CANCER GENETICS RESEARCH METHODS IN THE NEXT-GENERATION SEQUENCING ERA

Riku Katainen

Department of Medical and Clinical Genetics, Medicum
Applied Tumor Genomics Research Program
Doctoral Programme in Biomedicine (DPBM)
Faculty of Medicine
University of Helsinki
Finland

ACADEMIC DISSERTATION

To be presented for public discussion with the permission of the Faculty of Medicine of the University of Helsinki, in Haartman Institute, Lecture hall 2, Haartmaninkatu 3, Helsinki, on the 20th of March, 2020 at 12 noon

Helsinki 2020

Supervised by Academy Professor Lauri A. Aaltonen, M.D., Ph.D.
Department of Medical and Clinical Genetics, Medicum
Applied Tumor Genomics Research Program
Faculty of Medicine, University of Helsinki, Finland

&

Docent Esa Pitkänen, Ph.D.
Institute for Molecular Medicine Finland (FIMM)
Applied Tumor Genomics Research Program
University of Helsinki, Helsinki, Finland

Reviewed by Docent Merja Heinäniemi, Ph.D.
Institute of Biomedicine
University of Eastern Finland, Kuopio, Finland

&

Docent Sofia Khan, Ph.D.
Turku Bioscience Centre
University of Turku, Turku, Finland

Official opponent Jussi Paananen, Ph.D.
Institute of Biomedicine
University of Eastern Finland, Kuopio, Finland
Blueprint Genetics Oy

ISBN 978-951-51-5898-7 (paperback)
ISBN 978-951-51-5899-4 (PDF)
Unigrafia Oy
Helsinki 2020

> We are at the very beginning of time for the human race. It is not unreasonable that we grapple with problems. But there are tens of thousands of years in the future. Our responsibility is to do what we can, learn what we can, improve the solutions, and pass them on.
>
> Richard P. Feynman

TABLE OF CONTENTS

ORIGINAL PUBLICATIONS

This thesis is based on the following original publications:

I **Katainen R**, Donner I, Cajuso T, Kaasinen E, Palin K, Mäkinen V, Aaltonen L.A., Pitkänen E. Discovery of potential causative mutations in human coding and noncoding genome with the interactive software BasePlayer. *Nature Protocols* **13**(11), 2018.

II **Katainen R***, Dave K*, Pitkänen E*, Palin K*, Kivioja T, Välimäki N, Gylfe AE, Ristolainen H, Hänninen UA, Cajuso T, Kondelin J, Tanskanen T, Mecklin JP, Järvinen H, Renkonen-Sinisalo L, Lepistö A, Kaasinen E, Kilpivaara O, Tuupanen S, Enge M, Taipale J, Aaltonen L.A. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics* **47**(7):818-21, 2015.

III Donner I, **Katainen R**, Tanskanen T, Kaasinen E, Aavikko M, Ovaska K, Artama M, Pukkala E, Aaltonen L.A. Candidate susceptibility variants for esophageal squamous cell carcinoma. *Genes, Chromosomes and Cancer* **56**(6):453-459, 2017.

*Equal contribution

1.1 Author's contributions

I Designed, developed and tested the software. Designed the use cases and processed additional annotation files for end users. Wrote the manuscript together with other authors.

II Participated in designing the study. Performed primary sequence, somatic mutation and sequence motif analyses. Developed methods to calculate mutation clusters, analyze mutations in transcription binding motifs, and integrate regulatory genome and gene annotation with mutation data. Wrote the manuscript together with other authors.

III Participated in the variant analyses and designing the study. Produced control data sets. Developed methods to integrate case and control data for enrichment analysis. Wrote the manuscript together with other authors.

ABBREVIATIONS

ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
BAM/SAM	Sequence alignment map (B indicates binary format)
BED	Browser Extensible Data (file format)
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
CBS	Cohesin Binding Site
ChIP-seq/exo	Chromatin Immunoprecipitation Sequencing / Exonuclease Digestion
CRC	Colorectal Cancer
DNA	Deoxyribonucleic Acid
ESCC	Esophageal Squamous Cell Carcinoma
FFPE	Formalin-Fixed Paraffin-Embedded
GATK	Genome Analysis Toolkit
ICGC	International Cancer Genome Consortium
MSI / MSS	Microsatellite instability / stable
NGS	Next-Generation Sequencing
PSSM	Position Specific Scoring Matrix
RNA	Ribonucleic Acid
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SNV	Single Nucleotide Variant
TF	Transcription Factor
UTR	Untranslated Region
VCF	Variant Calling Format
WES	(Whole) Exome Sequencing
WGS	Whole-Genome Sequencing

Genes

APC	Adenomatous Polyposis Coli
BRAF	V-Raf Murine Sarcoma Viral Oncogene Homolog B
BRCA1 & 2	Breast Cancer Type 1 & 2 Susceptibility Protein
CTCF	CCCTC-Binding Factor
EPCAM	Epithelial Cell Adhesion Molecule
KRAS	Kirsten Rat Sarcoma virus (oncogene)
MLH1	MutL Homolog 1
MSH2 / *MSH6*	MutS Homolog 2 & 6
MYC	V-Myc Avian Myelocytomatosis Viral Oncogene Homolog
PMS2	PMS1 Homolog 2, Mismatch Repair System Component
POLE	DNA Polymerase Epsilon Catalytic Subunit
TP53	Tumor Protein p53

ABSTRACT

The research in cancer genetics aims to detect genetic causes for the excessive growth of cells, which may subsequently form a tumor and further develop into cancer. The Human Genome Project succeeded in mapping the majority of the human DNA sequence, which enabled modern sequencing technologies to emerge, namely next-generation sequencing (NGS). The new era of disease genetics research shifted DNA analyses from laboratory to computer screens. Since then, the massive growth of sequencing data has been facilitating the detection of novel disease-causing mutations and thus improving the screening and medical treatments of cancer. However, the exponential growth of sequencing data brought new challenges for computing. The sheer size of the data is not only expensive to store and maintain, but also highly demanding to process and analyze. Moreover, not only has the amount of sequencing data increased, but new kinds of functional genomics data, which are instrumental in figuring out the consequences of detected mutations, have also emerged. To this end, continuous software development has become essential to enable the utilization of all produced research data, new and old.

This thesis describes a software for the analysis and visualization of NGS data (publication I) that allows the integration of genomic data from various sources. The software, BasePlayer, was designed for the need of efficient and user-friendly methods that could be used to analyze and visualize massive variant, and various other types of genomic data. To this end, we developed a multi-purpose tool for the analysis of genomic data, such as DNA, RNA, ChIP-seq, and DNase. The capabilities of BasePlayer in the detection of putatively causative variants and data visualization have already been used in over twenty scientific publications. The applicability of the software is demonstrated in this thesis with two distinct analysis cases - publications II and III.

The second study considered somatic mutations in colorectal cancer (CRC) genomes. We were able to identify distinct mutation patterns at the CTCF/ Cohesin binding sites (CBSs) by analyzing whole-genome sequencing (WGS) data with BasePlayer. The sites were observed to be frequently mutated in CRC, especially in samples with a specific mutational signature. However, the source for the mutation accumulation remained unclear. On the contrary, a subset of samples with an ultra-mutator phenotype, caused by defective polymerase epsilon (*POLE*) gene, exhibited an inverse pattern at CBSs. We detected the same signal in other, predominantly gastrointestinal, cancers as well. However, we were not able to measure changes in gene expressions at mutated sites, so the role of the CBS mutations in tumorigenesis remained and still remains to be elucidated.

The third study considered esophageal squamous cell carcinoma (ESCC), and the objective was to detect predisposing mutations using the Finnish Cancer Registry (FCR) data. We performed clustering analysis for the FCR data, with additional information obtained from the Population Information System of Finland. We detected an enrichment of ESCC in the Karelia region and were able to collect and sequence 30 formalin-fixed paraffin-embedded (FFPE) samples from the region. We reported several candidate genes, out of which *EP300* and *DNAH9* were considered the most interesting. The study not only reported putative genes predisposing to ESCC but also worked as a proof of concept for the feasibility of conducting genetic research utilizing both clustering of the FCR data and FFPE exome sequencing in such studies.

INTRODUCTION

The concept of cancer is easy to understand; there are too many cells in the wrong place. This simplification may evoke a false notion that cancer is a simple subject to research and straightforward disease to cure. Decades of cancer research have, however, revealed the diverse nature of tumors and cancer; the understanding of the process, in which cells of a healthy tissue have become harmful to its host, requires research from the molecular to the tissue level. This thesis introduces challenges and methods of modern human cancer genetics research, the primary goal of which is to detect early events leading to tumors by analyzing the code written in the largest naturally occurring molecules - *chromosomes*.

Almost all human cells hold 46 chromosomes, large DNA molecules that contain instructions to build and maintain the essential functions and structures in and between all the trillions of cells, which form our bodies. Alterations in these instructions, *mutations*, may lead to abnormalities in the complex life-sustaining mechanisms and to an excessive reproduction of abnormal cells. The methods in cancer genetic research have been developed to detect these DNA alterations that may predispose to, or drive a particular cancer.

The key technique is the sequencing of DNA, where the information inside chromosomes is translated into an analyzable form. Next-generation sequencing enables the sequencing of all chromosomes or the genome of the tissue sample, which then allows researchers to compare DNA sequences between healthy and diseased samples, and thus detect abnormal alterations. The interpretation of NGS data is performed with computers, containing challenges such as: are the detected genetic alterations correctly read or sequencing artefacts? Has the alteration an effect on the studied disease? What is the function of the alteration? How to combine or integrate data from different sources to improve the interpretation? How to handle the massive sequencing data sets? The aim of this thesis is to introduce novel methods and solutions to these challenges and to clarify the concepts that are needed in everyday analysis of NGS data. The main focus is on cancers originating from solid tissues, however, presented techniques and principles are applicable to hematological malignancies as well. The biological concepts are described in the level of detail needed to understand the big picture of modern cancer genetics research and to follow the publications of this thesis.

REVIEW OF THE LITERATURE

5.1 Cancer as a disease

Healthy tissues of our bodies are composed of networks of collaborating, specialized cells, which all contain practically identical genetic material [1]. Normal tissue renewal, for instance, in skin or epithelium of intestine, is maintained by controlled cell divisions that occur continuously throughout the lifetime of an organism [2]. Tissue grows, when the rate of cell divisions exceeds the number of controlled cell death events [3, 4]. While the growth can be desired in cases such as the development of muscle mass or wound healing, it can be undesirable when it occurs unsuppressed, for instance, in internal organs. The basic concept of cancer is easy to understand - malfunctioning cells have started to divide excessively, forming a tumor, and have subsequently gained malignant abilities to spread to other parts of the body, leading to cancer. The tricky part, however, is to determine the underlying cause of the uncontrolled growth of cells - the genetics of cancer [5].

Every individual is different in terms of DNA composition, so is every tumor. Moreover, a single solid tumor may be a combination of multiple cell populations harboring distinct pathogenic mutations and tissue environments, further complicating cancer treatment and research [6]. Tumors generally arise from a cell or cells of the healthy tissue of an individual through decades of accumulated mutations in DNA and changes in the tissue environment. During the development, benign tumor cells can gain additional, stem cell-like properties, which enable the primary tumor to invade into foreign tissues (**Figure 1**) [7]. These properties can be gained through several distinct features, or hallmarks, listed below [8].

1. Autonomous growth and proliferative stimulation. Growth factors are useful when individuals are maturing and when damaged tissues need to be healed [9]. However, the cell controls its division rate in normal conditions by suppressing the growth factor mediating pathways [10]. One of the essential features of tumor cells is the sustained growth stimulation, and this is achieved by disrupting growth factor mediating pathways through mutations. In addition, tumor cells can gain the ability to stimulate the division of surrounding cells through, e.g., tumor-promoting inflammation [8].

2. Evasion of growth suppressing signals. The continuous proliferation of single-celled organisms, such as bacteria, is restrained almost merely by the depletion of nutrients and ecological competition. Multicellular organisms, however, are a combination of differentiated cells, the seamless interplay of which is vital in sustaining the growth balance in the ensemble of various tissues and organs to form a viable body [11, 12]. Not only are cells

limiting their individual growth, but they also receive suppressive signals from surrounding cells. By ignoring the suppressive signals, the cell can gain a growth advantage compared to surrounding cells and ignite tumorigenesis [8].

3. Avoidance of programmed cell death (apoptosis) and immune destruction. Cells have an internal guard system for the detection of malfunctions [13]. For instance, checkpoint proteins can send a cell death signal if excess damage in the DNA is detected. However, a broken checkpoint protein may give the cell permission to continue with the cell cycle and divide despite the disturbed homeostasis in the nucleus, leading to growth of damaged cells [4]. Also, abnormalities can alert the immune system, which is poised to deal with misbehaving cells. The ability to be hidden from the immune system and avoid an immune response have been proposed to be an additional hallmark of cancer (**Figure 1**) [8].

4. Ability to replicate the DNA indefinitely. Every cell division requires the replication of all chromosomes. Chromosome ends have repetitive sequences, telomeres, that work as a buffer to maintain the integrity of functional DNA and prevent the conjoining of chromosome ends [14]. Replication mechanisms operate such that chromosome ends get shorter during every division process. Hence, the number of cell divisions is limited to approximately 50 times [15]. In normal conditions, cells do not reconstruct the telomeres, but through activation of the telomerase protein, the function of which is to lengthen telomeres, the cell can divide perpetually in terms of DNA replication [8].

5. Maintaining genome instability and mutation accumulation. A tumor can be seen as a microenvironment with its own evolutionary system, where individual cells are reproductive units under constant selective pressure [16]. Tumors encounter multiple natural and unnatural barriers during their evolution, such as malnutrition, immune responses, and possible cancer therapies [17, 18]. Like in any evolutionary system, tumor cells can adapt to environmental changes through genetic alterations; fittest tumor cells survive, and by proliferation of these mutated cells (clonal expansion), they can grow a new, more resilient tumor mass [18]. Through, for instance, increased sensitivity to mutagenic agents and defective guard systems (see 3rd feature on this list), a tumor cell can maintain and accelerate instabilities in its genome [8].

6. Ensuring the availability of extra energy and nutrients for tumorigenesis. Solid tumors can generally not grow larger than ~2 mm in diameter, without a system to provide nutrients and oxygen to peripheral cells [19]. The ability to generate blood vessels (angiogenesis) enables the tumor to grow beyond that limit (**Figure 1**). Also, the tumor needs extra energy

for its excessive cell proliferation, which is provided through reprogrammed metabolism.

7. Ability to invade adjacent and distal tissues (metastasis). At later stages of tumorigenesis, the tumor cells can gain the potency to sustain growth or even thrive within foreign environments. Cancer of solid tissues originates from a primary tumor, which starts to invade adjacent tissue and disseminate its cells to the bloodstream (**Figure 1**). These circulating cells can then after a long period of dormancy invade other tissues, causing the tumor to metastasize [8]. These events, invasion of adjacent tissue and metastasis, are what differentiate a benign tumor from a malignant, that is, cancer.

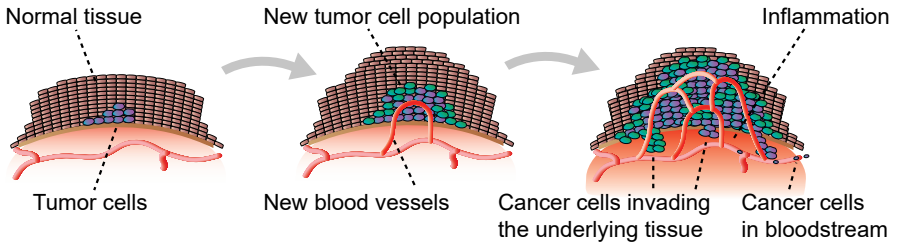

Figure 1: Multi-step evolutionary process of tumor. Tumor evolution from benign to malignant.

Even though only a handful of hallmarks is required for cancers to develop, there are numerous different paths to gain these features and form a unique microenvironment that allows tumors to grow and spread [20]. This microenvironment of billions of specialized cells harbors numerous genetic and epigenetic aberrations and abnormalities in extra- and intracellular signaling that make cancers particularly challenging to study and cure [21]. However, recent advancements in disease genetics and medical research have enhanced the survival of cancer patients through improved screening and targeted treatments. Although cancer is considered to be a common disease, one could argue that the formation of a cancerous tumor is, in fact, an infrequent and unlike event in terms of scale and time. An average adult human body is an assemblage of roughly 40 trillion cells, which are dividing and accumulating mutations while fighting against viruses and bacteria, and still, it commonly takes decades before a population of cells which have gained all the sufficient hallmarks to become cancer emerge.

5.2 Cancer as a research subject

Cancer genetics research aims to detect cancer-driving or predisposing alterations in the genome. Typically, cancer drivers can be detected by comparing *somatic mutations* present only in tumors of the same type, whereas predisposing alterations are studied by comparing *germline variants* between patients carrying the same disease [22–24]. Both approaches have their own challenges and procedures, however, they share the same research questions: what is the function of the found alteration, and how does it contribute to the studied disease? Genetic research begins with the detection of an alteration or defect in a certain genomic region, that is enriched in cancer cases. Next, the function of the alteration is assessed by studying which gene or genes it affects. Thus far, over 500 genes have been linked to cancer based on numerous cancer genetics studies [22, 25]. Often in general-audience publications, the term "cancer gene" is used to describe the results of cancer genetics research. While this is not entirely false or misleading, the gene itself is not cancer-causing when functioning normally. On the contrary, the "cancer gene" *BRCA1* (named after breast cancer), for instance, protects the cell or tissue from becoming cancerous, but when damaged by mutation, it can lose this protective function [26, 27].

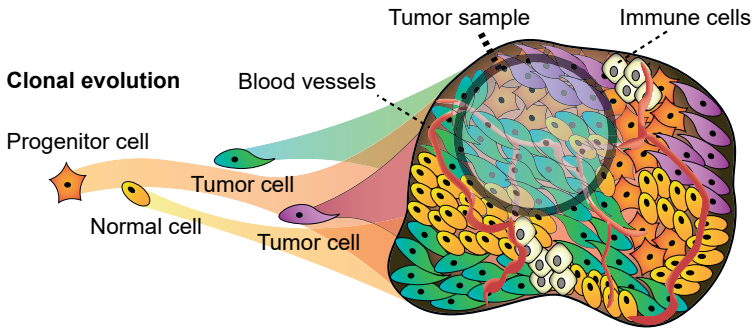

Figure 2: Tumor heterogeneity and purity. Heterogeneous tumors contain multiple tumor cell populations, possibly harboring different driver mutations. Tumor samples may contain cells only from one of the many populations. Impure tumor samples contain healthy cells (e.g., from blood vessels), which do not necessarily contribute to tumor growth and do not harbor pathogenic driver mutations.

Research on somatic mutations requires the DNA from diseased cells. Most current technologies require DNA material from a large mass (up to millions) of cells in order to produce accurate measurements. Moreover, a tumor sample constitutes only a small part of the whole tumor, and the sample is commonly a bulk of multiple cell populations (normal and tumor), that further complicates the analysis of tumor specific alterations and events [28–30]. Tumor heterogeneity (multiple cell populations in a single tumor) and

purity (the contamination of normal cells in a tumor sample) are factors, which affect almost all phases of the research from sample selection and preparation to computational processing and genetic analysis (**Figure 2**) [31, 32].

Studies on cancer susceptibility do not necessarily require utilization of tumor samples, hence heterogeneity and purity are not issues in these analyses. Practically all variants that can be detected from healthy tissue samples are inherited from the parents of the donor. By comparing these inherited variants between patients (*cases*) and healthy individuals (*controls*), it is possible to detect potential predisposing alterations to a particular disease [33]. For example, the variant present only in cases (i.e., segregating) within a family of three affected siblings, could be causing the disease in the family. However, analysis of small pedigrees is challenging, especially when studying common diseases, due to the possible presence of *phenocopies* (**Figure 3**). Phenocopies are individuals affected with the same disease but without the same inherited component [34]. The presence of phenocopies hampers the predisposing variant detection as they do not share the inherited variant with the "real" familial cases [33, 35]. Also, penetrance may be incomplete, meaning that some seemingly healthy individuals may be carriers of the inherited pathogenic variant (**Figure 3**). Thus, small-scale familial studies are often intended for the detection of rare variants in monogenic diseases.

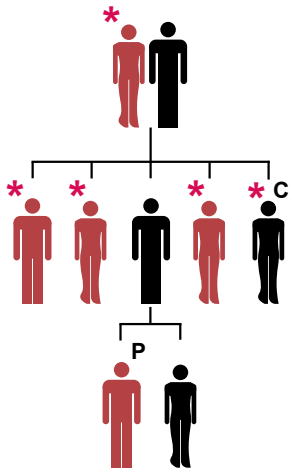

Figure 3: Familial cancer. Affected individuals (red) in a family with an inherited pathogenic mutation (asterisk). Non-affected carrier and phenocopy is denoted with **C** and **P**, respectively.

The availability of large biobanks and variant databases has enabled large-scale genome-wide association studies (GWAS) of more common DNA alterations (SNPs) and more complex traits on population level. GWAS utilizes statistical models to detect associations between diseases and SNPs [36, 37]. A decade worth of GWAS with thousands of samples and sample sets have revealed more than 16,000 trait associations; however, the causativity and functions of the reported loci are still vastly unknown [38–40]. The majority of these cancer predisposing SNPs reside in the noncoding genome, particularly in the *enhancer* rich regions, which are discussed further in the "Regulatory and the noncoding genome" chapter [25].

5.3 Cancer types relevant in this thesis

This thesis describes two distinct cancer genetics studies. Publication II focuses on somatic mutations in colorectal cancer, whereas publication III is a study of predisposing alterations in esophageal squamous cell carcinoma. Somatic mutations in the noncoding genome had not been thoroughly characterized, which prompted us to sequence over two hundred CRC samples genome-wide. Likewise, the role of inheritance in ESCC had not been extensively studied, and with the help of the Finnish Cancer Registry, we were able to collect familial cases for research. The two cancer types are described in more detail below.

5.3.1 Colorectal cancer

CRC is the most common type of gastrointestinal tract cancers arising from the inner lining (epithelium) of the large intestine (colon) or rectum [41]. It is also the third most common cancer worldwide and one of the leading causes of cancer-related deaths [42, 43]. While CRC prevention and survival have improved, the global CRC burden has been increasing alongside economic growth and the increasing life expectancy of the human population [42, 43]. The incidence rate is highest in wealthy countries with the western lifestyle, and the rate is increasing most rapidly in countries that have recently made the transition from low-income to high-income economy [42, 43]. The major lifestyle risk factors are excessive consumption of red meat (especially processed), alcohol, smoking, obesity, and physical inactivity [43]. Other risk factors include inflammatory bowel disease (IBD), and family history of CRC or adenomatous polyps. Family history has been estimated to account for up to 30% of CRC cases, where the proportion of inherited monogenic disorders such as Lynch syndrome, Familial Adenomatous Polyposis, and *MYH*-associated polyposis is estimated to be 5%. At least 70% of all CRC cases are sporadic (i.e., without family history). Colon and rectum are under strong mutagenic pressure due to nutritional exposures and rapid renewal of the epithelium tissue. Mutation patterns and mechanisms in CRC, including the ones found in Lynch syndrome, are described in the later sections.

5.3.2 Esophageal squamous cell carcinoma

ESCC is the most common cancer of the esophagus, and like CRC, it arises from the epithelial cells of the gastrointestinal tract. Albeit being one of the lesser-studied cancer types, ESCC is one of the most aggressive ones with a five-year survival rate of 15-25%. It is the sixth most common cause of cancer-related death and the eighth most common cancer worldwide [44]. Incidence rates of ESCC vary greatly internationally; the highest rates are found in Eastern Asia, China in particular, and in Eastern and Southern Africa, whereas the lowest rates are found in Western Africa. As with

CRC, the incidence of ESCC is increasing. However, the incidence rate of esophageal adenocarcinoma, the other main histological subtype of esophageal cancer, has exceeded the incidence rate of ESCC in some western countries such as the UK, USA, Finland, and France. Risk factors for ESCC include smoking, consumption of alcohol, poor oral hygiene, and nutritional deficiencies. While considerable geographical differences and strong correlations with smoking and alcohol imply that external factors cause the vast majority of ESCC cases, several studies have suggested that genetic factors may also contribute to the susceptibility of the disease [45, 46].

5.4 Genetics in cancer research

Cancer is fundamentally a disease of the genome [47]. The research on pathogenic alterations requires knowledge about the functional sites of the genome, which can drive cancer when defective. This section describes functionally relevant regions of the human genome and various types of alterations in the context of cancer genetics.

5.4.1 Structure of the human genome

The genome is a complete set of information coded with *nucleotides*, which are the units that form the large DNA molecules called chromosomes. Nucleotides hold four different bases: adenine, cytosine, guanine, and thymine (A, C, G, T), and they constitute the alphabet of our genetic code. The bonds between *base pairs* (bp) A-T and C-G maintain the double-helical structure of DNA [48]. The term "base pair" is often used as a length measurement unit of DNA sequences; for instance, the human genome is approximately 3 billion bp, and includes 16 kbp mitochondrial DNA located outside the nucleus in the cytoplasm. The nuclear DNA of a human is composed of 23 different sized chromosome pairs (one from both parents), which are packed into an extremely tight *chromatin* structure in the nuclei of almost all cells of our bodies [49]. In comparison, the genomes of a carrot and a donkey are composed of 9 and 31 chromosome pairs, respectively. Chromatin is a functional assembly of chromosomes and *histone* proteins, which provide dynamic structural modifications within the nucleus (**Figure 4**) [50].

Inside all nuclei, there are approximately two meters worth of DNA, with different combinations of open (*accessible*) and closed, tightly packed regions depending on the cell type [51]. The accessibility of DNA can determine the activity of genomic regions, for example, whether a particular gene is expressed or not in a cell [50]. The alterable structure of chromatin is an example of a mechanism responsible for *gene regulation* [52, 53]. DNA contains sections, which have distinct functions and purposes. Some parts of the DNA sequence, the genes, contain a code, that can be translated into proteins.

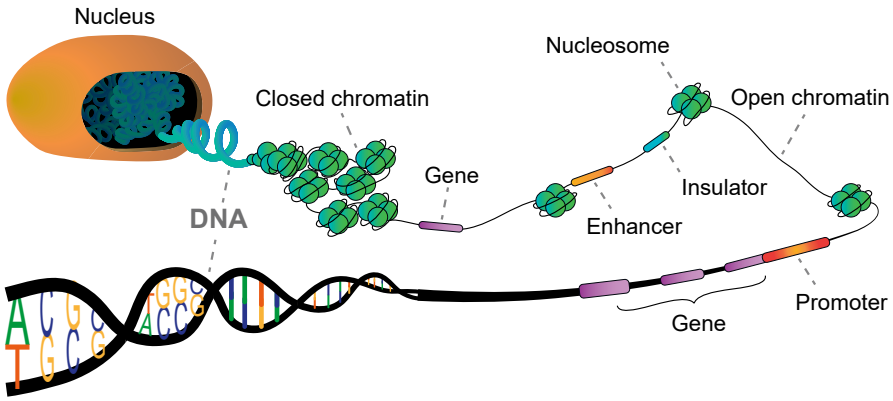

Figure 4: DNA, chromatin, genes and regulatory regions.

Other parts contain regions, which determine or regulate which proteins are produced and to what extent (**Figure 4**). Although the regulatory regions constitute the "second genetic code", only the protein-coding regions are considered to be coding and the rest of the genome is referred to as the noncoding genome.

5.4.2 Coding and noncoding genome

The division of the genome into coding and noncoding regions is rationalized by the distinct functions of these two; coding regions (*exons* in genes) can be translated into proteins, and they constitute ~1.5% of the human genome. The regulatory parts of noncoding regions determine which genes are expressed and their expression level at given conditions. The vast majority of the noncoding genome contains regions possessing unknown or seemingly redundant functionality. Moreover, only a small part of functional regions and genes are active in a specific tissue or cell type [54].

5.4.2.1 Genes and the coding genome

The human genome holds, according to current estimations, approximately 22,000 protein-coding genes, which encode all functional and structural proteins in our bodies. In comparison, the genomes of a grape and a fruit fly contain ~30,000 and ~15,000 genes, respectively [55]. Genes are segments of the DNA sequence, that are seemingly randomly dispersed throughout the genome. A typical gene contains untranslated regions (UTRs, **Figure 5**) and multiple protein-coding sequences, exons, which are separated by noncoding sections (introns). The size of a gene (sum of exon lengths) varies from ~200 bp to 100,000 bp. The total length of a gene (sum of exon, intron, and UTR lengths) can span over two million base pairs of the chromosome.

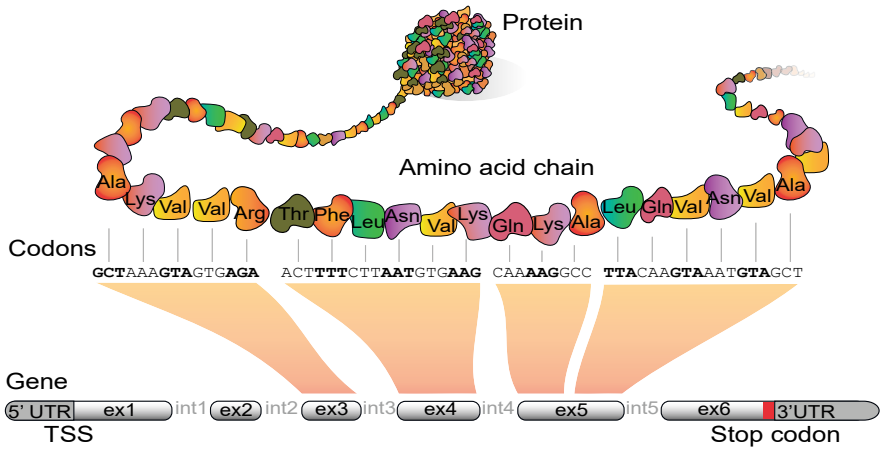

Figure 5: Structure of a gene and its relation to protein. Three-dimensional protein is formed from the amino acid chain. Amino acid chain is translated from the codon sequence of mRNA molecule. Codon sequence of mRNA is transcribed from the exons of a gene.

Exons contain the protein-coding sequence, divided into base triplets, *codons*, which correspond to specific amino acids (**Figure 5**). The protein synthesis, in short, goes as follows: the base pair sequence of exons are read (*transcribed*) by the transcription machinery, which forms the messenger RNA (mRNA) molecule. The mRNA is transferred outside of the nucleus, where the codon sequence of mRNA is *translated* into an amino acid chain, which is then able to fold into a three dimensional, functional protein. Cancer-driving mutations occur often inside exons, as they have the potential to directly change the protein sequence and break the homeostasis of a cell [56]. These and other mutations are discussed further in the "Genetic alterations" chapter.

Introns are noncoding sequences between exons, which are spliced out of the mRNA during and after transcription. However, despite this exclusion, introns have a multitude of functions in the process of mediating gene expression [57]. The most prominent and well-known feature of introns is the enabling of *alternative splicing*, which is a mechanism to produce different exon combinations, *isoforms* from a single gene, thus expanding the protein diversity of an organism. Introns and their splicing have also been measured to affect the initial transcription, pre-mRNA modification, nuclear export, and even translation of a gene [58]. Mutations in introns, especially in proximity to exons (splice sites), are known to hamper splicing and change the normal function of the protein product in tumor genomes [59].

UTRs are end sections of mRNA, which do not code amino acids, but are involved in various gene regulatory processes. Genes are transcribed in the 5' (5-prime) to 3' (3-prime) direction (**Figure 5**), so UTRs are referred to as 5' and 3' UTR depending on their location in the mRNA. MicroRNAs (miRNAs) are short (~20 bp) sequences, which predominantly bind to the 3' UTRs and repress the protein synthesis of the target gene [60]. This is the best-known regulatory function of UTRs, which have also been reported to be damaged in some cancers [61]. For instance, a point mutation in the 3' UTR can break the binding site of a miRNA and thus prevent the repression of the otherwise repressed gene [62].

In the context of cancer genetics, genes can be classified as either *tumor suppressors* or *proto-oncogenes*. As discussed in the hallmarks of cancer, one of the critical features of tumorigenesis is sustained growth stimulation. In normal conditions, proteins coded by proto-oncogenes participate in the regulation of cell growth and differentiation or prevention of apoptosis. Proto-oncogenes are silenced or suppressed when not needed, for example, by the binding of miRNAs or by a suitable DNA conformation, as discussed above [63]. Proteins coded by tumor suppressor genes, on the other hand, work as repressors of cell growth and may promote apoptosis or both. DNA repair genes are also classified as tumor suppressors. There are distinct ways in which these two types of "cancer genes" are damaged by gain or loss of function mutations in favor of tumorigenesis. The characteristics of proto-oncogene versus tumor suppressor mutations are further discussed in the "Genetic alterations" chapter.

Genes, or the proteins that they encode, can belong to a specific family or be part of biological pathways or protein complexes. Gene family is a term referring to a group of genes with a similar function and DNA sequence. Genes in a family have a common ancestor gene, which has been duplicated and altered by mutations during evolution [64]. In cancer genetics, for instance, genes in Ras and Raf proto-oncogene families, have been widely studied and are among the most mutated genes in tumors, colorectal in particular [65, 66]. The name of a gene does not necessarily reveal which family the gene belongs to; for example, the *BRCA1* tumor suppressor gene, which was discussed earlier, does not belong to the same gene family as the *BRCA2* gene, although they operate in the same pathway and have similar functions in the maintenance of genome integrity [67]. Neither does the gene name always relate to the protein function, as is the case with for example *BRCA1*. The name often merely corresponds to a disease or organism that the gene was found or studied in [68].

Table 1: The most relevant genes in this thesis.

Proto-oncogenes	**Function**
MYC	Encodes a protein (transcription factor) that can activate multiple pro-proliferative genes. Overexpressed in multiple cancers.
KRAS	Controls cell proliferation. Pathogenic mutations cause sustained proliferative signaling in a cell.
BRAF	Controls cell growth. Activating mutations result in excessive cell growth. Often mutually exclusively mutated with Ras family genes.
Tumor suppressors / DNA repair genes	
TP53	"The guardian of the genome". Has multiple essential functions in prevention of tumorigenesis. Highly mutated in various cancers.
APC	The most commonly mutated gene in colorectal cancer (~80% of cases).
POLE	Involved in DNA repair and replication. A single point mutation can cause an ultra-mutator phenotype.
MLH1, MSH2, MSH6, PMS2, EPCAM	Mismatch repair genes. Germline mutation can predispose to Lynch syndrome. Causes microsatellite instability (MSI) when defective.
Other	
CTCF	Protein commonly associated with insulators and TAD borders. Binds cohesin complex to DNA.
RAD21	Part of the cohesin complex. Used as a measurement marker of cohesin in the publication II of this thesis.

5.4.2.2 Regulatory and the noncoding genome

The majority of, a typical bacterial genome is composed of protein-coding regions while, in contrast, around 99% of the human genome is noncoding [69]. The noncoding genome contains regions that determine when, where, and how actively every gene in the genome is expressed in a particular cell or tissue type at given conditions (**Figure 6a**). These regulatory regions can be roughly classified as *promoters*, *enhancers*, and *insulators*, which together account for ~10-20% of the whole human genome sequence (**Figure 4**) [70]. Human DNA also contains hundreds of noncoding RNAs (e.g., miRNAs), which do not encode proteins but are involved in gene regulation by binding to the UTRs of freshly transcribed mRNAs, for example [71]. Regulatory regions contain DNA sequences which are recognized and bound by dozens or hundreds of transcription factors (TFs). The occupation of TFs can affect gene regulation indirectly, by granting or denying a particular transcription machinery access, or directly, by changing DNA conformation, thus enabling or preventing transcription [49, 72].

Promoters are located in the proximity (within 1000 bp) of the transcription start sites (TSSs) of genes (**Figure 6b**). They provide the foundation for the binding of TFs, assembly of the transcription machinery and, subsequently, the initiation of transcription [73]. A gene can have multiple promoter regions, which are activated differently based on, e.g., the cell type. Hence, both alternative splicing and the usage of different promoters can determine the expressed isoforms or transcripts of a gene. Genes that are part of complex and cell-type-specific mechanisms such as tissue renewal or DNA repair are generally activated through an interplay between their promoters and distal enhancer element(s). In contrast, some promoters, such as those responsible for the transcription of housekeeping genes or other continually expressed genes, can contain an integrated enhancer or in some cases not require any external factors whatsoever to be activated [25]. In cancer, the best-known and most frequently mutated regulatory hotspots are located at the promoter of the *TERT* gene (**Table 1**). The mutations generate novel binding sites for TFs, which elevate the expression of *TERT*, and through complex mechanisms, promote tumorigenesis [74, 75]. Another example of a pathogenic promoter defect is hypermethylation of the *MLH1* mismatch repair gene promoter, which leads to an excessive accumulation of specific mutations (**Table 1**) [76].

Methylation of DNA is a chemical, genome-wide process, which can epigenetically change the activity of regulatory regions [77]. Typically, methylation of a promoter has a silencing effect, like in the example above, where *MLH1* is silenced. Methylation generally occurs in the CpG sequence context (cytosine is followed by guanine). It changes the physical properties of DNA but not the sequence itself, and affects for instance TF binding of all three classes of regulatory regions [77, 78]. Promoters and proximal regions of

genes commonly contain CpG rich DNA stretches - CpG islands - which are differently methylated depending on the cell type. CpGs and other sequence contexts are further discussed in the "Mutational signatures" chapter.

Enhancers share common structural and functional features with promoters [79]. However, they regulate the expression of their target gene(s) from a longer distance than promoters. In fact, enhancers typically actualize their function by interacting physically with the promoter site of a target gene by DNA conformation changes or looping (**Figure 6b**) [80, 81]. In the human genome, the majority of enhancers are located within a 100 kbp distance (~15 kbp median) from the promoters of their target genes, however, in some cases enhancers have been detected to regulate genes located on a different chromosome even [82, 83]. The open, or accessible, enhancer DNA sequences are recognized and bound by a large group of collaborating TFs and *mediators*, which determine the expression levels of the target gene(s). At the same time, enhancers themselves can form large collaborating groups, *super-enhancers*, which have strong effects on gene regulation and have been associated with genes involved in cell differentiation [84]. In various cancers, super-enhancers have been measured to be enriched, especially at the chromosomal loci of proto-oncogenes, such as *MYC* (**Table 1**) [85, 86]. Also, at the same locus, a single SNP in an enhancer element has been reported to increase CRC risk ~1.5 fold, when present in both inherited chromosome copies of an individual (*homozygosity*) [87].

Insulators function as genome organizers that enable or disable putative enhancer-promoter interplay, i.e. initiation of gene expression. The key players in chromatin looping are the cohesin complex, which holds two separate DNA segments together, and CTCF, which physically binds the cohesin to DNA (**Table 1**) [88, 89]. In addition to insulation, cohesin binding sites have been associated with various other essential genomic functions, such as DNA repair and maintenance of epigenetic homeostasis. Also, the boundaries between active and silent chromatin domains, or topologically associating domains (TADs), are bound by these ancient and highly conserved proteins of the cohesin complex (**Figure 6a**) [90, 91].

TADs are varied sized (tens of kbps up to 2 Mbp) regions in chromosomes, commonly spanning multiple genes and regulatory regions. The chromatin of these domains is either open or closed, which contributes to the expression of all the genes within. The exact mechanisms of how TADs are formed and contribute to gene regulation are still unclear [92, 93]. However, both insulators and TADs manifest their regulatory functions through DNA conformation changes by looping, which is carried out by the cohesin complex and often with CTCF [92, 94, 95]. In tumor genetics, aberrant CTCF binding due to hypermethylation (as in the MSI case) was detected in a subset of gliomas [96]. Methylation-sensitive CTCF binding was shown to break the TAD

boundary by the hypermethylation of a specific CBS, and as a result, disrupt the gene insulation function at the known glioma oncogene, *PDGFRA*. In publication II of this thesis, we reported an accumulation of mutations at CBSs in multiple cancers [23]. In addition to gene regulation, TADs have been associated with regulation of replication timing, that is, when different regions of the genome are replicated during cell division [93]. In tumor genomes, replication timing has been detected to correlate strongly with the regional mutation frequencies and the forming of mutational landscapes across the genome. This phenomenon is further discussed in the "Somatic mutations" chapter.

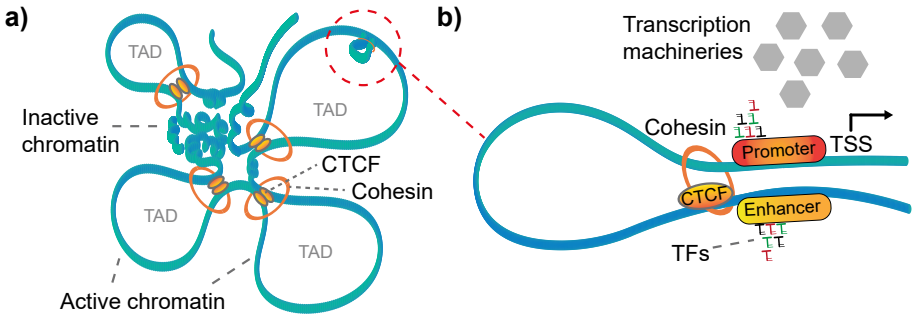

Figure 6: Regulatory regions. **(a)** CTCF and Cohesin work as TAD boundaries. **(b)** CTCF and Cohesin work as an insulator and loops enhancer to the target promoter.

5.4.3 Genetic alterations (mutations and variation)

The exact meaning of the terms "mutation", "variation", "variant", and "polymorphism" varies depending on context [97]. In this thesis, the following definitions are used: **mutation** is a DNA alteration, which affects a single individual or cell, and has been acquired spontaneously during one's lifetime. Mutations can be divided into germline and somatic, where the former occurs in germ cells and can be transferred to the next generation. Somatic mutations accumulate in all other (somatic) cells. As they are only passed on to the daughter cells of the mutated cell which by definition is somatic, they can not be inherited. Despite the negative connotation of the term, mutations can be completely harmless or even beneficial. Hence, the usage of the term "mutation" is usually avoided especially in medical context [97]. **Variation** is a population level term, which describes genetic differences between individuals, populations, and organisms. In bioinformatics context, a **variant** is used to describe both mutation and variation, and generally means any measurable aberration or substitution in DNA. In population-level context, a variant is a single unit of variation, and it can be either common, rare, or very rare. **Polymorphism** is a common variant, which is present in over 1% of individuals in a specific population. Rare and very rare variants are present in less than 1% and 0.1% of the population, respectively.

5.4.3.1 Mutation types and effects

The type, size, and location of a mutation determine its effect on genomic functions. Point mutations are single nucleotide variants (SNVs), where a base has been altered to another (e.g., T > C; **Figure 7**). Also 1 bp insertions and deletions (indels) are considered as point mutations. Larger events, from ~1 kbp up to chromosomal level, are considered structural variants. These include duplications, inversions, translocations, and large insertions and deletions (**Figure 8**).

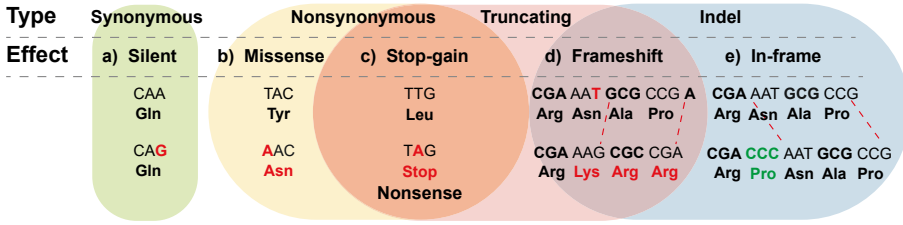

Figure 7. The effects of point mutations on a protein. **(a)** Silent mutation changes the base but not the amino acid. Glutamine (Gln) is encoded by CAA and CAG codons. **(b)** Missense mutation changes both the base and the amino acid. **(c)** Base substitution causes premature Stop codon (i.e., nonsense mutation), which is encoded by TAG, TAA, and TGA. **(d)** The deletion of T base causes following codons to change the reading frame. **(e)** The insertion of CCC shifts following codons, but does not change the reading frame.

Point mutations and short indels (e.g., 10 bp) can directly affect the protein product of a gene by altering the protein-coding sequence or by breaking sequences (intronic/exonic) regulating splicing. Coding SNVs can be either *synonymous* and *nonsynonymous*, where the former changes the codon triplet but not the amino acid, and the latter changes both (**Figure 7a, b & c**). Nonsynonymous mutations can be *missense* or *nonsense*, where the former changes the amino acid to another and the latter changes the amino acid to a premature stop codon. A nonsense mutation can prevent translation altogether or truncate translation prematurely, which may lead to a damaged or destroyed protein. Point mutations in splice sites (often located a few bases from the exon boundary) can lead to exon skipping during RNA splicing. Coding indels have the same effects as SNVs and can additionally shift the reading frame of the whole codon sequence if the length of the inserted or deleted sequence is not divisible by three (**Figure 7d & e**). Frameshifts lead to an aberrant amino acid sequence [98]. In the context of tumor suppressors and proto-oncogenes, mutations are classified as either loss or gain of function. Loss-of-function mutations are typically *truncating* (nonsense and frameshift) and break the protein products of tumor suppressor genes (**Figure 7c & d**). Gain-of-function mutations are often missense type mutations that hit specific domain(s) of proto-oncogenes [99].

Structural variants (SVs) have an effect on a larger portion of the chromosome from the length of hundreds of bps to the whole chromosome arm (**Figure 8**). A single deletion or duplication can affect the expression of one or multiple genes by spanning regulatory regions or the genes themselves [100]. For instance, the proto-oncogene *MYC* is activated by amplification of its enhancer region, as discussed earlier [85, 86]. While duplications usually increase and deletions decrease the expression of affected genes, the consequences can be the opposite [101]. In cancer, the other copy (allele) of tumor suppressors such as *TP53* is often lost by a large deletion accompanied by a point mutation in the remaining allele (**Table 1**) [102].

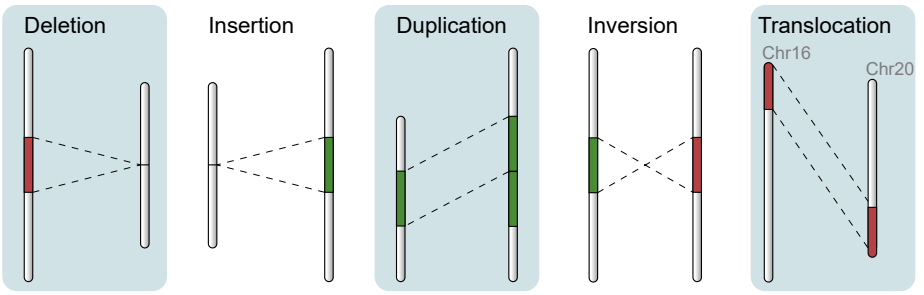

Figure 8: Structural variants. Schematic of the most common types of structural variants occurring in and between chromosomes.

The deletion causes loss of heterozygosity (LOH) at the germline variant locus, which is one mechanism to actualize the pathogenic potential of predisposing variants [103]. Insertions, inversions, and translocations can break regulatory regions and genes by having breakpoints at specific loci. For instance, an inversion or translocation can transfer an active enhancer element to the proximity of an otherwise silenced gene, and thus ignite its expression [104]. This mechanism is observed, for instance, in myometrium tumors (myomas), where a translocation between genes *HMGA2* and *RAD51b* has been detected [105].

5.4.3.2 Somatic mutations

The genomes of normal and cancerous cells harbor mutations (point mutations, short indels, and structural variants), which have accumulated during the lifetime of the individual. These somatic mutations are transferred to daughter cells during cell divisions, but are not inherited by the children of the individual. In cancer and tumor cells, the vast majority of somatic mutations have not been selected for during cancer evolution, but are merely harmless passengers, which have no or minimal effect on cell viability [106]. However, some mutations have been beneficial to cell growth, and have hence been retained in the tumor cell lineage. These growth-promoting mutations, or drivers, take part in tumorigenesis, as was described in the hallmarks of "Cancer as a disease" chapter.

Somatic mutations can occur due to internal (endogenous) or external (exogenous) factors. Exogenous factors such as radiation and tobacco smoke are known to be mutagenic in the cells of exposed tissue. Endogenous factors, such as DNA replication errors during cell division, have the most significant effect on tissues with high cell division rates, e.g., the epithelium. In an average adult human body, cell divisions account for over a light-year distance worth of DNA replication, requiring viable repair mechanisms to avoid accumulation of somatic mutations [107]. Dysfunctional repair mechanisms cause the affected cells to take on a *mutator phenotype*. Such cells have a higher than usual genomic mutation frequency. The most striking mutator is a damaged exonuclease domain in the polymerase epsilon gene (*POLE*), which may lead to a mutation load which is over a hundredfold that of an average CRC cancer cell (**Table 1**) [108]. In CRC, POLE mutants constitute ~1-2% of all cases. The more common mutator phenotype is MSI, which is characterized by small indels at short repeated sequences (microsatellites). The mutation load in MSI can be tenfold compared to the average CRC genome [108].

Somatic mutation frequencies vary between different regions of the genome (**Figure 9**). Generally, more active and accessible regions have fewer mutations than inactive due to factors such as earlier replication timing,

transcription-coupled repair, and differences in sequence context [109, 110]. As was brought up earlier, replication timing affects the mutation frequency so that later replicated regions have an increased mutation load compared to regions replicated in early S-phase (the DNA replication phase of the cell cycle) [111, 112]. This may be due to less effective mismatch repair and depleted nucleotide pools in the late S-phase [113, 114]. The mutational landscape of different cell and cancer types also reflects the underlying mutational mechanisms, which often prefer distinct sequence contexts. These characteristic mutational patterns are called signatures [115, 116].

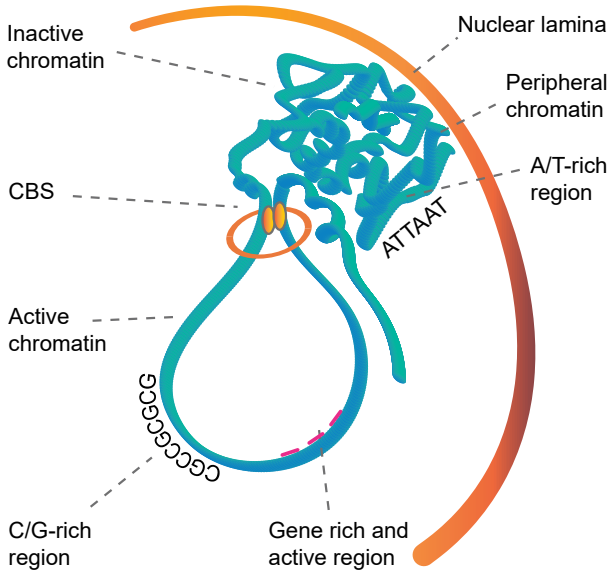

Figure 9: Genomic features forming mutational landscapes. Different parts of the genome are prone to distinct mutations and mutation frequencies. Closed, inactive, and peripheral chromatin are replicated later (more mutations) than open and active regions (fewer mutations). A/T and C/G rich sequence contexts can affect both mutation types and frequencies. CBSs accumulate mutations under specific mutational signatures.

5.4.3.3 Mutational signatures

Mutation processes and agents, such as mismatch repair deficiency, replication errors, and radiation, generate distinct mutational patterns - *signatures*. For instance, a common CRC tumor genome contains 10-20,000 somatic mutations, of which only a handful are selected for during tumor evolution. The rest, the passengers, have not been under selective pressure, and can thus be used as a historical footprint of the mutational processes that have been operative in a cell lineage from the embryo to the full-grown tumor [117, 118]. SNVs can be classified as transversions (e.g., C > A) and transitions (e.g., C > T) yielding six distinct mutation types C > A, C > G, C > T, T > A, T > C and T > G, where C and T also represent G and A on the opposing strand (i.e., C > A equals C:G > A:T). The simplest way to extract mutation patterns would be calculating the frequency of these six mutation types in a given tumor. While mutation type counts alone can be used as a rough projection of underlying processes, mutations have been discovered to occur in specific sequence contexts, which more accurately reflect processes operative in the nucleus [116]. For example, tobacco smoke has been shown to generate an excess of C > A transversions. Oxidation during DNA sample preparation has been shown to cause the same, in this case artefactual C > A mutations [117]. Separating these two phenomena in downstream analyses is impossible if only mutation types are considered. However, the mutation contexts of these two mutagenic agents are different. Tobacco smoke-induced mutations occur predominantly in the ApCpG and GpCpG context, whereas artefactual oxidation most frequently mutates CpCpG triplets [117–119]. Mutations can be classified according to adjacent bases (e.g., Cp[T>A]pG). This classification system results in 96 distinct mutation types. To this day, over sixty distinct signatures have been extracted from multiple cancer genomes by utilizing this sequence triplet context in signature detection [118, 120, 121].

The extraction of signatures from NGS data, as was done in Alexandrov et al. 2013, was performed using non-negative matrix factorization, which is a method developed to detect “hidden” features or associations from data matrices [118]. In this case, rows in the original matrix represent all 96 mutation types, and columns are individual samples. Each cell of the matrix thus holds the count of a given mutation type in a given sample. The challenge is to detect which mutations are the result of the same mutational process. Most cancer classes have multiple mutational processes active in a single tumor, and each process manifests mutations at varying magnitudes (i.e., exposures), further complicating analysis. Signature extraction results in two separate matrices, the product of which should match the original matrix as closely as possible. One of the matrices holds the extracted signatures and weights of all the mutation types in a particular signature. The other matrix holds signature exposure values for each sample, i.e., information on how strongly a given signature is present in the sample [116].

Different cancer types harbor distinct and shared combinations of mutational signatures. In the scope of this thesis, both CRC and ESCC exhibit at least signatures 1, 6, and 17, as classified in Alexandrov et al. 2013 (**Figure 10**) [121]. **Signature 1** has been measured in the majority of cancer classes, as well as in normal cells, and it has been shown to correlate with the age at diagnosis [117–119, 122]. This signature is characterized by an excess of C > T transitions, and is probably related to the spontaneous deamination process of methylated cytosines in the DNA, especially in the NpCpG context (**Figure 10a**). This process is related to the methylation of CpG islands discussed in the "Promoters" paragraph. However, CpG islands are found throughout the genome and their methylation is a very frequent (majority of CpGs are methylated in human cells) and genome-wide epigenetic phenomenon [123]. Signature 1 is an example of an endogenous process, which causes mutation accumulation in cells during an individual's lifetime. However, recent findings suggest that the mutation accumulation slows down as a consequence of a decreased division rate as humans age [124]. **Signature 6** is caused by a deficiency in the mismatch repair machinery, which leads to an excess of indels in microsatellites (i.e., MSI). However, signature 6 can be extracted using only SNVs, despite it having a similar mutation spectrum as signature 1 (**Figure 10b**). **Signature 17** is characterized by an excess of T > G and T > C mutations, predominantly in the CpTpT context, the source of which is unknown (**Figure 10c**). These mutations were shown to accumulate particularly at the CBSs in publication II of this thesis. **Signature 10**, caused by a damaging mutation in *POLE*, has been measured to generate mutation frequencies that are a hundredfold higher than the frequency of spontaneous mutations in CRC and other cancers (**Figure 10d**). The mutations are almost exclusively C > T substitutions in TpCpG and C > A substitutions in TpCpT context. Signature 10 was discovered to display an inverse pattern at CBSs in Publication II. Genome-wide mutation signature analyses have been made possible by next-generation sequencing technologies, which are described in the next chapter.

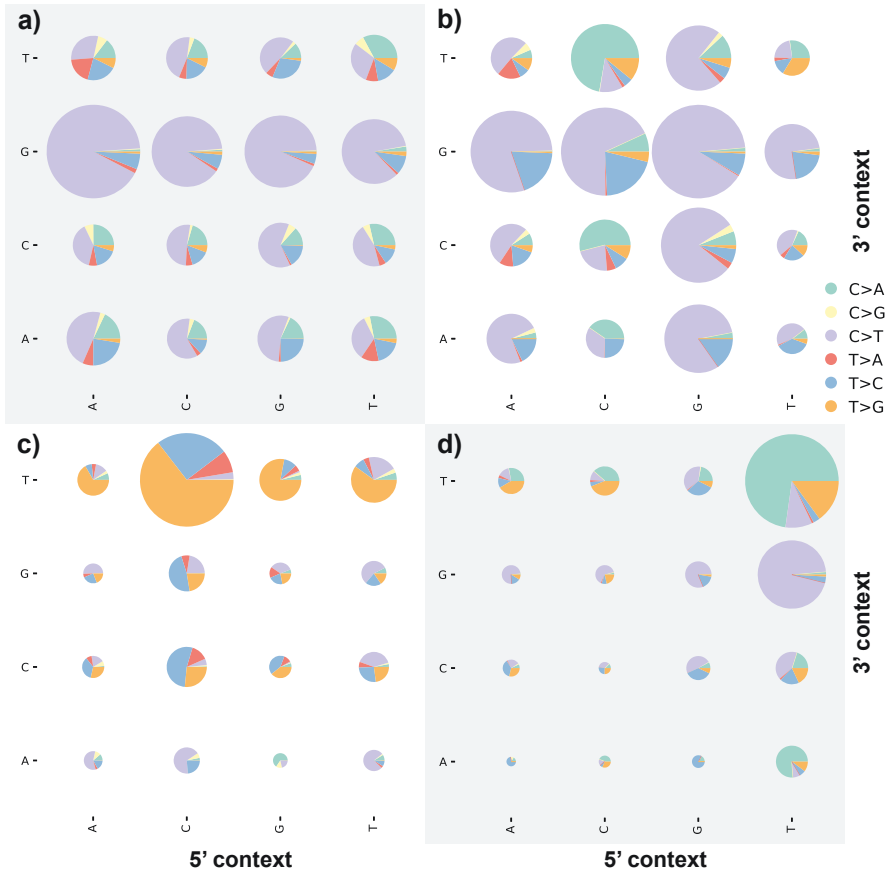

Figure 10: Mutational signatures and contexts. **(a) Signature 1** exhibiting predominantly C > T mutations in NpCpG context. **(b) Signature 6** (MSI) is characterized by indels at microsatellites but also by the excess of C > T mutations in various contexts with adjacent G, and C > A mutations in CpCpT context. **(c) Signature 17** is characterized by the excess of T > G and T > C mutations in NpTpT context. **(d) Signature 10** (POLE mutant) exhibit an excess of C > A and C > T mutations in TpCpT and TpCpG contexts, respectively.

5.5 The next-generation of cancer genetics

Modern cancer genetics analyses heavily resort to computing power and inventive algorithms that can manage vast amounts of data and process it for various research purposes. Indeed, computer-assisted biological data analysis, bioinformatics, has become instrumental in today's genetic research. The successful effort in mapping the majority of the whole human DNA sequence provided the foundation for the next-generation sequencing techniques to arise and thus caused a paradigm shift in genetics research.

5.5.1 Human reference genome

The map of the whole human DNA sequence, the human reference genome, forms an essential foundation for modern medical genetics research. The first draft of the complete human reference genome was published in 2001 by the Human Genome Project, which was formally launched in 1990 [125]. The result is a freely available, 3 gigabyte, or 3 billion characters long text file consisting of only the letters A, C, T, and G (the letter N is used to represent e.g., a gap or unannotated region in the sequence) in a specific order. More specifically, the human reference genome is a compilation of DNA sequences, divided into chromosomes (1-22, X, Y, and mitochondrial), and alternative or unlocalized sequences, often stored as a FASTA file. FASTA is a standard format for storing biological sequence data. The human reference genome was initially constructed using DNA from several donors, but the sequence has been subsequently updated to more precisely match the general population. Even though the reference genome does not represent the human population as such, but rather a small number of donors, the sequence is still to most part shared between all individuals.

The first official human reference genome (build 34) was published in 2003 by the Genome Reference Consortium (GRC), and since then, it has gone through four major updates till the current build 38 (GRCh38) [126]. There are smaller update patches between the major ones, these address minor issues without changing the genome size and thus the overall topology. Changes between genome builds, however, have a more significant impact on the reference through larger, structural updates that break the compatibility with earlier builds [126, 127]. The reference genome is used as a model when individual genomes are sequenced and mapped; if the model is different between studied individuals, the comparison of the sequencing data becomes highly cumbersome. Hence, the usage of a common genome build is essential in genomic analysis, when the data from multiple sources are integrated into the same study.

The raw reference sequence data provides no information about different genomic regions. Genes, regulatory regions, variation, and other information is added on top of the reference sequence by genome annotations, which enable the research on disease-causing mutations, for instance.

5.5.1.1 Genome annotation

The human reference genome sequence contains specific patterns that encode thousands of genes, regulatory regions, and structural features. However, the locations of these patterns are not encoded to the raw reference sequence itself, consequently requiring layers of data, genome annotation, on top of the reference genome. At the very simplest, a genome annotation contains a chromosomal position for a single base feature in the genome, or start and end positions to define a particular region. For instance, a gene annotation contains start and end positions for all exons and coding regions in the reference genome, which can be utilized in the detection of protein-altering mutations. Regulatory annotation could, for instance, hold all binding positions for a specific transcription factor.

Thus far, over 20,000 genes have been mapped to the human reference by the combination of automatic and manual methods [128]. The mapping is still ongoing, and new gene annotation versions are published with an interval of several months, mostly adding new isoforms and noncoding RNAs to the annotation. For instance, the comprehensive gene annotation by the GENCODE project, initiated in 2003, can be obtained from the Ensembl and UCSC genome browsers [129, 130]. The mapping of noncoding elements was revolutionized by methods such as ChIP-seq and DNase-seq, which enabled e.g., the detection of functionally active regions and sites of TF binding. In 2012, the ENCODE project released comprehensive regulatory genome annotation sets, which are available on their website (https://www.encodeproject.org).

5.5.2 Next-generation sequencing

DNA sequencing is a procedure, where the bases of the DNA are read from the molecule and translated into a human-readable form. Before modern sequencing methods were introduced in 2005, DNA sequencing was commonly performed with the Sanger sequencing method, which is able to sequence a few hundred bases (~0.00001% of the human genome) per single run. The Sanger method is still used for various purposes, such as validation or screening of specific mutations, but the usage is not feasible for large-scale genomic studies [131]. The ultimate goal in the development of sequencing technologies is to produce an entirely accurate representation of a given DNA molecule, for instance, the whole chromosome. The ideal sequencer would be able to take DNA from a single cell as an input and then produce a FASTA file from it (i.e., reference genome of an individual cell). However, there are various obstacles in reading accurately long strands of DNA, necessitating the usage of DNA amplification, fragmentation, and commonly (excluding single-cell sequencing) bulk DNA from thousands or millions of cells.

It is currently impossible to sequence whole chromosomes from start to end. However, it is possible to shear the DNA into smaller fragments (e.g., 500 bp long) and sequence ~100-250 bp from both ends (**Figure 11**) [130, 132]. Next-generation sequencing, also called massively parallel sequencing, is a high-throughput technique, where bulk DNA from thousands of cells is fragmented and read simultaneously, resulting in millions of short sequence reads. The raw read data contains genetic information from the donor sample in small, unorganized pieces, which as such are unusable in genetic analyses. Thus, the next challenge is to organize the data so that it can be used to call genetic alterations of a given individual and make it compatible with the genome annotation and other studied samples.

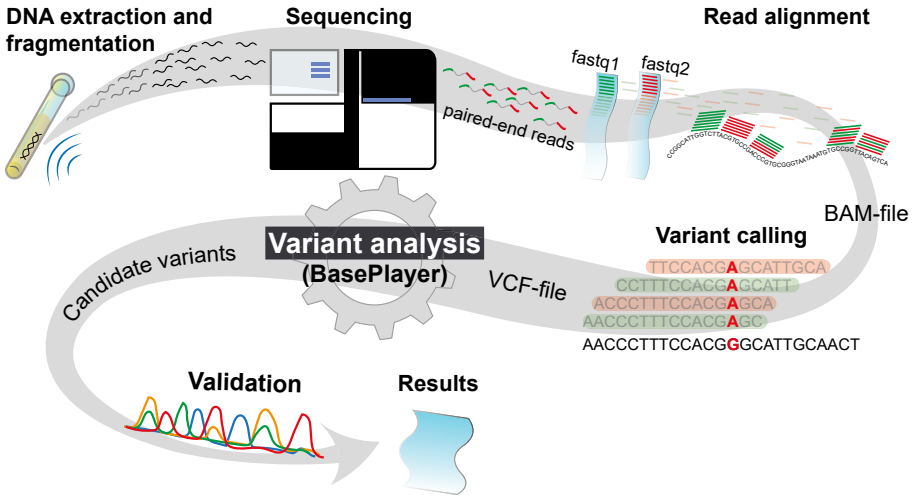

Figure 11: NGS data processing and analysis workflow in cancer genetics research.

The raw data is organized by the read alignment algorithm, which searches positions for all reads by using the reference genome as a model (**Figure 11**). Read alignment produces a sequence alignment map (SAM/BAM) file, sorted by chromosomal position, which enables the pileup of short reads "on top" of the reference sequence. The pileup is used to construct a consensus sequence of the donor genome, which is then used to call differences between the individual and reference genome i.e., the variants. The challenge in read alignment is to find the correct locations in the reference genome for millions of reads in a reasonable time. In principle, the alignment algorithm compares short read sequences with the reference and reports a position where the read sequence is present (e.g., finding "ign" from "alignment" would output position 3). However, the read sequence does not necessarily match the reference precisely due to genetic alterations and sequencing errors (or sequencing artefacts), which complicates the alignment.

Sequencing errors also complicate *variant calling*, which is the phase where the alignment file is scanned for differences between the reads and the reference, and genomic positions of variants are reported (**Figure 11**). Sequencing errors are the reason why multiple overlapping sequences from the same location are required in variant calling. By comparing multiple sequences from the same location, it is possible to create a consensus sequence, in which random mistakes in the sequencing have been ignored. The amount of overlapping sequences is called *coverage* - the higher the coverage at a particular location, the higher the confidence of the consensus sequence and the called variants. Each overlapping read represents an individual cell and its DNA at a given locus. Also, as a reminder, a single cell contains a copy of the DNA from both parents, and thus it produces two distinct consensus sequences. In the following chapters, read alignment and variant calling procedures are described in more detail.

5.5.2.1 Read alignment for the next-generation sequencing data

The most popular NGS technologies, which are also applied in all projects of this thesis, are provided by Illumina [131]. Sequencer instruments, such as the NextSeq, HiSeq, and NovaSeq, produce millions of short reads per sample. A single DNA fragment can be sequenced from both ends, resulting in two 100 bp paired-end read sequences, which are printed to separate FASTQ files (**Figure 11**). FASTQ is a standard file format for storing a sequence and corresponding base quality data. Read aligners, such as BWA and Bowtie, use these files as input and produce a single file (SAM/BAM) that includes chromosomal positions for each read in the reference genome [133, 134]. The compressed, most commonly used version of the SAM file is called BAM and used hereafter when referring to read alignment files. BAM files contain the positional information of reads and their pairs, and also FASTQC derived base quality scores. The information can be utilized in subsequent analysis phases, such as variant calling and read data visualization.

After the introduction of NGS, aligning the produced short read data formed a bioinformatic challenge - how to make the alignment for millions of reads in a reasonable time [135]. Various approaches to tackle the alignment problem were introduced, commonly resorting to read or genome sequence hashing techniques, which were soon discovered to be too slow and memory inefficient in practice. Moreover, one of the major challenges in read alignment is to allow varied number of mismatches (potential variants) and gaps/insertions (potential indels) in search of the best matching genomic location for any given read. The hashing based methods were not lenient enough to overcome this challenge. All these alignment challenges were solved with fast string search operations, which were enabled by the properties of *suffix trees*. In computer science, tree data structures are often used in algorithms that require fast searching. A suffix tree is a data structure, which contains all possible suffixes (end parts) for a word (e.g., CCATTG, CATTG, ATTG, TTG, TG, G). The tree is constructed so that matching of the string against the suffix tree can be done in log(n) time. For the whole genome, the suffix tree structure without compressing would consume too much memory for modern computers to store (in the order of n^2, where n is genome length). The challenge was to construct a suffix tree for the whole human genome sequence to enable highly efficient sequence matching and at the same time compress the structure to be small enough to adhere to memory restrictions. The solution was introduced in 1994 with the Burrows-Wheeler transformation, which is used to compress the suffix tree to an even smaller size than the original genome sequence and to enable fast suffix tree searches [136]. The most commonly used read aligners, such as BWA (Burrows-Wheeler Aligner), are based on this method. For example, BWA can align roughly a billion reads (whole-genome data) against the human reference genome in a single day in a modern server environment.

5.5.2.2 Variant calling

After read alignment, the variants of the donor individual can be called and printed to a variant call format (VCF) file. The variant caller reads through the alignment result file (e.g., BAM) and finds occurrences, where overlapping reads have the same mismatch or indel in the same position relative to the reference sequence (**Figure 12**). A mismatch in a read is either an artefact that has emerged during sample preparation or processing phases, or a real variant/mutation that is present in the DNA of the individual or cell. The challenge is to separate true variants from artefacts using the information from reads that overlap the putative variant position. In general, the more overlapping reads at a particular location, the more reliably a specific variant can be called. In practice, the quality of the variant call is assessed using a combination of determinants, which can be used to filter out false-positive calls during variant analysis.

Germline variant callers, such as HaplotypeCaller used in the Genome Analysis Toolkit (GATK) best practices pipeline, report multiple quality-related values for the variant call [137]. Determinants, including base quality scores, coverage, allelic fraction, strand bias, and sequence context, are considered in the quality assessment. The depth of coverage (DP) and allelic depth (AD) denote the number of all overlapping reads and the number of reads, which call the reference and alternative allele at the variant locus. The AD ratio, or allelic fraction, can be used to determine the genotype of the variant; the ratio of one and less than one would denote a homozygous and heterozygous variant, respectively (**Figure 12**). However, due to imperfections in NGS data preparation, such as sequencing errors and

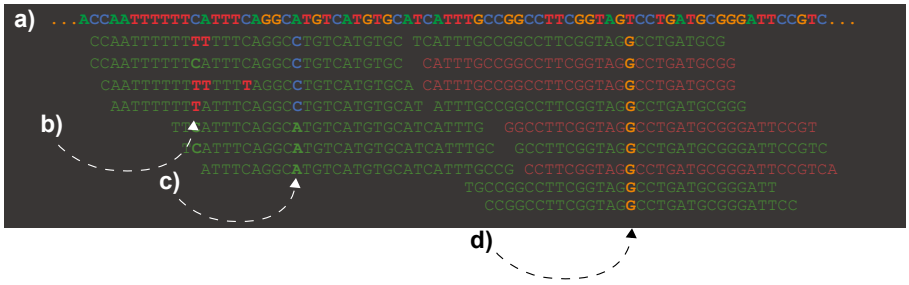

Figure 12: Variant calling. **(a)** Reference sequence. Green and red tinted sequence fragments (30 bp) are aligned reads in forward and reverse strand, respectively. Mismatched bases in reads are bolded. **(b)** Possible sequencing error due to adjacent mononucleotide repeat. **(c)** Putative heterozygous variant, where reference base is A and alternative C. Allelic fraction of the alternative base is 4/7 (0.57) and the read coverage at this locus is 7. Mismatched bases are present only in forward strand reads, which lowers the confidence of the variant call. However, there are no reverse strand reads at this locus, so the confidence still remains relatively high. **(d)** Putative homozygous variant. Allelic fraction for the alternative base is 1 and read coverage is 9. Mismatches are present in both read strands, which strengthens the confidence of the call.

duplications, the HaplotypeCaller uses Bayesian statistics to determine the most likely genotype for a given variant locus and reports the genotype quality (GQ). The number of false-positives can be reduced by applying variant calling to multiple samples simultaneously. This joint calling method has been measured to increase the accuracy of variant calling significantly [138, 139]. Somatic variant calling typically utilizes two BAM files as input - one from a normal/healthy and the other from a tumor tissue sample. Basically, somatic variant callers, such as Mutect2 [140], compare putative variant loci found in the tumor BAM against the corresponding normal BAM, and output variant calls only present in the tumor. Similarly to germline variant calling, an additional set of normal samples, or a pool of normals, can be utilized to reduce the number of false-positive calls, especially in error-prone loci.

5.5.2.3 Exome vs whole genome sequencing

Exome sequencing (WES) is an NGS technique, in which only transcribed or protein-coding regions are sequenced by exome capturing protocols (**Figure 13**). Although WES involves more sample preparation steps (capturing phase) than WGS, overall costs are substantially lower. Not only is the data easier to store but also faster to analyze due to smaller file sizes. Naturally, WES does not allow analysis of the noncoding regulatory genome, for instance. Still, it is an attractive option for most disease genetics research groups as ~85% of causative mutations with large effects have been estimated to lie in coding regions [56].

Table 2: Comparison of the targeted and non-targeted sequencing. Pros and cons of WES and WGS. The choice between these two depends on the research question and available resources.

Whole-genome sequencing	Exome sequencing
Reads cover the coding and noncoding genome	Reads cover only targeted (coding) sites
Large file sizes	Smaller file sizes
Expensive (to produce and to handle)	Relatively cheap
Comprehensive structural variant analysis	Limited SV analysis
Comprehensive signature analysis	Limited signature analysis
Mutation studies on the coding and noncoding genome	Mutation studies almost exclusively on the coding genome

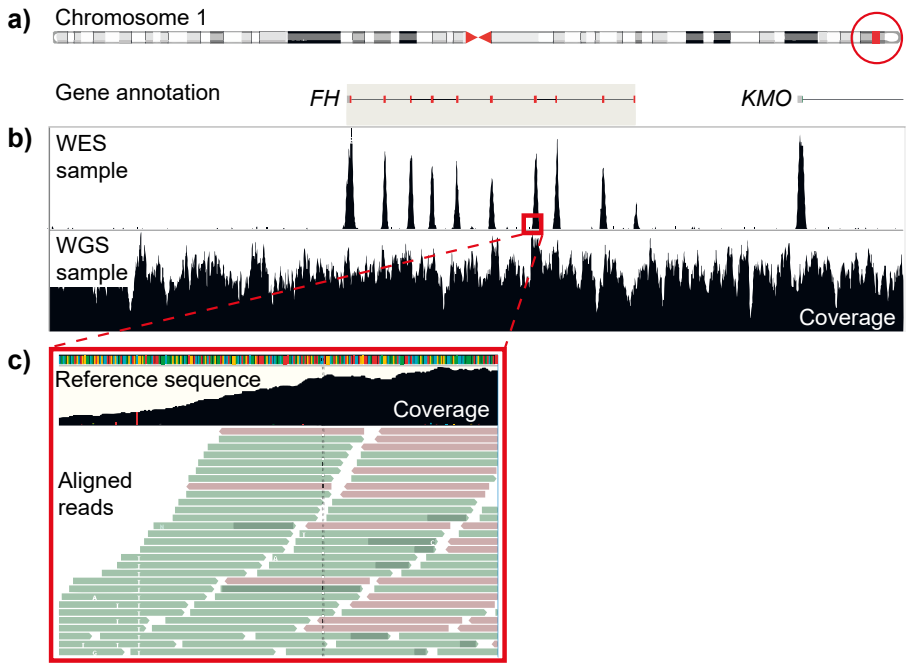

Figure 13: WES and WGS data in a gene region. Visualization (BasePlayer) of real sequencing data at a gene locus. **(a)** Chromosome 1 and the locus of the *FH* gene (red circle). Red sections are exons of *FH*. The gray section is UTR of the *KMO* gene. **(b)** Coverages of WES and WGS samples at the *FH* locus. Exome data covers solely the exons. **(c)** Zoomed in view shows aligned reads (green and red bars), the reference sequence, and coverage at the locus. Colors in the reference sequence are green, blue, orange, and red for A, C, G, T, respectively.

5.5.3 Noncoding genome mapping

The noncoding genome was long referred to as dark matter, or even junk DNA, due to unknown functions of intergenic regions [38]. However, novel biochemical methods coupled with NGS have shed light on this unknown territory. Methods such as ChIP-seq/exo/nexus and DNase-seq enable the mapping of protein-DNA interactions. They reveal active sites of the coding and noncoding genome by determining the exact positions where DNA interacts with proteins such as transcription factors, the transcription machinery, and histones. For instance, knowing the interaction sites of transcription related proteins makes it possible to determine which genes are transcribed/active in a particular cell or tissue type. Indeed, protein-DNA interactions can be highly tissue-specific; so only results from the same or a similar tissue type are usually comparable.

5.5.3.1 ChIP-seq/exo

DNA is bound by numerous proteins, which are involved in maintaining chromatin structure and gene regulation. Chromatin immunoprecipitation (ChIP) is a method to determine the locations where a specific protein binds DNA, and further estimate the abundance of that protein in a given genomic region [135, 141]. ChIP was introduced in the 1980s, when it was used on selected genomic regions. Microarray techniques (ChIP-chip) enabled mapping of the whole genome. Coupled with NGS techniques, ChIP became ChIP-seq, which could be used without the selection of regions to cover the entire genome with a considerably higher resolution than with the microarray-based methods [142]. In ChIP-seq, protein-bound DNA fragments are enriched and NGS technologies are then used to sequence and align the sequence reads to the reference genome. The resolution of ChIP-seq determined binding sites is in the 100-1000 bp range. A more recent method, ChIP-exo, is more precise, as it provides the exact position of the binding site. This has enabled an enhanced resolution of noncoding genome mapping [143, 144]. The exo method also reduces background noise, and thus requires fewer sequence reads to be produced [144].

The ChIP-seq procedure simplified is as follows. First, DNA from a tissue sample of ~10 million cells is fragmented with, for instance, sound waves (sonication). At this point, functional and structural proteins are still attached to the DNA. There are hundreds of different DNA binding proteins and protein complexes, which are found across the genome. The goal is to capture the interaction sites of a single protein. The number of binding sites depends on the function of the protein or protein complex; for example, the number of CTCF binding sites in the human genome is tens of thousands, whereas histone proteins are bound to millions of sites [145–147]. In the second phase of ChIP-seq, a protein-specific antibody (which is also a protein) is used to bind the protein of interest. Now, the solution contains DNA fragments, of which a fraction are bound by TF proteins, of which a fraction are bound by the antibody. In the third phase, all the DNA fragments without a bound protein-antibody complex are washed away. Proteins are removed before sequencing, leaving only the captured DNA fragments in the solution. In practice, washing and capturing phases are not perfect, adding varied levels of background noise and missing regions to the produced data. Similarly to exome sequencing, the captured fragments are enriched, sequenced, and aligned to the reference genome, ultimately revealing the binding locations of the studied protein [148].

5.5.3.2 SELEX for transcription factor binding sites

An *in vitro* selection method, systematic evolution of ligands by exponential enrichment (SELEX), was introduced in 1990 [149]. The technique is used to determine short DNA or RNA molecules (oligonucleotides) that a given molecule (e.g., protein) can bind. As in ChIP, the specific DNA fragments bound by the studied protein are probed *in vitro* in SELEX. The method can be applied to e.g., TFs. TFs recognize short (~6-20 bp) sequences in the DNA, which they use as binding sites. SELEX is used to determine the binding sequence(s) or binding motif that a specific TF recognizes [142, 150] (**Figure 14**). However, a single TF can bind to a series of similar sequences with varying affinity scores; thus, there is no exact binding sequence for a TF but rather a most preferred one.

Binding motifs are often represented as position-specific scoring or frequency matrices (PSSM/PSFM) and graphically as a sequence logos, which have been constructed from the sequences that a specific TF or other molecule binds (**Figure 14**). The higher the score in the matrix, the more essential the base is for binding. The SELEX-seq method uses NGS to determine multiple TF binding motifs parallelly [150]. The alignment of PSSM to the reference results in a map of possible TF binding sites across the genome. The challenge is to determine which of the aligned loci represent the real biological binding sites of the TF. The alignment affinity score gives the similarity between the motif and the reference sequence. It is an estimate of the binding probability for each site and can thus be used as indirect evidence of actual binding. The real binding positions of TFs can be determined at base-pair precision by integrating SELEX data with ChIP-seq/exo mappings and picking sites of overlap, as was done in the publication II of this thesis.

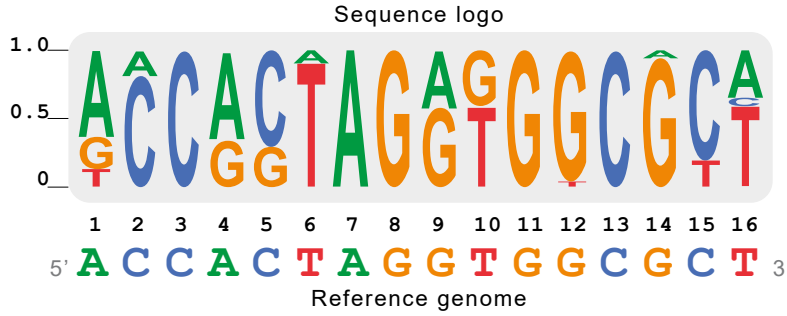

Figure 14: Sequence logo example. The sequence logo (CTCF in this case) is a graphical representation of a TF binding motif. Each position illustrates the relative frequency (y-axis) of bases preferred by CTCF. E.g., at positions 7 and 8, A and G bases have been measured in 100 % of binding sequences. The affinity score of a particular locus can be calculated by comparing the frequency matrix to the reference sequence. In this case, the motif would get the highest possible score, since the most frequent bases correspond to those in the reference.

5.5.4 Next-generation sequencing powered cancer genetics research

Successful detection of pathogenic variants by NGS typically requires layers of genetic knowledge, annotation data, control material, computation, and validation, as well as proper sample selection and preparation. Thus, modern disease genetics research is a multidisciplinary process, requiring skills in medicine, biology, and computer science. The literature review part of this thesis has introduced the basics of these fields in the context of cancer genetics. Finally, all this knowledge and available data require integration, which enables genetic discoveries in both germline and somatic settings.

5.5.4.1 Data integration in cancer genetics

The "Genome annotation" chapter described the integration of the reference sequence with gene annotation data. As discussed earlier, without this data integration, the reference genome would be useless in cancer genetics. Similarly, all the data types introduced in this thesis can utilize previous data integrations and build new layers of information on top of the old [151, 152]. For example, integration of germline variant and control data, gene annotation, and enhancer element mapping can be used to recognize cancer risk increasing SNPs in the regulatory genome, as was described in the "Regulatory and the noncoding genome" chapter [87]. The main advantage and the main challenge is the interpretation of the integrated data [152]. Data integration enables the interpretation of, for instance, the effect of a mutation on TF binding or a protein product. The challenge is to make the correct interpretations of the data types used; are the files from different sources compatible? Does the integration make biological sense? Compatibility issues can arise as a consequence of the use of different reference genome versions in the analyses (see "Human reference genome" chapter) or different tissue types in data production.

Visualization is often essential when assessing data compatibility and interpreting the integration results [153]. It is for example possible to detect the increase in somatic mutation frequency (discussed in the "Somatic mutations" chapter) of late replicating regions by visualizing replication time data and somatic mutation data in multiple samples simultaneously (**Figure 15a**). The conservation of a TF binding site can be determined by visualizing integrated TF motif and conservation data (**Figure 15b**). However, more sophisticated data integration in cancer genetics requires the use of mathematical models, which combine relevant biological components and strive to determine their relations [151]. The ultimate goal of biological data integration is to construct a complete model of how living systems function. The detection of cancer predisposing or driving mutations is performed by integrating germline variants or somatic mutations with genome annotation

data as was introduced in the “Cancer as a research subject” chapter. The two final sections of this literature review describe these research settings in practice.

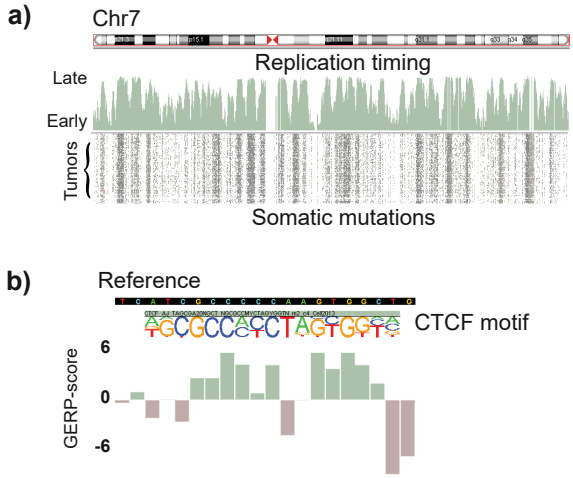

Figure 15: Data integration examples. **(a)** Replication timing data visually integrated with somatic mutations from 100 tumor samples. Regions with late replication timing show higher mutation frequencies. **(b)** Visual data integration of the reference sequence, an aligned CTCF motif, and conservation scores. High conservation scores overlapping the motif suggest that the aligned motif locus is a true biological binding site for CTCF. Also, mutations occurring at the highly conserved bases could be considered putatively harmful.

5.5.4.2 Germline variant analysis

Germline studies in cancer genetics research aim to detect variants that cause a particular cancer or predispose to it. The research method used is selected based on the assumed frequency of the causative variant in a population. Also, penetrance and trait complexity (i.e., whether the studied disease is mono- or multigenic) determine which parameters are used in a particular research setting [36]. The basic idea in these analyses is to compare the frequency of germline variants between cases and controls. In the search for common causative variants, large datasets (thousands of samples) of cases and controls are analyzed by genome-wide association. GWASes have typically been performed with SNP arrays, while smaller, rare variant studies on monogenic (Mendelian) diseases have been feasible to be performed using NGS data. A workflow for the detection of rare predisposing variants in an NGS dataset could go as follows: (1) sample selection, (2) NGS, (3) data processing, including alignment and variant calling, and finally, (4) variant analysis.

Variant analysis in germline studies is the data comparison, integration, and interpretation phase, in which the goal is to identify a shared predisposing variant from the massive amount (thousands to millions) of variants found in multiple individuals. The most critical asset is sufficient and suitable control material, which is used to exclude common, likely benign variants from the analysis [34, 36]. Publicly available gnomAD population control data, currently (v3) containing variants from over 70k genomes and over 100k exomes, has been instrumental in causative variant detection [154]. The gnomAD data set consists of the variants of individuals from a limited amount of different populations, and is most efficient in studies concerning these. Indeed, studies performed on samples from individuals of Finnish heritage leverage most from gnomAD, as Finnish individuals are overrepresented in the database [155]. There have been large international collaborative efforts (e.g., Sequencing Initiative Suomi, SISu) to sequence Finns, partly due to the particularly homogeneous gene pool of the Finnish population. Currently, gnomAD contains variant data from over 5,000 and over 10,000 Finnish genomes and exomes, respectively [154]. Moreover, Finland has an exceptionally comprehensive cancer registry, which facilitates genetic research even further [156]. The registry contains information on over a million cancer cases. The data can be utilized, for instance, in identification of families with increased cancer risk. Publication III is an example of a germline study, where these assets are utilized in both the sample selection and variant analysis phases.

Variants, which are enriched in cases (putatively pathogenic), are identified by comparing case and control data (e.g., gnomAD). The strength of enrichment can be reported as an odds ratio (OR) value, the significance of which heavily depends on the sizes of the case and control data (**Table 3**). Basically, the higher the sample count, the higher the probability that the association is correct.The fact that there are commonly several variants that predispose to the same phenotype constitutes a challenge in enrichment analyses. For example, let us consider a study on five familial cases of CRC with unknown heritable components. In case all individuals have different predisposing variants, enrichment would be impossible to detect and prove significant. Enrichment could, however, be studied on gene, gene family, protein complex, or genetic pathway level. In case of the other extreme, that all five individuals share the same rare inherited variant, the sample-wise comparison of all rare variants would probably be sufficient to pinpoint the putative causative variant. This scenario is much more probable in homogenous populations. Hence, e.g., Finnish and Icelandic samples have been of considerable interest in genetic research.

Table 3: Odds Ratio.

	Affected	Healthy
Variant	**Av**	**Hv**
No variant	**An**	**Hn**

$$\mathbf{OR} = \frac{\mathbf{Av} / \mathbf{An}}{\mathbf{Hv} / \mathbf{Hn}}$$

Enrichment analysis can yield multiple significant variants, some of which may be the result of sequencing or calling errors. Moreover, the probability of finding significant results increases with the number of variant comparisons or tests, necessitating multiple testing corrections [157]. The quality-based filtering of variants can be used if false positive calls contaminate the results. Variant callers, such as HaplotypeCaller, typically report quality and other metrics for each variant in the VCF file. Parameters such as coverage, allelic fraction (number of reads supporting the variant call per coverage), base quality derived score, and strand bias, can be utilized to filter out false positive-calls. "Blind" filtering of variants based on given quality metric values does not always give the best results, visual inspection of variant loci is often necessary in the variant validation phase [158]. Also, Sanger sequencing is routinely performed to verify variant calls [24, 159]. However, the pathogenicity of candidate variants is commonly determined by various prediction algorithms to separate benign variants from damaging ones before validation and verification.

Protein-truncating mutations (e.g., nonsense) are in general considered damaging by definition. However, the pathogenicity predictions of missense, synonymous, and in particular noncoding mutations require more scrutiny. Variant effect predictors often exploit genomic and proteomic features such as conservation of the variant site, exonic position (e.g., putative splicing effect), and protein structure. Prediction tools, such as Provean or Ensembl Variant Effect Predictor (VEP), which combines Polyphen-2 and SIFT, can be used for pathogenicity assessment of candidate variants [160–162]. The fact that some genes are more tolerant to loss of function mutations than others can also be used to predict the harmfulness of variants in genes [163]. In addition to these predictive algorithms, there are a multitude of processed data that can be used to predict the effects of mutations in both the coding and noncoding genomes, such as CADD, GERP, and aligned TF binding motifs [164–166]. The accuracy of the predictions increases as more processed data emerges and is integrated with the old. The challenge in applying these data to the variant analyses is the interpretation of integrated data, as was discussed in the previous chapter.

5.5.4.3 Somatic variant analysis

Somatic mutation data can be used to determine pathogenic drivers, mutational patterns, signatures, and landscapes in cancer genomes [167]. Detecting driver mutations follows the same basic principles as the search for predisposing variants where the aim is to detect recurrently mutated genes or genomic regions [159, 168]. The quality-based filtering, sample-wise comparison, and variant effect predictions are applied in somatic studies as well, with minor adjustments to the used parameters. For instance, tumor purity, heterogeneity, and possible subclonality of variants affect the appropriate allelic fraction thresholds in variant filtering (**Figure 16**).

The major challenge in somatic variant analyses is separating driver mutations from a mammoth amount of passengers. As discussed in the "Somatic mutations" chapter, the properties of different genomic regions and mutators operative in the nucleus induce different local mutation frequencies across the genome. Hence, recurrently mutated hotspots are more likely to be a byproduct of frequently mutating genomic regions or genes than evidence of selection. Integration of background mutation frequency estimations and functional genomic data can be utilized to identify mutations in genes or other functional sites, which have been putatively selected for during tumorigenesis [168–170]. In coding regions, the rate of synonymous and nonsynonymous mutations can reveal a functional bias towards the accumulation of damaging mutations in genes [168, 170]. However, the functional impact of noncoding somatic mutation enrichment is considerably more challenging to predict. To this end, machine learning efforts to integrate functional and contextual genomic properties and produce base level predictions for the whole genome have been made. The CADD score data, for instance, has been created by integrating genomic features such as conservation, DNase-seq, TF motif disruption, and sequence contexts [164]. Furthermore, tools such as OncodriveFML can utilize CADD scores to identify genomic regions exhibiting evidence of selection [169]. The basic principle in Oncodrive analysis is to determine whether the observed variants in a certain region have hit functionally important bases more frequently than by chance alone.

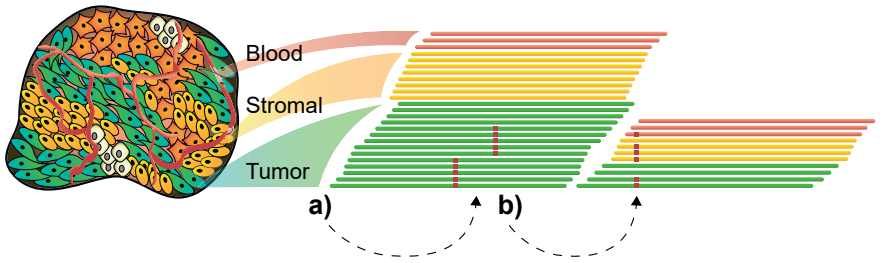

Figure 16: Heterogeneity and purity at read level. Tumor purity and heterogeneity affect the allelic fraction of variant calls. **(a)** Mismatches (putative somatic mutations) in the reads of two distinct tumor populations. **(b)** Mismatches in both tumor and normal reads (probably a germline variant).

Whether the research concerns coding or noncoding regions, or is focused on germline or somatic variants, variant analysis commonly requires multiple steps from sample-wise comparison to data filtering, controlling, interpretation, and finally visualization, as described in this chapter. Tools such as IGV, bcftools, bedtools, and Annovar have been developed for visualization, filtering, integration, and annotation tasks [153, 171]. Designing and developing a software that can handle all these variant analysis tasks efficiently and visually, has been and still is the first aim of this thesis work.

AIMS OF THE STUDY

The aim of this thesis was to develop novel methods for next-generation sequencing data analysis and apply them in various cancer research settings. The method that we developed allowed us and others to characterize and detect putatively causative mutations in various cancer and tumor types. The specific aims of studies I-III were:

I To develop a versatile and user-friendly software for next-generation sequencing data analysis in disease genetics

II To analyze and characterize regulatory somatic mutations in colorectal cancer to identify novel drivers or patterns in tumorigenesis

III To detect candidate susceptibility genes behind esophageal cancer using the Finnish cancer registry and archival tissue material

MATERIALS AND METHODS

7.1 Software requirements and availability

7.1.1 Requirements

BasePlayer runs on Windows, Linux, and macOS systems with Java runtime version 1.8 or later installed. At least 1 GB of memory is recommended to be allocated for the software, and for human studies, at least 3 GB of hard disk space is required for the reference genome.

7.1.2 Additional Java packages

BasePlayer utilizes the following packages for reading indexed file types:

- HTSJDK (https://github.com/samtools/htsjdk) for BAM, CRAM and Tabix indexed files
- BigBed and BigWig file readers by M. Decautis and J. Robinson for the Integrative Genomics Viewer (Broad Institute) obtained from https://github.com/lindenb/bigwig

7.1.3 Software and code availability

The software, additional materials, and an online manual are available at https://baseplayer.fi. The source code can be found at the BasePlayer project page on https://github.com/rkataine/BasePlayer.

7.2 Study materials and ethics approvals

7.2.1 Colorectal cancer samples

A total of 213 matched normal-tumor pairs were used in publication II. The samples constitute a subset of CRC material collected from Finnish hospitals starting from 1994 [172], and include 198 MSS, 12 MSI, 3 POLE mutant tumors, and the respective normals. The study was reviewed and approved by the Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS). Signed informed consent or authorization from the National Supervisory Authority for Welfare and Health were obtained for all used materials.

7.2.2 Esophageal cancer samples

The FFPE samples from selected ESCC cases were collected from Finnish hospitals, which were located using the FCR data. Sample selection was done

by clustering FCR data by cancer type, family name at birth, and municipality at birth. See publication III for a more detailed description of the clustering analysis. We were able to successfully extract and sequence the DNA from 30 individuals, out of which 24 were born in the region of ceded Karelia. When available, both tumor and normal samples were obtained.

Studies on FFPE samples collected in this way have been approved by the National Supervisory Authority for Welfare and Health (Valvira; 1423/06.01.03.01/2012), National Institute for Health and Welfare (THL; 1071/5.05.00/2011 and THL; 151/5.05.00/2017), and the ethics committee of the Hospital District of Helsinki and Uusimaa (HUS; 408/13/03/03/09). All living patients signed informed consent for genetic studies on tumor susceptibility.

7.3 Sequencing methods and data processing

Paired-end WGS of the 213 CRC and respective normal samples, used in publication II, was performed with an Illumina HiSeq 2000. Sequenced reads (100 bps each) were aligned to the 1000 Genomes Project Phase 2 reference assembly hs37d5 with BWA. We removed duplicates with the SAMtools rmdup and performed local realignment around suspected indel sites. Base score quality recalibration was performed using GATK IndelRealigner and BaseRecalibrator. After these steps, the genome-wide median coverages were over 40x in all samples. We used MuTect version 1.1.4 with default parameters for somatic SNV calling. Indels were called with the GATK SomaticIndelDetector. The UCSC genome browser tracks “Duke excluded regions” and “HiSeqDepth top 5%” were used to exclude poorly mappable regions from SNV, indel, and SV calling. Structural variants were called from the same data using DELLY [173]. Somatic SVs were produced by comparing calls from tumor and corresponding normal samples.

The DNA from 34 ESCC FFPE samples was extracted with the phenol-chlorophorm method. Samples were prepared for exome sequencing with the KAPA Hyper Prep and SeqCap EZ Exome + UTR kits. Normal tissue blocks were preferred in the extraction process to prevent somatic mutation contamination of the data in the subsequent germline analyses. 30 samples were successfully prepared for NGS, which was performed with either Illumina HiSeq 2000 or 4000. Reads were aligned and processed essentially as described above with the exception of duplicate removal; Samtools rmdup was applied to both paired- and single-end reads due to highly fragmented DNA caused by time and procedures related to tissue archiving (FFPE) [174]. Germline variants were called with the GATK HaplotypeCaller [137].

7.3.1 ChIP-seq / exo

High-Throughput ChIP-seq experiments were performed for 239 TFs in the MSI-CRC LoVo cell-line (American Type Culture Collection, CCL229TM). Experiments were carried out essentially as described in [87]. Sequencing was performed with Illumina GAIIx and HiSeq 2000. The resulting reads were aligned to the human reference genome (hg18) with BWA. ChIP-seq peaks and peak summits were called with MACS software [175]. Peak calling files (BED) were converted to the newer reference genome version (hg19) to make them compatible with other data used in the study.

The ChIP-exo data for RAD21, CTCF, KLF5, HNF4A, REST, MYC, and MAX TFs was first published in publication II. As in the ChIP-seq procedure, the LoVo cell-line was used in data production. Short-reads were aligned to hg18 with BWA. Coverage peaks were called with GEM using default parameters and converted to a newer reference version as described above [176]. Raw data from ChIP-exo experiments are deposited in the European Nucleotide Archive (ENA) under accession PRJEB9477.

7.3.2 Transcription factor binding sites

The SELEX data used in publication II was produced with a high-throughput SELEX method described in [150]. The data set contained in total 239 distinctly different TF binding motifs [177]. We used aligned SELEX data to determine the binding sites for HNF4A, KLF5, MAX, and REST TFs. The binding motif of CTCF was generated separately from the ChIP-exo peak summit data. The colocalization of ChIP-exo or ChIP-seq peak and corresponding binding motif was used as an indication of a real binding site of a particular TF. CBSs were determined by colocalization of CTCF and RAD21 ChIP-exo peaks as well as the alignment position of the CTCF motif.

7.4 Variant analyses

Both in publication II and III, an unpublished version of BasePlayer was utilized in variant analysis, visualization, and data integration. In publication II, we integrated somatic variant, LoVo ChIP-seq/exo, and ENCODE data to detect mutation clusters in the regulatory genome. Variant clusters were determined by using a 100 bp sliding window across the whole genome and between somatic variant calls from 198 MSS CRC samples. Mutation signature analysis was performed as described in the “Mutational signatures” chapter. In publication III, we filtered false positive calls based on qualities and read counts. Common variants (allele frequency >= 1 %) were excluded by utilizing ExAC data [154]. Enrichments were verified in a region specific control set of 186 exomes from ceded Karelia and nearby regions, obtained from the Finland-United States Investigation of NIDDM Genetics (FUSION)

study (https://fusion.sph.umich.edu/).

7.5 Statistical analyses

In publication II, genome-wide and CBS mutation counts were modeled using negative binomial regression with the covariates including mutation statuses in key CRC genes, clinical features (e.g., sex, age of diagnosis and tumor location in the colon), as well as the relative exposures of the three extracted mutational signatures (1, 17, and undefined). The mutations occurring at a particular CBSs were modeled with covariates including strand orientation, affinity score of CTCF binding motif, and replication timing. In publication III, Fisher's exact test was used in case-control enrichment analyses ($P < 0.01$).

RESULTS

8.1 The development of an analysis software for next-generation sequencing data

We published an analysis and visualization platform, BasePlayer, for research in various disease genetics settings utilizing NGS data. BasePlayer had already been used in over 20 scientific publications before its official release. We demonstrated the usage of the software with two distinct studies in germline and somatic cancer research. The first case demonstrated the detection of the predisposing variant in a family with inherited meningioma. The second case is described in the next chapter. The novelty of the software arises from its ability to integrate a wide variety of NGS data visually and, at the same time, provide efficient variant analysis features on an ordinary desktop computer. Also, the capability to visualize and analyze variants of hundreds or thousands of samples simultaneously is one of the unique features of BasePlayer among existing bioinformatic tools [178]. The analysis phases, such as quality filtering, variant annotation, data integration, and visualization, which are typically performed with different scripts and tools, possibly even using multiple platforms, can be done with one graphical user interface (**Figure 17**). Also, challenges related to analysis of the noncoding genome and tumor sample heterogeneity and impurity, have been given attention during the process of designing and developing the software.

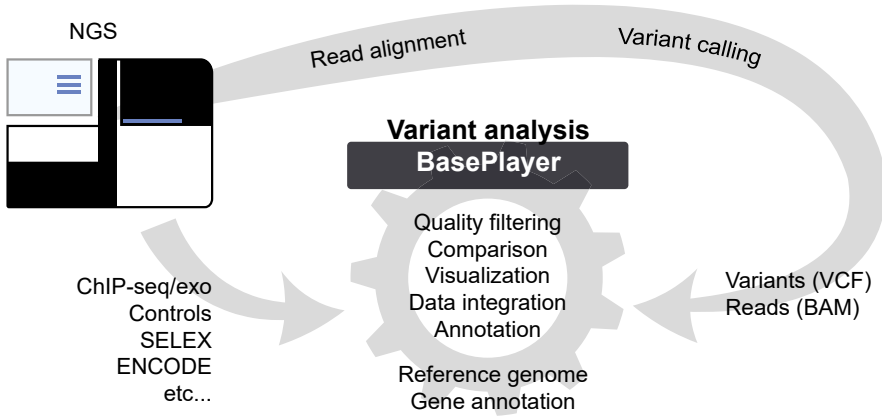

Figure 17: BasePlayer in the NGS data analysis workflow. BasePlayer integrates variant, read, and genome annotation data in variant analysis.

In this thesis, I have outlined reference genomes, genome annotations, sequencing and variant data, quality and population control filtering, and comparative analyses in germline and somatic analysis settings. In publication II, we demonstrated that the integration of these resources, data types, and procedures could be performed graphically in a matter of minutes or hours, rather than days. To facilitate the deployment of BasePlayer, we have shared annotation files essential for human genome analysis, such as gnomAD, CADD, and ENCODE regulatory annotations, on the BasePlayer website (https://baseplayer.fi). Even though BasePlayer has been designed and tested mainly using the human genome, it can be used with any mapped reference genome, the number of which is constantly increasing.

8.2 The discovery of a specific somatic mutation accumulation in the regulatory genome present in multiple cancers

We examined somatic mutation clusters in the regulatory genome by analyzing variants from 213 CRC WGS samples (198 MSS, 12 MSI, and 3 POLE mutants). By utilizing the variant clustering and regulatory data integration features of BasePlayer, we identified a somatic mutation accumulation at CTCF/cohesin binding sites, especially in gastrointestinal cancers (**Figure 18**). Approximately 50% of all dense (i.e., adjacent variants within 100 bp) mutation clusters in the regulatory genome (annotated by ChIP-seq and ENCODE data) overlapped with the 39 bp wide CBSs (n=28,331). In contrast, POLE mutants showed an inverse pattern at these sites, suggesting that POLE does not replicate these sites, which in turn indicates that the clustered mutations arise during replication (**Figure 18**). We also reported, for the first time, the presence of signature 17 in CRC genomes. We observed that variants compatible with signature 17 are more frequently than others subclonal, which suggests that they accumulate during tumorigenesis. Furthermore, CBS mutation count correlated with signature 17 exposure. The finding was initially made with our CRC data, and then validated in the publicly available ICGC data [179]. We found that especially gastrointestinal cancers had high levels of CBS mutations, which was somewhat expected, as signature 17 had been previously reported in esophageal and gastric cancers. However, MSI samples did not show mutation accumulation in CBSs. The effect of the mutations remained unknown, since we were not able to detect changes in gene expression at the flanking regions of mutated sites with our limited RNA sequencing data set. Also, we did not observe accumulation of SV breakpoints at these sites. However, by analysing the precise SELEX motif hits, we showed that the SNVs tended to weaken the affinity of the CTCF binding sites.

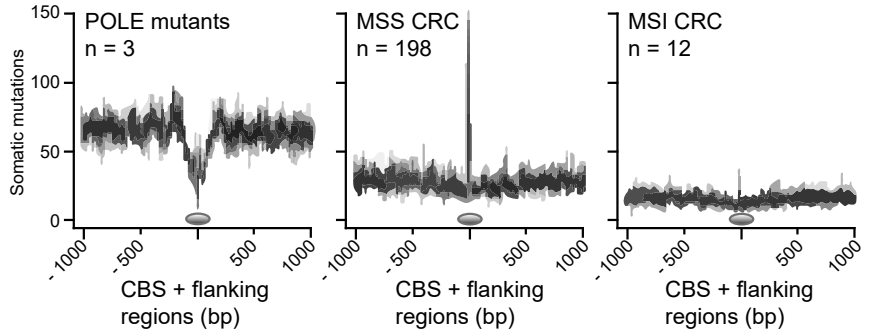

Figure 18: Mutation accumulation in CBSs in different CRC classes.

The mutation accumulation was not explained solely by the sequence context of the motif under signature 17 exposure, which was noted by observing the ratio of observed versus expected mutation frequencies. We also examined whether the mutation signal was present in CTCF binding sites that were not occupied by the cohesin complex (determined as RAD21 ChIP-exo signals; **Table 1**). Indeed, there was no mutation accumulation at these sites, which strengthens the conclusion that the sequence context is not the source of the signal. The result also suggests that the mutations are the consequence of a specific genomic function that requires the presence of both CTCF and cohesin.

8.3 The detection of putative predisposing mutations in esophageal squamous cell carcinoma

We reported several candidate susceptibility genes for ESCC by analyzing the variants of 30 cases with BasePlayer. Based on variant enrichment and gene function, the strongest candidates were *DNAH9* and *EP300*. The study material was selected based on the results of clustering cancer cases in the Finnish Cancer Registry data regionally by cancer type and last name at birth. Municipality and name at birth were obtained from the Population Information System of Finland. We identified an enrichment of clustered ESCC cases in the region of ceded Karelia, from where most of the samples were collected. We also included six samples from the highest-ranking (with regard to observed / expected values) ESCC clusters from within the whole country. Finally, we were able to collect 34 FFPE samples in total, out of which 30 had a sufficient amount of tissue material for exome sequencing.

DNAH9 was found to harbor a very rare nonsense (stop gain) mutation in 15% of our samples. We were also able to detect LOH in the tumor of a carrier, the only one from whom a tumor tissue sample was available. Four

individuals from four distinct clusters shared the variant. The frequency of the variant in Finland and the area specific control set was ~0.09% and ~0.27%, respectively. The gene itself has not been previously linked to ESCC. It has, however, been shown to be frequently mutated in invasive micropapillary carcinomas of the breast and aberrantly methylated in non-small cell lung cancer [180, 181]. Furthermore, the chromosomal locus of *DNAH9* (17p12) has been previously detected to be frequently lost in ESCC genomes [182]. The enrichment of the variant in *DNAH9* showed the highest odds ratio in the comparative analysis of all samples. We also detected rare missense variants in the *GKAP1*, *BAG1*, *NFX1*, *DDOST*, and *FCSK* genes. The variant in *GKAP1* was shared by three and the rest by two individuals.

EP300 is a well-established ESCC related gene. It was found to be mutated in the largest sampled cluster consisting of four cases, which was also confirmed to be a true ESCC pedigree. However, we could not identify a convincing variant shared by all individuals in the WES covered regions. Hence, there were three probable options; (1) the causative variant was located outside the targeted regions, (2) there was a phenocopy in the family, or (3) the ESCC in the family was sporadic rather than inherited. To this end, we analyzed these four samples in different combinations of three to detect a possible phenocopy. The *EP300* (and some other) variants were shared by two affected brothers and their cousin, which suggested the mother of the cousin to be the putative phenocopy. The pedigree structure supports this hypothesis, as the cousin is related to the brothers through his father (**Figure 19**). The affected men also shared likely damaging variants in the *DCDC2B*, *ANK2*, and *CABIN1* genes.

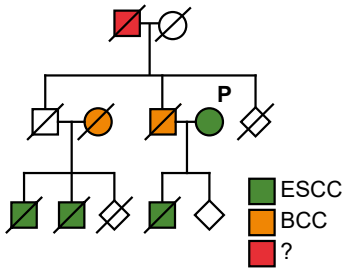

Figure 19: ESCC pedigree. Putative phenocopy is marked with **P**. There are four ESCC and two basal cell carcinoma (BCC) cases in the family.

DISCUSSION

This thesis gives a glimpse of modern sequencing and computing technologies used in search of disease-causing genetic defects. As computing power, biological knowledge, and genomic data types continue to increase exponentially, the constant development and maintenance of bioinformatic tools are necessary. The methods presented in this thesis describe an exploratory or knowledge-based approach to analyze genomic variant data. In contrast to purely statistical analyses in genetics, such as GWAS, epidemiology, and differential expression analysis, exploratory analyses aim to utilize, as much as possible, accumulated knowledge to explain the role of genetic alterations within complex biological functions and structures [183–185]. However, the majority of biological systems are still too complex to be analyzed using a single approach or method; thus, statistics, data exploring, visualization, and machine learning are often required to produce sensible results from the data at hand.

In context of variant analysis with NGS data, a researcher should be aware of (**1**) possible error sources in sample preparation, data processing and used datasets, (**2**) the genetics behind the study, such as inheritance patterns, known driver mutations and predisposing variants as well as genes associated with the studied disease, (**3**) the types and qualities of studied samples, such as tissue type, estimated tumor percentages, average coverages and ancestries of donors, (**4**) required size for the sample set in order to produce significant results, and (**5**) tissue-specific genomic regions and genes of interest, and how to integrate data from multiple sources to narrow down the genomic search space. The challenge in NGS workflows is to run optimal processing steps for given samples, which requires knowledge of the issues listed above. Variant processing (e.g., quality filtering) and genetic analysis (e.g., detection of somatic driver mutations) are often performed by different people, for instance, a bioinformatician/computer scientist and geneticist, respectively. This kind of arrangement may lead to suboptimal variant calls for the genetic study setting. For example, when analyzing somatic mutations of tumor samples of unknown purity, the variant caller should have high sensitivity parameters to detect low-frequency variants. Although this leads to a high false-positive rate of variant calls, the comparison between multiple samples can be used to identify the correct causative mutation signal. BasePlayer was designed to tackle this challenge by providing built-in interactive and visual quality filters and various sample comparison features. In addition, fast variant visualization by the read inspection was developed to facilitate the exclusion of false positive-calls.

There is a plethora of tools available for visualization of various biomedical data types. It is essential to select the right visualization tool for the data at hand in order to recognize patterns and outliers in often massive and wide-ranging biomedical data sets. Different tools are in general used for browsing and plotting the data; interactive and linear genome browser-style tools such as BasePlayer, IGV, and UCSC genome browser are used when sequencing data is viewed alongside genes and other annotations [178, 186]. Tools and tool packages specialized for scientific data plotting, such as Chipster, Biopython, and clusterProfiler, are utilized when multidimensional and mainly numerical data such as differential gene expression, large biological networks, and protein interactions are being analyzed [187–189]. A circular view (e.g., Circos plot) of the genome can be instrumental when studying for e.g., genome-wide SVs [190]. In the context of this thesis, mainly linear genome browsers were used in NGS data visualization. A recent article in which different tools for SV visualization were compared highlighted the capabilities of BasePlayer in long-read sequencing data analysis. The authors also stated that it may be the only tool available for managing and visualizing variant data for thousands of samples simultaneously [178]. Although BasePlayer is essentially a variant analysis and data integration software, it can thus also be considered a state-of-the-art tool for NGS data visualization.

Germline variant studies commonly rely on a case-control setting, where a high number of control samples is essential [36]. The public release of ExAC, and later gnomAD, variant data, was a game-changer in germline variant analysis in disease genetics, in particular for research groups without control data of their own. Also, somatic calls could be produced with relatively high precision in cases where corresponding normal samples were not available [191]. Germline analyses previously utilized in-house control material and public variant databases such as dbSNP, which, however, was shown to be suboptimal for the purposes of disease genetics as it was contaminated by somatic and false-positive variants [192]. The ExAC and gnomAD data were rapidly adopted in BasePlayer, increasing its applicability in human genetics research immensely.

Similarly to the ExAC release, the comprehensive noncoding genome annotation data sets, provided e.g., by the ENCODE project, were soon made usable in BasePlayer, which facilitated in particular human regulatory genome variant analysis. The value of variant data increases in time with the increased availability of genome annotation data through the added meaning of the data. BasePlayer was designed for straightforward integration of new annotations with variant data so that researchers could make the best use of their data as new annotations emerge. However, the search of causative variants often requires ranking or prediction tools to narrow down a massive list of candidates. The tools, such as Oncodrive, Mutsig, and DeepSEA, facilitate the data mining process and can significantly improve the BasePlayer analysis workflow [169, 193, 194].

Various noncoding genome annotations were applied in the second work of this thesis, where we integrated somatic variants with ENCODE, ChIP-seq/exo, SELEX-seq, and replication timing data. We used BasePlayer to detect somatic mutation clusters in the noncoding genome using variant data from 198 MSS CRCs and noncoding genome annotations. We had ChIP-exo data from five TFs, which we were able to add to the study for more precise binding site measurements. We had previously detected the mutation frequency decrease at RAD21 binding sites (a proxy for the cohesin complex occupation) in POLE mutants and saw a slight increase in MSS CRCs at the same sites. However, the increase in mutation frequency was not significant when the RAD21 ChIP-exo loci were used alone.

To measure mutation frequencies at the specific TF-DNA interaction sites relative to the reference genome, we first used SELEX-seq motif data to determine the precise loci of putative binding sites. Then we used the ChIP-exo data from the corresponding TF to select motif sites with biological evidence of the TF protein binding (**Figure 20**). This data integration enabled us to inspect mutation accumulation across thousands of TF binding sites. Thus, we were able to merge binding sites of specific TFs and report mutations relative to the binding motifs and flanking regions at base-level precision. The mutation accumulation at CBSs was not observed as significant, as the RAD21 measurements were not precise in relation to the reference sequence positions, which leveled out mutation frequencies when all measured sites were observed at once. However, strong mutation peaks were observed, when the CTCF binding motif was used as an anchor point for CBSs. This is an example of a case where the value of mutations and measurement data is increased through data integration.

The next challenge was to determine whether the CBS mutation accumulation was significantly higher than expected or just a product of the specific sequence context of the CTCF motif and underlying mutational signatures. Despite the CBS mutations being almost exclusively signature

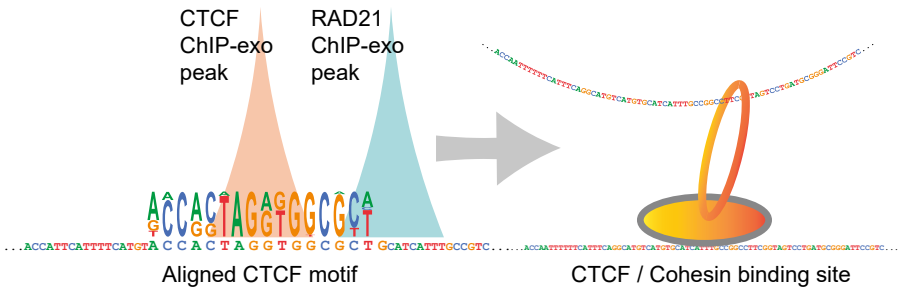

Figure 20: Determining cohesin binding sites. CBSs were selected by using ChIP-exo measurements from RAD21 (part of cohesin protein complex) and CTCF (physical binder of cohesin to DNA). In addition, we required CTCF SELEX-motif alignment within 100 bp of the ChIP-exo peaks.

17 compatible and the CTCF binding sequence including signature 17 preferred contexts, the mutation frequency was significantly higher than expected under the given signature exposures. We also showed that cohesin occupation was required for the excessive mutation frequency, which suggested the function of these sites to be involved in the observed mutation signal. Subsequent studies by others showed CBSs to be mutated more frequently at TAD borders and active genomic sites [195, 196]. These studies confirmed that mutations accumulate at sites with a specific function. Active TF binding sites show increased mutation frequencies due to hindered DNA repair [197]. For instance, a bound TF can block the repair machinery and hamper its function during replication. TAD borders (often occupied by cohesin and CTCF) have been linked to the initiation of replication and one of the main functions of the cohesin complex is to maintain sister chromatid cohesion during replication. Thus, CBSs may act as repair machinery blockades during replication and cause the mutation accumulation at these sites [197, 198].

Although we detected recurrently mutated CBSs, we did not detect changes in the expression levels of the proximal genes. The expression analyses were not conclusive as we had only a limited number of expression data available for the corresponding samples. Also, low tumor percentages and heterogeneity may have hampered the results as well [199]. In other studies, indel mutations in CTCF binding sites have been confirmed to change the conformation and replication timing of the mutated region, but the effects of SNVs are still unclear [200, 201]. In terms of detecting pathogenic point mutation hotspots in the regulatory genome, the results reported in this manuscript were somewhat disappointing, as they indicated that there are no highly recurrently mutated regulatory regions with strong effects to be found in the CRC genome, such as *TERT* promoter hotspot mutations found in melanoma and other cancers.

The subclonality of the detected CBS mutations also indicated that they have no or only a minor effect in promoting tumorigenesis. Also, the mutation signal was present in only a small percentage of the studied sample set. These results suggest that point mutations that damage coding regions are more favored during colorectal tumorigenesis than mutations breaking, for instance, a certain TF binding site of a certain regulatory region. Indeed, a single regulatory region acts as a binding site for numerous TFs, the overall dosage of which determines the activity of the region [202]. Moreover, a single gene may be regulated by multiple enhancers possessing similar activity. This enhancer redundancy and the dosage effect of TFs prevent single mutations in regulatory regions from having strong effects on gene expression [203]. Finally, a damaged regulatory region would still require a second hit on the other allele in haplosufficient cases, where one copy of a gene is enough to maintain a healthy phenotype.

We aimed to investigate annotated regulatory elements for mutation hotspots rather than to focus on specific genes and their flanking and intronic regions. This approach allowed us to search densely mutated regulatory regions genome-wide in an unbiased manner, with no particular set of genes or gene regions having been cherry-picked. The major drawback of this method is the incompleteness of regulatory annotation; in general, regulatory annotations are mere approximations of putative regions, which have been measured to have an activity in given cell-lines *in vitro*. Thus, for example, intronic mutations affecting splicing would most likely have been missed in the study. Also, a major challenge is to link a particular regulatory element to its target gene or genes. Methods such as Hi-C have been developed for this task, but the understanding and complete characterization of regulatory targets are still largely incomplete [204, 205]. Most regulatory regions in the human genome are located within 100 kbp up- or downstream of their target genes, so another approach would be to focus on the flanking regions of genes that have been linked to a certain cancer type. This method would make it possible to scrutinize raw sequences around established "cancer genes" so that no particular regions are favored (i.e., annotated as regulatory regions). Also, the putative target gene for a discovered hotspot would be known. However, the challenges are similar to the ones of the previous approach; mutation hotspots may not have actual functional relevance and the true target gene would still be uncertain. Moreover, restricting the search to gene flanks would exclude numerous real regulatory elements from the study. Finally, both approaches provide only putatively causative sites, and thus functional studies are required for conclusive confirmation of the effects of found hotspots.

In the third study, we utilized the Population Information System and Finnish Cancer Registry data to detect familial ESCC cases with an in-house developed clustering method. Case clusters were produced based on surname and municipality at birth. ESCC clusters from mainly ceded Karelia were selected for further analysis, and we managed to successfully exome sequence 30 FFPE samples. The challenge was to detect predisposing variants from the sample set, which in addition to cases with actual ESCC was likely to include an unknown number of sporadic cases. Indeed, lifestyle related risk factors contribute to this phenotype in particular. Also, cases were collected from a specific region (ceded Karelia), and the sample set is thus likely to show enrichment of a number of variants unrelated to ESCC. These factors should be taken into account when assessing variant enrichment in the studied sample set. In our study, we determined the enrichment of all variants that were shared by two or more individuals and used a Karelia-specific control set to exclude regionally enriched variants from the candidate list. Formalin fixation and paraffin-embedded storing of tissue samples cause excessive fragmentation and random artefactual mutations to the DNA, which hamper subsequent variant analyses through

inconsistent coverages and false-positive variant calls. We were, however, able to identify candidate predisposing mutations from the ESCC sample set by utilizing the case-control features of BasePlayer together with gnomAD data. In addition to the gnomAD controls, we used a Karelia region-specific control set as mentioned earlier. This study strengthened our confidence in being able to take advantage of the Finnish Cancer Registry in finding familial cancer clusters, and convinced us that NGS of FFPE tissue-derived DNA and subsequent analysis of the data is feasible in studies on genetic predisposition.

CONCLUDING REMARKS AND FUTURE PROSPECTS

The impact short-read sequencing-based NGS technologies have had on cancer or any other genetics research can not be overestimated. The third-generation sequencing platforms have taken steps towards reading much longer sequencing fragments. Long-read sequences (up to dozens of kbps) enable studies on genomic regions, which were challenging to cover with previous methods (e.g., long repeats). While short-reads are indisputably powerful in the detection of point mutations due to relatively low sequencing error rates, there are major limitations in determining, for instance, certain size SVs (~50-300 bp), transcript isoforms, and haplotypes. These and other challenges can be tackled with long reads, despite relatively high sequencing error frequencies currently contaminating the data. New sequencing and bioinformatic methods are also needed for single-cell sequencing, which has revolutionized the analyses, e.g., in developmental biology and tumor heterogeneity [29]. Continually changing research questions and novel data types have created challenges in software development; the more features are added, the more complex the maintenance of the software becomes, which then may lead to software instability. Thus, modularity and documentation should be high-priority during software development. To this end, there have been efforts, such as the Chan Zuckerberg Initiative, to support "essential open-source software for science" and thus ensure the maintenance of such platforms.

The constantly growing data mass of DNA, RNA, methylation, ChIP, ATAC, HI-C, and other multi-omic NGS data types have provided unprecedented views to cancer genomes and genomic landscapes. Indeed, due to the increase of functional genomic data, cancer genetics research has been moving from causative variant detection towards more functionally oriented analyses, which aim to determine the physical effects of studied variants. Hence, data integration platforms, such as BasePlayer, have become even more essential than before for modern cancer research. However, definitive functional evidence typically requires laboratory experiments *in vitro*, which require significantly more work than *in silico* analyses. The applications of CRISPR-Cas9 methods have been instrumental in functional genomic experiments at noncoding regions in particular, and will most likely dominate genetic research also in the coming years. At the same time, machine- and deep learning techniques have become increasingly more popular in the field and have already brought novel insights into cancer genomes.

Cancer genetics research provides the theory of the biological properties underlying tumorigenesis and malignancy. Translational medicine applies

this theory to practice in developing novel treatments to cancer. However, base studies in cancer genetics not only provide information about diseased cells, but also elucidate the mechanisms sustaining all life.

> There is a computer disease that anybody who works with computers knows about. It's a very serious disease and it interferes completely with the work. The trouble with computers is that you 'play' with them!
>
> Richard P. Feynman

ACKNOWLEDGEMENTS

This work was carried out at the Department of Medical and Clinical Genetics and Applied tumor Genomics Research Program, Faculty of Medicine, University of Helsinki, between 2013 and 2019. I acknowledge the present and former directors for the excellent research facilities. I am truly grateful for all the financial support from the Doctoral Programme in Biomedicine (DPBM) and for personal grants from The Biomedicum Helsinki Foundation, The Cancer Foundation, The Emil Aaltonen Foundation, The Juhani Aho Foundation for Medical Research, The Ida Montin Foundation, The Orion Research Foundation, and The Instrumentarium Science Foundation.

Thank you Lauri, for allowing me to work in the super-group of yours all these years, and for trusting my abilities to learn and grow as a scientist. I could not have managed to get this far anywhere else, really. And what a fun and challenging ride it has been! Thank you Esa, for supervising my bachelor's, master's, and Ph.D. theses, and whatnot. I also thank you for being an inspirational colleague and mentor during this endeavor. I want to thank my thesis committee, Anu Loukola and Ari Löytynoja, for being highly supportive and encouraging since our first meeting. I acknowledge the pre-examiners of this thesis, Merja Heinäniemi and Sofia Khan, for accepting the task of reviewing and grading my work, and for giving valuable comments and suggestions. I want to express my gratitude to the collaborators and co-authors outside of my lab who have contributed to the research of this thesis: Anna Lepistö, Eero Pukkala, Heikki Järvinen, Jukka-Pekka Mecklin, Jussi Taipale, Kashyap Dave, Kristian Ovaska, Martin Enge, Miia Artama, Laura Renkonen-Sinisalo, Teemu Kivioja, and Veli Mäkinen.

I want to thank my colleagues in our wet lab, who year after another have managed to produce top-quality samples and materials. Inkku, Iina, and Alison, you truly are the right people and personalities in the right place at the right time. I dearly thank our office team, who have taken care that everybody has everything they need and feel comfortable in our lab. Sini, Sirpa, and Marjo, you have contributed greatly to making our lab feel special and a pleasant place to work and be. Thank you Sini in advance for the wool socks I am probably getting in the near future. My deepest gratitude goes to "the Original Incubator": Eevi, Mervi, Miika, and Yilong. I will never forget our enthusiasm, parties, and discussions as we were starting our scientific careers. You all taught me (a ragged computer scientist and wannabe bioinformatician) everything I needed to know about the work in the field of cancer genetics. I obviously also wish to thank all subsequent incubator members—Emmu, Heikki M, Heikki R, Heli K, Iikki, Jiri, Klaus, Krisse, Lauri S, and Päivi—for the great times in that crowded room. Thank you Kimmo for the invaluable help in various analyses, especially concerning

mutational signatures. Thank you Auli for all the great scientific and totally non-scientific discussions we have had along these years. I thank Johanna for being so helpful with the matters concerning the final stages of this doctoral effort. Thank you, Tatiana and Ulrika, for your spirit and knowledge in our great joint research projects. Also, thanks Tati for showing us Menorca. I thank dearly all our current and previous group members: Anna, Alexandra, Aurora, Davide, Elina, Hande, Heli L, Iina T, Jaana, Janne, Javier, Justyna, Kati, Linda F, Linda vdB, Maija, Mairi, Maritta, Manuel, Netta, Niko, Outi, Petra, Pia, Poojitha, Rainer, Roosa, Saija, Samantha, Sanna, Sanni, Sari, Silva, Simona, Sini K, Sofie, Tomas, Tuomas, and Verna.

Thank you "Rinki" in alphabetical order: Ari, Hannu, Heikki, Keisteri, Kuisma, Oope, and Teppo. We really form an exceptional group wherever we go! Thank you Heikki for being patient enough to be my flatmate, colleague (as well as incubator roommate), and friend during these lab years.

Thank you Iikki for being there for me in and outside of work. You are my dearest collaborator. Without your support, love, goofy dogs, and invaluable proofreading (including even these acknowledgements), writing this thesis would have been an unpleasant ordeal.

I thank my family for letting me walk my own path and for providing support, which enabled me to try different things without worrying too much. It did not turn out too bad, after all.

I owe my deepest gratitude to all the patients and families who donated samples and made these studies possible. Finally, I acknowledge all fellow researchers who dedicate and have dedicated their lives to research. Let us not forget that we are all standing on the shoulders of giants, as the great Isaac Newton once said.

Helsinki, February 2020

Riku Katainen

REFERENCES

1. Bissell, M. J. & Labarge, M. A. Context, tissue plasticity, and cancer: are tumor stem cells also regulated by the microenvironment? *Cancer Cell* **7**, 17–23 (2005).

2. Frank, S. A. *Dynamics of Cancer: Incidence, Inheritance, and Evolution*. (Princeton University Press, 2007).

3. Shraiman, B. I. Mechanical feedback as a possible regulator of tissue growth. *Proc. Natl. Acad. Sci. U. S. A*. **102**, 3318–3323 (2005).

4. Brown, J. M., Martin Brown, J. & Attardi, L. D. The role of apoptosis in cancer development and treatment response. *Nature Reviews Cancer* **5**, 231–237 (2005).

5. Li, F., Tiede, B., Massagué, J. & Kang, Y. Beyond tumorigenesis: cancer stem cells in metastasis. *Cell Res*. **17**, 3–14 (2007).

6. Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., Esteller, M., Fitzgerald, R., Korbel, J. O., Lichter, P., Mason, C. E., Navin, N., Pe'er, D., Polyak, K., Roberts, C. W. M., Siu, L., Snyder, A., Stower, H., Swanton, C., Verhaak, R. G. W., Zenklusen, J. C., Zuber, J. & Zucman-Rossi, J. Toward understanding and exploiting tumor heterogeneity. *Nat. Med*. **21**, 846 (2015).

7. Vogelstein, B. & Kinzler, K. W. The multistep nature of cancer. *Trends Genet*. **9**, 138–141 (1993).

8. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

9. Brown, L. F., Yeo, K. T., Berse, B., Yeo, T. K., Senger, D. R., Dvorak, H. F. & van de Water, L. Expression of vascular permeability factor (vascular endothelial growth factor) by epidermal keratinocytes during wound healing. *J. Exp. Med*. **176**, 1375–1379 (1992).

10. Aaronson, S. A. Growth factors and cancer. *Science* **254**, 1146–1153 (1991).

11. Kerr, J. F., Wyllie, A. H. & Currie, A. R. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br. J. Cancer* **26**, 239–257 (1972).

12. Puliafito, A., Hufnagel, L. & Neveu, P. Collective and single cell behavior in epithelial contact inhibition. *Proceedings of the National Academy of Sciences* **109.3**, 739-744 (2012).

13. Lopes, M., Cotta-Ramusino, C., Pellicioli, A. & Liberi, G. The DNA replication checkpoint response stabilizes stalled replication forks. *Nature* **412**, 557-561 (2001).

14. Gilson, E. & Géli, V. How telomeres are replicated. *Nat. Rev. Mol. Cell Biol*. **8**, 825–838 (2007).

15. Verdun, R. E. & Karlseder, J. Replication and protection of telomeres. *Nature* **447**, 924–931 (2007).

16. Nowell, P. C. The clonal evolution of tumor cell populations. Science 194, 23–28 (1976).

17. Vesely, M. D., Kershaw, M. H., Schreiber, R. D. & Smyth, M. J. Natural Innate and Adaptive Immunity to Cancer. *Annu. Rev. Immunol*. **29**, 235–271 (2011).

18. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306 (2012).

19. Brem, H. & Folkman, J. Inhibition of tumor angiogenesis mediated by cartilage. *J. Exp. Med.* **141**, 427–439 (1975).

20. Van Dyke, T. & Jacks, T. Cancer Modeling in the Modern Era: Progress and Challenges. *Cell* **108**, 135–144 (2002).

21. Marusyk, A. & Polyak, K. Tumor heterogeneity: Causes and consequences. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* **1805**, 105–117 (2010).

22. Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T. & Campbell, P. J. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*. **45**, D777–D783 (2017).

23. Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A. E., Ristolainen, H., Hänninen, U. A., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J.-P., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., Enge, M., Taipale, J. & Aaltonen, L. A. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet*. **47**, 818–821 (2015).

24. Aavikko, M., Li, S.-P., Saarinen, S., Alhopuro, P., Kaasinen, E., Morgunova, E., Li, Y., Vesanen, K., Smith, M. J., Evans, D. G. R., Pöyhönen, M., Kiuru, A., Auvinen, A., Aaltonen, L. A., Taipale, J. & Vahteristo, P. Loss of SUFU Function in Familial Multiple Meningioma. *Am. J. Hum. Genet*. **91**, 520–526 (2012).

25. Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat. Rev. Cancer* **16**, 483–493 (2016).

26. Moynahan, M. E., Chiu, J. W., Koller, B. H. & Jasin, M. Brca1 Controls Homology-Directed DNA Repair. *Mol. Cell* **4**, 511–518 (1999).

27. Soussi, T. & Béroud, C. Assessing TP53 status in human tumours to evaluate clinical outcome. *Nat. Rev. Cancer* **1**, 233–239 (2001).

28. Yadav, V. K. & De, S. An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Brief. Bioinform*. **16**, 232–241 (2015).

29. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med*. **50**, 96 (2018).

30. Qin, Y., Feng, H., Chen, M., Wu, H. & Zheng, X. InfiniumPurify: An R package for estimating and accounting for tumor purity in cancer methylation research. *Genes Dis* **5**, 43–45 (2018).

31. Di Palma, S. & Bodenmiller, B. Unraveling cell populations in tumors by single-cell mass cytometry. *Curr. Opin. Biotechnol*. **31**, 122–129 (2015).

32. do Valle, Í. F., Giampieri, E., Simonetti, G., Padella, A., Manfrini, M., Ferrari, A., Papayannidis, C., Zironi, I., Garonzi, M., Bernardi, S., Delledonne, M., Martinelli, G., Remondini, D. & Castellani, G. Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. *BMC Bioinformatics* **17**, 341 (2016).

33. Snape, K., Ruark, E., Tarpey, P., Renwick, A., Turnbull, C., Seal, S., Murray, A., Hanks, S., Douglas, J., Stratton, M. R. & Rahman, N. Predisposition gene identification in common cancers by exome sequencing: insights from familial breast cancer. *Breast Cancer Res. Treat*. **134**, 429–433 (2012).

34. Kilpivaara, O. & Aaltonen, L. A. Diagnostic cancer genome sequencing and the contribution of germline variants. *Science* **339**, 1559–1562 (2013).

35. Klein, C., Chuang, R., Marras, C. & Lang, A. E. The curious case of phenocopies in families with genetic Parkinson's disease. *Mov. Disord*. **26**, 1793–1802 (2011).

36. Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R. & Lander, E. S. Searching for missing heritability: designing

rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E455–64 (2014).

37. Begum, F., Ghosh, D., Tseng, G. C. & Feingold, E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Research* **40**, 3777–3784 (2012).

38. Elkon, R. & Agami, R. Characterization of noncoding regulatory DNA in the human genome. *Nat. Biotechnol*. **35**, 732–746 (2017).

39. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. & Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet*. **101**, 5–22 (2017).

40. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. **42**, D1001–6 (2014).

41. Keighley, M. R. B. Gastrointestinal cancers in Europe. *Alimentary Pharmacology and Therapeutics* **18**, 7–30 (2003).

42. Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A. & Bray, F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut* **66**, 683–691 (2017).

43. Center, M. M., Jemal, A., Smith, R. A. & Ward, E. Worldwide variations in colorectal cancer. *CA Cancer J. Clin*. **59**, 366–378 (2009).

44. Enzinger, P. C. & Mayer, R. J. Esophageal cancer. *N. Engl. J. Med*. **349**, 2241–2252 (2003).

45. Chen, T., Cheng, H., Chen, X., Yuan, Z., Yang, X., Zhuang, M., Lu, M., Jin, L. & Ye, W. Family history of esophageal cancer increases the risk of esophageal squamous cell carcinoma. *Sci. Rep*. **5**, 16038 (2015).

46. Akbari, M. R., Malekzadeh, R., Nasrollahzadeh, D., Amanian, D., Sun, P., Islami, F., Sotoudeh, M., Semnani, S., Boffeta, P., Dawsey, S. M., Ghadirian, P. & Narod, S. A. Familial risks of esophageal cancer among the Turkmen population of the Caspian littoral of Iran. *Int. J. Cancer* **119**, 1047–1051 (2006).

47. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13**, 795–806 (2012).

48. Watson, J. D., Crick, F. & Others. A structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

49. Alberts, B. *Molecular Biology of the Cell*. (Garland Science, 2017).

50. Thurman, R. E., Rynes, E., Humbert, R. & Vierstra, J. The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

51. Grosberg, A. Y. How two meters of DNA fit into a cell nucleus: Polymer models with topological constraints and experimental data. *Polym. Sci. Ser. C: Rev.* **54**, 1–10 (2012).

52. Swygert, S. G. & Peterson, C. L. Chromatin dynamics: interplay between remodeling enzymes and histone modifications. *Biochim. Biophys. Acta* **1839**, 728–736 (2014).

53. Bouwman, B. A. M. & de Laat, W. Getting the genome in shape: the formation of loops, domains and compartments. *Genome Biol.* **16**, 154 (2015).

54. Meshorer, E. & Misteli, T. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat. Rev. Mol. Cell Biol*. **7**, 540–546 (2006).

55. Pertea, M. & Salzberg, S. L. Between a chicken and a grape: estimating the number of

human genes. *Genome Biol.* **11**, 206 (2010).

56. Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S. & Lifton, R. P. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A*. **106**, 19096–19101 (2009).

57. Bicknell, A. A., Cenik, C., Chua, H. N., Roth, F. P. & Moore, M. J. Introns in UTRs: why we should stop ignoring them. *Bioessays* **34**, 1025–1034 (2012).

58. Le Hir, H., Nott, A. & Moore, M. J. How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci*. **28**, 215–220 (2003).

59. Venables, J. P. Aberrant and Alternative Splicing in Cancer. *Cancer Research* **64**, 7647–7654 (2004).

60. Lytle, J. R., Yario, T. A. & Steitz, J. A. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proceedings of the National Academy of Sciences* **104**, 9667–9672 (2007).

61. Imielinski, M., Guo, G. & Meyerson, M. Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. *Cell* **168**, 460–472.e14 (2017).

62. Jiang, S., Zhang, H.-W., Lu, M.-H., He, X.-H., Li, Y., Gu, H., Liu, M.-F. & Wang, E.-D. MicroRNA-155 functions as an OncomiR in breast cancer by targeting the suppressor of cytokine signaling 1 gene. *Cancer Res*. **70**, 3119–3127 (2010).

63. Huang, Y.-W., Liu, J. C., Deatherage, D. E., Luo, J., Mutch, D. G., Goodfellow, P. J., Miller, D. S. & Huang, T. H.-M. Epigenetic repression of microRNA-129-2 leads to overexpression of SOX4 oncogene in endometrial cancer. *Cancer Res*. **69**, 9038–9046 (2009).

64. Bardelli, A. & Velculescu, V. E. Mutational analysis of gene families in human cancer. *Curr. Opin. Genet. Dev*. **15**, 5–12 (2005).

65. Bos, J. L. The ras gene family and human carcinogenesis. *Mutat. Res*. **195**, 255–271 (1988).

66. Deng, G., Bell, I., Crawley, S., Gum, J., Terdiman, J. P., Allen, B. A., Truta, B., Sleisenger, M. H. & Kim, Y. S. BRAF mutation is frequently present in sporadic colorectal cancer with methylated hMLH1, but not in hereditary nonpolyposis colorectal cancer. *Clin. Cancer Res*. **10**, 191–195 (2004).

67. Venkitaraman, A. R. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* **108**, 171–182 (2002).

68. Narod, S. A. & Foulkes, W. D. BRCA1 and BRCA2: 1994 and beyond. *Nat. Rev. Cancer* **4**, 665–676 (2004).

69. Hou, Y. & Lin, S. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PLoS One* **4**, e6978 (2009).

70. Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., Guigo, R., Hubbard, T., Kent, J., Lieb, J. D., Myers, R. M., Pazin, M. J., Ren, B., Stamatoyannopoulos, J. A., Weng, Z., White, K. P. & Hardison, R. C. Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U. S. A*. **111**, 6131–6138 (2014).

71. Morris, K. V. *Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers*

of Natural Selection. (Horizon Scientific Press, 2012).

72. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. & Weirauch, M. T. The Human Transcription Factors. *Cell* **175**, 598–599 (2018).

73. Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., David Hawkins, R., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E. & Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39**, 311–318 (2007).

74. Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L. & Garraway, L. A. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).

75. Heidenreich, B., Rachakonda, P. S., Hemminki, K. & Kumar, R. TERT promoter mutations in cancer development. *Curr. Opin. Genet. Dev.* **24**, 30–37 (2014).

76. Simpkins, S. B., Bocker, T., Swisher, E. M., Mutch, D. G., Gersell, D. J., Kovatich, A. J., Palazzo, J. P., Fishel, R. & Goodfellow, P. J. MLH1 promoter methylation and gene silencing is the primary cause of microsatellite instability in sporadic endometrial cancers. *Hum. Mol. Genet.* **8**, 661–666 (1999).

77. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

78. Ziller, M. J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., Boyle, P., Epstein, C. B., Bernstein, B. E., Lengauer, T., Gnirke, A. & Meissner, A. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet*. **7**, e1002389 (2011).

79. Meng, H. & Bartholomew, B. Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II. *J. Biol. Chem*. **293**, 13786–13794 (2018).

80. Zhang, Y., Wong, C.-H., Birnbaum, R. Y., Li, G., Favaro, R., Ngan, C. Y., Lim, J., Tai, E., Poh, H. M., Wong, E., Mulawadi, F. H., Sung, W.-K., Nicolis, S., Ahituv, N., Ruan, Y. & Wei, C.-L. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306–310 (2013).

81. Kim, T.-K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**, 948–959 (2015).

82. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer–promoter interactome in human cells. *Proc. Natl. Acad. Sci. U. S. A*. **111**, E2191–E2199 (2014).

83. Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.-L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K. I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M. J., Cheung, E., Liu, E., Sung, W.-K., Snyder, M. & Ruan, Y. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).

84. Pott, S. & Lieb, J. D. What are super-enhancers? *Nat. Genet*. **47**, 8 (2014).

85. Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I. & Young, R. A. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).

86. Sur, I. K., Hallikas, O., Vähärautio, A., Yan, J., Turunen, M., Enge, M., Taipale, M., Karhu, A., Aaltonen, L. A. & Taipale, J. Mice lacking a Myc enhancer that includes human

SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360–1363 (2012).

87. Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Niittymäki, I., Mecklin, J.-P., Järvinen, H., Ristimäki, A., Di-Bernardo, M., East, P., Carvajal-Carmona, L., Houlston, R. S., Tomlinson, I., Palin, K., Ukkonen, E., Karhu, A., Taipale, J. & Aaltonen, L. A. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet*. **41**, 885–890 (2009).

88. Rubio, E. D., Reiss, D. J., Welcsh, P. L., Disteche, C. M., Filippova, G. N., Baliga, N. S., Aebersold, R., Ranish, J. A. & Krumm, A. CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. U. S. A*. **105**, 8309–8314 (2008).

89. Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K. & Peters, J.-M. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796–801 (2008).

90. Ong, C. T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet*. **15**, 234-246 (2014).

91. Davidson, I. F., Bauer, B., Goetz, D., Tang, W., Wutz, G. & Peters, J.-M. DNA loop extrusion by human cohesin. *Science* **366**, 1338-1345 (2019).

92. Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M. J., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P. & Peters, J.-M. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J*. **36**, 3573–3599 (2017).

93. Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gülsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. A., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B. & Gilbert, D. M. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402–405 (2014).

94. de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., Splinter, E., Wijchers, P. J., Krijger, P. H. L. & de Laat, W. CTCF Binding Polarity Determines Chromatin Looping. *Mol. Cell* **60**, 676–684 (2015).

95. Ghirlando, R. & Felsenfeld, G. CTCF: making the right connections. *Genes Dev*. **30**, 881–891 (2016).

96. Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., Suvà, M. L. & Bernstein, B. E. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).

97. Condit, C. M., Achter, P. J., Lauer, I. & Sefcovic, E. The changing meanings of 'mutation:' A contextualized study of public discourse. *Hum. Mutat*. **19**, 69–75 (2002).

98. Baralle, D. & Baralle, M. Splicing in action: assessing disease causing sequence changes. *J. Med. Genet*. **42**, 737–748 (2005).

99. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr & Kinzler, K. W. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

100. Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M., Nord, A. S., Kusenda, M., Malhotra, D., Bhandari, A., Stray, S. M., Rippey, C. F., Roccanova, P., Makarov, V., Lakshmi, B., Findling, R. L., Sikich, L., Stromberg, T., Merriman, B., Gogtay, N., Butler, P., Eckstrand, K., Noory, L., Gochman, P., Long, R., Chen,

Z., Davis, S., Baker, C., Eichler, E. E., Meltzer, P. S., Nelson, S. F., Singleton, A. B., Lee, M. K., Rapoport, J. L., King, M.-C. & Sebat, J. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).

101. Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A. & Mundlos, S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

102. Jones, M. H. & Nakamura, Y. Detection of loss of heterozygosity at the human TP53 locus using a dinucleotide repeat polymorphism. *Genes Chromosomes Cancer* **5**, 89–90 (1992).

103. Feuk, L., Marshall, C. R., Wintle, R. F. & Scherer, S. W. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet*. **15** Spec No 1, R57–66 (2006).

104. Chesi, M., Nardini, E., Brents, L. A., Schröck, E., Ried, T., Kuehl, W. M. & Bergsagel, P. L. Frequent translocation t(4;14)(p16.3;q32.3) in multiple myeloma is associated with increased expression and activating mutations of fibroblast growth factor receptor 3. *Nat. Genet.* **16**, 260–264 (1997).

105. Mehine, M., Kaasinen, E., Mäkinen, N., Katainen, R., Kämpjärvi, K., Pitkänen, E., Heinonen, H.-R., Bützow, R., Kilpivaara, O., Kuosmanen, A., Ristolainen, H., Gentile, M., Sjöberg, J., Vahteristo, P. & Aaltonen, L. A. Characterization of uterine leiomyomas by whole-genome sequencing. *N. Engl. J. Med*. **369**, 43–53 (2013).

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

107. Ha, T. Probing nature's nanomachines one molecule at a time. *Biophys. J*. **110**, 1004–1007 (2016).

108. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

109. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* **33**, 136-143 (2012).

110. Liu, L., De, S. & Michor, F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun*. **4**, 1502 (2013).

111. Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M. & Sunyaev, S. R. Human mutation rate associated with DNA replication timing. *Nat. Genet*. **41**, 393–395 (2009).

112. Du, Q., Bert, S. A., Armstrong, N. J., Caldon, C. E., Song, J. Z., Nair, S. S., Gould, C. M., Luu, P.-L., Peters, T., Khoury, A., Qu, W., Zotenko, E., Stirzaker, C. & Clark, S. J. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nat. Commun*. **10**, 416 (2019).

113. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).

114. Papadopoulou, C., Guilbaud, G., Schiavone, D. & Sale, J. E. Nucleotide Pool Depletion

Induces G-Quadruplex-Dependent Perturbation of Gene Expression. *Cell Rep*. **13**, 2491–2503 (2015).

115. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet*. **15**, 585–598 (2014).

116. Maura, F., Degasperi, A., Nadeu, F., Leongamornlert, D., Davies, H., Moore, L., Royo, R., Ziccheddu, B., Puente, X. S., Avet-Loiseau, H., Campbell, P. J., Nik-Zainal, S., Campo, E., Munshi, N. & Bolli, N. A practical guide for mutational signature analysis in hematological malignancies. *Nat. Commun*. **10**, 2969 (2019).

117. Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., Fostel, J. L., Friedrich, D. C., Perrin, D., Dionne, D., Kim, S., Gabriel, S. B., Lander, E. S., Fisher, S. & Getz, G. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res*. **41**, e67 (2013).

118. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3**, 246–259 (2013).

119. Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., Gomez, C., Degasperi, A., Harris, R., Jackson, S. P., Arlt, V. M., Phillips, D. H. & Nik-Zainal, S. A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821–836.e16 (2019).

120. Mutational Signatures Working Group, P. The repertoire of mutational signatures in human cancer. *BioRxiv,* 322859 (2018).

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

122. Welch, J. S., Ley, T. J., Link, D. C., Miller, C. A., Larson, D. E., Koboldt, D. C., Wartman, L. D., Lamprecht, T. L., Liu, F., Xia, J., Kandoth, C., Fulton, R. S., McLellan, M. D., Dooling, D. J., Wallis, J. W., Chen, K., Harris, C. C., Schmidt, H. K., Kalicki-Veizer, J. M., Lu, C., Zhang, Q., Lin, L., O'Laughlin, M. D., McMichael, J. F., Delehaunty, K. D., Fulton, L. A., Magrini, V. J., McGrath, S. D., Demeter, R. T., Vickery, T. L., Hundal, J., Cook, L. L., Swift, G. W., Reed, J. P., Alldredge, P. A., Wylie, T. N., Walker, J. R., Watson, M. A., Heath, S. E., Shannon, W. D., Varghese, N., Nagarajan, R., Payton, J. E., Baty, J. D., Kulkarni, S., Klco, J. M., Tomasson, M. H., Westervelt, P., Walter, M. J., Graubert, T. A., DiPersio, J. F., Ding, L., Mardis, E. R. & Wilson, R. K. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).

123. Yong, W.-S., Hsu, F.-M. & Chen, P.-Y. Profiling genome-wide DNA methylation. *Epigenetics & Chromatin* **9**, (2016).

124. Tomasetti, C., Poling, J., Roberts, N. J., London, N. R., Jr, Pittman, M. E., Haffner, M. C., Rizzo, A., Baras, A., Karim, B., Kim, A., Heaphy, C. M., Meeker, A. K., Hruban, R. H., Iacobuzio-Donahue, C. A. & Vogelstein, B. Cell division rates decrease with age, providing a potential explanation for the age-dependent deceleration in cancer incidence. *Proc. Natl. Acad. Sci. U. S. A*. **116**, 20482–20488 (2019).

125. Collins, F. S. Implications of the Human Genome Project for Medical Science. *JAMA* **285**, 540 (2001).

126. Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., Harden, G., Hubbard, T., Pelan, S., Simpson, J. T., Threadgold, G., Torrance, J., Wood, J. M., Clarke, L., Koren, S., Boitano, M., Peluso, P., Li, H., Chin, C.-S., Phillippy, A. M., Durbin, R., Wilson, R. K., Flicek, P., Eichler, E. E. & Church, D. M. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res*. **27**, 849–864 (2017).

127. Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L. & Scherer, S. W. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol*. **11**, R52 (2010).

128. Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J. & Clamp, M. The Ensembl automatic gene annotation system. *Genome Res*. **14**, 942–950 (2004).

129. Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R. & Hubbard, T. J. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. **22**, 1760–1774 (2012).

130. Mount, S. M. Genomic Sequence, Splicing, and Gene Annotation. *The American Journal of Human Genetics* **67**, 788–792 (2000).

131. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet*. **34**, 666–681 (2018).

132. Stranneheim, H. & Lundeberg, J. Stepping stones in DNA sequencing. *Biotechnol. J*. **7**, 1063–1073 (2012).

133. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

134. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

135. Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V. & Gibrat, J.-F. Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. *Journal of Computational Biology* **19**, 796–813 (2012).

136. Burrows, M., Wheeler, D. J. A block-sorting lossless data compression algorithm. (1994).

137. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V.,

Altshuler, D., Gabriel, S. & DePristo, M. A. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–33 (2013).

138. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).

139. Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., Littin, R., Nohzadeh-Malakshah, S., Rathod, M., Ware, D., Trigg, L. & De La Vega, F. M. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J. Comput. Biol*. **21**, 405–419 (2014).

140. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol*. **31**, 213–219 (2013).

141. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* **10**, 669–680 (2009).

142. Nelson, J. D., Denisenko, O. & Bomsztyk, K. Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nat. Protoc*. **1**, 179–185 (2006).

143. Rossi, M. J., Lai, W. K. M. & Pugh, B. F. Simplified ChIP-exo assays. *Nat. Commun*. **9**, 2842 (2018).

144. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).

145. Cuddapah, S., Jothi, R., Schones, D. E., Roh, T.-Y., Cui, K. & Zhao, K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res*. **19**, 24–32 (2009).

146. Poulos, R. C., Thoms, J. A. I., Guan, Y. F., Unnikrishnan, A., Pimanda, J. E. & Wong, J. W. H. Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the *Motif. Cell Rep*. **17**, 2865–2872 (2016).

147. Teif, V. B., Vainshtein, Y., Caudron-Herger, M., Mallm, J.-P., Marth, C., Höfer, T. & Rippe, K. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol*. **19**, 1185–1192 (2012).

148. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).

149. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).

150. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E. & Taipale, J. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*. **20**, 861–873 (2010).

151. Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A. & Tegnér, J. Data integration in the era of omics: current and future challenges. *BMC Syst. Biol. 8 Suppl* **2**, I1 (2014).

152. Palsson, B. & Zengler, K. The challenges of integrating multi-omic data sets. *Nat. Chem. Biol*. **6**, 787–789 (2010).

153. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. Integrative genomics viewer. *Nat. Biotechnol*. **29**, 24–26 (2011).

154. Karczewski, K. J., Francioli, L. C., Tiao, G. & Cummings, B. B. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, 531210 (2019).

155. Lim, E. T., Würtz, P., Havulinna, A. S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., Chan, Y., Salem, R. M., Lek, M., Flannick, J., Sim, X., Manning, A., Ladenvall, C., Bumpstead, S., Hämäläinen, E., Aalto, K., Maksimow, M., Salmi, M., Blankenberg, S., Ardissino, D., Shah, S., Horne, B., McPherson, R., Hovingh, G. K., Reilly, M. P., Watkins, H., Goel, A., Farrall, M., Girelli, D., Reiner, A. P., Stitziel, N. O., Kathiresan, S., Gabriel, S., Barrett, J. C., Lehtimäki, T., Laakso, M., Groop, L., Kaprio, J., Perola, M., McCarthy, M. I., Boehnke, M., Altshuler, D. M., Lindgren, C. M., Hirschhorn, J. N., Metspalu, A., Freimer, N. B., Zeller, T., Jalkanen, S., Koskinen, S., Raitakari, O., Durbin, R., MacArthur, D. G., Salomaa, V., Ripatti, S., Daly, M. J., Palotie, A. & Sequencing Initiative Suomi (SISu) Project. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet*. **10**, e1004494 (2014).

156. Leinonen, M. K., Miettinen, J., Heikkinen, S., Pitkäniemi, J. & Malila, N. Quality measures of the population-based Finnish Cancer Registry indicate sound data quality for solid malignant tumours. *Eur. J. Cancer* **77**, 31–39 (2017).

157. Gui, J., Tosteson, T. D. & Borsuk, M. Weighted multiple testing procedures for genomic studies. *BioData Min*. **5**, 4 (2012).

158. Wöste, M. & Dugas, M. VIPER: a web application for rapid expert review of variant calls. *Bioinformatics* **34**, 1928–1929 (2018).

159. Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I. B., Calderaro, J., Bioulac-Sage, P., Letexier, M., Degos, F., Clément, B., Balabaud, C., Chevet, E., Laurent, A., Couchy, G., Letouzé, E., Calvo, F. & Zucman-Rossi, J. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet*. **44**, 694–698 (2012).

160. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).

161. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet*. **Chapter 7**, Unit7.20 (2013).

162. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. **31**, 3812–3814 (2003).

163. Balasubramanian, S., Fu, Y., Pawashe, M., McGillivray, P., Jin, M., Liu, J., Karczewski, K. J., MacArthur, D. G. & Gerstein, M. Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat. Commun.* **8**, 382 (2017).

164. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. **47**, D886–D894 (2019).

165. Cooper, G. M., Stone, E. A., Asimenos, G., NISC Comparative Sequencing Program, Green, E. D., Batzoglou, S. & Sidow, A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. **15**, 901–913 (2005).

166. Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E. & Taipale, J. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).

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

168. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Research* **40**, e169–e169 (2012).

169. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol*. **17**, 128 (2016).

170. Evans, P., Avey, S., Kong, Y. & Krauthammer, M. Adjusting for background mutation frequency biases improves the identification of cancer driver genes. *IEEE Trans. Nanobioscience* **12**, 150–157 (2013).

171. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. **38**, e164 (2010).

172. Aaltonen, L. A., Salovaara, R., Kristo, P., Canzian, F., Hemminki, A., Peltomäki, P., Chadwick, R. B., Kääriäinen, H., Eskelinen, M., Järvinen, H., Mecklin, J. P. & de la Chapelle, A. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N. Engl. J. Med*. **338**, 1481–1487 (1998).

173. Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V. & Korbel, J. O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

174. Watanabe, M., Hashida, S., Yamamoto, H., Matsubara, T., Ohtsuka, T., Suzawa, K., Maki, Y., Soh, J., Asano, H., Tsukuda, K., Toyooka, S. & Miyoshi, S. Estimation of age-related DNA degradation from formalin-fixed and paraffin-embedded tissue according to the extraction methods. *Exp. Ther. Med*. **14**, 2683–2688 (2017).

175. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

176. Guo, Y., Mahony, S. & Gifford, D. K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* **8**, e1002638 (2012).

177. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T. & Taipale, J. DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).

178. Yokoyama, T. T. & Kasahara, M. Visualization tools for human structural variations identified by whole-genome sequencing. *J. Hum. Genet.,* 1-12 (2019).

179. Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., Wong-Erasmus, M., Yao, L. & Kasprzyk, A. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database* **2011**, (2011).

180. Kusakabe, M., Kutomi, T., Watanabe, K., Emoto, N., Aki, N., Kage, H., Hamano, E., Kitagawa, H., Nagase, T., Sano, A., Yoshida, Y., Fukami, T., Murakawa, T., Nakajima, J., Takamoto, S., Ota, S., Fukayama, M., Yatomi, Y., Ohishi, N. & Takai, D. Identification of G0S2 as a gene frequently methylated in squamous lung cancer by combination of in silico and experimental approaches. *Int. J. Cancer* **126**, 1895–1902 (2010).

181. Gruel, N., Benhamo, V., Bhalshankar, J., Popova, T., Fréneaux, P., Arnould, L., Mariani, O., Stern, M.-H., Raynal, V., Sastre-Garau, X., Rouzier, R., Delattre, O. & Vincent-

Salomon, A. Polarity gene alterations in pure invasive micropapillary carcinomas of the breast. *Breast Cancer Res.* **16**, R46 (2014).

182. Huang, J., Hu, N., Goldstein, A. M. & Emmert-Buck, M. R. High frequency allelic loss on chromosome 17p13. 3–p11. 1 in esophageal squamous cell carcinomas from a high incidence area in northern China. *Carcinogenesis* **21**, 2019-2026 (2000).

183. Datta, S., Datta, S., Kim, S., Chakraborty, S. & Gill, R. S. Statistical Analyses of Next Generation Sequence Data: A Partial Overview. *J. Proteomics Bioinform*. **3**, 183–190 (2010).

184. Moore, J. H., Asselbergs, F. W. & Williams, S. M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **26**, 445–455 (2010).

185. Bessarabova, M., Ishkin, A., JeBailey, L., Nikolskaya, T. & Nikolsky, Y. Knowledge-based analysis of proteomics data. *BMC Bioinformatics* **13** Suppl 16, S13 (2012).

186. O'Donoghue, S. I., Baldi, B. F., Clark, S. J., Darling, A. E., Hogan, J. M., Kaur, S., Maier-Hein, L., McCarthy, D. J., Moore, W. J., Stenau, E., Swedlow, J. R., Vuong, J. & Procter, J. B. Visualization of Biomedical Data. *Annu. Rev. Biomed. Data Sci*. **1**, 275–304 (2018).

187. Kallio, M. A., Tuimala, J. T., Hupponen, T., Klemelä, P., Gentile, M., Scheinin, I., Koski, M., Käki, J. & Korpelainen, E. I. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12**, 507 (2011).

188. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

189. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

190. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. & Marra, M. A. Circos: an information aesthetic for comparative genomics. *Genome Res*. **19**, 1639–1645 (2009).

191. Hiltemann, S., Jenster, G., Trapman, J., van der Spek, P. & Stubbs, A. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Res*. **25**, 1382–1390 (2015).

192. Musumeci, L., Arthur, J. W., Cheung, F. S. G., Hoque, A., Lippman, S. & Reichardt, J. K. V. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum. Mutat*. **31**, 67–73 (2010).

193. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S. & Getz, G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).

194. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

195. Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genet*. **12**, e1006207 (2016).

196. Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J. & Young, R. A. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).

197. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).

198. Sima, J., Chakraborty, A., Dileep, V., Michalski, M., Klein, K. N., Holcomb, N. P., Turner, J. L., Paulsen, M. T., Rivera-Mulia, J. C., Trevilla-Garcia, C., Bartlett, D. A., Zhao, P. A., Washburn, B. K., Nora, E. P., Kraft, K., Mundlos, S., Bruneau, B. G., Ljungman, M., Fraser, P., Ay, F. & Gilbert, D. M. Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. *Cell* **176**, 816–830.e18 (2019).

199. Markowitz, S. D. & Bertagnolli, M. M. Molecular Basis of Colorectal Cancer. *N. Engl. J. Med.* **361**, 2449–2460 (2009).

200. Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E. O. & Reinberg, D. CTCF-mediated topological boundaries during development foster appropriate gene regulation. *Genes Dev*. **30**, 2657–2662 (2016).

201. Bergström, R., Whitehead, J., Kurukuti, S. & Ohlsson, R. CTCF regulates asynchronous replication of the imprinted H19/Igf2 domain. *Cell Cycle* **6**, 450–454 (2007).

202. Spivakov, M. Spurious transcription factor binding: Non-functional or genetically redundant? *Bioessays* **36**, 798–806 (2014).

203. Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B. J., Afzal, S. Y., Lee, E. A., Zhu, Y., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Akiyama, J. A., Afzal, V., Lopez-Rios, J., Dickel, D. E., Visel, A. & Pennacchio, L. A. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).

204. Orlando, G., Law, P. J., Cornish, A. J., Dobbins, S. E., Chubb, D., Broderick, P., Litchfield, K., Hariri, F., Pastinen, T., Osborne, C. S., Taipale, J. & Houlston, R. S. Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer. *Nat. Genet*. **50**, 1375–1380 (2018).

205. Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., LeProust, E., Follows, G. A., Fraser, P., Luscombe, N. M. & Osborne, C. S. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet*. **47**, 598–606 (2015).