Table 6. High confidence non-coding candidates regions ranked by the statistics values (full list of recurrently mutated regions in Supplementary Table 1).

The highest scoring intergenic candidate region, with a mutational recurrence of 16, was the region harboring the well-characterized *TERT* promoter mutations (Figure 17). This region was found to be mutated in 5 liver cancer and 11 medulloblastoma samples.

We aimed to identify other interesting examples among the recurrently mutated regions. We observed that most of the candidates identified were located from 7 Mb to 200 Mb away from any genes, with the closest ones encoding lncRNAs or pseudogenes (Table 6). The lack of knowledge in the existing literature on functional roles of these genes in cancer made it difficult for us to propose any mechanism upon which the mutations could act in tumorigenesis. In addition, the regions were located relatively far from their gene "targets", that were identified based on their location as the closest genes. We assumed that these candidate regions might be involved in tumorigenesis through more complex mechanisms than in the case of *TERT* mutations. For example, these could be mechanisms involving distal *cis*-effects, *trans*-effect or even secondary targets. In order to study these mechanisms data on higher-order organization of the genome, such as Hi-C data, may be necessary. Unfortunately, such data was not available for the samples that were used in this study.



Human hg19

chr11

chr11:61,624,376-61,647,023

Go

p15.4
p15.2
p14.3
p13
p12
p11.2
p11.11
q12.2
q13.2
q13.5
q14.2
q21
q22.2
q23.1
q23.3
q24.2
q25
22 kb
b
61,626 kb
61,628 kb
61,630 kb
61,632 kb
61,634 kb
61,636 kb
61,638 kb
61,640 kb
61,642 kb
61,644 kb
61,646 kb
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LUAD-2GUGK_som
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU..._somatic.vcf
LungAdeno_LU-A08-43_somatic.
RefSeq Genes
FADS2
FADS3
enhancers.bed

Figure 18. *FADS2* candidate region from the Integrative Genomics Viewer.

By ranking the candidate intergenic regions by the distance to the closest gene, we identified one region with the closest distance from its "target", after the *TERT* region. This was a region on chromosome 11 located approximately 700bp upstream of the *FADS2* gene (Figure 17, highlighted). The region intersected with an enhancer and had a mutational recurrence of eight out of 698. Notably, all of the mutated samples corresponded to the same cancer type – lung adenocarcinoma. The mutations in the corresponding samples all occurred within a very narrow region of only 98 bp size (Figure 18). *FADS2* encodes for Δ6-desaturase, which is a critical enzyme in the biosynthesis of long-chain polyunsaturated fatty acids. It has been shown that in several cancer entities desaturation of fatty acids catalyzed by the enzyme does not occur (Park *et al.*, 2011). For example, in MCF7 breast cancer cells molecular defects in *FADS2* were found to cause loss of the enzyme activity (Park *et al.*, 2011). Compensation for *FADS2* loss of function by *FADS1* is known to lead to the production of two independent fatty acid products that likely act as competitive inhibitors of the eicosanoid cascade. This leads to depletion and alteration of the normal eicosanoid and docosanoid cell signaling milieu, with presumed consequences for cellular communication that to date are poorly understood. It is tempting to speculate that the mutated region close to *FADS2* may be involved in deactivation of the *FADS2* gene and therefore in the loss of the Δ6-desaturase activity in the mutated cancer cells. Since, unfortunately, no gene expression data were available for these lung cancer dataset, we could not test this hypothesis using the correlation analysis.

Discussion

We designed a workflow for cancer genomic studies that uses a window-based approach to screen for recurrently mutated intergenic regions with an additional cross-validation step to assess robustness and allow for identification of optimal parameters. Our approach uncovered intergenic regions with a mutation recurrence as low as 1% (7/698 samples).

Using our approach, we detected known cancer-relevant targets, both exonic (*TP53*, *MYC*, *SMARCA4* etc) and intergenic, including the previously identified recurrent mutations in the *TERT* promoter, in an unbiased genome-wide manner. The recurrent *TERT* promoter mutations is the only recurrent non-coding event identified in common between previous studies that aimed to uncover regulatory mutations in cancer (Fredriksson *et al.*, 2014; Melton *et al.*, 2015; Weinhold *et al.*, 2014). We believe that this is because this event is exceptional in terms of its frequency of recurrence, close distance to the target gene and mutation clustering.

We identified a novel intergenic region upstream of the *FADS2* gene as a source of candidate driver mutations in lung cancer.

In conclusion, cross-validation enables parameter selection and robustness assessments in the context of challenging searches for intergenic somatic point mutation recurrence in cancer genomes.

4. Summary, conclusions and future directions

In summary, in this work we studied patterns of accumulation of somatic mutations along the genome in cancer in order to establish an appropriate background model to correct for the regional mutational heterogeneity (chapter 2). We identified a genetic feature that showed the highest correlation with somatic mutation rates alone at different resolution ranging between 50 bp and 10 Mb. Additionally we identified an even better correlate with the somatic mutation rate – the 1st PC loading vector based on five individual features – and proposed to use it as a covariate to control for the background mutation rates in cancer genomics studies aimed to identify frequently mutated regions.

As a scope for the future improvement, the use of tissue-specific genetic and epigenetic features for modeling of the background mutation rates should be preferred. Recently, epigenetic features derived from the most likely cancer cell type of origin of the corresponding tumor were shown to be the best predictors of local somatic mutational density, even when compared to features derived from the matched cancer cell lines (Paz Polak, Rosa Karlic, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence and Eric Rynes, Kristian Vlahovicˇek, 2015). A study performed on 173 cancer genomes from eight different cancer types and a list of 424 epigenetic features derived from 45 different tissue types (Kundaje *et al.*, 2015), demonstrated that chromatin accessibility and modifications together with replication timing, explained up to 86% of the variance in mutation rates along the cancer genome. The highest correlation values between the observed mutation densities and the densities predicted using Random Forest regression were reached at 1-Mb resolution. Even though it is difficult to compare these results with our observations (Pearson correlation coefficient of max 0.8 between the observed mutational densities and the 1st PC loading vector at 1-Mb resolution, Figure 2b), we conclude that using features obtained from the corresponding cell-of-origin tissues might indeed improve the mutational model we established here even further. Hence, allowing correcting for the background mutation rates even more effectively in the following analysis steps and uncover additional candidates.

Similarly, implementing data on sample-specific mutational processes could be used to acquire more accurate estimates of the background mutation rates. Indeed, passenger mutations accumulate as the outcome of mutational processes that occurred thought the development of tumor (Helleday *et al.*, 2014). Therefore, accounting for them within a computational approach could help to distinguish passenger from driver mutations.

Motivated by several recent studies that identified TERT promoter driver mutations as a driver event in several cancer types, we tested different computational approaches in order to identify other potential intergenic drivers with similar properties. Among the approaches we developed here were gene-expression correlation analysis (chapter 3.2.1), TFBS changes prediction analysis (chapter 3.2.2) and two approaches focusing on the mutational recurrence of the candidate regions (chapter 3.2.3); we applied our methods to one of the largest appropriate cancer dataset available at the time this work was performed.

By systematically addressing the problem of computational detection of driver events within the intergenic regions in cancer, we concluded that the most successful approach was to focus on mutational recurrence. By combining the abovementioned background model to correct for the regional mutational heterogeneity (chapter 2.2) together with the cross-validation-based scheme aimed to identify the optimal statistical parameters (chapter 3.2.3). We established a pipeline for genome-wide, unbiased identification and prioritization of intergenic regions recurrently mutated across multiple cancer entities.

We believe, that our pipeline can be successfully used for various applications, including the studied in pan-cancer setup or with large datasets on the same cancer type; but not limited to the cancer genomics field. Potential wider applications of our pipeline outside of the field could include, for example, identification of actively acting enhancers relevant in disease from the ChIP-Seq experiments. In a cancer setup, the pipeline could be applied to a dataset consisting of multiple cancer entities or a dataset of different subtypes of the same cancer, given that the dataset is large enough to perform cross-validations.

Coming back to the original application of our pipeline, the potential future improvements could include integration of the pipeline's output (list of recurrently mutated intergenic regions) with matching gene expression data, as described in section 3.2.1 of this thesis (Figure 18).

An additional level of information can be obtained from the germline SNV calls if such data are available. First of all, these calls can be used to control for population structure prior to the correlation with gene expression analysis as described in section 3.2.1. Moreover, in some cases, germline variants may account for the genetic predisposition to cancer. For instance, about 20% of all known oncogenes are altered by the germline mutations (Martincorena and Campbell, 2015). Given the important role of the non-coding regions in gene regulations, it is tempting to speculate, that some of the intergenic germline mutations recurrent among cancer patient, while observed at low frequencies within the general population, may play important role in determining the genetic predisposition of the affected individuals to cancer. It is also important to mention, that being able to identify such predisposing germline drivers will be beneficial for cancer screening; and could also find wide application in the emerging field of personalized medicine.

Therefore, as a potential improvement of our approach, it would be possible to run our pipeline from section 3.2.3 on the list of germline SNVs to detect genomic regions that are recurrently mutated in the germline of cancer patients. Next, by contrasting the identified candidate mutations with the common SPNs identified in the 1000 Genomes Project, similarly to the idea used in Khurana, Fu, Colonna, Mu, Kang, T. Lappalainen, *et al.*, 2013, one could identify mutations that experience selective constrains and are likely to be predisposing to cancer. One the could add these mutations to the list of high confidence recurrently mutated regions identified in the previous step (Figure 18).

The candidate regions identified in the previous steps could be computationally followed-up in order to identify transcription factor binding sites changes using the approach that we designed in section 3.2.2. The combined output of the different strategies could provide the researcher with insight into the biological mechanisms of the newly identified intergenic candidates.

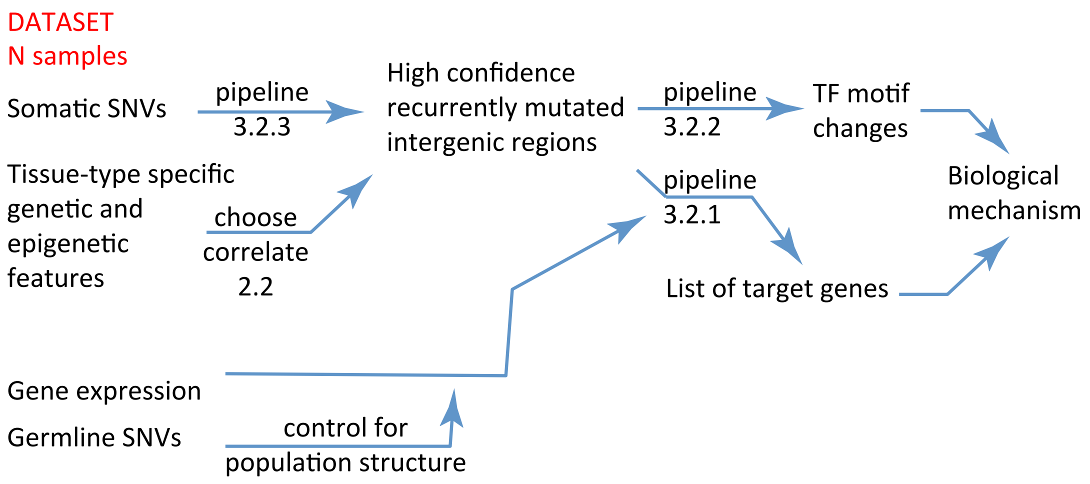


Figure 19. A combined approach for identification of intergenic somatic driver events in cancer based on the integration of the methods designed in previous chapters of this thesis.

Appendix

A Supplementary information

A.1 Supplementary tables for Chapter 3

Genomic region	Mutational recurrence	Closets Gene (Distance in bp)	Test statistic	Category
chr3_41266001_41266200	27	CTNNB1(0)	4,34E-52	5_UTR,Exon
chr8_128748801_128749000	26	MYC(0)	2,71E-50	Enhancer,5_UTR,Exon
chr17_7578201_7578400	23	TP53(0)	1,22E-43	Enhancer,Exon
chr17_7578401_7578600	22	TP53(0)	1,85E-41	Enhancer,5_UTR,Exon
chr3_187462801_187463000	22	BCL6(0)	2,62E-40	Enhancer,Exon
chr17_7577401_7577600	21	TP53(0)	2,69E-39	Enhancer,Exon
chr2_89160201_89160400	21	MIR4436A(48234)	5,73E-37	Promoter,Exon
chr8_128749201_128749400	19	MYC(0)	4,94E-35	Enhancer,Exon
chr8_128749001_128749200	19	MYC(0)	4,94E-35	Enhancer,Exon
chr2_89159401_89159600	19	MIR4436A(47434)	6,44E-33	Promoter,Exon
chr3_187462601_187462800	18	BCL6(0)	5,54E-32	Enhancer,Exon
chr12_25398201_25398400	18	KRAS(0)	3,51E-31	Enhancer,5_UTR,Exon
chr2_89160401_89160600	17	MIR4436A(48434)	5,92E-29	Promoter,Exon
chr5_1295201_1295400	16	TERT(18)	8,38E-29	Enhancer,Promoter
chr17_7577001_7577200	16	TP53(0)	8,38E-29	Enhancer,Exon
chr3_187463001_187463200	16	BCL6(0)	5,92E-28	Enhancer,5_UTR,Exon
chr2_89159601_89159800	16	MIR4436A(47634)	5,23E-27	Promoter,Exon
chr8_128749401_128749600	15	MYC(0)	8,92E-27	Enhancer,Exon
chr8_128749601_128749800	15	MYC(0)	8,92E-27	Enhancer,Exon
chr6_91005401_91005600	15	BACH2(0)	2,65E-25	Enhancer,Exon
chr2_89159801_89160000	15	MIR4436A(47834)	4,36E-25	Promoter,Exon
chr8_128750401_128750600	14	MYC(0)	8,93E-25	Enhancer,5_UTR,Exon
chr14_69259201_69259400	13	ZFP36L1(0)	8,37E-23	Enhancer,Exon
chr1_23885601_23885800	13	ID3(0)	8,37E-23	Enhancer,Exon

chr3_187462401_187462600	13	BCL6(0)	4,18E-22	Enhancer,Exon
chr8_128748601_128748800	12	MYC(0)	7,32E-21	Enhancer,5_UTR,Exon
chr14_96180001_96180200	12	TCL1A(0)	4,90E-20	Enhancer,Exon
chr16_3786601_3786800	11	CREBBP(0)	3,49E-19	Enhancer,Exon
chr8_128750601_128750800	11	MYC(0)	5,93E-19	Enhancer,Exon
chr11_102188401_102188600	11	BIRC3(0)	1,78E-18	Enhancer,Exon
chr14_96179801_96180000	11	TCL1A(0)	3,42E-18	Enhancer,Exon
chr2_89160801_89161000	11	MIR4436A(48834)	1,08E-17	Promoter,Exon
chr6_31549601_31549800	10	LTB(0)	2,73E-17	Enhancer,Exon
chr12_122458801_122459000	10	BCL7A(0)	4,43E-17	Enhancer,5_UTR,Exon
chr17_1021001_1021200	10	ABR(0)	4,43E-17	Enhancer,Exon
chr3_187461801_187462000	10	BCL6(0)	1,56E-16	Enhancer,Exon
chr16_33953201_33953400	13	LINC00273(7653)	2,18E-16	Enhancer
chr2_89159201_89159400	10	MIR4436A(47234)	6,34E-16	Promoter,Exon
chr21_15554801_15555000	12	LIPI(0)	1,27E-15	Exon
chr19_11134201_11134400	9	SMARCA4(0)	1,95E-15	Enhancer,Exon
chr6_41903601_41903800	9	CCND3(0)	3,03E-15	Enhancer,5_UTR,Exon
chr2_133018801_133019000	12	ANKRD30BL(3260)	6,50E-15	Promoter
chr16_33953001_33953200	12	LINC00273(7853)	6,50E-15	Enhancer
chr1_246395801_246396000	10	SMYD3(0)	6,70E-15	Enhancer,Exon
chr11_102188601_102188800	9	BIRC3(0)	7,55E-15	Enhancer,Exon
chr3_187463201_187463400	9	BCL6(0)	9,49E-15	Enhancer,Promoter,5_UTR,Exon
chr15_89096601_89096800	9	DET1(6696)	9,49E-15	NA
chr2_89160001_89160200	9	MIR4436A(48034)	3,39E-14	Promoter,Exon
chr2_89160601_89160800	9	MIR4436A(48634)	3,39E-14	Promoter,Exon
chr2_89165401_89165600	9	MIR4436A(53434)	3,39E-14	Exon
chr2_89165201_89165400	9	MIR4436A(53234)	3,39E-14	Exon
chr21_15555001_15555200	11	LIPI(0)	3,99E-14	Exon
chr5_2145601_2145800	9	IRX4(258252)	8,96E-14	NA
chr3_18830001_18830200	10	SATB1(342922)	1,13E-13	Enhancer,Exon
chr11_61635601_61635800	8	FADS2(776)	1,27E-13	Enhancer
chr16_503001_503200	8	RAB11FIP3(0)	1,27E-13	Enhancer,Exon
chr22_27169001_27169200	9	MIAT(96564)	1,30E-13	Exon
chr6_119558601_119558800	9	MAN1A1(0)	1,30E-13	Enhancer,Exon

chr11_51570001_51570200	11	OR4C46(53791)	1,80E-13	NA
chr11_51580401_51580600	11	OR4C46(64191)	1,80E-13	NA
chr12_122459201_122459400	8	BCL7A(0)	1,88E-13	Enhancer,5_UTR,Exon
chr8_128748401_128748600	8	MYC(0)	1,88E-13	Enhancer,5_UTR,Exon
chr12_122463001_122463200	8	BCL7A(0)	1,88E-13	Enhancer,Exon
chr3_49413001_49413200	8	RHOA(0)	1,88E-13	Enhancer,5_UTR,Exon
chr8_128750801_128751000	8	MYC(0)	1,88E-13	Enhancer,Exon
chr16_10973001_10973200	8	CIITA(0)	2,95E-13	Enhancer,Exon
chr1_235692001_235692200	8	GNG4(18788)	2,95E-13	Enhancer
chr2_112595201_112595400	8	ANAPC1(0)	5,25E-13	Enhancer,Exon
chr16_10746601_10746800	8	TEKT5(0)	5,25E-13	Enhancer,Exon
chr3_187462201_187462400	8	BCL6(0)	5,25E-13	Enhancer,Exon
chr4_49316401_49316600	10	CWH43(252304)	5,56E-13	NA
chr6_27792201_27792400	8	HIST1H4J(0)	6,96E-13	Enhancer,Promoter,3_UTR,Exon
chr3_187660801_187661000	8	BCL6(197287)	6,96E-13	Enhancer,Exon
chr8_129131001_129131200	8	PVT1(17503)	6,96E-13	NA
chr13_23151201_23151400	10	BASP1P1(320081)	1,16E-12	NA
chr2_89163401_89163600	8	MIR4436A(51434)	1,65E-12	Exon
chr2_89157601_89157800	8	MIR4436A(45634)	1,65E-12	Exon
chr2_89164401_89164600	8	MIR4436A(52434)	1,65E-12	Exon
chr2_89164801_89165000	8	MIR4436A(52834)	1,65E-12	Exon
chr2_88791201_88791400	8	TEX37(32770)	2,50E-12	Enhancer
chr3_197825801_197826000	8	ANKRD18DP(18211)	2,50E-12	Enhancer
chr7_64574001_64574200	8	CCT6P3(38911)	3,94E-12	NA
chr2_107016201_107016400	8	RGPD3(5047)	3,94E-12	NA
chr16_46412201_46412400	10	ANKRD26P1(90854)	4,59E-12	NA
chr2_133019401_133019600	10	ANKRD30BL(3860)	4,59E-12	Promoter
chr7_298201_298400	8	FAM20C(0)	5,52E-12	Exon
chr19_46151601_46151800	7	EML2(2715)	7,38E-12	Enhancer
chr7_75616801_75617000	7	TMEM120A(0)	7,38E-12	Enhancer,Exon
chr19_11144001_11144200	7	SMARCA4(0)	7,38E-12	Enhancer,Exon
chr7_982001_982200	7	COX19(0)	7,38E-12	Enhancer,Exon
chr8_69933401_69933600	9	C8orf34(202145)	8,06E-12	Exon

Supplementary Table 1. High confidence recurrently mutated regions identified in an unbiased, genome-wide analysis setup.

B Details on the methods used in the Thesis

B.1 Methods for Chapter 2

B.1.1 Somatic mutations dataset

The somatic SNV calling for the dataset were performed by authors of the original studies (Table 1). We downloaded whole-genome lists of somatic mutations in VCF format.

B.1.2 Principal component analysis

Lists of genetic and epigenetic features used for principal component analysis were obtained from the following sources: replication timing and expression levels for each region of the genome at 100-kb-resolution were taken from the supplementary data of Lawrence *et al.*, 2013; GC content dataset was downloaded from the UCSC genome browser, H3K9me3 data was used from Barski *et al.*, 2007 dataset similarly to the work by Schuster-Böckler and Lehner, 2012; HiC compartment data for lymphoblastoid cell line GM06990 at 100-kilobase resolution was obtained from Lieberman-Aiden *et al.*, 2009, only the first eigenvector was used. All features were averaged to compute their values for different window sizes.

To perform principal component analysis we constructed a matrix, in which rows corresponded to genomic and columns corresponded to five features (Table 2). Rows with missing values for any of the features were omitted; then the matrix was scaled and centered. The principal component analysis was performed in the R statistical environment using the *prcomp* function and the corresponding PC loading vectors for the 1st and 2nd PCs were obtained.

We studied correlation between five individual features, as well as the 1st and the 2nd PC loading vectors and somatic mutation rates in our cancer dataset at different resolution *i.e.* window sizes ranging from 50 kb to 10 Mb using Pearson correlation coefficients. Somatic mutation rates were calculated for every window size as the total number of mutations observed within a genomic window. Pearson correlation coefficients were computed using the *cor* function from the R statistical environment.

B.2 Methods for Chapter 3

B.2.1 Lymphoma dataset

The lists of somatic mutations for 23 lymphoma samples were provided by Tobias Rausch. The list of somatic and germline SNVs in VCF format as well as raw RNA-Seq data in FASTQ format were available in the lab.

B.2.2 Somatic SNV filtering and annotation of lymphoma samples

To obtain a list of high confidence somatic SNVs, we filtered out all SNVs that occurred in so-called "unreliable" regions. This definition included: 1) DAC Blacklisted Regions created by the ENCODE project (https://genome.ucsc.edu/ENCODE/); 2) low mappability regions; 3) highly repetitive regions according to the RepatMasker(http://www.repeatmasker.org/).

All data were downloaded from the UCSC Genome browser website https://genome.ucsc.edu/.

Additionally, we filtered out mutations that were listed in the dbSNP132 database (ftp://ftp.ncbi.nlm.nih.gov/snp/) as well as common polymorphisms identified by the 1000 Genomes Project (http://www.1000genomes.org/).

B.2.3 RNA-Seq data analysis

Raw RNA-Seq data in FASTQ format were provided by the GeneCore. FASTQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for the quality assessment and Trimmomatic was used to trim sequences based on their quality (Bolger *et al.*, 2014). The resulting sequences were then mapped to the reference genome hg19 annotated using the Gencode_v14. Mapping was performed using the STAR aligner (Dobin *et al.*, 2013) The resulting BAM files were processed using the RSeQC tool (Wang *et al.*, 2012) .

B.2.4 Population structure analysis

For this analysis germline SNV calls for 23 samples were combined with germline variants from 11 HapMap populations (Supplementray Table 2), only common SNPs were used. Pairwise identity-by-state (IBS) and identity-bydescent (IBD) values for every sample were calculated using PLINK tool (Purcell *et al.*, 2007) with default settings. The general guidelines on how to perform this analysis are described in this online tutorial: http://www.cureffi.org/2012/10/15/population-covariates-using-1000-genomes/

B.2.5 Genotype-phenotype correlations analysis

Gene expression correlation analysis with genotypes was performed using the modified version of a pipeline developed by Andreas Schlattl. The pipeline originally performs GC corrections and normalization by the read depth on BAM files and later integrates the expression values with copy-number state data. It was customized so that it could utilize the genotype (state of 1 kb region as described in the main text) and a burden test. It calculates Pearson correlation coefficients between the genotype and observed gene expression levels represented as RPKMs, followed by a multiple testing correction.

B.2.6 Transcription factor binding site changes

The somatic mutation data obtained from a dataset on eleven cancer types (Table 1) was filtered to prior to the analysis. First, somatic SNVs that were mutated at the exact same position in two individuals were selected and called recurrent. Then among the recurrent mutations were selected those that had another neighboring recurrent mutation within 100 bp window. Next, anti-correlation principle was applied (two recurrent mutations can not be present in the same individual). And finally on the remaining list of candidate regions the computational prediction of TFBS changes was performed.

Motif data for this analysis were obtained from the two sources: experimentally obtained data from the ENCODE project and a wide collection of TF motifs from various sources provided by HOCOMOCO (Kulakovskiy *et al.*, 2013).

PWMs for the analysis were first converted into the MEME format (Bailey *et al.*, 2009). Mutated sequences of 100 bp length for each individual were constructed by computationally introducing the mutations observed the samples into the reference sequence. For every sample we considered three sequences: a reference sequence, a mutated sequence and a reverse complimented version of the mutated sequence. Then using FIMO tool (Grant *et al.*, 2011) we computationally predicted all TFBS in the given sequences. Using the custom script we compared FIMO outputs between the sequences for each of the patients first and among all samples harboring mutations with an assumption that the observed somatic mutations should change in the same direction (e.g. either create a TFBS or disrupt it). As a result we identified genomic regions in which somatic mutations lead to changes in TFBS.

B.2.7 Windows-based approach

To identify genomic regions with single recurrent mutations or clusters of recurrent mutations we used a windows-based approach and binned the human genome in non-overlapping windows of various sizes ranging between 50 bp to 10 Mb. For each window we calculated the number of patients having at least one mutation in the given window and called this the mutational recurrence of the region.

For the restricted analysis we considered only those regions that had an overlap of more than 1 bp or 50% with the regions of interest (*i.e.* promoters and enhancers; regions with RegulomeDB score values 1-5), while for the genome-wide setup of the analysis we used all genomic windows without any filtering.

B.2.8 Annotations of recurrently mutated regions

We used gene annotations from Ensembl (v75) for the transcripts of all protein-coding genes. 5' UTRs and 3' UTRs were used as defined by Ensembl.

For restricted analysis promoter regions were defined as in the work by Weinhold *et al.*, 2014: the genomic intervals ranging from 2,000 bp upstream to 200 bp downstream of all transcription start sites; 27,493 enhancer regions (7,550 merged unique regions) were downloaded from the FANTOM5 website (Lizio *et al.*, 2015). Using bedtools we identified regions having more than 1 bp or 50% overlap with the combined list of promoters and enhancers. Likewise, we selected regions that had more than 1 bp or 50% overlap with regulatory genomic regions according to the RegulomeDB classification. RegulomeDB is a resource that provides functional annotation for any region in human genome based on multiple levels of evidence and classifies genomic regions based on the evidence into 7 categories, where 1-5 categories correspond to regulatory regions (Boyle *et al.*, 2012). Similar to work by Melton *et al.*, 2015 we required for genomic windows to overlap with regions of 1-5 RegulomeDB categories in this study.

Unfortunately, the number of genomic windows overlapping with at least 50% of their size with RegulomeDB annotated regulatory elements was not sufficient for further analysis including cross-validations.

B.2.9 Identification of recurrently mutated regions

To identify which of the genomic windows are recurrently mutated while controlling for the regional mutational heterogeneity we used the following strategy.

Let n be the total number of samples. For a given region i, k_i is the number of individuals that have at least one mutation in the region. To estimate the background mutational rate μ_i we used a “global” model: we stratified the genome into 25 equally-sized groups of genomic windows with similar genetic and epigenetic background based on the 1st PC loading vector values for each window. This way for each region i we could identify a list of m genome-wide regions and therefore estimate its background mutational rate μ_i from the list of regions as an average number of individuals with mutations, average(k_1,..,k_m), as well as its variance v_i.

For each region i we computed its Enrichment Score as k_i/μ_i ; one-tailed Binomial p-values using k_i and μ_i ; Negative Binomial test p-values using k_i, μ_i and a dispersion parameter calculated as $\frac{{\mu_i}^2}{v_i-\mu_i}$.

B.2.10 Cross-validations

To choose the significance cut-off that would give us reproducible results we performed cross-validations. Samples were segregated by cancer type and one half of samples of each cancer type were selected as set S_1, while the other half was referred to as set S_2.

We performed the identification of recurrently mutated regions independently for S_1 and S_2 sets and then compared how reproducible the results were in terms of p-values and enrichment scores.

Based on the results of cross-validations, we chose a combination of the window size, test statistic and a cut-off value that ensured high precision and recall values based on the precision-recall analysis. We then use the chosen parameters to run the pipeline on the complete ($S_1 \cup S_2$) dataset.

B.2.11 Precision-recall analysis

For each window size and cut-off combination precision and recall values were computed as follows. Recall was calculated as a number of regions that satisfy the cut-off in both S_1 and S_2 sets results. Precision was calculated as a fraction of the recalled regions to the total number of regions satisfying the cut-off in both datasets.

The combination of the window size with the p-value cut-off that allows for the highest precision given large recall was selected as the optimal choice of the parameters for the data.

B.2.12 Gene expression analysis

We used gene expression data obtained using RNA-seq technology for 128 medulloblastoma samples (Jones *et al.*, 2012). TERT gene expression values were compared for samples wsith and without mutations in the region of interest, 6 and 122 samples respectively. Samples with high-level TERT amplification were excluded. RPKM values were used.

C List of publications

Rudneva VA., Anders S., Huber W., Korbel JO. Robust identification of recurrent intergenic somatic mutations in cancer genome. (in review in *International Journal of Cancer*).

Bibliography

Ajay,S.S. *et al.* (2011) Accurate and comprehensive sequencing of personal genomes. *Genome Res.*, **21**, 1498–505.

Aleman,A. *et al.* (2014) A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. *Nucleic Acids Res.*, **42**, W88–W93.

Alexandrov,L.B. *et al.* (2015) Clock-like mutational processes in human somatic cells. *Nat. Genet.*, **47**, 1402–1407.

Alexandrov,L.B. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–21.

Bailey,T.L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–8.

Baker,M. (2012) Structural variation: the genome's hidden architecture. *Nat. Methods*, **9**, 133–7.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–37.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Boyle,A.P. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, **22**, 1790–1797.

Burns,M.B. *et al.* (2013) Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.*, **45**, 977–983.

Denisova,E. *et al.* (2015) Frequent *DPH3* promoter mutations in skin cancers. *Oncotarget*, **6**, 35922–30.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Feuk,L. and Carson,A. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 1169–1171.

Forbes,S.A. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–11.

Frazer,K.A. *et al.* (2009) Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.*, **10**, 241–51.

Fredriksson,N.J. *et al.* (2014) Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.*, **46**, 1258–1263.

Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. *Science*, **342**, 357–60.

Grant,C.E. *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–8.

Helleday,T. *et al.* (2014) Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, **15**, 585–598.

Horn,S. *et al.* (2013) TERT promoter mutations in familial and sporadic melanoma. *Science*, **339**, 959–61.

Huang,F.W. *et al.* (2013) Highly recurrent TERT promoter mutations in human melanoma. *Science*, **339**, 957–9.

Jager,N. *et al.* (2013) Hypermutation of the Inactive X Chromosome Is a Frequent Event in Cancer. *Cell*, 567–581.

Jones,D.T.W. *et al.* (2012) Dissecting the genomic complexity underlying medulloblastoma. *Nature*, **488**, 100–105.

Khurana,E., Fu,Y., Colonna,V., Mu,X.J., Kang,H.M., Lappalainen,T., *et al.* (2013) Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science (80-.).*, **342**, 1235587–1235587.

Khurana,E., Fu,Y., Colonna,V., Mu,X.J., Kang,H.M., Lappalainen,T., *et al.* (2013) Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science (80-.).*, **342**, 1235587–1235587.

Kool,M. *et al.* (2014) Genome Sequencing of SHH Medulloblastoma Predicts Genotype-Related Response to Smoothened Inhibition. *Cancer Cell*, **25**, 393–405.

Kulakovskiy,I. V. *et al.* (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.*, **41**, D195–202.

Kundaje,A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

Lada,A.G. *et al.* (2012) AID/APOBEC cytosine deaminase induces genome-wide kataegis. *Biol Direct*, **7**, 47; discussion 47.

Lawrence,M.S. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–8.

Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-.).*, **326**, 289–293.

Lizio,M. *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, **16**, 22.

Loots,G.G. (2008) Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis. *Adv. Genet.*, **61**, 269–293.

Mansour,M.R. *et al.* (2014) An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (80-.).*, **346**, 1373–1377.

Martincorena,I. and Campbell,P.J. (2015) Somatic mutation in cancer and normal cells. *Science (80-.).*, **349**, 1483–1489.

Melton,C. *et al.* (2015) Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.*, **47**, 710–716.

Meyerson,M. *et al.* (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Publ. Gr.*, **11**, 685–696.

Mills,R.E. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Mills,R.E. *et al.* (2006) Recently Mobilized Transposons in the Human and Chimpanzee Genomes. *Am. J. Hum. Genet.*, **78**, 671–679.

Montgomery,S.B. *et al.* (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.*, **23**, 749–61.

Nik-Zainal,S. *et al.* (2012) Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell*, **149**, 979–993.

Northcott,P. a. *et al.* (2014) Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature*, **511**, 428–434.

Northcott,P.A. *et al.* (2012) Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*, **488**, 49–56.

Nowell,P. and Hungerford,D. (1960) National Academy of Sciences. *Science (80-.).*, **132**, 1488–1501.

Onishi-Seebacher,M. and Korbel,J.O. (2011) Challenges in studying genomic structural variant formation mechanisms: The short-read dilemma and beyond. *BioEssays*, **33**, 840–850.

Park,W.J. *et al.* (2011) FADS2 function loss at the cancer hotspot 11q13 locus diverts lipid signaling precursor synthesis to unusual eicosanoid fatty acids. *PLoS One*, **6**, e28186.

Paz Polak, Rosa Karlic, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence,A.R. and Eric Rynes, Kristian Vlahovicˇek,J.A.S.& S.R.S. (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, **518**, 360–364.

Peifer,M. *et al.* (2015) Telomerase activation by genomic rearrangements in high-risk neuroblastoma. *Nature*, **526**, 700–704.

Pleasance,E.D. *et al.* (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, **463**, 191–6.

Poulos,R.C. *et al.* (2015) The search for cis -regulatory driver mutations in cancer genomes. *Oncotarget*, **6**.

Puente,X.S. *et al.* (2015) Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **526**, 519–524.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–75.

Rausch,T. *et al.* (2012) Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. *Cell*, 59–71.

Richter,J. *et al.* (2012) Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.*, **44**, 1316–20.

Schuster-Böckler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–7.

Shaw,C.J. (2004) Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.*, **13**, 57R–64.

Shen,J.-C. *et al.* (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.*, **22**, 972–976.

Stankiewicz,P. and Lupski,J.R. (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet.*, **18**, 74–82.

Stein,L.D. *et al.* (2015) Data analysis: Create a cloud commons. *Nature*, **523**, 149–151.

Tomlins,S. a (2005) Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. *Science (80-.).*, **310**, 644–648.

Vinagre,J. *et al.* (2013) Frequency of TERT promoter mutations in human cancers. *Nat. Commun.*, **4**, 2185.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Wang,L. *et al.* (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185.

Ward,L.D. and Kellis,M. (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, **337**, 1675–8.

Weinhold,N. *et al.* (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.*, **46**, 1160–1165.

Weischenfeldt,J. *et al.* (2013) Article Integrative Genomic Analyses Reveal an Androgen-Driven Somatic Alteration Landscape in Early-Onset Prostate Cancer. 159–170.

Yang,H. *et al.* (2010) Important role of indels in somatic mutations of human cancer genes. *BMC Med. Genet.*, **11**, 128.

Zhang,Z. and Gerstein,M. (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res.*, **31**, 5338–48.