Transcriptomics analysis and its applications in cancer

Alejandra Cervera Taboada

Research Program in Systems Oncology
Research Programs Unit
Biochemistry and Developmental Biology
Medicum
Faculty of Medicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
on the 14th of December 2020, at 15 o'clock.
The defence is open for audience through remote access.

Helsinki 2020

Supervisor

Sampsa Hautaniemi, DTech, Professor
Research Program in Systems Oncology
Research Programs Unit
Biochemistry and Developmental Biology
Medicum
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Reviewers appointed by the Faculty

Rolf Skotheim, PhD
Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet
Oslo, Norway

Joaquin Dopazo, PhD, Director Clinical Bioinformatics Area
Fundación Progreso y Salud
Universidad de Valencia
Sevilla, Spain

Opponent appointed by the Faculty

Carla Daniela Robles Espinoza, PhD
International Laboratory for Human Genome Research, National Autonomous University of Mexico
Juriquilla, Mexico

Faculty of Medicine
Doctoral Programme in Biomedicine
ISBN 978-951-51-6865-8 (paperback)
ISBN 978-951-51-6866-5 (PDF)
http://ethesis.helsinki.fi
Unigrafia Oy
Helsinki 2020

The Faculty of Medicine uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

"We reveal ourselves in the metaphors we choose for depicting the cosmos in miniature."
- Stephen Jay Gould

To Antonio,

Abstract

Cancer is a collection of diseases that combined are one of the leading causes of deaths worldwide. Although great strides have been made in finding cures for certain cancers, the heterogeneity caused by both the tissue in which cancer originates and the mutations acquired in the cell's DNA results in unsuccessful treatments for some patients. The genetic alterations caused by carcinogenics or by random mutations acquired during normal cell division promotes changes in the cell's metabolism. These changes are usually reflected in abnormal gene expression that can be studied to understand the underlying mechanisms giving rise to cancer as well as suggest treatments that can exploit each tumor's specific vulnerabilities.

RNA-Seq is a technology that allows the identification and quantification of the genes that are being expressed inside the cell in a given moment. RNA-Seq has several characteristics and advantages that allow a diversity of applications to exist. For example, apart from quantifying gene expression, it can be used to detect different variants of the same gene, has base pair resolution which is informative of the gene sequence, and can also be used to quantify other RNA molecules besides messenger RNA (mRNA), such as microRNAs.

The two main aims of this work are to provide computational methods for data analysis of RNA-Seq and to show specific applications of RNA-Seq that can shed light into cancer mechanisms. In Publications I and IV we developed the Sequence Processesing Integration and Analysis (SePIA) and the Fusion Gene Integration (FUNGI) toolsets that facilitate the creation of reproducible pipelines for investigating different aspects of the cancer transcriptome. SePIA's utility is showcased with the analysis of datasets from two public data repositories. One of the analysis shows a standard RNA-Seq analysis, while the second one produced a pipeline for mRNA-microRNA integration. The second toolset, FUNGI, is aimed specifically at finding reliable gene fusions with oncogenic potential. To demonstrate FUNGI's features, we analyzed 107 in-house samples and processed over 400 public samples from a public data repository. FUNGI allowed us to detect fusions in ovarian cancer with a higher prevalence than previously recognized. Additionally, we identified a fusion gene that has not been reported before in ovarian cancer, but that can be targeted with a drug currently in clinical trials. In Publication II we investigated the role of alternative splicing in diffuse large B-cell lymphoma and were able to show that isoform-level instead of gene-level is better at discriminating between subtypes. Additionally, specific isoforms, such as *APH1A*, *KCNH6*, and *ABCB1*, were correlated with survival. In Publication III, we used RNA-Seq to complement the phasing of genetic variants with somatic mutations in tumor suppressor genes. In this study we found enrichment of haplotype combinations that suggest that

haploinsufficiency of tumor suppressor genes is enriched in cancer patients.

SePIA and FUNGI are tools that can be used by the community to explore their datasets and contribute to the acquisition of knowledge in the field of cancer genetics with next generation sequencing. The applications of RNA-Seq studies included in this dissertation showed that RNA-Seq can be effectively used to aid in the classification of cancer subtypes, and that RNA-Seq can be used in combination with DNA sequencing to explore gene expression mediated by genetic variation in cancer.

Contents

Abbreviations

ABC	Activated B-cell
AS	Alternative splicing/spliced
bp	Base pair
CADD	Combined annotation dependent depletion
cDNA	Complimentary DNA
CGCI	Cancer Genome Characterization Initiative
DEEs	Differentially expressed exons
DEGs	Differentially expressed genes
DNA	Deoxyribonucleic acid
ENCODE	Encyclpedia of DNA elements
eQTL	Expression quantitative trait loci
FFPM	Fusion fragments per million total fragments
FPKM	Fragment per kilobase of transcript per million mapped reads
FUNGI	Fusion gene intergration
SePIA	Sequence processing integration and analysis
GATK	Genome Analysis Toolkit
GCB	Germinal center B-cell
GEO	Gene Expression Omnibus
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
GWAS	Genome wide association studies
JRC	Junction read count
MEAP	Multiple exon array preprocessing
mRNA	Messenger RNA
miRNA	MicroRNA
NYGC	New York Genome Center
OS	Overall survival
PFS	Progression-free survival
RNA	Ribonucleic acid
PCR	Polymerase chain reaction
qPCR	Quantitative PCR
RNA-Seq	RNA sequencing
snRNP	Small nuclear ribo nucelear proteins
siRNA	Silencing RNA
TCGA	The Cancer Genome Atlas
TPM	Transcripts per million
TSG	Tumor suppressor genes
TSS	Transcription start sites
WGS	Whole genome sequencing

Genes

AGO	Argonaute
BCL-2	B-cell lymphoma gene-2
BRCA	Breast cancer protein
CCNE1	Cyclin E1

FAK	Focal adhesion kinase
Rb	Retinoblastoma
RISC	RNA-induced silencing complex
TNFR	Tumor necrosis factor receptor

Cancer types

APL	Acute myeloid leukemia
BCLA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CML	Chronic myelogenous leukemia
DLBCL	Diffuse large B-cell lymphoma
EOC	Epithelial ovarian carcinoma
ESCA	Esopharingeal carcinoma
GBM	Glioblastoma multiforme
HGSOC	High-grade serous ovarian cancer
HNPCC	Hereditary nonpolyposis colorectal cancer
HNSC	Head and neck squamous cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUSC	Lung squamous cell carcinoma

Publications and author's contributions

Publication I Icay, K.*, Chen, P.*, **Cervera, A.***, Rantanen, V., Lehtonen, R., & Hautaniemi, S. (2016). SePIA: RNA and small RNA sequence processing, integration, and analysis. BioData Mining, 9(1), 20.

Publication II Leivonen, S.-K., Taskinen, M., **Cervera, A.**, Karjalainen-Lindsberg, M.-L., Delabie, J., Holte, H., Lehtonen, R., Hautaniemi, S., & Leppä, S. (2017). Alternative splicing discriminates molecular subtypes and has prognostic impact in diffuse large B-cell lymphoma. Blood Cancer Journal, 7(8), e596–e596.

Publication III Castel, S. E., **Cervera, A.**, Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., & Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nature Genetics, 50(9), 1327–1334.

Publication IV **Cervera, A.***, Rausio, H.*, Kähkönen, T., Andersson, N., Partel, G., Rantanen, V., Paciello, G., Ficarra, E., Hynninen, J., Hietanen, S., Carpén, O., Lehtonen, R., Hautaniemi, S., Huhtinen, K., Comprehensive gene fusion analysis pipeline reveals novel fusion genes in ovarian cancer. Submitted to Bioinformatics.

* equal contribution

Publications included in other thesis

Publication I was included in the thesis of Katherine Abigail Icay-Rouhiainen (Computational analysis of microRNAs in biomedicine, Helsinki 2018)

Author's contributions

Publication I Designed and developed components and pipelines for (mRNA) RNA-seq processing and analysis. Performed analysis for the manuscript. Contributed to the manuscript.

Publication II Collected and performed the differential exon usage analysis and the quantification of the validation data from CGCI. Contributed to the manuscript.

Publication III Performed the analysis for the cancer case of the study. Contributed to the manuscript.

Publication IV Designed the workflow and developed components for the fusion gene analysis toolset. Performed the analysis. Wrote the manuscript.

1 Introduction

Cancer is collection of diseases that combined are responsible for $\sim$10 million deaths per year in the world [1]. One of the main characteristics of cancer is the abnormal growth of cells that lead to death when vital organs' functions are compromised by the tumors. The standard hystopathological classification of cancer into different subtypes is based on the tissue of origin. Cancer is a highly heterogenous disease caused by both the difference in type of cell of origin and the array of distinct processes occurring inside cells within the same tumor.

The work presented in this thesis comprises the study of a variety of cancer datasets, but focused studies to better understand disease mechanisms were performed in two cancer types: diffuse large B-cell lymphoma (DLBCL) and high-grade serous ovarian cancer (HGSOC). The difference between these two cancers is considerable since DLBCL arises from blood cells and it can be considered a liquid tumor together with leukemias. In lymphomas the outgrowth of cells frequently concentrates at lymph nodes forming abnormal masses, which makes lymphoma also a particular case of solid tumors. On the other hand, HGSOC is a classical example of solid cancer called carcinoma, which means that the cell of origin is in the epithelial tissue. Despite the differences in cell of origin, the mechanisms behind the transformation of healthy cells into cancerous ones are often shared among cancers [2]. This allows us to study different cancers using similar techniques such as the identification of germline predispositions, somatic mutations, or abnormal gene expression.

The inherent heterogeneity of different cancers types augmented by the diversity of genetic abnormalities that accumulate in individual tumors have given rise to the field of personalized medicine in cancer. The hope of personalized or precision medicine is to tailor cancer treatments to specific genetic abnormalities of each individual's tumors. To be able to identify these possible drug targets, cancer research requires the study of increasingly larger datasets. Over the past two decades, advancements in high-throughput technologies have facilitated the creation of repositories of genetic data of both cancer samples and normal tissues. Examples of these repositories are The Cancer Genome Atlas (TCGA) [3] which houses over 20,000 datasets, Genotype-Tissue Expression (GTEx) [4] that provides expression quantification of 54 non-cancer tissues from 1000 individuals, Cancer Genome Characterization Initiative [5] focused on rare cancers, the Encyclpedia of DNA Elements (ENCODE) [6], the Human Cell Atlas [7], among many others. The amount of information that has been made accessible to cancer researchers is enormous and has required the development of computational methods that are efficient and reproducible to analyze these vast information resources.

Bioinformatics is the field of research that produces the infrastructure and algorithms that allow cancer genetic studies. The studies included in this dissertation are focused on the analysis of a particular technology developed for exploring sequences of expressed genes: RNA sequencing. Its main application is quantifying gene expression, but apart from allowing the measurement of other RNA species such as microRNAs, and not circumscribing the measurements to our current knowledge of the transcriptome, RNA-Seq has the enormous advantage of providing base pair resolution of long stretches of DNA. The multiple applications of RNA-Seq has spurred a development of algorithms and tools for the processing, interpretation and exploitation of this resource. In that sense, with Publication I, we contribute to the development of a set of workflows that facilitate the standard analysis of RNA-Seq datasets, such as differential gene expression, but also analysis of microRNAs and integration with their possible mRNA targets. In Publication II we used two different technologies, microarrays and next-generation sequencing, to study alternative splicing in DLBCL. In Publication III, we leveraged the advantages of RNA sequencing reads to complement phasing of germline genetic variation and somatic mutations and test if these haplotype combinations are enriched in cancer. In Publication IV we developed a workflow for studying a specific type of structural variation known as gene fusions and applied it to both an in-house and a large public ovarian cancer datasets.

2 Cancer

Every function taking place inside our cells is carefully regulated by a series of signals that tell the cell when to divide, what processes to initiate, what proteins to produce, and when to die [8][9]. The information the cell requires to synthetize the proteins that undertake all of the cells functions is encoded in the DNA. Cancer occurs when this master blueprint gets corrupted and the signals become scrambled.

Given that cells have many mechanisms in place to protect their genetic code, it is usually precisely a mutation in genes involved in safe guarding the DNAs integrity the first step in cancer development. Once the gate has been opened for a cell to start accumulating mutations, carcinogenesis can take place, and a normal cell evolves into a cancerous one. The main abilities that cells need to gain to give rise to tumors have been grouped into well defined categories, or hallmarks.

2.1 Hallmarks of cancer

In 2000, Hanahan and Weinberg published their landmark paper, *The Hallmarks of Cancer* [2], where they included six characteristics essential for cancer development. A decade later, four more hallmarks that are also considered fundamental aspects from cancer initiation to metastasis [10] were added to the list. A short description of each hallmark is included below.

Genome instability and mutation Alterations in the DNA damage response pathways permit mutations, caused by carcinogenics or by random errors during replication, to evade detection and prevail. The accelerated mutation rate that can happen when DNA repair is not operating allows the cells to evolve rapidly and start acquiring other tumor promoting characteristics.

Sustaining proliferation signal Normally cells receive cues to begin cell division from outside through transmembrane receptors. When a cell acquires the capacity of producing its own growth factors then a cell can start dividing without the need of external stimuli, which is a necessary step for tumor formation.

Evading growth suppressors To maintain tissue homeostasis, mechanisms exist to stop cells from proliferating without control. In cancer, cells acquire mutations that allow them to evade the growth suppressor signals and continue replicating.

Enabling replicative immortality Cell's telomeres[1] get shortened with every cell division. When telomeres become too short replicative senescence is induced

[1] A telomere is a region of repetitive nucleotide sequences at each end of a chromosome, which protects the end of the chromosome from deterioration

in the cell. Telomerase has been found to be upregulated in 85-90% of tumors [11] allowing the cell to extend telomeres and evade senescence. Cells that manage to evade senescence without increasing telomere lengths further induce genomic instability since chromosome ends become unprotected and fusions between genomic regions can occur.

Resisting cell death Apoptosis can be triggered in response to DNA damage, elevated levels of oncogenes, or cancer therapy. Cancerous cells accumulate mutations in pro-apoptotic pathways early in their development to be able to escape apoptosis which also renders them resistant to several cancer therapies.

Inducing angiogenesis Tumors require a steady nutrient supply to keep growing therefore creation of blood vessels to feed the tumor is necessary. Angiogenesis can be triggered by hypoxia inside the tumor or by alterations in genes controlling production of angiogenic regulators.

Activating invasion and metastasis Metastasis is a multistep process in which first cells need to be able to detach from their local environment, get into blood vessels (intravasation), travel and survive the pressure intact, exit the vessels at a different site (extravasation), and start forming new tumors by adapting to the new environment and sustaining proliferation again.

Avoiding immune destruction The immune system is capable of detecting tumors and eliminate them. Cancer cells can evade immune destruction by disabling components of the immune system that have been dispatched to neutralize them by secreting immunosuppressive factors.

Tumor promoting inflammation Tumors benefit from immune cells drawn to the tumor microenvironment. Inflammation can contribute by supplying the tumors with bioactive molecules such as growth factors, survival factors, proangiogenic factors, extracellular matrix-modifying enzymes that facilitate angiogenesis, invasion, and metastasis.

Deregulating cellular energetics Cancer cells rewire their metabolism to sustain proliferation. It has been observed that tumors preferentially over-utilize glucose to obtain energy, a pathway preferred by normal cells under anaerobic conditions (Warburg effect). The benefits to the tumor are still not completely understood but the rationale has shifted from thinking it was due to defective mitochondria or hypoxia to believe subproducts of glycolysis benefit the tumor in other ways [12].

A reference to each cancer hallmark with the main cellular process that are associated to them can be seen in Figure 1. The cancer enabling characteristics summed up by the so-called hallmarks are not exclusive of cancer. Many benign tumors share some of these abilities. Even metastasis is not unique to cancer,

since endometriosis[2], for example, also colonizes other organs [13]. Nevertheless, all of the functions described by the hallmarks are a necessary aspect of cancer development and their study help us better understand the cellular transformations that lead to cancer.

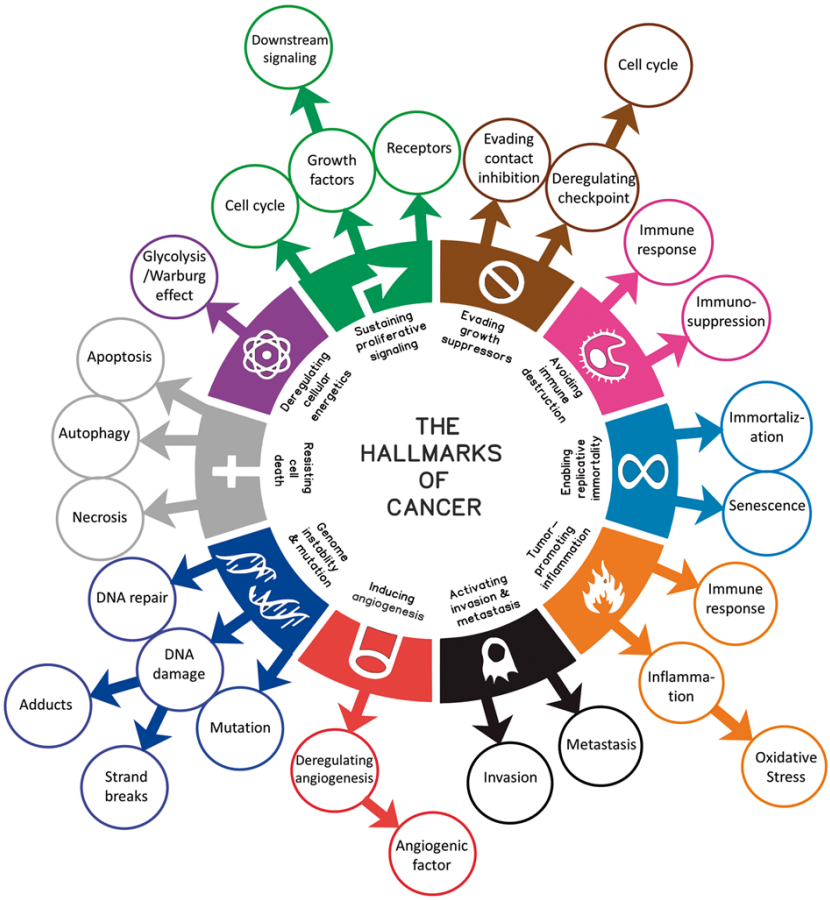

Figure 1: The Hallmarks of Cancer taxonomy [14]. The inner circle represents the main ten cancer hallmarks and the outer circles indicate the cellular processes associated with each cancer hallmark.

2.2 Tumor suppressor genes

Cancers are polygenic disorders; each of the hallmarks described above are regulated by different groups of genes. These genes can be broadly divided in two types based on their role being passive or active in cancer development: tumor suppressor

[2]Endometriosis is a condition in which cells similar to those in the endometrium, the layer of tissue that normally covers the inside of the uterus, grow outside it.

genes (TSGs) and oncogenes. TSGs are involved in DNA damage repair, inhibition of cell division, induction of apoptosis, and suppression of metastasis. Inactivation of TSGs is necessary for cancer progression. On the other hand, oncogenes belong to pathways that promote cellular growth. Oncogenes are genes that normally regulate cell cycle and proliferation, but when mutated or overexpressed lead to tumor formation.

Prototypic tumor suppressor genes are recessive, requiring "two-hit" inactivation of both alleles [15] for the cell to bypass the protection against cancer transformation conferred by TSGs. Studies have shown that haploinsufficiency, when not enough copies of a gene are produced to permit the cell to achieve the standard phenotype, also contributes to the development and progression of many cancers [16]. In Publication III we observed that cancer patients show enrichment of haplotype combinations—germline genetic variants that regulate expression in the same allele as somatic mutations—that can lead to haploinsufficiency of TSGs.

The main functions of tumor suppressor genes are 1) suppressing proliferation, 2) repair DNA damage, 3) induction of apoptosis and 4) inhibition of metastasis. An overview of canonical TSGs and their role in cancer is given below.

Suppression of cell division Both retinoblastoma (*Rb*), the first discovered TSG, and *p53*, the most mutated TSG in cancer, act as regulators of the cell cycle [17]. During replication, transcription factor E2F binds to the DNA to activate DNA replication enzymes and allow the cell to progress from G1 to S phase. Rb suppresses DNA replication by binding directly to E2F. Mutations can deactivate Rb by impeding its binding to E2F [18]. On the other hand, p53 activates a gene that halts the cell cycle, if p53 is non-functional due to mutations the cell cycle is not arrested.

DNA damage repair Most of the genes involved in detecting and correcting mistakes or copying errors in the DNA are considered TSGs. MutS (MSH2, MSH3, MSH6) and MutL (MLH1, MLH3, PMS1, PMS2) proteins recognize and repair point mutations or small indels in the genome. ATM is a sensor of break damage and its function is to phosphorilate p53 to impede the MDM2 binding to it [19]. Breast cancer protein (BRCA) with NBS1 play key functions in homologous recombination and nonhomologous end joining [20]. Homologous recombination is the mechanism used by the cell to repair double strand breaks in which the sequence from the other chromosome is used as template to repair the damaged one. Nonhomologous end joining is also used to repair double strand breaks, but in this case an homologous sequence is not used; in nonhomologous end joining the overhanging sequences from each side of the break are simply joined back together. Inherited BRCA1 and BRCA2 mutations are associated with familial breast and ovarian cancers

[21]. Lynch syndrome is the most common genetic colorectal carcinoma syndrome in which 70-85% of cases are caused by mutations in MLH1 or MSH2 [22].

Induction of apoptosis Programmed cell death, or apoptosis, is regulated by many pathways, two of which are mediated by p53 [23]. Apoptosis triggered by stress conditions, such as cytokine deprivation, ER stress or DNA damage is called the intrinsic pathway and it depends on p53 promoting upregulation of pro-apoptotic members of the BCL-2 protein family. The extrinsic pathway, or death receptor, activates the caspase cascade by ligation of members of the tumor necrosis factor receptor (TNFR) family bearing an intracellular death domain. PTEN utilizes an alternative mechanism to promote apoptosis. PTEN is involved in cell cycle arrest and apoptosis through negatively regulating the survival signaling mediated by PIP3 kinase (PI3K) and its down-stream target, a serine/threonine kinase AKT (also called protein kinase B) [24]. Activation of AKT regulates the function, by phosphorylation activation or suppression, of a broad array of proteins involved in cell growth, proliferation, motility, adhesion, neovascularization, and cell death. Inhibition of the PI3K-AKT signalling pathway is an active area of clinical development [25]. In Publication IV we identified several fusion genes in ovarian cancer samples where one of the genes involved in the fusion belongs to the PI3K-AKT pathway.

Metastasis Two important TSGs that prevent metastasis are metastin and breast cancer metastasis suppressor 1 (BRMS1). Metastatin inhibits metastasis by increasing the activity of focal adheshion kinase (FAK/PTK2) which prevents cells from migrating and is usually overexpressed in ovarian cancer. In Publication IV we identified two patients with a fusion gene involving PTK2. BRMS1 suppresses metastasis in multiple tumor types including ovarian, bladder, melanoma and non-small cell lung carcinoma.

2.3 Oncogenes

Proto-oncogenes are genes involved cell growth and proliferation pathways in normal cells, but when mutated become oncogenes. Usually the mutations are dominant in nature; a mutation in one of the alleles is enough to disregulate their functions and promote cancer [26]. The mutant proteins often retain some of their capabilities but are no longer sensitive to the controls that regulate the wild-type protein form. One of the most notorious examples of oncogenes is B-cell lymphoma gene-2 (BCL-2) which is a regulator of cell death and acts both as inhibitor or promotor of apoptosis. When overexpressed, BCL-2, allows continued division of

cancerous cells [27].

2.4 Diffuse large B-cell lymphoma

Lymphoma is a form of cancer that affects the lymphocytes. Lymphocytes are small white cells called leukocytes of which two main types exist: B cells and T cells. Both types originate from stem cells in the bone marrow, but B cells stay in the bone marrow while T cells travel to the thymus. Lymphocytes are a main component of the immune system, and the job of B cells is to produce antibodies, while T cells directly destroy bacteria or cells infected by viruses. Both B and T cells can become cancerous, but B cells account for 90% of all lymphoma cases [28].

Lymphoma is classified in two major groups based on the presence or lack of Reed-Sternberg cells into Hodgkin or non Hodgkin lymphoma. Diffuse large B-cell lymphoma (DLBCL) is the most common type of non-Hodgkin lymphoma and the most common lymphoid neoplasm in adults, with an average age of diagnosis around 70 [28]. DLBCL can develop in the lymph nodes or in "extranodal sites" (areas outside the lymph nodes) being most common in the gastrointestinal tract, testes, thyroid, skin, breast, bone, and brain. It may present localized (in one spot) or generalized (spread throughout the body). DLBCL is an aggressive (fast-growing) disease, where $\sim$60% of patients respond favorably to a rituximab and anthracycline-based CHOP or CHOP-like chemoimmunotherapy[3] [29], which is the standard therapy.

DLBCL can be further classified into two molecular subtypes: germinal center B-cell-like (GCB) and activated B-cell-like (ABC). These subtypes were identified after a systematic characterization of DLBCL tumors on a panel of genes involved in lymphocyte development and activation showed that gene expression patterns were indicative of different stages of B-cell differentiation [30]. Gene-expression profiling can define ABC and GCB subgroups of DLBCL, leaving approximately 10 to 20% of cases unclassified or unknown subtype [30]. Main differences between the subtypes include translocations in BCL-2 which are more frequent in GCB, while ABC tumors show activation of B-cell receptor-dependent nuclear factor κB (NF-κB). The identification of the subtype is relevant since it has been observed that the ABC subtype has far worse survival prognosis than GCB subtype.

[3] Rituximab is a type of targeted cancer drug called a monoclonal antibody that targets CD20 protein, which is found on B cells. The immune system can recognise the marked cells and kill them.The other drugs in CHOP are cyclophosphamide, doxorubicin, vincristine, and prednisone which stop cells from dividing.

2.5 High-grade serous ovarian cancer

Ovarian cancer is a generic term used to classify cancers involving the ovaries though they can arise from many different cells. Ovarian cancers of epithelial cell origin account for more than 85% of all ovarian tumors [31]. Epithelial ovarian carcinoma (EOC) has the highest mortality rate among gynecologic malignancies. Typically, EOC is classified into five different histological subtypes: high-grade serous, low-grade serous, endometrioid, clear cell and mucinous.

High-grade serous ovarian carcinoma (HGSOC) is the most common ovarian cancer subtype (more than 70%) which presents at an advanced stage with a 5-year survival below 50%. HGSOC therefore accounts for the majority of both ovarian cancer cases and deaths [31]. HGSOC is usually diagnosed at late stages, less than 5% are found at Stage I when the tumor is confined to the ovaries. The first line of treatment for HGSOC is debulking surgery, but complete resection of the tumor is difficult once the cancer has progressed from Stage I and has invaded the abdominal cavity. Neoadjuvant therapy[4], usually consisting of carboplatin plus paclitaxel, is given to patients not eligible for debulking surgery.

HGSOC tumors are associated with genomic instability since almost all ($>$95%) have somatic p53 mutations and over half have homologous DNA repair pathway deficiencies predominantly in BRCA1, BRCA2, or related proteins. Germline mutations in BRCA1 or BRCA2 confer a lifetime risk of up to 44% by 80 years of age [21]. Genomic instability can lead to the inactivation of other TSGs through gene breakage, for example, loss of PTEN has been associated to poor patient survival. On the other hand, genomic instability can also cause amplifications, for example amplification of cyclin E1 (CCNE1) is associated again with poor prognosis and platinum resistance. Given that about 50% of all high-grade serous patients have mutations in DNA repair pathways, including BRCA1/2, PARP inhibitors appear to improve progression-free survival in women with recurrent platinum-sensitive ovarian cancer. PARP1 is a protein involved in DNA repair, in tumours with mutations in other DNA repair pathways, PARP inhibitors can be used to cause synthetic lethality[5] in the cell when used in combination with radiation to induce cell damage [32].

[4] Neoadjuvant therapy refers to treatment given before surgery with the aim of shrinking the tumor.

[5] Synthetic lethality arises when a combination of deficiencies in the expression of two or more genes leads to cell death, whereas a deficiency in only one of these genes does not. In the case of tumors with BRCA mutations, PARP inhibitors would leave the cells with no mechanism to repair DNA double strand breaks and die if the damage is induced by radiation, for example.

2.6 Survival analysis in cancer

Survival analysis is a branch of statistics used to estimate the expected duration of time until an event happens. In medicine, and in particular in oncology, the event of interest is usually either death or disease recurrence; respectively known as overall survival or progression free survival.

The time to relapse or death is not normally distributed. For example, in high risk patients, the majority of relapses will occur early with very few occurring towards the end of the follow-up time. In the case of a cancer with excellent prognosis, most deaths will happen at the end of the follow-up time, or will be censored—the precise time is not known, but it exceeds the time of last follow-up. Therefore, statistical methods that assume normality of distributions cannot be used in survival analysis.

Kaplan-Meier is a non-parametric method to estimate the survival function, a function that gives the probability that an individual will survive beyond any specified time [33]. The Kaplan-Meier survival curve is defined as the probability of surviving in a given length of time where time is split in small intervals; each of these time intervals starts with the date of occurrence of at least one event. The Kaplan-Meier estimator uses the following formula to calculate the total probability of survival considering all the probabilities of survival at all time intervals preceding that time:

$$\hat{S}(t) = \prod_{i=t_i \leq t} (1 - \frac{d_i}{n_i}),$$

where t_i is a time where at least one event happened, d_i is the number of events that happened at time t_i and n_i is the number of individuals that have survived until time t_i. The plot of the survival function is a step wise curve in which the probability of surviving is constant between adjacent death or recurrence times and only decrease at each event. An example of the Kaplan-Meier curves of two groups of individuals is shown in Figure 2.

There are several tests for comparing two survival curves, but the most often used is the log rank test [34]. The formula for the log rank test is given by

$$Log\ rank\ test\ statistic = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2},$$

where E is the expected number of events and O is the number of observed events. The expected number of events for each group is the risk of death or recurrence considering that there is no difference between the groups, in other words, E_1 is the risk of event of the total number of individuals multiplied by the size of group 1. The statistic obtained is revised against a χ^2 table to determine significance.

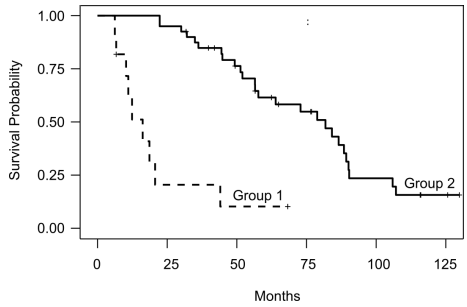

Figure 2: A plot of the Kaplan-Meier curve of two groups of patients showing different survival probabilities. Examples of different groups of patients that can produce different survival curves are type of treatment, presence or absence of a genetic variant, or stage of diagnosis.

The log rank test is used to decide if the survival estimates between two groups is significant, but does not allow us to test if other independent variables have an effect in survival.

The Cox proportional hazards model is commonly used in medicine to determine the association between patients survival times and one or more predictors. The Cox model works well with both quantitative and categorical data, and allows testing of several risk factors, or covariates, which is not possible to do with Kaplan-Meier. A hazard function, which can be interpreted as the risk of dying at time t, is calculated in the Cox model with the following formula:

$$h(t) = h_0(t) \cdot exp(b_1 x_i + b_2 x_2 + ... + b_p x_p),$$

where t is the survival time, $h(t)$ is the hazard function determined by a set of p covariates, h_0 is the baseline hazard which is the value of the hazard when all the covariates equal zero. Each of the $exp(b_i)$ terms are called hazard ratios. When b_i is greater than zero, the value of the ith covariate increases. A hazard ratio above 1 means that the covariate is positively associated with the event, and therefore negatively associated with survival. When the hazard function evaluates to 1, there is no effect in survival, below 1 is interpreted as reduction in hazard, and higher than 1 as increase in hazard. The Cox model allows to test if the covariates have an effect in survival and therefore it is widely used to test if there are significant differences between sample groups based on different covariates.

Kaplan-Meier and log rank test were used in Publication II to compare survival curves between groups with different gene expression, while Cox proportional hazards model was utilized to test if significant differences exist between the covariates of the different cohorts from the same publication.

3 Transcriptomics

The set of all RNA molecules produced by cells comprise the transcriptome. This set of molecules includes not only the messenger RNA which is the link between DNA and proteins, but also smaller RNA species that participate in gene regulation. In the next section the basics of transcription are covered followed by a description of some aspects of gene regulation that are part of the studies included in this dissertation. Namely, how different isoforms of a protein can be produced from the same gene sequence through alternative splicing, how messenger RNA (mRNA) expression can be affected by other RNA molecules, such as microRNAs or by other sequences in the genome called expression quantitative trait loci (eQTL), and how cancer-specific processes can affect the genes through genomic instability resulting in gene fusions.

3.1 Transcription

Most of the signals that govern cell functions are transmitted through the expression of genes. Genes are stretches of nucleotide sequences in the DNA that define each one of the proteins that cells need to operate. Genes consists of both regions that code for the protein, exons, and interleaving regions called introns that will not be part of the protein and are involved in regulation [35]. Transcription is the process of making gene copies from the DNA that can be later translated into amino acid chains and folded into proteins. The main components of the transcription process are shown in Figure 3.

The process of transcribing a gene begins when the RNA polymerase binds to the promotor region of the gene. The promotor is a short sequence of 100 to 1000 base pairs (bp) upstream of genes that have sequences that are binding sites for RNA polymerase and other proteins such as transcription factors[6]. This stage of the processes is called initiation. The RNA polymerase binds to the 3' to 5' strand of the DNA and starts traversing the template forming a complementary sequence in the same fashion that DNA polymerase does during DNA replication, with the exception that uracil is added instead of thymine to match adenine [37]. The nascent pre-mRNA sequence is synthesized in the 5' to 3' direction, in a step called elongation. Termination of pre-mRNA transcription occurs when the RNA polymerase reaches a terminator region and the pre-mRNA is released. A cap is added at the $5'$ end of the pre-mRNA and a polyA-tail at the $3'$ end to protect the

[6]Transcription factors work alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes

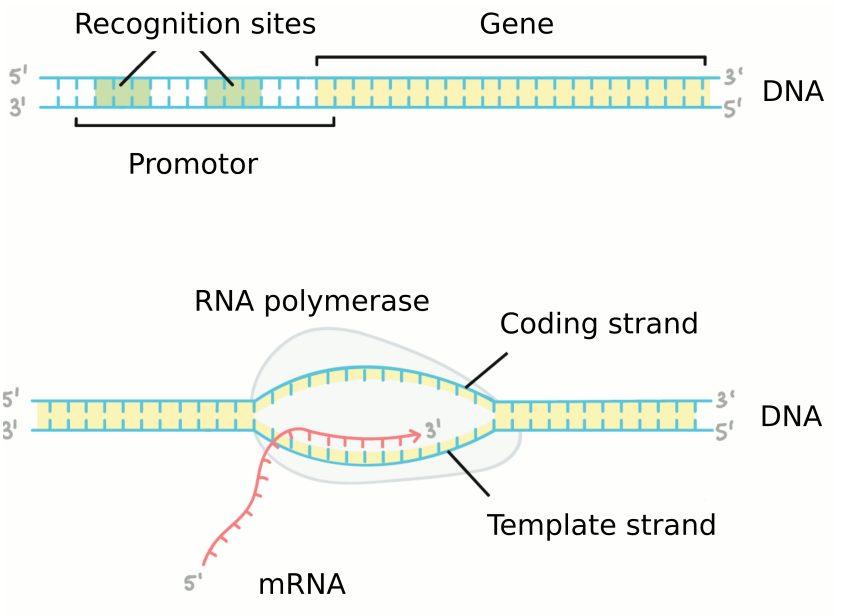

Figure 3: Transcription is the process of copying the information contained in genes into mRNA molecules that can be translated into proteins. Once recruited to the promotor region of a gene, the RNA polymerase opens the double-stranded DNA and traverses the template strand synthesizing an RNA molecule, until a terminator signal is reached and transcription ends. Modified from [36].

transcript from enzymes. The pre-mRNA is transformed into a mature mRNA when the introns are removed or spliced out and at that point the mRNA can be transported outside the nucleus to the ribosome for translation [37].

3.2 Alternative splicing

Splicing is a post transcriptional process in which introns are removed and exons are joined together to produce a mature mRNA. Some introns can self splice, while others require a spliceosome complex formed by small nuclear RNAs attached to small nuclear ribo nuclear proteins (snRNP) that attach to introns' splicing sites and cleave them out from the transcript. Alternative splicing (AS) is a common regulatory mechanism generating multiple RNA transcripts from a single gene, an example is depicted in Figure 4. AS is a highly regulated process, modulated by activator and repressor proteins, that increases the diversity of protein products from the genome; it has been estimated that about 95% of multiexonic genes are alternatively spliced [38]. The most common way in which AS is achieved is through exon skipping. In eukaryotes, five mechanism have been observed [39]:

Exon Skipping An exon can be either spliced out or retained in different isoforms of the same gene.

Mutually Exclusive Exons In this case two exons cannot co-exist in the same isoform, so when one is retained the other one is spliced out.

Alternative Donor Site Different $5'$ junction sites (donor sites) are used for different isoforms, which changes the $3'$ boundary of the upstream exon.

Alternative Receptor Site Different $3'$ junction sites (acceptor sites) are used for different isoforms, which changes the $5'$ boundary of the downstream exon.

Intron Retention Introns are by definition not coding parts, so what makes intron retention different from exon skipping, is that a retained intron is not flanked by other introns. Retained introns encode amino acids in frame with the neighboring exons. This is the least common type of alternative splicing.

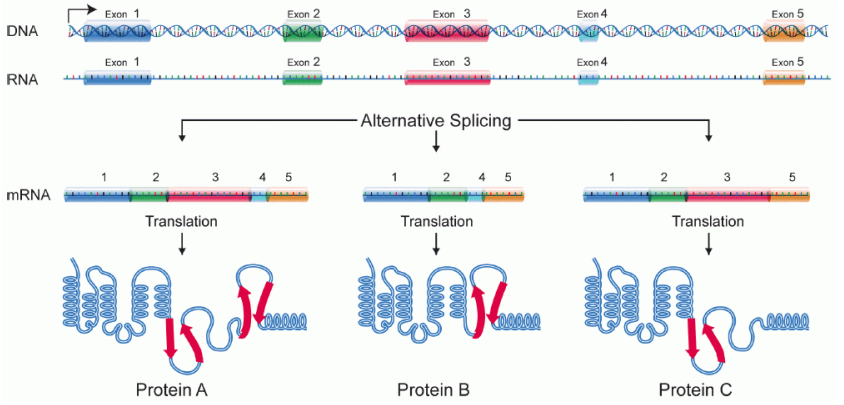

Figure 4: Alternative Splicing [40]. Three different protein isoforms are produced from the same mRNA transcript by keeping or splicing out different exons.

Disruptions in the regulation of AS are not uncommon in cancer. Genes from all of the hallmarks of cancer are known to be affected by aberrant alternative splicing [41]. For example, the active form of p53 is a tetramer of four identical units. A fully spliced mRNA from p53 consists of eleven canonical exons that encode a functional p53 protein. Alternative isoforms of p53 that retain intron 9 cause the loss of the oligomerization domain which precludes the formation of tetramers and in turn results in inactivation of the isoform [42]. For BCL-2 inclusion of intron 2 produces a BCL-2 isoform that is anti-apoptotic, while its exclusion results in BCL-2 missing an important domain and becoming pro-apoptotic.

3.3 MicroRNAs

RNA molecules can be classified in two major types: coding and non-coding. The messenger RNA, which codes for proteins, accounts for about 20% of all the RNA molecules in the cell. The remaining 80% is composed of 1) ribosomal RNA, which together with ribosomal proteins forms the ribosomes, 2) transfer RNA which transports amino acids to the ribosome, and 3) regulatory non-coding RNA. MicroRNAs (miRNAs) belong to the latter class along with many other molecules of which the most well known ones are silencing RNAs (siRNAs), small nuclear RNAs (snRNAs), and long non-coding RNA.

A miRNA is transcribed in the same way as mRNA from sequences that can be up to several hundreds bases long. The ensuing transcript (pri-miRNA) is made from single stranded RNA that twists on itself to make a hairpin structure. Drosha, an enzyme, cleaves the bases outside the hairpin structure to form the pre-miRNA. Exportin-5 then escorts the pre-miRNA outside the nucleus and releases it into the cytoplasm. The Dicer complex (Dicer and TRBP2) removes the stem loop from the pre-miRNA to transform it into a mature miRNA [43]. This new structure is an asymmetrical double stranded RNA molecule of 20 to 25 nucleotides in length. One strand of the miRNA is incorporated into the RNA-induced silencing complex (RISC) and together locate the conserved sites in the target mRNA. RISC uses the miRNA as a template for recognizing complementary mRNA. Once found, one of the proteins in RISC, Argonaute, activates and cleaves the mRNA [43]. Members of the Argonaute (Ago) protein family are central to RISC function. They bind the mature miRNA and orient it for interaction with a target mRNA. AGO2 belongs to the argonaute family of genes and it is capable of cleaving target transcripts directly; other argonautes may also recruit additional proteins to achieve translational repression.

MiRNAs main function is to prevent transcribed mRNAs from being translated into proteins. Many miRNAs are evolutionary conserved and it is estimated that around 60% of genes are targets for miRNAs. A single gene can be regulated by more than one miRNA and a given miRNA may have hundreds of different mRNA targets [44][45]. Considering the breadth of genes that can be regulated by miRNAs, it is not surprising to find that miRNAs play a role in cancer as well. For example, in 28% of glioblastoma multiforme (GBM), an aggressive brain cancer, it has been observed that the level of microRNA miR-181d is inversely correlated with the expression of its direct target, MGMT, a DNA repair enzyme [46].

3.4 Expression quantitative trait loci

Gene expression is constantly influenced by the dynamics inside and outside the cell, but it can also be regulated by genetic variation. Genome Wide Association studies (GWAS), which assay genetic variants in large number of samples, have identified over 70,000 variants associated to an array of phenotypes [47]. The effect size of most of these variants, due to the nature of complex diseases in which many genes contribute to the phenotype in different measure, has been found to be very small. Furthermore, over 80% of the variants identified through GWAS are located outside coding regions of the genome, which has complicated the understanding on how the variants influence the phenotype. A plausible mechanism that can explain the effect of non-coding variants is genetic regulation of an intermediate phenotype such as gene expression, splicing or methylation.

Expression quantitative trait loci (eQTLs) are genomic regions that correlate with the mRNA levels of genes. When these loci are single nucleotide polymorphisms each locus is called an eSNP. The eSNP can be located in the same allele as the gene it acts upon, in which case it receives the name of cis-eQTL. On the other hand, trans-eQTL refer to eSNPs located at a considerable distance of the target, including a different chromosome altogether, in which case their influence on the expression is usually through a transcription factor.

Cis-eQTLs are easier to identify since the effect is usually larger than with trans-eQTLs, and therefore less amount of samples is needed to detect the correlation. The effects are usually additive, the eSNPs are located near transcription start sites (TSS), and it has been observed that the shorter the distance between TSS and the eSNP the effect tends to be more pronounced [48]. Furthermore, eQTLs also are affected by cell type and tissue specificity. A comparison of B-cells and monocytes showed an overlap of 21.8% of the detected cis-eQTLs and 7% of the detected trans-eQTLs between both cell types, finding that suggests that genetic regulation in trans is more cell-type-specific than cis regulation [49]. Figure 5 shows how the mechanism of expression regulation mediated by eQTLs would impact the amount of a protein carrying a mutation.

Examples of inherited germline variation that increase cancer risk abound. Recent estimates suggest that 20-25% of ovarian cancers are due to a germline loss-of-function variant in one of several genes that confer moderate-to-high risk [51][52]. Expression quantitative trait loci (eQTLs) have been mapped in many tumor types, including glioma [53], colon [54], breast [55], ovarian [56], and prostate cancer [57]. Additionally, studies on cancer eQTLs in the context of pharmacogenomics have found associations between miR-30d and *ABCD2* expression that correlated with both carboplatin- and cisplatin- specific sensitivity in lymphoblastoid cell lines

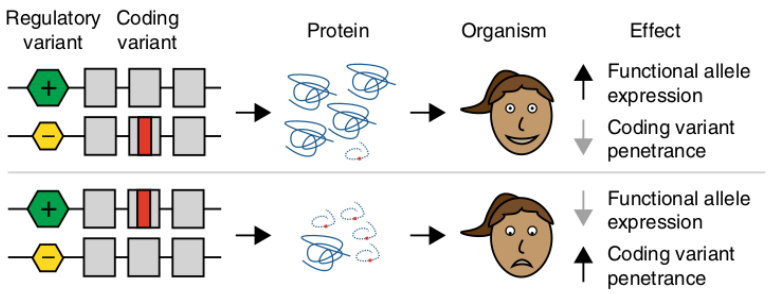

Figure 5: Regulatory variants as modifiers of coding variance penetrance. In this example an heterozygous individual for both a regulatory variant and a pathogenic coding variant is shown. The two possible haplotype configurations would result in either decreased penetrance of the coding variant, if it was on the lower expressed haplotype, or increased penetrance of the coding variant, if it was on the higher-expressed haplotype [50].

[58], a model system commonly used in studies of genetics of drug response. In the same study, it was observed that reduction of *ABCD2* expression was correlated with increase in apoptotic activity after treatment with cisplatin in an ovarian cancer cell line, while the microRNA miR-30d is associated with poor clinical outcomes in ovarian cancer patients [59].

3.5 Gene fusions

Genomic instability often results in translocations, deletions, or chromosomal inversions that can merge two distinct genes into a chimeric one. Each type of chromosomal rearrangement can produce fusion genes in a different manner. Figure 6 shows some examples of how fusion genes are formed.

The most well known example is the Philadelphia translocation which forms the BCR–ABL1, which juxtaposes the ABL1 tyrosine kinase from chromosome 9 into the region of BCR gene in chromosome 22. BCR–ABL1 was the first fusion gene to be identified, discovered in 1960, and it is present in more than 96% of patients with chronic myelogenous leukemia (CML) [61]. PML–RARα is another example of translocation that produces a fusion occurring in 90% of acute myeloid leukemia (APL) which is correlated with relapse after therapy [62]. Deletions can join together nearby genes that transcribe in the same direction. Examples of this type of fusion genes are ATG7–RAF1 in pancreatic cancer and EIF3E–RSPO2 in colon cancer. Fusions involving RET or NTRK1, both tyrosine kinases, are common in papillary thyroid carcinomas. Some of these fusions are generated by chromosome translocations, but the three most prevalent, including CCDC6/H4–

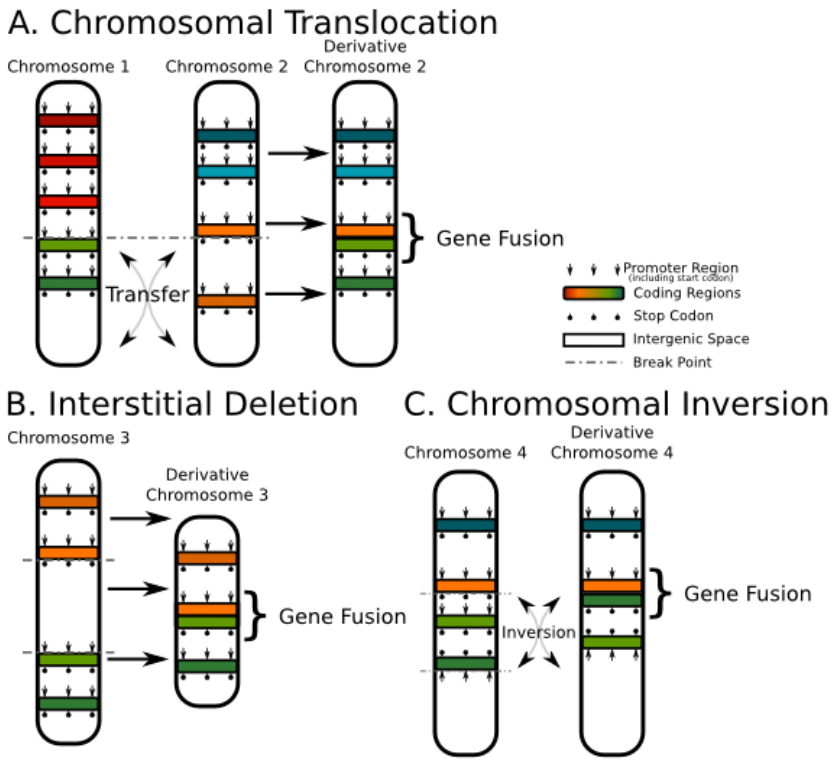

Figure 6: A schematic showing the ways a fusion gene can occur at a chromosomal level [60]. Three main mechanisms are depicted.

RET, are generated by inversion of a large section of chromosome 10. Another fusion found in papillary carcinomas, AKAP9–BRAF which is known to activate BRAF, a proto-oncogene, is the product of an inversion [63].

Fusion genes can be either inter- or intra-chromosomal depending if the genes involved reside in the same chromosome or different one. They can also be balanced, if the amount of genetic information exchanged between to chromosomes is the same, or unbalanced if losses or gains occur. The junction point where the DNA breaks in each gene segment is called the breakpoint and it can be located in intronic regions leaving whole exons intact, within exons with the possibility of causing frameshifts, or in promotor or intergenic regions.

Most of the gene fusions are likely passenger events caused by genomic instability, but they can also lead to 1) amplification of expression of an oncogene, 2) disruption of a TSG, or 3) creation of a new protein. Examples of each one of these cases include TMPRSS2–ERG fusion in prostate cancer, in which androgen regulatory elements of TMPRSS2 drive ERG overexpression. ERG activation present in 50-

70% of prostate tumors, represents one of the most common oncogenic alteration in prostate cancer [64]. PPP2R2A-CHEK2 fusion is found in several solid tumors where CHEK2 is a TSG that regulates cell division by preventing cells from entering mitosis or arresting cell cycle in G1 phase, in response to DNA damage [65]. The translocation of chromosome 9 at ABL1 gene to a part of BCR gene on chromosome 22 results in an unusually short chromosome that encodes an hybrid protein with abnormal tyrosine kinase activity [66]. It is also possible that fusions occur due to RNA polymerase missing a termination signal and continuing transcribing to make a readthrough of two genes. This type of event is not very common, out of 4,344 fusions recently identified in a pan-cancer study, only 351 were read-throughs. About 44% of the read-through events were concentrated in only three out of 33 cancers analyzed: ovarian, esophageal, and acute myeloid leukemia [67].

Fusion genes are not only strong driver mutations in cancer, providing insight in disease mechanisms, but also serve as specific targets for treatment. Table 1 includes a list of recurrent fusions with drugs that are either already approved and in use in the clinic or in different stages of clinical trials. The most shared fusion among cancers from the table is FGFR3–TACC3 which has been identified in lung squamous cell carcinoma (LUSC), esopharingeal carcinoma (ESCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), head and neck squamous cell carcinoma (HNSC), bladder urothelial carcinoma (BCLA), liver hepatocellular carcinoma (LIHC), and in GBM, and it serves as biomarker of sensitivity to Erdafitnib [68].

Karyotyping techniques allowed the discovery of the BCR–ABL1, fusion and of the first fusions to be detected in solid tumors, CTNNB1–PLAG1 in salivary gland adenoma in 1980. This discovery was soon followed by the identification of PAX3–FOXO1 and EWSR1–FLI1 in sarcomas in 1982 and 1983, respectively. The first fusions to be identified in carcinomas were PRCC–TFE3 and MYB–NF-IB in 1986. The development of fluorescence in situ hybridization (FISH) has substantially facilitated localization of chromosomal breakpoints, but the explosion of the field arrived with the development of sequencing techniques. Over 90% of the 10,000 fusions that have been identified in solid tumors, although very few of them recurrent, were detected by either whole genome sequencing (WGS) or RNA sequencing (RNA-Seq).

Target	Partners	Drug
ALK	EML4, STRN	Alectinib, Brigatinib, Crizotinib, Ceritinib, Entricitinib, Lorlatinib, MEK inhibitors, novel ALK inhibitors, PF2341066, TAE684
BCR	ABL1	Imatinib, Dasatinib, Nilotinib, Bosutinib, Ponatinib, Venetoclax
BRAF	MRPS33, SND1	Cobimetinib, MEK inhibitors, Selumetinib, Sorafenib, Trametinib
EGFR	SEPT14	Afatinib, EGFR TKIs, Elotinib, first generation and irreversible TK1s, Geftinib, HSP90 inhibitors, ZD6474
ESR1	CCDC170, TMEM212	Anti-estrogens
FGFR2/3	ATE1, BICC1, SHTN1; TACC3	AZD4547, BGI398, Debio1347, Erdafitinib, FGFR inhibitors
MET	CAV1, ST7	Criozitinib
NOTCH2	SEC22B	Gamma secretase inhibitors
NTRK1	IRF2BP2, TPM3	Crizotinib, Entrectinib, IGF1R inhibitors, Larotrectinib, pan-TRK inhibitor
NTRK1	ETV6	Entrectinib, Larotrectinib, Midostaurin
PDGFRA/B	COL1A1; FIP1L1	Imatinib
PML	RARA	Arsenichtrioxide, Tretinoin, Bolasertib
RET	CCDC6, ERC1, NCOA4	BLU-667, Crizotinib, LOXO-292, Nintedanib, Sunitinib, Vandetanib
TMPRSS2	ERG, ETV4	DNA-PKc inhibitors

Table 1: List of fusion genes for which drugs have been already approved or are in clinical trials. A slash (/) is used to indicate two genes from the same family for which the same drugs are being tested and a semicolon (;) for separating the respective partners for each gene. This table has been modified from [67].

4 High-throughput technologies

High-throughput technologies, namely microarrays and next generation sequencing, are methods that allow the study of whole genomes or transcriptomes in a single experiment. Before the development of microarrays and later of next-generation sequencing it was practically impossible to quantify several thousand genes simultaneously or compare large number of samples in a rapid and cost efficient manner. This section gives a short overview of microarray technology for gene expression profiling and next-generation sequencing of both whole genomes and RNA molecules, followed by a description on common data analysis steps.

4.1 Microarrays

A microarray is a chip usually made of glass, silicon or plastic, depending on the manufacturer, which has fluorescently labeled target DNA probes attached to it. The probes are 25 bp sequences that can be mapped to specific gene regions. To quantify expression, the RNA of interest is first converted into cDNA and then washed on the plate causing complimentary sequences to hybridize. A second wash is needed to remove residual non-hybridized DNA. A fluorescent image of the array is acquired by a laser scanning confocal microscope. The array thus obtained can be matched to the probe information to determine the presence or absence of the sequence or to quantify expression by measuring the intensity of the probe signal [69].

Several types of microarrays exist to quantify expression: gene, exon, tiling, $3'$, and human transcriptome. The main difference among the models is the number of probes and the location in the gene region that map to each gene. Figure 7 shows the principal types of microarrays that are used for gene expression profiling. The basic advantage of exon arrays over gene arrays is that exon arrays have at least four probes for each exon, while gene arrays do not necessarily cover all exons. Tiling arrays have better coverage of all exons than exon arrays, but are not able to cover mammalian genomes completely, and at least six different arrays are needed to cover the human genome [70]. Exon arrays initially were the only microarrays that allow quantification at both gene and exon level of the human genome with a single kit [71], but further developments in microarray technology resulted in the Human Transcriptome Array which includes coverage of exon junctions as well [72].

The main limitation of expression profiling with microarrays that can be circumvented with sequencing are 1) limitations in detection range, 2) lack of base pair

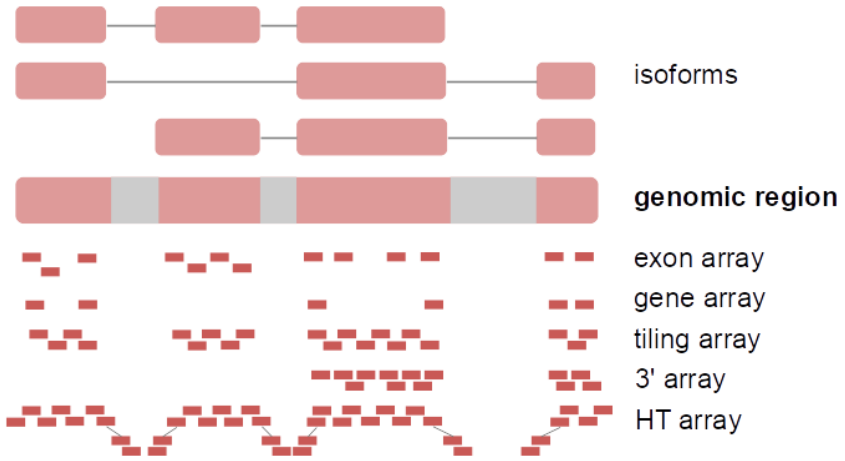

Figure 7: Different types of microarrays. The top panel shows an example of a gene with different isoforms and in the bottom panel the arrangement of the probes depending on the microarray type. Human Transcriptome (HT) Array is the only one to include exon junction probes.

resolution, and 3) reliance on reference sequences [73]. Microarrays can fail to detect very lowly expressed genes and can also reach saturation with highly expressed genes. Since hybridization depends on the sequence being complementary to the probe, is not possible to detect mutations or post transcriptional modifications with microarrays for which a probe does not exist. Even with Human Transcriptome Arrays, alternative splicing is difficult to detect if the right exon junction does not have a probe. Quantification with microarrays is always limited to the sequences in the probes. Initially, an important advantage of microarrays over sequencing was its lower cost, but continuous improvements in sequencing technology have permitted the prices to drop enough for RNA-Seq to overtake microarrays for gene expression profiling.

4.2 Next-generation sequencing

Sequencing is the process of determining the nucleotide order of a given DNA fragment. Next-generation sequencing is the term used for the technology capable of massively sequence thousands of DNA fragments in parallel. Illumina, currently the most popular maker of next-generation sequencers for short read sequencing (150-300 bp long), uses a methodology known as sequencing by synthesis. The basic process of sequencing by synthesis entails the following steps [74]:

Library preparation First DNA is extracted from the sample followed by fragmentation, size selection of the fragments and addition of adapters at both

ends of the fragments.

Cluster generation The first stage of sequencing is cluster generation by bridge amplification. In this process the adapters hybridize to oligosequences in the flow cell of the sequencer. A polymerase adds nucleotides to the template forming a double strand. The DNA is denatured and the original strand is washed away leaving the complementary strand tethered to the oligo. This single strand folds and attaches to another oligo in the flow cell, forming a bridge structure. Here, a polymerase replicates the template, beginning a new cycle of DNA denaturation, washing away, and a new template folding again on a new oligo in the flow cell. In this manner, the original DNA is clonally amplified in clusters. Finally, reverse strands are cleaved and washed off leaving forwards strands only.

Sequencing Fluorescently tagged nucleotides are added to the template strand. After the addition of each nucleotide the clusters are excited by a light source and a characteristic fluorescent signal is emitted. The number of cycles of nucleotide additions determines the length of the read, the wavelength and the intensity of the fluorescent signal determines the base call. After the completion of the first read the sequence is washed away and the sequence is allowed to form a bridge again. A polymerase replicates the template, forward strands are discarded and now the sequencing steps will be repeated on the reverse strand. When the protocol includes sequencing both ends of the fragment the process is called paired-end sequencing (depicted in Figure 8).

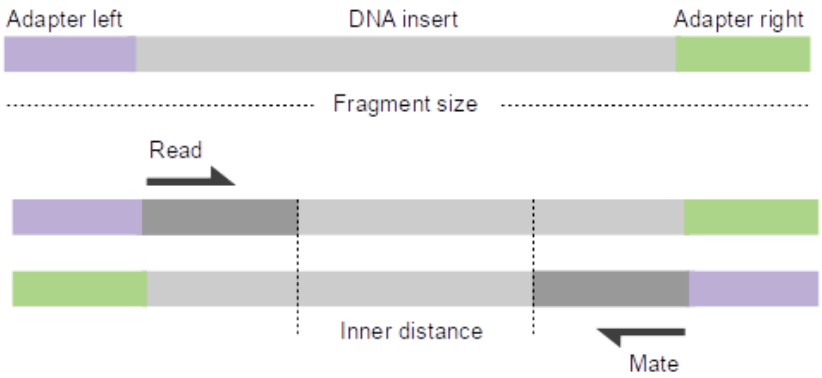

Figure 8: Paired-end sequencing. During library preparation adapters are added to both ends of the DNA insert. Both the read and the mate are sequenced, and although the inner distance is not sequenced, paired-end sequencing provides information on the whole fragment.

Variations of the protocol described above allow different applications to exist. For example, with the aim of reducing costs in sequencing it is possible to only sequence exonic regions instead of whole genomes. Exome sequencing is achieved by an additional step of target, in this case exons, selection during library preparation. For quantifying gene expression, instead of extracting DNA from the sample, RNA is extracted either selecting molecules with polyA tails or by depleting ribosomal RNA. Size selection is used to separate mRNA from small RNA in case of ribosomal depletion. The RNA is then reverse transcribed into cDNA. The cDNA can then be sequenced following the same protocol as whole genome sequencing [75].

4.3 Data analysis

Regardless of the protocol, next-generation sequencing produces millions of short reads that require computational methods to reconstruct the reads into genomes or expression profiles. The files produced by both, RNA-Seq and whole genome sequencing (WGS), contain the read fragments of up to 300 bp with no other information than the base call and score that represents the estimated likelihood of the base call being correct. This format is called fastq (or fasta when the quality score is missing) and the size of the file depends on the sequencing depth, which is given by the number of cycles during sequencing. Tailored methods have been developed to assess the quality of the reads and implement the different analysis steps required to make sense of the sequencing reads. The initial data analysis steps for next-generation sequencing datasets are the same for whole genome or RNA-Seq and are described below.

Preprocessing The reads obtained by the sequencer may still include adaptors incorporated in library preparation that need to be trimmed to improve the mapping of the reads to the reference genome. Also, it is advisable to remove low quality bases that are slightly common at the end of a sequencing run. FastQC [76], multiQC [77] and AfterQC [78] are tools widely used for assessing the quality of sequencing reads. For trimming or clipping low quality bases or adapters Trimmomatic [79] and CutAdapt [80] are popular tools still in use, FastX-Toolkit (`http://hannonlab.cshl.edu/fastx_toolkit/`) used in Publication I is not longer maintained.

Assembly/Mapping When a reference genome is available for the organism of study, the mapping can be done either directly to the genome or transcriptome. If the reference genome is unknown or incomplete then de novo assembly needs to be done. For WGS or exome-sequencing the most popular tools and that have been widely used for several years are Bowtie [81] and BWA [82]. For RNA-Seq, in the case of mRNA, splice-aware aligners capable

of mapping sequences to non-continuous genomic regions such as exons separated by introns are needed. TopHat [83] was for many years the preferred aligner for RNA-seq, and although it remains popular, it has largely been replaced by STAR [84]. One of the main advantages of STAR over TopHat is the speed, STAR is much faster than TopHat but has a bigger memory requirement, although as an offset STAR allows parallel runs to access the same loaded genome in memory. Additionally, TopHat (or HISAT2 which is the successor of both TopHat and HISAT) are less robust to mismatches in the sequences which results in less read mappings [85]. For de novo assembly, Velvet [86] and Trinity [87] can be used for WGS and RNA-Seq respectively. Mapping and assembly produce files in the Sequence Alignment/Map (SAM) format which is text format that encodes the position in the genome the read aligns to, information on the mate in case of paired end sequencing, an alignment score and presence of mismatches. The binary counterpart of a SAM file is called BAM.

Typical data analysis workflow steps after mapping are highly dependent on the type of study yielding different pipelines for whole genome and RNA-Seq. Downstream analyses for each dataset are explained below.

4.3.1 Whole genome sequencing data analysis

In the case of WGS common downstream analyses in cancer studies are variant calling and structural variation identification. Variant calling refers to the identification of somatic mutations which usually require a comparison of normal samples against tumor samples [88]. Some of the most popular algorithms are the four used in the TCGA somatic variant calling pipeline: Mutect/2 [89], Muse [90], Varscan2 [91] and SomaticSniper [92]. Strelka is also widely used and well evaluated in recent benchmarks [93][94] along with EBCall [95] and Virmid [96]. Variant calling is also possible to do with RNA-Seq data but it is not commonly performed. For WGS, structural or copy number variation analysis is also part of a standard workflow. The aims of this analysis is to identify regions in the genome that have been amplified, deleted, or translocated, for which again many methods exist, each with their own sensitivities to different types of rearrangements that can be consulted in a very recent and comprehensive review of 69 methods [97], which includes Pindel [98] the method used in the TCGA pipeline.

Fusion genes can be detected through structural variation analysis, but information on expression cannot be obtained through genome analysis alone. Methods that combine WGS and RNA-Seq for fusion calling exist such as INTEGRATE [99] and Comrad [100]. Efficiently identifying fusion genes from WGS data is difficult

due to tumor heterogeneity that can result in low coverage of fusion breakpoints. Furthermore, WGS information do not inform on actual transcription of the fusion gene, and integration with RNA-Seq leads to false negatives since junction coordinates cannot be matched when located outside exonic regions.

4.3.2 RNA-Seq data analysis

Next-generation sequencing for RNA can be applied to both mRNA and smaller RNA molecules such as miRNAs. The initial analysis steps are the same for mRNA and miRNA, both require pre-processing and mapping, although for miRNAs splice-aware aligners are not necessary. Common RNA-Seq processing and downstream analysis steps are described below.

Quantification The number of reads mapping to each gene feature are counted. The counts can be reported at exon, transcript or gene level. Several metrics are used to report the expression values: raw counts (used in downstream differential expression analysis), fragments per kilobase of million reads (FPKM) and TPM (transcript per million reads). Both, FPKM and TPM, account for the library size (how many reads were sequenced) and the gene length. HTSeq [101] and RSEM [102] are quantifiers that rely on alignment files. Newer algorithms for quantification such as Kallisto [103] and Salmon [104] do not require pre-aligned reads and can perform pseudo alignment on the fly.

Differential expression A classical RNA-Seq analysis usually compares the gene expression of samples under different conditions. The most popular tools for differential expression are DESeq2 [105], EdgeR [106] and limma+voom [107][108] due to favorable reviews and their extensive documentation which facilitates user adoption [109] [110]. The basic workflow of differential expression analysis of RNA-Seq data is to first account for library size, then perform statistical tests and finally correct for over dispersion observed in the data. EdgeR and DESeq both assume a binomial distribution, while limma uses a non parametric approach. For differential exon usage or alternative splicing existing tools are DEXSeq [111], IsoformSwitchAnalyzeR [112], DreamSeq [113] and LeafCutter [114]. Our SePIA workflow presented in Publication I, allows differential expression analysis using several methods including DESeq2 and EdgeR for gene-level analysis and DEXSeq for exon-level. Both DESeq2 and DEXSeq were utilized in Publication II in our study of alternative splicing in DLBCL.

Many other applications of RNA-Seq that deviate from this standard pipeline exist. A brief overview of some of this applications is included below:

Fusion genes For example the identification of gene fusions usually uses tailored algorithms that work with the raw reads directly. Tens of methods for fusion calling exist and no clear gold standard is available yet. A recent, non-comprehensive benchmark, tested 25 methods for fusion calling [115]. The main idea behind detection for all of them is the identification of discordant read pairs. Two type of discordant reads exist, junction or split reads which cover the fusion breakpoint in a single read, and spanning reads in which the read aligns to one gene while the mate aligns to a different gene. Each method uses different heuristics to report fusions which can result in more or less sensitivity to different types of fusion events. Depending on the cancer type and stage, and the algorithm utilized, the number of fusions detected in each sample can range from zero to even a few thousands. False positives are common due to sequence similarity among genes from the same family. Furthermore, in cancers with high genomic instability is very likely that most of the fusions detected are passenger events that have no role in cancer progression. Prioritizing fusions for further analysis or functional experiments is a necessary downstream step. Two main strategies exist for discarding fusion genes: mining databases for known artifacts and scoring the oncogenic potential of fusions based on prior knowledge of relevant cancer fusions. For the former strategy, many databases exist, of which the Mitelman database of Chromosome Aberrations and Gene Fusions in Cancer (https://mitelmandatabase.isb-cgc.org) since its creation in 1983 continues to be a key resource. Many more databases, usually specific to certain projects exist, as well as webservices such as FusionHub (https://fusionhub.persistent.co.in) which queries at least 23 databases for reporting fusion matches. In the second strategy for fusion prioritization, oncogenic scoring methods such as Pegasus [116], Oncofuse [117], and more recently DeepPrior [118] utilize either protein domains or amino acid sequences to train algorithms to identify oncogenic fusions. In Publication IV, we developed a pipeline for integrating gene fusions from any of the available method and standardized the post-processing steps of the analysis. Two of the three top methods from the latest review have been included in our fusion gene analysis pipeline, as well as two scoring methods.

Phasing One of the main advantages of RNA-Seq over microarrays is that the reads provide much more information than just expression quantification. RNA-Seq in particular, also provide an advantage over WGS in terms of coverage over longer stretches of DNA. Since RNA-Seq reads contain only exons, and although their read length is about 300 bp long, the region of DNA that is covered by reads that span exon junctions can be of several kilobases

long. In Publication III we exploit this characteristic of RNA-Seq reads to complement haplotype phasing of germline variants and mutations utilizing phASER [119] which combines WGS and RNA-Seq reads.

MiRNA analysis A variety of tools to analyze miRNAs exist, for example `https://tools4mirs.org/` a platform that does target prediction for miRNAs, but also curates a list of miRNA methods currently has 203 listed methods. Several of the tools are capable of performing all of the steps in the miRNA analysis pipeline. The first step in miRNA analysis from sequencing reads is mapping to known miRNAs. For this step, the miRNA annotations are retrieved from databases such as miRBase [120] which contains information on sequences for both mature miRNA and hairpin structures. The alignment can be done with sequence aligners such as Bowtie or BWA. Novel miRNA identification is usually performed with sRNAbench [121] (formerly miR-Analyzer [122]) and miRDeep [123]. Both tools also provide quantification which previously was obtained by read counting methods such as HTSEq. Interesting miRNAs are usually selected from differential expression tests which can be performed with DESeq or EdgeR. For predicting targets for the selected miRNAs TargetScan [124], miRanda [125], and miRDb [126] are commonly used [127]. One of the main challenges for miRNAs studies is that one sequence can map to multiple sites in the genome. Furthermore, isoforms of miRNAS exist and modern pipelines have to account for it [128].

eQTL Mapping eQTLs requires finding associations between genetic variants and gene expression levels. Considering that the effects on expression of the variants is usually small, large amount of data is usually required to reliably catalog these associations. Millions of associations tests are required to identify eSNP and eGene combinations that result in gene expression regulation. Software packages that can be used to search for eQTLs are Matrix EQTL [129] used initially by the GTEx, now replaced by FastQTL [130]. The GTEx consortium also provides calculator in their webportal `https://gtexportal.org/home/` where candidate eQTLs can be tested for association in different human tissues. For Publication III, we utilized GTEx own list of pre-identified eQTLs and matched heterozygous combinations of eSNPs and somatic mutations of tumor suppressor genes from TCGA data genomic data.

The analysis steps described in this section apply only to RNA extracted from tissue samples that contain thousands of cells, known as bulk RNA-Seq. Advancements in sequencing techniques allow today to disaggregate cells prior to sequencing and produce read-outs from several hundreds to a few thousands of single cells. Although bulk RNA-Seq is very much still in use and will continue to be, the

possibility of studying individual cells has opened the doors for a diverse array of studies and has become a very active field of research and method development.

5 Aims of the study

The general aim of this work is to study different aspects of cancer using RNA-Seq.

The specific goals of each publication are:

1. Facilitate the integration of multiple tools for processing and analysis of RNA-Seq datasets with a focus on cancer research (Publication I).
2. Study alternative splicing in DLBCL (Publication II).
3. Study the effect of eQTLs on variant penetrance using cancer as a case study (Publication III).
4. Facilitate the integration of multiple tools for fusion gene detection and standardize the analysis of the combined results to help prioritize fusions for further analysis (Publication IV).

6 Materials and methods

A summary of the datasets and the methods utilized in the four publications is included in this section. Detailed description on the sample material and each specific analysis can be found in each corresponding publication.

6.1 Biological sample material (Pub I-IV)

In Publications I and III we make use of the invaluable resource which are public repositories of sequencing data from TCGA, the Cancer Genome Characterization Initiative (CGCI) and Gene Expression Omnibus (GEO). Both Publication II and IV combine the use of in-house patient samples with public datasets to validate or strengthen the results obtained in the study of our own cohorts. Publications I, II and III combine the use of different technologies to study cancer more comprehensively with miRNA and mRNA integration (Pub I), to validate results from exon array with RNA-seq (Pub II), and to improve whole-genome sequencing (WGS) haplotype phasing with RNA-seq reads (Pub III). Table 2 briefly describes the cancer samples processed and analyzed in the completion of this work.

Pub	Tumor	Normal	Cancer	Technology	Source
I	17	3	BRCA	RNA-Seq	GEO
I	120	15	BRCA	RNA-Seq	TCGA
I	133	16	BRCA	small RNA-seq	TCGA
II	38	0	DLBCL	Exon array	In-house
II	92	0	DLBCL	RNA-seq	CGCI
III	925	925	15 types	WGS	TCGA
III	925	0	15 types	RNA-Seq	TCGA
IV	107	0	HGSOC	RNA-Seq	In-house
IV	424	0	HGSOC	RNA-Seq	TCGA

Table 2: Cancer sample datasets used in Publications I-IV.

Complete information on sample material can be found in each publication, but a short description of each sample set follows. In Publication I the data was analyzed in two separate case studies. For Case I, we downloaded from GEO (GSE52194) fastq reads from total RNA extracted, and sequenced with Illumina HiSeq 2000, from 17 breast primary tumor samples and 3 normal human breast organoids. For Case II, we used TCGA level 1 data from poly(A)-extracted mRNA and small RNA sequenced with Illumina Genome Analyzer II. For Publication II, for the in-house DLBCL samples, we extrated total RNA from primary tumors,

the prepared libraries were then hybridized to Affymetrix Human Exon 1.0 ST arrays. The RNA-seq reads from CGCI used as validation cohort in Publication II are from poly-A extracted mRNA sequenced with Illumina-GAIIx (study accession: phs000532.v3.p1). In Publication III, TCGA level 1 WGS and RNA-seq aligned sequences (reference genome b37) from 925 patients of 15 different cancers[7] were obtained by the New York Genome Center (NYGC) through dbGap (study accession phs000178.v9.p8). For Publication IV in-house samples, total RNA was extracted from 68 primary (before chemotherapy), 32 interval (after neoadjuvant platinum-taxane chemotherapy) and seven relapsed tumors (after being diagnosed as recurring) from 36 HGSOC patients and sequenced at BGI. In Publication IV, in addition to the in house samples, 424 RNA-Seq TCGA level 1 fastq reads from ovarian cancer samples were analyzed.

6.2 RNA-seq processing (Pub I & II)

Sequence quality was assesed with FastQC [76] and low quality bases and adaptors were removed with Trimmomatic [79] (Pub I & II). Reads were aligned with STAR [84] (Pub I) and Tophat [83] (Pub II). STAR continues to be state-of-the-art in RNA-Seq alignment and processing. Although that is not the case with Tophat, which has been disrecommended by its own authors in deference of better methods, at the time when the data for Pub II was analyzed, TopHat was still considered a top method in RNA-Seq alignment. Variants were called with Bambino [131] (Pub I) and pre- and post-processing was done using the Genome Analysis Toolkit (GATK). Read counts and normalized expression values were obtained from HTSeq (Pub I & II) and Cufflinks [132] (Pub I). For gene-level differential expression several methods were used: Cuffdiff [133] (Pub I), DESeq2 [105](Pub I & II), edgeR [106] (Pub I), and upper quartile normalized t-test (Pub I). Differential exon usage analysis was performed with DEXSeq [111] (Pub I & II).

6.3 MiRNA-seq processing (Pub I)

RNA-seq from small RNA were first trimmed with FASTX-Toolkit to remove adaptors and low quality bases and then aligned with Bowtie. MiRNA sequences were then annotated to known human miRNAs from miRBase [134]. Target prediction was done with miRanalyzer [122].

[7] Bladder Urothelial Carcinoma (BLCA), breast ductal carcinoma (BRCA), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck (oral) squamous cell carcinoma (HNSC), kidney chromophobe (KICH), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamos cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM), thyroid carcinoma (THCA), uterine corpus endometrial carcinoma (UCEC), uveal melanoma (UVM).

6.4 Exon array processing (Pub II)

Quantification of both gene- and exon-level expression from the 38 exon arrays of our DLBCL samples was performed with MEAP [135] using MEAP probe annotation version 70. Differentially expressed exons (DEEs) were annotated by their genomic locations (5′ untranslated region (UTR), 3′ UTR, coding, noncoding and unknown). Domain analysis was done for all coding DEEs by translating coding exonic regions into peptide sequences using Ensembl API [136] (version 70) and fetching domain information (Pfam, SMART, SignalP and TMHMM)[8] for all peptide sequences with InterProScan [137] (version 5). The phosphorylation sites identified within peptide sequences were compared against all known phosphorylation motifs downloaded from PhosphoSitePlus [138].

6.5 Survival analysis (Pub II)

For survival analysis we used Cox proportional hazards model and Kaplan-Meier method, the Kaplan-Meier curves were compared using log rank test. For overall survival (OS) the time interval considered was from the date of study entry or diagnosis to the date of the last follow-up or death from any cause; for progression-free survival (PFS) the time interval began with date of registration or diagnosis and ended with date of progression or death of any cause, and for disease-specific survival the date of registration to the date of death due to lymphoma were used. The analysis was done using IBM SPSS Statistics 22.0 (IBM, Armonk, NY, USA).

6.6 Phasing (Pub III)

Variants were called with Bambino v1.06 on matched tumor and normal WGS TCGA alignments from 925 patients. The resulting germline genotypes were population phased with EAGLE2 [139] v2.3 using the 1000 Genomes Phase 3 panel. Both WGS and RNA-seq bam files were used for read-back phasing with phASER [119] v1.0.0 with default parameters except for reads mapping quality set to (MAPQ) $\geq$30 and with a base quality $\geq$10. Only overlapping heterozygous sites were used. The resulting phased genotypes were imputed into 1000 Genomes Phase 3 with Minimac3 [140] v2.0.1. The bottom 30% of samples by number of variants

[8] Pfam is a database of protein families that includes their annotations and multiple sequence alignments generated using hidden Markov models. SMART (Simple Modular Architecture Research Tool) is a web resource (http://smart.embl.de) for the identification and annotation of protein domains and the analysis of protein domain architectures. SignalP annotates signal peptides, which are short peptide (usually 16-30 amino acids long) present at the N-terminus of the majority of newly synthesized proteins that are destined toward the secretory pathway. TMHMM is a method for prediction transmembrane helices based on a hidden Markov model.

called and median EAGLE phase confidence across autosomes were discarded, resulting in 615 individuals left for downstream analysis.

6.7 Test for variant penetrance (Pub III)

Our hypothesis is that a regulatory variant, or eQTL, can affect the penetrance of a pathogenic coding variant by increasing the dosage of the mutant protein in cancer patients. Therefore, the null hypothesis is that eQTL mediated expression has no effect in the penetrance of a mutation. In this study, as a proxy for variant penetrance, since penetrance itself is difficult to measure, we quantify the frequency at which the major allele is observed in the same haplotype as the lower expression causing eQTL. If the null hypothesis holds, a random mutation would occur on random haplotypes irrespective of eQTL genotype. The opposite, that the combination of mutation and higher expression causing eQTL, are observed more frequently together than expected by chance, would signify that eQTL mediated expression is playing a role in variant penetrance in cancer. To test our hypothesis we selected only heterozygous variants on tumor suppressor genes (compiled from Tumor Suppressor Gene Database, `https://bioinfo.uth.edu/TSGene/` on August 2017). Only variants with a Combined Annotation Dependent Depletion (CADD)[141] score $>$15 were considered. For each gene, when more than one eQTL was reported, we selected the most significant one from GTEx v6p [142] across all tissues. For each heterozygous somatic variant we counted the combination of wild type, or functional variant, and the lower-expressed allele as success for each sample. We applied a binomial test for each haplotype combination separately.

6.8 Fusion genes detection & prioritization (Pub IV)

FUNGI (v.1.0) was used for calling fusion genes with five different fusion callers and selecting the candidates for validation. FUNGI is a toolset for identifying, annotating, filtering, and scoring fusions which is described in detail in Results section 6.2. Using FUNGI's FusionCaller module, the five methods used for detecting fusions were 1) SoapFuse [143] (v1.27), 2) FusionCatcher [144] (v1.00), 3) EricScript [145] (v0.5.5), 4) ChimeraScan [146] (v.0.4.6) and 5) STAR-Fusion [115](v1.1.0). Fusions were filtered depending on the caller as follows: score $>$ 0.5 (EricScript), Counts of common mapping reads $<$ 30 (FusionCatcher), Overlapping Same = true (ChimeraScan), and LargeAnchorSupport = YES_LDAS (STAR-Fusion). Using FUNGI's FusionAnalyzer module, fusions were post-processed. First, fusions were combined into a standardized format and matched against Ensembl's database [147] for removing false positives (fusions between

paralog/homolog genes, not matching Ensembl gene coordinates or neither gene having a known function). For removing fusions reported previously as known artifacts or present in non-cancerous tissue, FusionCatcher's database annotation was used on the combined list of fusions called from all methods. As the last step of FusionAnalyzer, Pegasus [116] and Oncofuse [117] were used for estimating the probability of the fusions being oncogenic; and a respective score of 0.5 and 0.25 was used to select fusions for further analysis. Furthermore, fusions detected by only EricScript were discarded. The fusions that were reported in more than 10% of the patients were excluded after confirmation of previous reports of the fusions in healthy individuals, although they are still missing from databases. The remaining fusions were used as input for FUNGI's FusionVisualizer module that runs FusionInspector [115]. Confirmed fusions with a junction read count ≥ 3 and FFPM > 0.1 were manually inspected in IGV [148]. To compare frequency of fusion genes detected in our HGSOC sample set and other previously reported fusions we utilized FUNGI's FusionVisualizer module with FusionInspector (included in STAR-Fusion v2.7.0f_0328) for supervised fusion calling on 424 TCGA ovarian cancer samples.

6.9 Additional resources (Pub I-IV)

Pipelines for Publications I, II and IV were constructed in Anduril [149][150]. In addition to the software already included in this section, the following, non-exhaustive, list of resources were extensively used in the completion of the work presented in this dissertation: awk, annovar [151], bash, bcftools [152], GeneCards [153], KEGG [154], python, R, samtools [155], vcftools [156], and UCSC [157] and Ensembl websites and tools.

7 Results

The main results presented in this dissertation are two workflows for processing and analyzing RNA-seq data and three applications of transcriptomics for studying alternative splicing, fusion genes and possible effects of eQTLs in variant penetrance, all of them in the context of cancer studies.

7.1 SePIA a workflow for RNA-seq analysis (Pub I)

SePIA (Sequence Processing Integration and Analysis) is an open-source processing workflow for RNA-seq data implemented in Anduril. Anduril is an analysis and integration framework that facilitates the design, use, parallelization and reproducibility of bioinformatics pipelines. The first aim of SePIA is to make available a selection of state-of-the-art RNA-seq tools and methods with minimal effort to run. For this purpose, SePIA includes a standard RNA-seq analysis pipeline that performs alignment, quantification and differential expression analysis. More complex pipelines can be constructed using any of the close to 400 available Anduril components[9]. The second aim of SePIA is to automatically organize computational results in reproducible, presentable, and easy-to-use formats for downstream analysis. Figure 9(a-c) illustrates some of the web-reports created by SePIA on quality control of the raw reads (fastq), alignment and expression quantification. Finally, the third main aim of SePIA is to facilitate the integration of miRNA-mRNA analysis for which several components and an example pipeline are included with SePIA. To showcase the potential of SePIA for conducting standard RNA-Seq analysis and for integration of mRNA and miRNA data, we analyzed two cancer datasets from TCGA using SePIA. The first case study consists of tumor and normal RNA-seq, while the second one combines both mRNA and small RNA datasets. A summary of the analysis steps performed in each case study is shown in Table 3. SePIA's documentation is available at `https://anduril.org/sepia/`. A snapshot of the report created for our case study of top miRNA-mRNA anti-correlated pairs is shown in Figure 9(d).

[9] Reusable code units that encapsulate common pre-processing and analysis steps. Anduril's, and therefore SePIA's, pipelines are constructed by linking components together through their inputs and outputs.

Figure 9: A snapshot of the reports created by SePIA for the case studies. **a** Small RNA preprocessing report for Case II, including FastQC results organized by patient sample. **b**, **c** Alignment and expression statistics for Case I with some standard visualization. **d** The searchable miRNA-target mRNA report for Case II.

Analysis	Software	Case Study
Quality Control	FastQC	I & II
Trimming	Trimmomatic	I & II
	FASTX Toolkit	II
Alignment	STAR	I & II
	Bowtie	II
Expression	Cufflinks	I & II
	HTSeq	II
Differential expression	DESeq2, Cuffdiff, DEXSeq	I
	DESeq, edgeR	II
Variant Calling	Bambino, GATK	I
Prediction (miRNA)	mirAnalyzer	II

Table 3: Analysis and tools used in the two cancer case studies. Each of the tools used in the analysis has a corresponding component implemented in Anduril for use in SePIA. Component documentation is available in www.anduril.org.

7.2 FUNGI a toolset for identifying, integrating, and prioritizing fusion genes (Pub IV)

FUNGI (FUsioN Genes Integration toolset) consists of three modules that have been designed to undertake the following main tasks of a fusion gene analysis pipeline: 1) fusion calling, 2) prioritization of fusions and 3) supervised fusion calling and visualization.

The first of these modules, FusionCaller, facilitates the execution of seven different fusion calling algorithms: Arriba (https://github.com/suhrig/arriba/), STAR-Fusion, FusionCatcher, SoapFuse, Chimerascan, deFuse [158] and EricScript. FusionCaller takes raw RNA-Seq reads in compressed fastq format and outputs a list of detected fusions. This list of fusions can be used directly as input for the next module, FusionAnalyzer, or after filtering using tool-specific criteria first.

FusionAnalyzer takes as input fusions with minimum information: gene names, breakpoint coordinates, and number of junction and spanning reads. In this manner, FusionAnalyzer can process not only fusions called by any of the seven supported methods by FusionCaller, but also from fusions identified by other tools, given that breakpoints are provided. FusionAnalyzer verifies that the fusion reported coordinates match current Ensembl annotation and (optionally) filters fusions whose genes are homologous, belong to the same family or that neither of them have a reported function in Gene Ontology (GO). The valid fusions, annotated by FusionAnalyzer with Ensembl ids, are queried against over 20 databases of fusion-gene detection projects, provided by FusionCatcher, to identify fusions previously reported in both cancer and non-cancerous tissues. Users can discard fusions based on these database annotations and score the remaining fusions with Pegasus and Oncofuse. FusionAnalyzer's final output is an integrated list of scored and annotated fusions from all analyzed, methods and samples, which allows for the swift identification of recurrent fusions if present.

Finally, FusionVisualizer, the third of FUNGI's modules, takes a list of fusions as input together with RNA-seq reads. FusionVisualizer can either recreate the exact fusion (if breakpoints are provided) and then map the reads to the newly created virtual reference for the fusion, or it can be used to search for fusions, independent of breakpoint, in the provided samples. The latter is achieved using FusionInspector which allowed us to investigate if fusions from our discovery cohort were also found in TCGA ovarian cancer patients. An overview of FUNGI toolset is shown in Figure 10. FUNGI's modules are available as part of the Anduril framework (`https://anduril.org`) or as a standalone version at `https://bitbucket.org/alejandra_cervera/fungi`.

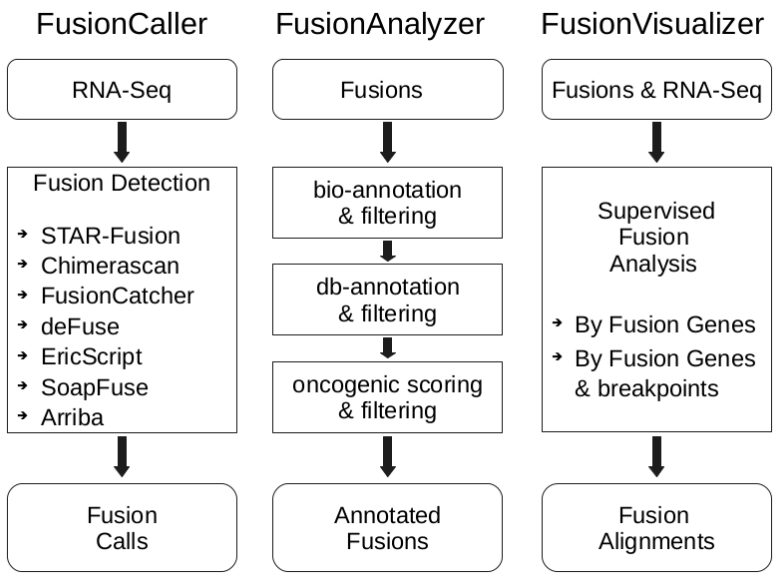

Figure 10: FUNGI is divided into three main modules for 1) calling, 2) filtering, annotation, and scoring, and 3) supervised detection and visualization of fusions. FusionCaller currently supports six fusion calling algorithms, takes RNA-seq reads as input, and outputs the detected fusions by the selected algorithm. FusionAnalyzer aids in filtering fusions that are likely false positives or that have been previously reported in databases, and also scores the oncogenic potential of the genes. The input for FusionAnalyzer is fusion calls taken directly from FusionCaller output or fusions detected by any other algorithm provided in a standard format. The output is a filtered table of annotated fusions scored with Pegasus and Oncofuse. FusionVisualizer can be used for inspecting a fusion by recreating the exact fusion and mapping the reads to the virtual reference for visualization in a genome browser, or by using FusionInspector to search for fusions between the same gene partners in the same or different samples. One of the inputs of FusionVisualizer, in the first case, is a list of fusions with the breakpoint coordinates for each gene, while for the second case only the gene names are needed. In both cases, FusionVisualizer requires also as input the RNA-seq reads to inspect. The output is a virtual reference and the alignments that can be loaded to a genome browser for inspection if the fusion was found.

7.3 Association of alternative spliced genes with survival in DLBCL (Pub II)

We compared good prognosis (n=29) versus bad prognosis (n=9) exon array data from 38 tumor samples of DLBCL patients. Good prognosis patients have been in remission over 24 months while poor prognosis relapsed after chemotherapy. Differential expression analysis revealed 220 genes of which 59% were down regulated in poor prognosis and 41% were overexpressed. Analysis of alternative

splicing yielded 3,888 genes that had at least one exon differentially expressed, but that the gene itself was not. We performed the same analysis using 92 tumor RNA-Seq samples from DLBCL patients from CGCI where we found 547 (from the 3,888) genes also alternatively spliced, of which 33 genes matched exactly (same exons and same direction) the results from the discovery cohort. In Table 11 we show the 29 out of the 37 DEEs that we found associated with PFS ($P \leq 0.05$) according to Cox univariate analysis and 20 DEEs associated with OS in the CGCI cohort. From the validated genes, APH1A (anterior pharynx defective-1α) a component of the $\gamma-$secretase complex that cleaves integral membrane proteins such as Notch receptors and $\beta-$amyloid precursor protein) was one of the top genes both in OS and PFS in the CGCI cohort. Exon domains and Kaplan-Meier curves for APH1A are showin in Figure 12.

Gene	*Discovery cohort*				*Validation cohort*			
	PFS		*OS*		*PFS*		*OS*	
	P-val exon	*P-val gene*	*P-val exon*	*P-val gene*	*P-val exon*	*P-val gene*	*P-val exon*	*P-val gene*
CUL3	0.070	0.975	0.167	0.728	**0.007**	0.393	**0.024**	0.399
DKK3	0.091	0.330	0.099	0.427	**0.037**	0.497	0.064	0.623
BCAR1	**0.022**	0.179	0.108	0.499	**0.005**	0.439	**0.002**	0.449
SLC9A3	0.112	0.154	0.084	**0.016**	**0.002**	0.164	**0.002**	0.139
GAL	0.051	0.938	0.468	0.475	**0.003**	0.687	**0.033**	0.659
ABCB1	0.116	0.525	0.204	0.806	**0.019**	0.343	**0.045**	0.895
NAMPT	**0.016**	0.757	0.138	0.401	**0.015**	**0.003**	0.121	**0.004**
CUBN	**0.035**	0.285	**0.038**	0.246	**0.047**	**0.024**	0.079	**0.018**
APH1A	**0.048**	0.123	**0.039**	0.043	**<0.001**	**0.011**	**<0.001**	**0.016**
SSUH2	**0.010**	0.201	**0.015**	0.113	**0.022**	0.396	0.400	0.094
RHOT1	0.099	0.597	0.126	0.610	**0.001**	**0.023**	**0.002**	**0.027**
NHSL1	**0.012**	0.732	**0.037**	0.700	**<0.001**	0.153	**0.007**	0.354
TACC2	0.091	0.973	0.177	0.728	**0.002**	0.183	0.087	0.245
FRAS1	**0.002**	**0.039**	**0.001**	0.051	**0.001**	0.107	**0.033**	0.085
PTPRQ	**0.013**	0.862	0.170	0.861	0.130	0.404	**0.018**	**0.032**
CYP4B1	0.051	0.327	0.054	0.150	**0.013**	0.269	0.136	0.268
HMGCLL1	**0.030**	0.060	0.597	0.299	**0.011**	0.834	**0.023**	0.395
FAM83A	**0.018**	0.207	**0.002**	0.348	**0.003**	0.865	**0.029**	0.671
CPB1	**0.020**	0.330	**0.024**	0.228	**0.006**	0.394	**0.002**	0.185
SORBS2	0.104	0.418	0.285	0.962	0.101	0.151	0.417	0.149
C2orf65	**0.042**	0.273	0.361	0.583	**<0.001**	0.238	**0.010**	0.173
CAMK2N1	**0.011**	0.433	**0.009**	0.794	0.489	**0.028**	0.986	0.094
NOS3	**0.022**	0.246	**0.031**	0.687	**0.034**	0.753	0.064	0.760
KCNH6	**0.015**	**0.019**	0.080	**0.008**	**0.001**	**0.050**	**0.001**	**0.002**
TRIML2	**0.005**	**0.005**	**0.008**	**0.003**	**0.015**	0.338	**0.019**	0.859
MAP3K15	**0.011**	**0.024**	0.142	**0.025**	**0.001**	0.578	**0.003**	0.217
TMEM232	**0.030**	0.359	0.172	0.298	**0.001**	**0.034**	**0.001**	**0.003**
CYHR1	**0.041**	0.167	0.086	**0.019**	**0.001**	0.724	**0.001**	0.586
VEPH1	0.125	0.092	0.282	**0.04**	**<0.001**	**0.025**	**0.006**	0.062
AC018705.5	**0.013**	0.061	**0.012**	**0.045**	**0.018**	0.920	0.078	0.876
SPANXA2_OT1	**0.029**	**0.048**	0.151	0.076	0.124	0.053	0.153	0.200
RP11_69C17.1	**0.007**	0.400	**0.029**	0.636	**0.012**	0.246	0.131	0.561
RP11_696N14.1	0.068	0.160	0.614	0.447	**0.017**	0.982	0.190	0.691

Table 11: Cox univariate analysis of the DEEs common in the discovery and validation cohorts (significant P<0.05 are in bold). Abbreviations: DEE, differentially expressed exon; OS, overall survival; PFS, progression-free survival.

7.4 Regulatory modifiers of coding variants contribute to cancer risk (Pub III)

When loss of function mutations occur in tumor suppressor genes, for cells to turn from normal to cancerous ones, usually the mutation needs to occur in both alleles (the two-hit hypothesis in cancer). We wanted to investigate if haploinsufficiency,

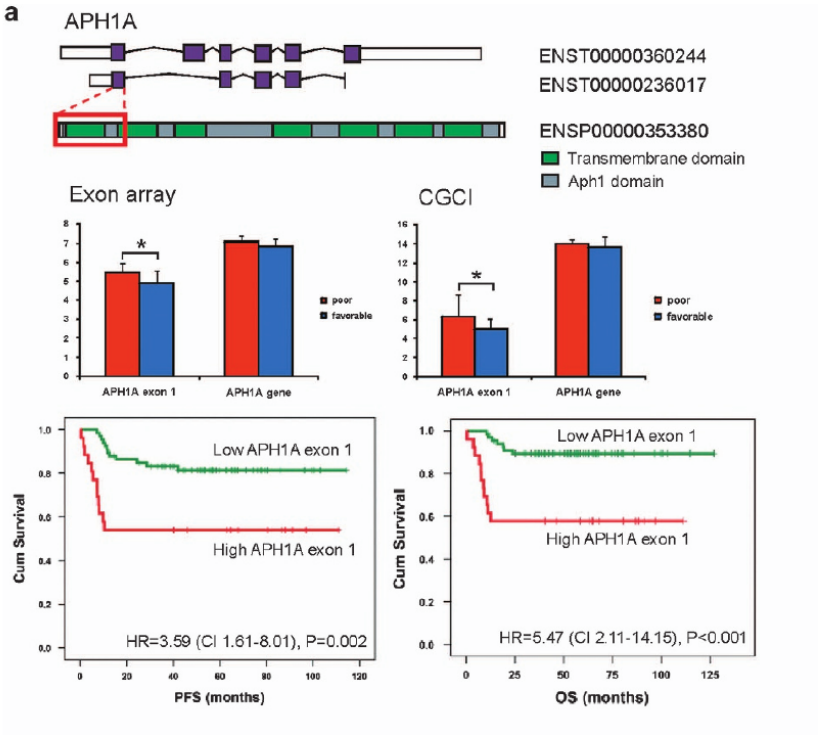

Figure 12: Differentially expressed exons may affect the functional properties of the protein and are associated with survival. The upper panel shows the domain information, the middle panel shows the exon and gene expression in the discovery and validation cohorts and the lower panel shows Kaplan–Meier survival plots of the exons in DLBCL patients (validation cohort).

as opposed to full gene inactivation, could have a role in cancer risk. In this case haploinsufficiency of a tumor suppressor gene can be caused by a pathogenic mutation occurring on a major expressed allele, whilst the wild type variant would lie on a minor expressed allele mediated by eQTLs. While in cancer we would expect to see an enrichment of deleterious mutations on higher expressed alleles, in the general population, purifying selection should deplete haplotype combinations that give higher penetrance to pathogenic variants. We first tested this hypothesis using the GTEx dataset, which does not include individuals with severe diseases. For each of the 44 tissues from the GTEx project we calculated the expression of coding variant minor alleles using allelic fold change and compared the expression of missense variants with allele frequency matched synonymous controls. Rare missense pathogenic variants (CADD $>$ 15) showed a significant difference in allele expression in comparison to synonymous controls, while rare but deemed benign (CADD $<$ 15) did not. Next, to study the role of regulatory modifiers in germline

risk of cancer, we compared the frequency of haplotype combinations between 615 TCGA cancer patients and 620 GTEx individuals. The analysis was stratified by considering first all rare pathogenic variants from both cases and controls, then the variants unique to each group, and finally variants shared between them, Figure 13. We found that the major allele was significantly more often found in combination with the lower-expression causing eQTL (P=0.00953), while the control specific variants showed no difference.

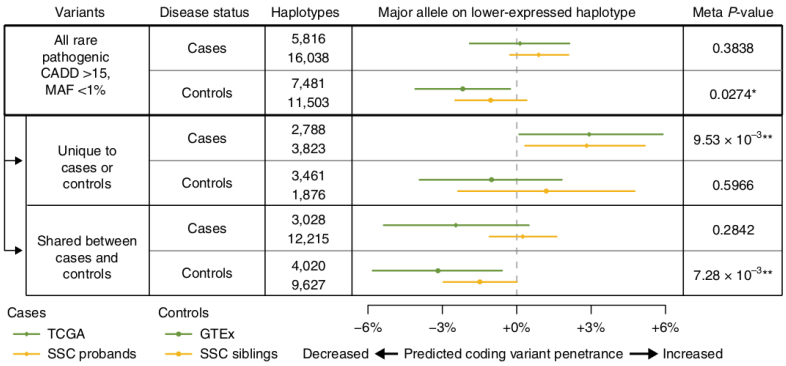

Figure 13: eQTL haplotype configurations that are predicted to increase pathogenic coding variant penetrance are enriched in individuals with cancer and autism spectrum disorder. Analysis of eQTL and coding variant haplotype configurations in cases and controls for autism spectrum disorder (ASD) and cancer, using the top GTEx v6p eQTL per gene by P-value across all tissues. For cancer analysis, haplotype configurations generated from population and read-back phased germline whole genomes of 615 TCGA individuals (cases) and 620 whole genomes of v7 GTEx individuals (controls) were used, and haplotypes were analyzed at tumor suppressor genes. To enrich for putatively disease-causing variants, results were stratified based on whether variants were restricted to cases or controls or shared between both. Median estimates and 95 confidence intervals were generated using 100,000 bootstraps, and two-sided empirical P-values were generated from these confidence intervals and combined between cohorts using Fisher′s method to produce meta p-values.

7.5 Fusion genes in HGSOC (Pub IV)

With the aim of identifying recurring fusion genes in HGSOC we applied FUNGI to 107 RNA-Seq samples from 36 patients. After selecting the most reliable candidates in terms of low probability of being false positive and high probability of oncogenic potential, we tested if the fusions can also be detected in a sample set of 423 TCGA ovarian samples. We were able to identify a previously reported recurrent fusion in HGSOC, CCDC6–ANK3, which was reported at 1% frequency in both Earp and TCGA. We identified the fusion in two of our patients, but we were able to also detect it in 11 TCGA samples, which yields a frequency of 5% and 2%, respectively.

We also found an enrichment on fusions involving a gene from the PI3K-Akt pathway. The PI3K-Akt pathway is an intracellular signal transduction pathway involved in metabolism, proliferation, cell survival, growth and angiogenesis in response to extracellular signals. A list containing the fusions identified with one gene from the pathway in our dataset is included in Table 4 with information of fusions involving the same genes in TCGA. From the PI3K-Akt pathway genes, we validated with Sanger sequencing AKT2–PBK4 and AKT2–ZNF546, PIK3R1–CCDC178, and PTK2–AGO2 fusions. Both PTK2 and AGO2 are on chromosome 8 only 22,000 bases apart in a region commonly amplified in ovarian cancer (reported amplified in 25 and 27% of samples, respectively). We also detected FGFR–TACC3, which is one of the most common fusion genes in TCGA, reported in cervical cancer, bladder carcinoma, GBM, squamous lung carcinoma, and HNSC, but not in ovarian cancer. Several FGFR inhibitors are currently in clinical trials.

FusionGenes	N	TCGA Partner	NT	OV
AKT2–ZNF546, **AKT2**–PBX4	6	SRRM5, UPF1, ELK3, NFYA, Y1BF1	5	NO
ATF4–MIEF1	1			
CHD9–**CCND3**	1	TRERF1, GUCA1B, TBC1D22B, C6orf89, VPS52, EXOC4, GMDS, EHMT2, SRPK1, LUZP1, MED20	14	NO
TRIP12–**CREB1**	1	FASTKD2, SUCNR1	2	NO
FGFR3–TACC3	1	TACC3	36	NO
GSK3B–ASTE1	3	C3orf15, FSTL1, GPR156, LSAMP, COL8A1, WNK2, ATP11B, PLA1A, SEMA5B, TSC22D2, BEST3, RASSF2	13	NO
HRAS–ANO9	1	RNH1	1	NO
UBE2E3–**ITGA4**	1	UBE2E2	1	NO
MED24–**ITGB3**	1			
PCSK5–**JAK2**	2	CDK12, KDM4C, SIK2, TTC13, C9orf46, RCL1, GLDC, DOCK8, CSTF3, C6orf204, PKNOX2, CTD-2021H9.3	13	NO
KRAS–SSPN	1	RERGL, IFLTD1	2	YES
LAMC1–NPL	3	C1orf21 ,UTRN, EIF4G3	3	NO
MAGI1–LRIG1	1	GAK, DPH3, LRIG1, SUCLG2, TBC1D30, MITF	6	NO
PIGU–**MAGI2**	2	PILRB, VWC2, SLC25A13	3	NO
PIK3R1–CCDC178	2	NDUFB7, RP11-404L6.2, MRPL42, MARVELD2	5	NO
PIK3R3–NFIA	1	LRRC41, ZSWIM5, NASP, MYO18A, SPATA6	6	NO
PPP2CB–ACLY	4	PURG, DLC1, PTPRJ	4	YES
PPP2R3A–PCCB	1	EIF4G3, MSL2, EEFSEC, HKR1,FRMD4B, OPA1, SPSB4	7	NO
PTK2–AGO2	2	UG2T2A3, VPS13B, EIF2C2, AC016722.1, AC016722.2, NCR3, PPP2R5E, TRAPPC9, PTDSS1, ANKRD11, CACNG8, RNF139, SLC45A4, UBE2H, CCDC91, DENND3, MKLN1	24	YES
FTO–**RBL2**	1	FTO	1	NO

Table 4: Fusion genes with a gene from the PI3K pathway (shown in bold). N = number of samples with the fusion, NT = number of TCGA samples with the fusion. If the PI3K gene has been identified in TCGA with a different partner, the partners are listed in column 3, and if in TCGA ovarian cancer it is marked in column 5.

8 Discussion

Next-generation sequencing has transformed cancer research into a data-rich field. Individual laboratories nowadays are capable of producing considerable amounts of data which can be deposited in public repositories for other scientists to explore. Furthermore, multi-institutional collaborations, at both national and international levels, have arisen with the specific aim of systematically sequence and analyze whole genomes and transcriptomes of both cancer and healthy tissues, as well as explore a variety of other functional components of the genome. This explosion of data resources requires constant development of computational frameworks and methods to help scientists integrate, analyze, and draw conclusions from this vast amount of data.

The work presented in this thesis in Publications I and IV, aims at aiding in the labor of analyzing large datasets. Both SePIA and FUNGI are toolsets that facilitate the creation of reproducible pipelines for investigating different aspects of the cancer transcriptome. SePIA allows differential expression analysis and mRNA-miRNA integration. Its utility is showcased with the analysis of datasets from GEO and TCGA. On the other hand, FUNGI is aimed specifically at finding reliable gene fusions with oncogenic potential. To demonstrate FUNGI's features, we analyzed 107 in-house samples and processed over 400 public samples from TCGA. We also integrated a variety of published methods for fusion calling, developed our own strategy for fusion visualizing, and combined everything in a workflow of our own design, that includes the use of over 20 databases for annotating fusions.

SePIA and FUNGI are tools that can be used by the community to explore their datasets and contribute to the acquisition of knowledge in the field of cancer genetics with next generation sequencing. A limitation of workflows such as SePIA and FUNGI is that although the aim is to automate to the maximum the process of data analysis, expertise is needed to interpret results and suggest follow up experiments. Furthermore, both FUNGI and SePIA rely on knowledge deposited in databases and the quality of the results is directly affected by the quality of the information deposited in those databases. For example, recently it has been observed that proven oncogenic fusions can sometimes be computationally detected in healthy tissue samples, which complicates the matter of automating filtering based on database annotations [159]. For miRNAs or pathway analysis of differentially expressed genes, the reliance is on the databases that identify the functions of the genes. It has been observed that results are affected by how often this databases are updated. Additionally, working pipelines are hard to replace even when more reliable methods are published. For this reason, it is important to create software that is modular and well documented to allow rapid integration and substitution of

tools.

The biological results presented in this work aim to help in the understanding of cancer mechanisms. In Publication II we identified that alternative splicing is better at discriminating between cancer subtypes of DLBCL than more traditional approaches of differential expression at gene-level. The classification of patients with a similar diagnosis into different subgroups can help physicians into deciding treatments and can improve outcomes of clinical trials. Additionally, we identified isoforms positively associated to survival, which can promote studies into specific gene variants or into the role of alternative splicing in DLBCL. A reason why differential exon usage is not as often explored as differential gene expression is that pathway enrichment analysis at isoform level is still lacking. Even if different isoforms are identified, not many databases have documented functional differences between splice variants. In Publication III, we investigated if haploinsufficiency of tumor suppressor genes could be caused by genetic variation. This study was more a proof of concept than a dedicated analysis to shed light on a specific cancer type. We found enrichment of haplotype combinations that up-regulate the expression of mutation-carrying tumor suppressor alleles. Hopefully, this work will motivate allele specific expression and eQTL analysis in cancer, which although several such studies exist, is not part of the standard analysis. Finally, in Publication IV, the combination of different methodologies for fusion gene calling showed a higher prevalence than previously reported of some fusion genes identified in other cancer datasets. Furthermore, we identified a fusion not previously reported in HGSOC for which targeted drug trials are currently on-going and can result in an increase of treatment options for carriers of the mutation irrespective of cancer type.

Transcriptomics is a very active field of research both for developments on technology and computational methods. Long read sequencing for transcriptomics will facilitate alternative splicing analysis and detection of fusion genes, but currently the low accuracy remains an important challenge that needs to be overcome before it can be effectively used. Another technological development, single cell sequencing, allows the identification of individual cell types and the characterization of cell states within tumors or tissues. An important limitation of single cell RNA-Seq is the reliance on poly-A selection for library preparation and the amount of genes that can be quantified at a time. Both long read sequencing and single cell have spurred the development of tailored computational algorithms to exploit these new technologies. Nevertheless, methods that combine the use of short and long reads or bulk and single cell RNA-Seq are still currently needed. The study of long non coding RNAs has been somewhat neglected due to lack of proper tools to understand their functions, but developments in both wet and dry lab methodology such as CRISPR/CAS9 and deep learning algorithms are already proving useful in

characterizing these transcripts and understanding their role in gene regulation in cancer.

Acknowledgements

This work was carried out in the Systems Biology of Drug Resistance in Cancer Laboratory at the Faculty of Medicine, University of Helsinki during 2013-2020. I thank my supervisor Sampsa Hautaniemi for his support and for the many opportunities provided during my PhD. I thoroughly enjoyed the multi-disciplinary meetings held with collaborators as part of the HERCULES-EU-project, which allowed me to better understand how my work fits in the overall picture. Coding camps, although very intense, were a great way of changing pace, explore ideas, and bond with lab mates. I am also very grateful for the encouragement and funding to do a research visit that helped me grow as a scientist. Finally, I thank Sampsa for always having an open door for me during the many years I had the privilege of working in his lab.

I thank the members of my thesis committee, Garry Wong and Mikko Frilander, for their time and their helpful feedback and suggestions. I am also grateful to the pre-examiners of my thesis, Rolf Skotheim and Joaquín Dopazo, for the time spent revising my thesis and their very kind comments on my work.

I thank the Orion Research Foundation and K. Albin Johanssons Stiftelse for their financial support. I also thank The Doctoral Programme in Biomedicine for both financial and academic support during my studies.

It has been a joy to work with most of the people I have crossed paths in the Hautaniemi lab. I felt very lucky when I joined the lab after getting to know my lab mates at the time: Tiia, Anna-Maria, Marko, Sirkku, Kristian, Riku, Ville, Ping, Chengyu, Erkka, Viljami, Rony, Lilli, Vladimir, Elena, and Javier; you made a great lab and I really enjoyed our discussions. I am also thankful to the students that joined shortly and not so shortly after me: Julia, Amjad, Chiara, Katherine, Emilia, Kaiyang; thanks for sharing with me in the experience of starting and completing a PhD. It was really fun to both work and hang out with you. Other students, postdocs and visitors to the lab that I am glad I had the opportunity to meet: Mikko Kivikoski, Mikko Kivelä, Lauri Lyly, Lauri Lapatto, Antti, Jaana, Ingrid, Yilin, Kari, Oskari, Veli-Matti, Valeria, Gabriele and Enrica. To all of you I am also thankful for the camaraderie during the many social events we attended together.

I am thankful for the collaborators in Helsinki I had the opportunity to work with throughout my PhD, I learned from all of you: Jukka Lehtonen, Harri Sihto, Kaisa Huhtinen, Heidi Rausio, Sirpa Leppä, Suvi-Katri Leivonen, Minna Taskinen, Leo Meriranta, Anniina Färkilä and Barun Pradhan.

I am also very grateful to Tuuli Lappalainen for receiving me in her lab; it was a very motivating experience to work in a project so different from my own projects,

but still within the field of cancer and gene expression. I was able to learn so much and experience a different way of doing science and being mentored. Special thanks to Stephane Castel for his guidance on our work together. Heartfelt thanks to Sarah Kim and the rest of the lab in NYGC for their support during my visit.

I thank Dra. Marta Menjívar for giving the opportunity of joining her group in Mérida where I have been able to delve into fascinating new research topics. To my new lab mates I am thankful for their support and for sharing with me the last stage of my PhD: Bárbara, Shérlin, Enrique, Rachel, Carla and Alfredo.

Special thanks to all the members of the rowing team, I can't say I always enjoyed rowing with you for six hours, but the trainings, the pizzas and beer, and the Sulkava trips were definitely a highlight of my time in Finland.

My warmest thanks to Tiia for her friendship all this years, for being there through the ups and the downs, and pulling me through when things got complicated. Ville thanks for always having my back and everyone else's, you inspire me to try to be a kinder and more patient person. Erkka thanks for the friendship and the many discussions and beers shared. Julia, Edu, Maria, Tatiana, thanks for being around. I miss you all.

To all my friends that supported me through my PhD studies and specially for coming to visit to Helsinki: Sergio, Karla, Nico, Victor, Bárbara, Cheko, Marianita, Alex, and Mariana Esther. I am glad Antonio and I were able to share with you our time in Finland and thanks for staying close. Kim, Vignesh, Daniel, Abi, and Daniela, thanks for being always there to catch up.

I thank my family, sisters, brothers, aunts, uncles, cousins, nieces and nephews for staying close despite the distance, and especially my mom María Antonieta for her unconditional love and support. To my *madre postiza* Martha I am thankful for all the years she was part of my life, she will be forever missed. Finally, I thank Antonio for always rooting for me, for inspiring me in science and in life, and lastly for providing me with coffee every morning; I would have never been able to do this without you.

The writing of this thesis happened during the COVID-19 lockdown in Mexico. I am very thankful to Dr. Hugo López-Gatell for his outstanding job as a science communicator and for the display of empathy and social concern during the daily pandemic press conferences. My deepest gratitude to all health personnel organizing the response and taking care of the sick during these trying times.

Alejandra Cervera Taboada
Mérida, 2020

References

[1] The Institute for Health Metrics and Evaluation (IHME) (2020) (http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2017-permalink/be6aa651170badea5e7a9124389f5737). Cited in page 1.

[2] Hanahan D, Weinberg RA (2000) The Hallmarks of Cancer. *Cell* 100(1):57–70. Cited in pages 1, 3.

[3] The Cancer Genome Atlas Program - National Cancer Institute (2018) (https://www.cancer.gov/tcga). Cited in page 1.

[4] (2013) The Genotype-Tissue Expression (GTEx) project. *Nature genetics* 45(6):580–585. Cited in page 1.

[5] Cancer Genome Characterization Initiative (2013) (https://ocg.cancer.gov/programs/cgci). Cited in page 1.

[6] Dunham I, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74. Cited in page 1.

[7] Regev A, et al. (2017) The Human Cell Atlas. *eLife* 6. Cited in page 1.

[8] Cooper GM (2000) Regulation of Protein Function. *The Cell: A Molecular Approach. 2nd edition*. Cited in page 3.

[9] Lodish H, et al. (2000) Cell Death and Its Regulation. *Molecular Cell Biology. 4th edition*. Cited in page 3.

[10] Hanahan D, Weinberg RA (2011) Hallmarks of Cancer: The Next Generation. *Cell* 144(5):646–674. Cited in page 3.

[11] Shay JW, Bacchetti S (1997) A survey of telomerase activity in human cancer. *European Journal of Cancer* 33(5):787–791. Cited in page 4.

[12] Fouad YA, Aanei C (2017) Revisiting the hallmarks of cancer. *American Journal of Cancer Research* 7(5):1016–1036. Cited in page 4.

[13] Yang M, et al. (2015) The involvement of osteopontin and matrix metalloproteinase-9 in the migration of endometrial epithelial cells in patients with endometriosis. *Reproductive biology and endocrinology: RB&E* 13:95. Cited in page 5.

[14] Baker S, et al. (2017) Cancer Hallmarks Analytics Tool (CHAT): A text mining approach to organize and evaluate scientific literature on cancer. *Bioinformatics* 33(24):3973–3981. Cited in page 5.

[15] Knudson AG (1971) Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* 68(4):820–823. Cited in page 6.

[16] Inoue K, Fry EA (2017) Haploinsufficient tumor suppressor genes. *Advances in medicine and biology* 118:83–122. Cited in page 6.

[17] Cooper GM (2000) Tumor Suppressor Genes. *The Cell: A Molecular Approach. 2nd edition*. Cited in page 6.

[18] Kurayoshi K, et al. (2018) The Key Role of E2F in Tumor Suppression through Specific Regulation of Tumor Suppressor Genes in Response to Oncogenic Changes. *Gene Expression and Regulation in Mammalian Cells - Transcription Toward the Establishment of Novel Therapeutics*. Cited in page 6.

[19] Cheng Q, Chen J (2010) Mechanism of p53 stabilization by ATM after DNA damage. *Cell cycle (Georgetown, Tex.)* 9(3):472–478. Cited in page 6.

[20] Powell SN, Kachnic LA (2003) Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. *Oncogene* 22(37):5784–5791. Cited in page 6.

[21] Kuchenbaecker KB, et al. (2017) Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA* 317(23):2402–2416. Cited in pages 7, 9.

[22] Barnetson RA, et al. (2006) Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. *The New England Journal of Medicine* 354(26):2751–2763. Cited in page 7.

[23] Chipuk JE, Green DR (2006) Dissecting p53-dependent apoptosis. *Cell Death & Differentiation* 13(6):994–1002. Cited in page 7.

[24] Chen CY, Chen J, He L, Stiles BL (2018) PTEN: Tumor Suppressor and Metabolic Regulator. *Frontiers in Endocrinology* 9. Cited in page 7.

[25] Yang J, et al. (2019) Targeting PI3K in cancer: Mechanisms and advances in clinical trials. *Molecular Cancer* 18(1):26. Cited in page 7.

[26] Lee EY, Muller WJ (2010) Oncogenes and Tumor Suppressor Genes. *Cold Spring Harbor Perspectives in Biology* 2(10). Cited in page 7.

[27] Lessene G, Czabotar PE, Colman PM (2008) BCL-2 family antagonists for cancer therapy. *Nature Reviews Drug Discovery* 7(12):989–1000. Cited in page 8.

[28] Smith A, et al. (2015) Lymphoma incidence, survival and prevalence 2004–2014: Sub-type analyses from the UK's Haematological Malignancy Research Network. *British Journal of Cancer* 112(9):1575–1584. Cited in page 8.

[29] Pfreundschuh M, et al. (2006) CHOP-like chemotherapy plus rituximab versus CHOP-like chemotherapy alone in young patients with good-prognosis diffuse large-B-cell lymphoma: A randomised controlled trial by the MabThera International Trial (MInT) Group. *The Lancet. Oncology* 7(5):379–391. Cited in page 8.

[30] Alizadeh AA, et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403(6769):503–511. Cited in page 8.

[31] Devouassoux-Shisheboran M, Genestie C (2015) Pathobiology of ovarian carcinomas. *Chinese Journal of Cancer* 34(1):50–55. Cited in page 9.

[32] Jannetti SA, Zeglis BM, Zalutsky MR, Reiner T (2020) Poly(ADP-Ribose)Polymerase (PARP) Inhibitors and Radiation Therapy. *Frontiers in Pharmacology* 11. Cited in page 9.

[33] Goel MK, Khanna P, Kishore J (2010) Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research* 1(4):274–278. Cited in page 10.

[34] Bland JM, Altman DG (2004) The logrank test. *BMJ* 328(7447):1073. Cited in page 10.

[35] Rose AB (2019) Introns as Gene Regulators: A Brick on the Accelerator. *Frontiers in Genetics* 9. Cited in page 12.

[36] Stages of transcription (2020) (https://www.khanacademy.org/science/biology/gene-expression-central-dogma/transcription-of-dna-into-rna/a/stages-of-transcription). Cited in page 13.

[37] Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM (2000) Transcription and RNA polymerase. *An Introduction to Genetic Analysis. 7th edition*. Cited in pages 12, 13.

[38] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40(12):1413–1415. Cited in page 13.

[39] Matlin AJ, Clark F, Smith CWJ (2005) Understanding alternative splicing: Towards a cellular code. *Nature Reviews Molecular Cell Biology* 6(5):386–398. Cited in page 13.

[40] Alternative splicing produces three protein isoforms. (2014) (http://www.genome.gov/Images/EdKit/bio2j_large.gif). Cited in page 14.

[41] Oltean S, Bates DO (2014) Hallmarks of alternative splicing in cancer. *Oncogene* 33(46):5311–5318. Cited in page 14.

[42] El Marabti E, Younis I (2018) The Cancer Spliceome: Reprograming of Alternative Splicing in Cancer. *Frontiers in Molecular Biosciences* 5. Cited in page 14.

[43] O'Brien J, Hayder H, Zayed Y, Peng C (2018) Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology* 9. Cited in page 15.

[44] Krek A, et al. (2005) Combinatorial microRNA target predictions. *Nature Genetics* 37(5):495–500. Cited in page 15.

[45] Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* 19(1):92–105. Cited in page 15.

[46] Zhang W, et al. (2012) miR-181d: A predictive glioblastoma biomarker that downregulates MGMT expression. *Neuro-Oncology* 14(6):712–719. Cited in page 15.

[47] Buniello A, et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* 47(Database issue):D1005–D1012. Cited in page 16.

[48] Brown CD, Mangravite LM, Engelhardt BE (2013) Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLOS Genetics* 9(8):e1003649. Cited in page 16.

[49] Fairfax BP, et al. (2012) Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nature Genetics* 44(5):502–510. Cited in page 16.

[50] Castel SE, et al. (2018) Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics* 50(9):1327–1334. Cited in page 17.

[51] Walsh T, et al. (2011) Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 108(44):18032–18037. Cited in page 16.

[52] Kanchi KL, et al. (2014) Integrated Analysis of Germline and Somatic Variants in Ovarian Cancer. *Nature communications* 5:3156. Cited in page 16.

[53] Kinnersley B, et al. (2015) Genome-wide association study identifies multiple susceptibility loci for glioma. *Nature Communications* 6. Cited in page 16.

[54] Ongen H, et al. (2014) Putative cis-regulatory drivers in colorectal cancer. *Nature* 512(7512):87–90. Cited in page 16.

[55] Li Q, et al. (2013) A novel eQTL-based analysis reveals the biology of breast cancer risk loci. *Cell* 152(3):633–641. Cited in page 16.

[56] Kuchenbaecker KB, et al. (2015) Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nature genetics* 47(2):164–171. Cited in page 16.

[57] Whitington T, et al. (2016) Gene regulatory mechanisms underpinning prostate cancer susceptibility. *Nature Genetics* 48(4):387–397. Cited in page 16.

[58] LaCroix B, et al. (2014) Integrative analyses of genetic variation, epigenetic regulation, and the transcriptome to elucidate the biology of platinum sensitivity. *BMC Genomics* 15(1):292. Cited in page 17.

[59] Li N, et al. (2012) A combined array-based comparative genomic hybridization and functional library screening approach identifies mir-30d as an oncomir in cancer. *Cancer Research* 72(1):154–164. Cited in page 17.

[60] A schematic showing the ways a fusion gene can occur at a chromosomal level. Leonard G (2012) (https://en.wikipedia.org/wiki/Fusion_gene#/media/File:Gene_Fusion_Types.png). Cited in page 18.

[61] Nowell PC, Hungerford DA (1960) Chromosome studies on normal and leukemic human leukocytes. *Journal of the National Cancer Institute* 25:85–109. Cited in page 17.

[62] Ferrucci PF, et al. (1997) Cell death induction by the acute promyelocytic leukemia-specific PML/RARα fusion protein. *Proceedings of the National Academy of Sciences of the United States of America* 94(20):10901–10906. Cited in page 17.

[63] Edwards PA (2010) Fusion genes and chromosome translocations in the common epithelial cancers. *The Journal of Pathology* 220(2):244–254. Cited in page 18.

[64] Tomlins SA, et al. (2005) Recurrent Fusion of TMPRSS2 and ETS Transcription Factor Genes in Prostate Cancer. *Science* 310(5748):644–648. Cited in page 19.

[65] Apostolou P, Papasotiriou I (2017) Current perspectives on CHEK2 mutations in breast cancer. *Breast Cancer : Targets and Therapy* 9:331–335. Cited in page 19.

[66] Mertens F, Johansson B, Fioretos T, Mitelman F (2015) The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer* 15(6):371–381. Cited in page 19.

[67] Vellichirammal NN, et al. (2020) Pan-Cancer Analysis Reveals the Diverse Landscape of Novel Sense and Antisense Fusion Transcripts. *Molecular Therapy - Nucleic Acids* 19:1379–1398. Cited in pages 19, 20.

[68] Karkera JD, et al. (2017) Oncogenic Characterization and Pharmacologic Sensitivity of Activating Fibroblast Growth Factor Receptor (FGFR) Genetic Alterations to the Selective FGFR Inhibitor Erdafitinib. *Molecular Cancer Therapeutics* 16(8):1717–1726. Cited in page 19.

[69] Barrett JC, Kawasaki ES (2003) Microarrays: The use of oligonucleotides and cDNA for the analysis of gene expression. *Drug Discovery Today* 8(3):134–141. Cited in page 21.

[70] (2005) Affymetrix, GeneChip human tiling arrays. Cited in page 21.

[71] (2005) Affymetrix, GeneChip exon array. Cited in page 21.

[72] Xu W, et al. (2011) Human transcriptome array for high-throughput clinical studies. *Proceedings of the National Academy of Sciences* 108(9):3707–3712. Cited in page 21.

[73] Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1):57–63. Cited in page 22.

[74] Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J (2016) High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian Journal of Microbiology* 56(4):394–404. Cited in page 22.

[75] Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: The teenage years. *Nature Reviews Genetics* 20(11):631–656. Cited in page 24.

[76] FastQC : A quality control tool for high throughput sequence data Babraham-Bioinformatics (2012) (Babraham Bioinformatics). Cited in pages 24, 32.

[77] Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32(19):3047–3048. Cited in page 24.

[78] Chen S, et al. (2017) AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics* 18(3):80. Cited in page 24.

[79] Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 30(15):2114–2120. Cited in pages 24, 32.

[80] Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10–12. Cited in page 24.

[81] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10(3):R25. Cited in page 24.

[82] Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25(14):1754–1760. Cited in page 24.

[83] Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* 25(9):1105–11. Cited in pages 25, 32.

[84] Dobin A, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21. Cited in pages 25, 32.

[85] Baruzzo G, et al. (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods* 14(2):135–139. Cited in page 25.

[86] Zerbino DR (2010) Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* CHAPTER:Unit–11.5. Cited in page 25.

[87] Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29(7):644–652. Cited in page 25.

[88] Xu C (2018) A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal* 16:15–24. Cited in page 25.

[89] do Valle ÍF, et al. (2016) Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. *BMC Bioinformatics* 17(12):341. Cited in page 25.

[90] Fan Y, et al. (2016) MuSE: Accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology* 17(1):178. Cited in page 25.

[91] Koboldt DC, et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22(3):568–576. Cited in page 25.

[92] Larson DE, et al. (2012) SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 28(3):311–317. Cited in page 25.

[93] Krøigård AB, Thomassen M, Lænkholm AV, Kruse TA, Larsen MJ (2016) Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLOS ONE* 11(3):e0151664. Cited in page 25.

[94] Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y (2014) Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* 15(1):244. Cited in page 25.

[95] Shiraishi Y, et al. (2013) An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Research* 41(7):e89. Cited in page 25.

[96] Kim S, et al. (2013) Virmid: Accurate detection of somatic mutations with sample impurity inference. *Genome Biology* 14(8):R90. Cited in page 25.

[97] Kosugi S, et al. (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* 20(1):117. Cited in page 25.

[98] Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25(21):2865–2871. Cited in page 25.

[99] Zhang J, et al. (2016) INTEGRATE: Gene fusion discovery using whole genome and transcriptome data. *Genome Research* 26(1):108–118. Cited in page 25.

[100] McPherson A, et al. (2011) Comrad: Detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. *Bioinformatics (Oxford, England)* 27(11):1481–1488. Cited in page 25.

[101] Anders S, Pyl PT, Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166–169. Cited in page 26.

[102] Li B, Dewey CN (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(1):323. Cited in page 26.

[103] Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34(5):525–527. Cited in page 26.

[104] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* 14(4):417–419. Cited in page 26.

[105] Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology* 11(10):R106. Cited in pages 26, 32.

[106] Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–140. Cited in pages 26, 32.

[107] Ritchie ME, et al. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43(7):e47. Cited in page 26.

[108] Law CW, Chen Y, Shi W, Smyth GK (2014) Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* 15(2):R29. Cited in page 26.

[109] Seyednasrollah F, Laiho A, Elo LL (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics* 16(1):59–70. Cited in page 26.

[110] Jiménez-Jacinto V, Sanchez-Flores A, Vega-Alvarado L (2019) Integrative Differential Expression Analysis for Multiple EXperiments (IDEAMEX): A Web Server Tool for Integrated RNA-Seq Data Analysis. *Frontiers in Genetics* 10. Cited in page 26.

[111] Anders S, Reyes A, Huber W (2012) Detecting differential usage of exons from RNA-seq data. *Genome Research* 22(10):2008–17. Cited in pages 26, 32.

[112] Vitting-Seerup K, Sandelin A (2019) IsoformSwitchAnalyzeR: Analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics* 35(21):4469–4471. Cited in page 26.

[113] Gao Z, Zhao Z, Tang W (2018) DREAMSeq: An Improved Method for Analyzing Differentially Expressed Genes in RNA-seq Data. *Frontiers in Genetics* 9. Cited in page 26.

[114] Li YI, et al. (2018) Annotation-free quantification of RNA splicing using LeafCutter. *Nature genetics* 50(1):151–158. Cited in page 26.

[115] Haas BJ, et al. (2019) Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biology* 20(1):213. Cited in pages 27, 34, 35.

[116] Abate F, et al. (2014) Pegasus: A comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC systems biology* 8:97. Cited in pages 27, 35.

[117] Shugay M, Ortiz de Mendíbil I, Vizmanos JL, Novo FJ (2013) Oncofuse: A computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics (Oxford, England)* 29(20):2539–2546. Cited in pages 27, 35.

[118] Lovino M, Ciaburri MS, Urgese G, Di Cataldo S, Ficarra E (2020) DEEPrior: A deep learning tool for the prioritization of gene fusions. *Bioinformatics* p. btaa069. Cited in page 27.

[119] Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T (2016) Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nature Communications* 7(1):12817. Cited in pages 28, 33.

[120] Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: From microRNA sequences to function. *Nucleic Acids Research* 47(D1):D155–D162. Cited in page 28.

[121] Aparicio-Puerta E, et al. (2019) sRNAbench and sRNAtoolbox 2019: Intuitive fast small RNA profiling and differential expression. *Nucleic Acids Research* 47(W1):W530–W535. Cited in page 28.

[122] Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM (2011) miRanalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research* 39(Web Server issue):W132–W138. Cited in pages 28, 32.

[123] Friedländer MR, et al. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nature Biotechnology* 26(4):407–415. Cited in page 28.

[124] Agarwal V, Bell GW, Nam JW, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4:e05005. Cited in page 28.

[125] Enright AJ, et al. (2003) MicroRNA targets in Drosophila. *Genome Biology* 5(1):R1. Cited in page 28.

[126] Chen Y, Wang X (2020) miRDB: An online database for prediction of functional microRNA targets. *Nucleic Acids Research* 48(D1):D127–D131. Cited in page 28.

[127] Zhao B, Xue B (2019) Significant improvement of miRNA target prediction accuracy in large datasets using meta-strategy based on comprehensive voting and artificial neural networks. *BMC Genomics* 20(1):158. Cited in page 28.

[128] Zhao S, et al. (2017) QuickMIRSeq: A pipeline for quick and accurate quantification of both known miRNAs and isomiRs by jointly processing multiple samples from microRNA sequencing. *BMC Bioinformatics* 18(1):180. Cited in page 28.

[129] Shabalin AA (2012) Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28(10):1353–1358. Cited in page 28.

[130] Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O (2016) Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 32(10):1479–1485. Cited in page 28.

[131] Edmonson MN, et al. (2011) Bambino: A variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics* 27(6):865–866. Cited in page 32.

[132] Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28(5):511–5. Cited in page 32.

[133] Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 7(3):562–78. Cited in page 32.

[134] Kozomara A, Griffiths-Jones S (2014) miRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* 42(D1):D68–D73. Cited in page 32.

[135] Chen P, Lepikhova T, Hu Y, Monni O, Hautaniemi S (2011) Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants. *Nucleic Acids Research* 39(18):e123. Cited in page 33.

[136] Yates A, et al. (2016) Ensembl 2016. *Nucleic Acids Research* 44(D1):D710–D716. Cited in page 33.

[137] Jones P, et al. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240. Cited in page 33.

[138] Hornbeck PV, et al. (2015) PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research* 43(Database issue):D512–520. Cited in page 33.

[139] Loh PR, Palamara PF, Price AL (2016) Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics* 48(7):811–816. Cited in page 33.

[140] Fuchsberger C, Abecasis GR, Hinds DA (2015) Minimac2: Faster genotype imputation. *Bioinformatics* 31(5):782–784. Cited in page 33.

[141] Kircher M, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* 46(3):310–315. Cited in page 34.

[142] Aguet F, et al. (2017) Genetic effects on gene expression across human tissues. *Nature* 550(7675):204–213. Cited in page 34.

[143] Jia W, et al. (2013) SOAPfuse: An algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biology* 14(2):R12. Cited in page 34.

[144] Nicorici D, et al. (2014) FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv* p. 011650. Cited in page 34.

[145] Benelli M, et al. (2012) Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics (Oxford, England)* 28(24):3232–3239. Cited in page 34.

[146] Iyer MK, Chinnaiyan AM, Maher CA (2011) ChimeraScan: A tool for identifying chimeric transcription in sequencing data. *Bioinformatics (Oxford, England)* 27(20):2903–2904. Cited in page 34.

[147] Cunningham F, et al. (2019) Ensembl 2019. *Nucleic Acids Research* 47(D1):D745–D751. Cited in page 34.

[148] Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14(2):178–192. Cited in page 35.

[149] Ovaska K, et al. (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Medicine* 2(9):65. Cited in page 35.

[150] Cervera A, et al. (2019) Anduril 2: Upgraded large-scale data integration framework. *Bioinformatics* 35(19):3815–3817. Cited in page 35.

[151] Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 38(16):e164–e164. Cited in page 35.

[152] Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993. Cited in page 35.

[153] Stelzer G, et al. (2016) The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics* 54(1):1.30.1–1.30.33. Cited in page 35.

[154] Kanehisa M (2019) Toward understanding the origin and evolution of cellular organisms. *Protein Science: A Publication of the Protein Society* 28(11):1947–1951. Cited in page 35.

[155] Li H, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079. Cited in page 35.

[156] Danecek P, et al. (2011) The variant call format and VCFtools. *Bioinformatics* 27(15):2156–2158. Cited in page 35.

[157] Lee CM, et al. (2020) UCSC Genome Browser enters 20th year. *Nucleic Acids Research* 48(D1):D756–D761. Cited in page 35.

[158] McPherson A, et al. (2011) deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. *PLOS Computational Biology* 7(5):e1001138. Cited in page 38.

[159] Oliver GR, Jenkinson G, Klee EW (2020) Computational Detection of Known Pathogenic Gene Fusions in a Normal Tissue Database and Implications for Genetic Disease Research. *Frontiers in Genetics* 11. Cited in page 45.