TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Life Sciences

Integrative approaches of DNA, non-coding RNAs, and protein data levels reveal molecular functions and biological system responses in multiple cancers

Adriana Pitea

Vollständiger Abdruck der von TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Frank Johannes

Prüfer der Dissertation:

1. Univ. Prof. Dr. Dr. Fabian J. Theis
2. Univ. Prof. Dr. Dmitrij Frishman
3. Univ. Prof. Dr. John Gordan

Die Dissertation wurde am 11.05.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 08.11.2021 angenommen.

Acknoledgments

I would like to thank my supervisors Fabian Theis, Nikola Müller and Kristian Unger for giving me the opportunity to evolve as a scientist and for guiding me through this wholesome journey. Being a PhD Student in your groups was an unforgettable growing experience.

One of the most exciting and flourishing experiences during my PhD program was my lab visit at the Ideker Lab, University of California San Diego. My gratitude goes to Wei Zhang and Trey Ideker for believing in me as a strong applicant and further offering me the opportunity for a postdoctoral program at UCSD. I am also grateful to Hannah Carter, Michelle Dow, Brian Y. Tsui, and the Ideker Lab, for all the constructive discussions and the scientific debates during our lab meetings.

I would also like to express my gratitude to John Gordan and Manon Eckhard, for an amazing collaboration experice with the Krogan Lab at UCSF. Our work together kept me motivated through the last year of my PhD. I am also grateful to all my collaboration partners within HMGU and at the DKFZ.

Many thanks go to my colleagues for our work together, our Advent tea breaks, our creative paper, and so many other wonderful and not so wonderful shared events.

Lastly, I can never be grateful enough for all the friendships and connections that I've made during my PhD. Atefeh, Michi, Norbi, Ivan, Jan, Jörg, Flo, Hansi, Valle, Lisa, Carsten, Mo. Thank you for all the constructive feedback, scientific challenges, lunch breaks, table tennis competitions, fun scientific retreats, friendly conversations and wonderful memories together. Thank you for caring. Thank you for being there for me.

Agata, Cati, and Steffen, you know I wouldn't be here if it weren't for you. Forever grateful, no matter where we are in the world and where we are in life. Before having you in my life, I used to believe authentic connection is something I can only earn by always, but always, striving to be "good enough". Thanks to you, I've learnt that it is actually something we forge just by being willing enough together.

Abstract

Cancer accounts for one in six deaths and is the second leading cause of death in the world. A thorough understanding of the mechanisms that control carcinogenesis will provide valuable insights for targeted studies, with applications in cancer therapy.

To achieve this, scientists have focused on single-level omics studies for a long time. However, investigating single-level omics only explains a certain extent of the molecular process of carcinogenesis. To comprehend the molecular changes on a system level, the scientific community requires multilevel omics studies.

Thus, we set up to design and develop studies on models that integrate multilevel omics. Our models revealed patterns that enabled us to comprehensively infer molecular mechanisms ranging from fundamental functions to disease-underlying events.

Motivated by a previous study that identified a genomic copy number gain of chromosome 16q24.3 in head and neck squamous cell carcinoma (HNSCC) patients treated with radiochemotherapy only, we set up to validate the Fanconi anemia group A protein (FANCA) at multiple levels in an independent cohort. This starting point led to a benchmarking study of copy number calling algorithms in the presence of cancer-specific confounding variables. Our results indicated that tumor purity and copy number aberration burden strongly influenced the performance of all the analyzed algorithms. Overall, we discovered that CGHcall* - our adjusted version of CGHcall, and OncoSNP showed reasonable performance, particularly in samples with high purity.

Next, we expanded our integrative models to non-coding genes: microRNA and long non-coding RNAs (lncRNAs).

To assign functionality to microRNAs in HNSCC and lung cancer, we used a penalized elastic net model that inferred microRNA - protein-coding gene interactions using transcriptomic data as well as prior knowledge. Furthermore, our pipeline exploited the local structure of the inferred network providing functional annotation of the targets. We identified two functional clusters predicted to mediate HPV-associated dysregulation in HNSCC. Our findings in lung cancer confirmed the involvement of miR-509 in cell cycle, and p53 signaling. Finally, we inferred microRNAs that were involved in cell adhesion, cell migration,

and epithelial-to-mesenchymal transition, suggesting their involvement in lung tumor migration and metastasis.

At last, we constructed a "guilt-by-association"-based multilevel framework to interrogate the lncRNA functionality. Specifically, for each lncRNA, we predicted associated protein-coding genes. This enabled us to find a tissue-specific lncRNA cluster including LINC01123 - part of a plasma lncRNA signature that distinguished malignant intraductal mucinous neoplasms. Moreover, model-based gene set analysis confirmed lncRNA involvement in translation regulation and revealed association with cellular maintenance and immune system signaling pathways.

To summarize, we proved that multi-level omics integrative models are essential in finding interactions between mutations, copy number changes, protein interactions, coding, and non-coding gene expression across multiple cancer types. Not only did they confirmed known cancer-specific molecular changes but our models also revealed knowledge that aided us in proposing new oncogenic targets.

Abstract

Krebs ist die Ursache für einen von sechs Todesfällen jährlich und damit die zweithäufigste Todesursache weltweit. Ein umfassendes Verständnis der Mechanismen, die die Karzinogenese steuern ist Voraussetzung für die Entwicklung effektiver Krebstherapien.
Bisherige Studien befassen sich lediglich mit einzelnen „-omics"-Levels als Datengrundlage und erklären daher nur ein beschränktes Ausmaß des molekularen Prozesses der Karzinogenese. Um die molekularen Veränderungen auf Systemebene zu verstehen, benötigt die wissenschaftliche Gemeinschaft Studien, welche die multiplen Omics-Level gemeinsam, anstatt voneinander isoliert, betrachtet. Daher werden in dieser Arbeit Studien und Modelle entwickelt, welche mehrere Omics-Level integrieren und damit eine umfassendere Charakterisierung der Tumorentwicklung auf molekularer Ebene ermöglichen.
Motiviert durch eine frühere Studie von Copy Number Variations des Chromosoms 16q24.3 bei Kopf-Hals Krebs Patienten, haben wir uns vorgenommen, das Fanconi-Aämie-Gruppe-A-Protein (FANCA) auf mehreren Ebenen zu validieren in einer unabhängigen Kohorte. Dies führte zu einer Benchmarking-Studie von Algorithmen zum „copy number calling" unter Berücksichtigung Krebsspezifischer Störgrößen. Unsere Ergebnisse zeigten, dass die Tumorreinheit und die Belastung durch veränderte Kopienzahl die Leistung aller analysierten Algorithmen stark beeinflussten. Insgesamt stellten wir fest, dass CGHcall* - unsere angepasste Version von CGHcall, und OncoSNP - eine angemessene Leistung zeigten, insbesondere bei Proben mit hoher Reinheit.
Des Weiteren analysierten wir zunächst nicht-kodierende Gene: microRNA und lange nicht-kodierende RNAs (lncRNAs). Um die Funktion von microRNAs im Kopf-Hals- und Lungenkrebs zu finden, wurde ein Regressionsmodell mit Elastic-Net Penalisierung entwickelt, dass mögliche microRNA-targets identifiziert. Hierzu wurde eine Kombination von microRNA- und Genexpressionsdaten, Vorkenntnisse und zudem, nutze unsere Methode die lokale Struktur des regulatorischen Netzwerks zur Annotation der Targets. Wir haben zwei funktionelle microRNA Cluster identifiziert, welche die Veränderungen in Kopf-Hals-Krebs Patienten mit Papillomvirus erklären. Unsere Ergebnisse bestätigten die Rolle von miR-509 in Zellzyklus und p53 Signaling im Lungenkrebs. Weiterhin

wurden microRNAs die an der Zelladhäsion, Zellmigration und Epithelialen-Mesenchymale Transition beteiligt waren, identifiziert. Diese Ergebnisse lassen den Schluss zu, dass die identifizierten microRNAs die Migration und Metastasierung von Lungentumoren regulieren.

Schließlich, haben wir ein auf „Schuld durch Assoziation“ Modell basierend auf multiplen molekularen Ebenen konstruiert, um die Funktion von lncRNA in der Zelle aufzuklären. Für jede untersuchte lncRNA wurden zugehörige Gene vorhergesagt, was uns ermöglicht gewebsspezifische lncRNA Cluster zu finden. Modell-basierte Gene-set Analyse bestätigte die Beteiligung von lncRNAs in der Regulierung der Translation und zeigt deren Assoziation mit Immunsystem Pathways und Prozessen zur „cellular maintainance“ auf.

Insgesamt konnten wir zeigen, dass integrative multi-level omics Modelle essenziell sind um die Interaktion von Mutationen, copy number changes, Protein-Interaktion und Geneexpression in verschieden Tumoren zu identifizieren und analysieren. Zum einen konnten schon bekannte Krebs-spezifische molekulare Veränderungen verifiziert werden, aber auch neue Erkenntnisse gewonnen werden, die uns erlauben neue Targets für die Behandlung von Krebs vorzuschlagen.

Scientific Publications and Patents

The results of this thesis are partly based on the previously published papers and papers, which are currently within the publication process. These together with further publications and patents are listed below:

- S. Sass*, **A. Pitea***, K. Unger, J. Hess, N. S. Mueller, and F. J. Theis. MicroRNA-target network inference and local network enrichment analysis identify two microRNA clusters with distinct functions in head and neck squamous cell carcinoma. *Int J Mol Sci*, 16(12): 30204–30222, Dec 2015.

- **A. Pitea**, I. Kondofersky, S. Sass, F. J. Theis, N. S. Mueller, and K. Unger. Copy number aberrations from Affymetrix SNP 6.0 genotyping data - how accurate are the commonly used prediction approaches? *Briefings in Bioinformatics*, 2018.

- J. Gordan*, **A. Pitea***, , M. Eckhardt, G. Jang, R. E. Turnham, A. Choi, J. Von Dollen, H. C. Lim, G. Chan, R. K. Kelley, D. Swaney, W. Zhang, F. J Theis, T. Ideker, N. J. Krogan. HBV alters HCC signaling and proliferation via physical remodeling of PP2A, to be sumitted.

- **A. Pitea***, L. Krause*, G. Eraslan, S. Sass, J. Arloth, M. Preusse, and N. S. Mueller. Holistic analysis of long non-coding RNAs identifes role in human metabolism. *Nucleic Acids Res*, to be submitted.

- M. González-Vallinas, M. Rodríguez-Paredes, M. Albrecht, C. Sticht, D. Stichel, J. Gutekunst, **A. Pitea**, S. Sass, F. Sánchez-Rivera, J. Lorenzo-Bermejo, J. Schmitt, C. De La Torre , A. Warth, F. J. Theis, N. S. Mueller, N. Gretz, T. Muley, M. Meister, D. F. Tschaharganeh, P. Schirmacher, F. Matthäus, and K. Breuhahn. Epigenetically Regulated Chromosome 14q32 miRNA Cluster Induces Metastasis and Predicts Poor Prognosis in Lung Adenocarcinoma Patients. *Mol Cancer Res.*, 16(3):390-402, Mar 2018.

- M. Niyazi, **A. Pitea**, M. Mittelbronn, J. Steinbach, C. Sticht, F. Zehentmayr, D. Piehlmaier, H. Zitzelsberger, U. Ganswindt, C. Rödel, K.

Lauber, C. Belka, and K. Unger. A 4-miRNA signature predicts the therapeutic outcome of glioblastoma. *Oncotarget.*, 7(29): 45764–45775, Jul 2016.

- C. M. Wilke, J. Hess, S. V. Klymenko, V. V. Chumak, L. M. Zakhartseva, E. V. Bakhanova, A. Feuchtinger, A. K. Walch, M. Selmansberger, H. Braselmann, L. Schneider, **A. Pitea**, J. Steinhilber, F. Fend, H. C. Bösmüller, H. Zitzelsberger, and K. Unger. Expression of miRNA-26b-5p and its target TRPS1 is associated with radiation exposure in post-Chernobyl breast cancer. *Int J Cancer.*, 142(3):573-583, Feb. 2018.

- L. Wintergerst, M. Selmansberger, C. Maihoefer, L. Schüttrumpf, A. Walch, C. Wilke, **A. Pitea**, C. Woischke, P. Baumeister, T. Kirchner, C. Belka, U. Ganswindt, H. Zitzelsberger, K. Unger, and J. Hess. A prognostic mRNA expression signature of four 16q24.3 genes in radio(chemo)therapy-treated head and neck squamous cell carcinoma (HNSCC). *Mol Oncol.*, Sep. 2018.

- C. Belka, K-M. Niyazi, K. Unger, H. Zitzelsberger, **A. Pitea**, M. Mittelbronn. Differential diagnosis in glioblastoma multiforme. *US PatentApp. 16/063,175.*

$*$ = equal contributions

Contents

To my beloved grandparents.

Chapter 1

Introduction

With the second-highest worldwide mortality rate and an incidence rate of 33% between 2005 and 2015 [1], cancer remains one of the main unsolved health problems of the present.

Cancer is a complex disease in which cells in a specific tissue respond faultily to cell cycle control signals. As a result, these cells present uncontrolled growth, impairment, and inflammation within the tissue, and, in advanced stages, they invade other tissues [2]. Cancer develops due to an excessive stepwise accumulation of changes reflected on multiple omics levels caused by hereditary or environmental factors – mutagens, chemicals that damage DNA, hormonal factors, and deficient diets, genetic predisposition, aging, or viral infections [3]. The complexity of this stepwise process results in over 200 different types of cancer [4].

Besides being a highly heterogeneous and complex disease, cancer evolves. Specifically, as cancer grows, cells accumulate mutations and chromosomal aberrations promoting proliferation and immune escape [5]. After this process, cancer cells adapt and emerge [6]. This makes it difficult to find efficient treatments to cure cancer: if a small population of cancer cells escapes the treatment, the remaining cancerous cells adapt to new conditions, develop resistance, and expand [6].

Initially, cancer studies focused on single omics levels: Bignell et al. analyzed DNA copy number changes and identified 2,428 somatic homozygous deletions in 746 cancer cell lines [7], Ma et al. identified gene expression alterations en-

abling invasive growth present already in the preinvasive stages of breast cancer [8], while Calin et al. identified a microRNA signature associating with prognostic markers and progression in chronic lymphocytic leukemia [9]. These studies showed that cancer manifested itself on different omics levels. Hence, in order to get a comprehensive picture of the underlying mechanisms, multilevel omics studies are required.

1.1 Cancer impacts different molecular levels

Studies on multilevel omics of tumor genomes use whole-genome sequencing technologies and various profiling techniques of DNA copy number changes, epigenome, transcriptome, proteome and microbiome [10]. These studies aim to identify the changes that occur during tumorigenesis in different omics levels. Ultimately, cancer research aspires at identifying genes and pathways that can be used in molecular-guided diagnosis and management of cancer.
To exploit the multilevel omics data, one first needs to understand how different omics function and how disrupting their function can potentially induce cancer. The following sections introduce the different omics levels together with examples of cancer impact on each level.

Changes at the DNA level

Direct changes at the genome level can be caused by errors occurring during DNA replication (e.g. double strand breaks, mismatch pairing, replication slippage, tautomeric shifts) or by environmental factors known as mutagens. Examples of mutagens include chemical mutagens and radiation, both of which result in DNA damage [3].

DNA repair mechanisms are among the most important and remarkable molecular processes. During DNA repair, cells engage numerous genome editing mechanisms to correct DNA sequence errors [11]. Dysregulation of the DNA repair mechanisms can result in DNA changes escaping the editing mechanisms. Next, the new generation of cells inherits these DNA changes. The repeated process

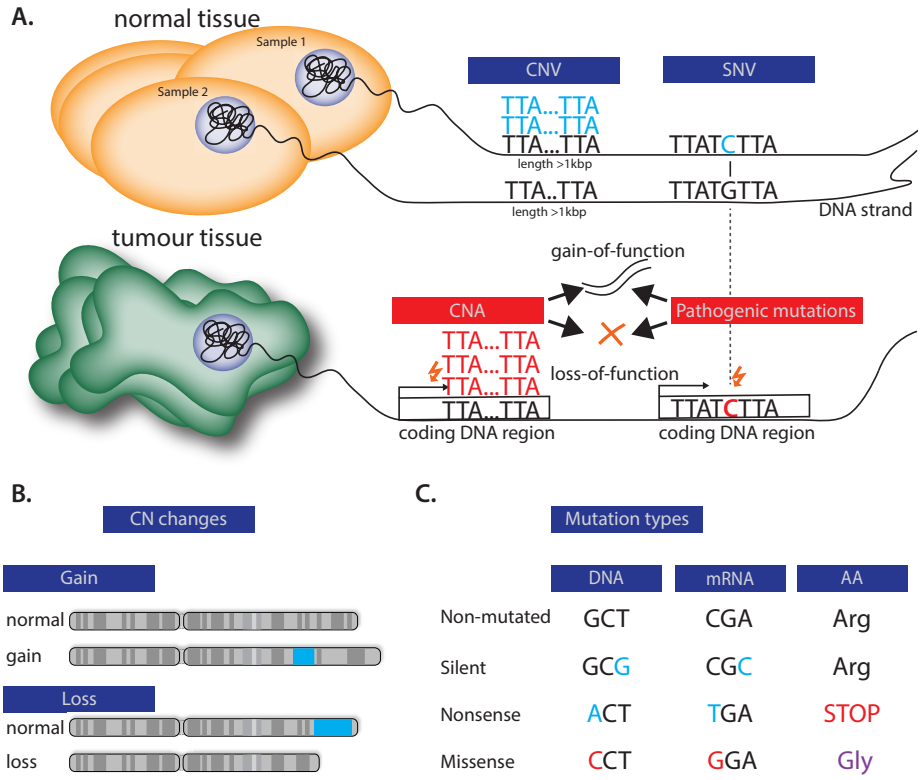

Figure 1.1: **Changes at the DNA level.** A. Genomic variations: CNVs and SNVs are shown in blue, while CNAs and pathogenic mutations are shown in red. CNVs and CNAs, here represented as stacked copies of the same region, are concatenated in the DNA strand. B. Copy number changes: gains and losses. In the upper panel of the figure the blue highlighted region represents the gained region when compared to the normal chromosome. The lower panel of the figure highlights the region in the normal chromosome that is then lost. C. Starting from the DNA sequence GCT, the figure exemplifies what kind of mutations are possible and what are the changes at mRNA level and further at aminoacid(AA) level.

leads to cells accumulating the wrong type or the wrong number of nucleotides introduced by DNA replication or mutagens. Depending on the type of introduced errors, the DNA changes arise at different scales. Wide-scale variations represent genomic regions of length > 1 kilobase pair (kbp), that show a different number of copies (gains or losses – Figure 1.1 B.) in comparison to a reference genome, while small-scale variations represent mutations that can cover from one to several base pairs (bps) in length.

Large-scale DNA changes

Several studies showed that large-scale copy number changes commonly take place at multiple locations in the human genome. For example, Iafrate et al. identified 255 loci across the human genome with genomic imbalances [12], while Conrad et al. identified an average of 1,098 validated altered genomic regions [13]. Throughout this thesis we will refer to the large-scale copy number changes that occur naturally and are not associated with human disease as copy number variation (CNVs) (Figure 1.1 A. upper row). Since their discovery, CNVs have been thoroughly studied in humans and have been shown to influence gene expression through gene regulatory molecular mechanisms – such as gene dosage [14, 15] or gene disruption [16]. This way, CNVs can cause random or Mendelian genetic traits leading to genetic diversity [16, 17].

Large-scale copy number changes that occur somatically, emerge after many selection events, and are specific to human disease development and progression, particularly to cancer, will be referred to as copy number aberrations (CNAs) (Figure 1.1 A. second row).

Due to their negative impact on human health, CNAs have acquired a growing interest in research and over time have been shown to play major roles in diverse human diseases: Pérez Jurado et al. showed that Williams syndrome (a neurodevelopmental disease) was associated with the partial deletion of the chromosomal band 7q11.23 [18]. The 20 kilobase deletion upstream the immunity-related GTPase family M protein (IRGM) was associated with both altered expression of the protein and Chron's disease [19]. Walters et al. discovered a frequent heterozygous deletion on chromosome 16p11.2 in patients with congenital malformations and/or developmental delay in addition to obesity [20]. The 15q11-q13 duplication acquired its own disease name – 15q Duplication Syndrome, and has also been associated with other disorders: Prader-Willi syndrome – patients present specific facial features, infantile hypotonia (abnormal limpness), excessive eating, hypogonadism (diminished functional activity of the testes or ovaries), mild intellectual disability, and obsessive-compulsive behavior [21, 22], Angelman syndrome – characterized by distinctive facial features, severe intellectual disability, severe language impairment, seizures, ataxia (lack of voluntary coordination of muscle movements), and an unusually happy

or excitable disposition [23, 24], autism spectrum disorder – children present impairment in motor, social and communication skills [25].

CNAs have been associated with a high diversity of cancers [26, 27] and are known to be present in all cancer genomes [28]: Bardeesy et al. showed that the deletion of the tumor suppressor gene SMAD4 plays a critical role in the progression and tumor biology of pancreatic cancer [29], while Witkiewicz et al. showed that amplification of the proto-oncogene MYC is uniquely associated with poor outcome in pancreatic ductal adenocarcinoma [30]. Leucci et al. showed that the survival-associated mitochondrial long non-coding RNA (lncRNA) – SAMMSON, is consistently co-gained with the melanocyte inducing transcription factor (MITF) in more than 90% of human melanomas [31].

In head and neck cancer, smoking-related carcinomas presented inactivation of the Cyclin-Dependent Kinase Inhibitor 2A – CDKN2A, with frequent amplifications of regions 3q26-q28 and 11q13-q22 [32]. A previous study on head and neck cancer patients treated with radiotherapy alone identified a genomic copy number gain of chromosome 16q24.3. This region overlaps the DNA repair gene Fanconi anemia complementation group A (FANCA). The 16q24.3 gain was associated with unfavourable outcome in radiation-treated patients [33]. The results of Bauer et al. motivated us to validate the FANCA gain in an independent cohort (analysis included in Hess et al. [34]).

The accumulation of DNA copy number changes affects the transcriptional process and can lead to activation [35], repression [36] or complete inactivation of a protein-coding gene [37]. When the affected protein-coding genes are oncogenes or tumor suppressor genes, CNAs become partly responsible for tumorigenesis [38, 39].

Furthermore, when DNA changes affect genes involved in the DNA repair mechanisms, the errors in the DNA sequence accumulate at higher rates, eventually leading to cancer [40].

Finding CNAs specific to oncogenes and tumor suppressor genes can pinpoint their impact on oncogenic pathways and ultimately can aid the development of personalized cancer therapies. It is thus particularly important to provide accurate estimates of DNA changes in tumor genomes (Chapter 4).

Small-scale DNA changes

Another class of genetic variants consists of small-scale sequence mutations – changes in the DNA sequence that affect only one or a small number of nucleotides. Small-scale mutations can be classified based on several aspects. Based on the type of sequence change, small-scale mutations are classified as deletions, substitutions and insertions. Substitutions of a single nucleotide are also known as single nucleotide variations (SNVs).

Furthermore, depending on their effect on the functionality of a gene, mutations are classified as silent mutations – they do not affect the function of a protein, missense mutations – they cause a change that results in different amino acids, and thus can affect the protein biosynthesis, and nonsense mutations – they do not code for any amino acid and thus no protein is produced (Figure 1.1 C.).

Another criterion for classifying mutations is considering the genomic region type they affect: splicing site, flanking site ($5'-$ and $3'-$ untranslated regions (UTRs), $5'-$ and $3'-$ flanks), start and stop codons, translational termination codons ('nonstop' mutations), in-frame mutations (changes that affect an integer number of codons – the genetic code can still be read in sequence), frameshift mutations (changes that affect only part of a codon and end the reading beyond the mutation) and noncoding mutations (mutations that are overlapping noncoding elements such as promoters, enhancers, microRNAs, lncRNAs).

Mutations contribute to human disease by inactivating or activating protein function: D'Souza et al. showed that missense, silent, and intronic tau mutations can increase or decrease splicing of tau exon 10 (E10) by acting on three different cis-acting regulatory elements [41], while Fedele et al. analyzed missense mutations in GRIN2B - the gene encoding the N-methyl-D-aspartate (NMDA) receptor GluN2B subunit, and revealed activation of the mutated NMDAs receptors known to control synaptic plasticity and memory function [42]. Mutations related to human disease will be further referred to as pathogenic mutations.

Environmental factors like viral infections together with gene regulation are frequently causing pathogenic mutations and inducing multifactorial diseases [43]. To be specific, viruses take control over the cellular machinery and promote forced cell division through pathogenic mutations that change the host protein

expression [43]. Consequently, viruses like Hepatitis C can initiate and promote hepatocellular carcinomas [44]. Therefore, knowing how human and viral proteins interact can aid in understanding how cancer progresses and how to disrupt it. With this in mind, we designed a study where we inquired whether and how viral infections impact protein-protein interactions through mutations in liver cancer (Chapter 5).

Epigenomics

The epigenome consists of all the genome-wide modifications that determine which genes are activated, for which cell type, and when [45]. Two mechanisms induce these modifications: DNA methylation takes and histone modifications. While DNA methylation enables specific proteins to attach methyl groups to the DNA strand in specific locations, histone modifications decide if a specific DNA region will be used in a specific cell type or not [45]..

Factors such as an unhealthy diet, lack of physical activity, smoking, stress, alcohol consumption and environmental pollutants, lead to changes in the epigenome. Changes in the epigenome can control the expression of genes that are involved in the immune response of a cell, apoptosis, or cell proliferation. Such changes can determine cells to become resistant to apoptosis and immune response. Some of these epigenetic modifications were shown to be involved in diseases as Alzheimer's disease [46] and autoimmune diseases [47]. Recently, studies have shown that epigenetic changes are prevalent in cancer: Fleischer et al. determined that DNA methylation in enhancers regions is distinct between breast cancer lineages [48], while Ju et al. found that one of the most common protein covalent modification in eukaryotes – NatD, promotes lung cancer migration and invasion by preventing the phosphorylation of histone H4 serine - known to be involved in cell proliferation [49]. Thus, mechanisms at the epigenome level play an important role in tissue-specific gene expression.

Transcriptomics

During transcription, the information contained in single DNA strands sequences from the genome of a cell is copied to RNA molecules. Approximately 30,000 of these RNA molecules encode the information required for protein synthesis

[50]. These RNA molecules are known as messenger RNAs (mRNAs) and represent the most extensively studied transcriptomics data level. Since mRNAs are single-stranded copies of a gene and represent the templates that form proteins, they influence protein synthesis. The overexpression of genes that host driver mutations and the repression of genes that repress cell growth are examples of transcriptomics alterations that can initiate cancer [51].

The remaining transcribed DNA sequences do not code for proteins, and are known as non-coding RNAs. The non-coding RNAs cover 95% of the transcriptional output [52]. Despite the fact that the human genome can now be sequenced with a reasonable degree of accuracy, we still cannot fully understand the mechanisms of non-coding RNAs. Based on their length, non-coding RNAs can be significantly short – 20-24 nucleotides (nts), for example microRNAs (miRNAs), or they can span over 200 nts – long non-coding RNAs (lncRNAs). MiRNAs target specific mRNAs and play an important role in the control of gene expression after transcription [53, 54].

Unlike miRNA, lncRNAs remain elusive due to their low conservation, low expression levels and their tissue specificity [55]. Until now, lncRNAs have been associated with gene expression regulation both during transcription and post-transcription [56, 57].Additionally, lncRNAs have been shown to regulate processes ranging from coordinating ribosomal RNAs (rRNAs) transcription and methylation, to mediation of epigenetic changes [57]. Recently, lncRNAs, just like miRNAs, have been associated with the regulation of oncogenic pathways across many cancer types [58, 59, 60]. Nonetheless, the functional mechanisms of lncRNAs and how miRNAs choose their mRNA targets remain yet poorly understood [53]. For this reason, this thesis includes a chapter that analyses miRNA-target networks in head and neck cancer and liver cancer (Chapter 6), and an overall functional characterization of lncRNAs (Chapter 7).

Proteomics and metabolomics

Following transcription, mRNAs are translated into proteins via the joint transfer of rRNA and transfer RNA (tRNA) [61]. Since they perform most of the biological activities, proteins are considered vital elements of a living organism. The correct synthesis and functioning of proteins are essential for the normal

development and maintenance of healthy cells, tissues, organs, and organisms. Dysregulation of protein co-regulation affects cellular fitness, and thus cellular signaling and homeostasis [62]. Ultimately, alterations in protein-protein interactions affect the phenotype of biological systems and can give rise to disease [62, 63]. For example, Eckhardt et al. showed that the human papillomavirus-host protein network promotes multiple routes to oncogenesis in head and neck and cervical cancers [64]. Within this thesis, we performed a similar analysis where we aim to understand the interactions between hepatitis B and C viral infections and human proteins in liver cancer (Chapter 5).

Protein activity regulates the metabolome of an organism. However, the regulation is mutual – protein activity is influenced by metabolites. Since the metabolome contains all the biochemical reactions that occur in living organisms, it provides an overview of the fundamental molecular interactions. Deviant cell metabolism is common across many cancer types [65]. Cancer-associated changes across the previously described omics levels can change the levels of and can initiate and promote tumorigenesis [66]. For example, Chiarugi et al. discussed the role of a key determinant of cancer biology – the NAD metabolome, known to be involved in both energy and signal transduction, and thus a putative target for new cancer therapeutic concepts [67].

Microbiomics

One distinct layer of omics that comprises all the genomes of the microorganisms living in an environmental niche within the human body is the microbiome. The human microbiome together with lifestyle and environmental factors shapes the human body phenotype [68]. Following, specific changes of the microbiome have been associated with diseases: Zhang et al. demonstrated that the expression of human proteins related to oxidative antimicrobial activities increased in pediatric inflammatory bowel disease cases and correlated with the alteration of microbial functions [69], while Jangi et. al uncovered associations between changes in the human gut microbiome and multiple sclerosis [70].

Since the microbiome affects a multitude of host functions as metabolic, immune and cellular response functions [71], recent studies focused on the role of microbiome in cancer. Matson et al. showed association between commensal mi-

crobiome and response to immunotherapy in metastatic melanoma patients [72], while Gopalakrishnan et al. showed that gut microbiome modulates response to the same immunotherapy in melanoma patients [73]. Another study discovered that the microbiome can affect anticancer immunosurveillance in multiple way [74]. Therefore, the microbiome can be targeted and used as an adjuvant treatment to improve the anticancer immune responses, and thus, the efficiency of standard cancer treatments.

Changes at every omics level contribute to multiple pathways and can reveal significant biomarkers. Nonetheless, findings at single omics levels are not sufficient to explain the underlying cellular mechanisms. Thus, to fully understand the biology of cancer, one needs to comprehend how cancer impacts multilevel omics.

1.2 Technology and multilevel omics

The technological progress that emerged after the human genome sequencing, enabled the profiling of multilevel omics. In addition to providing a variety of multilevel omics data, the advances in high-throughput measurement techniques together with the decrease in costs, enabled the development of large-scale sequencing projects like The Cancer Genome Atlas (TCGA), the Roadmap Epigenomics project and the Genotype-Tissue Expression (GTEx). The Roadmap Epigenomics project obtained comprehensive data for 111 consolidated epigenomes so far [75], while the GTEx project collected tissue and blood biospecimens from over 900 deceased donors [76]. While the Roadmap Epigenomics project focuses on providing a public collection of normal epigenomes to define changes in DNA activity and to predict function independent of any change of DNA sequence, GTEx is a resource that studies the relationship between genetic variation and gene expression across multiple human tissues.
TCGA began in 2005 intending to characterize a multitude of different cancer types at multiple levels of omics and, based on this, to reveal cancer-causing genome alterations in large cohorts of human tumors [77]. TCGA covers over 11,000 patient-derived samples across 33 different cancer types. Starting with

the availability of the resources provided by TCGA, numerous studies have dedicated considerable effort to use multiple omics profiles for predicting survival across a multitude of cancers: Verhaak et al. used gene expression to classify glioblastoma tumors into Proneural, Neural, Classical, and Mesenchymal subtypes and showed that response to aggressive therapy differs by subtype [78]. Noushmehr et al. analyzed DNA methylation in glioblastoma tumors and found of a CpG island methylator phenotype defining a distinct subgroup of glioma [79]. The following studies used genomics, DNA methylation, exome data, transcriptomics and proteomics to comprehensively characterize molecular landscapes of human breast tumors, colon and rectal cancer, and head and neck tumours [80, 81, 82].

1.3 Integrative approaches

Nonetheless, the amount of data produced by projects as TCGA is too complex to analyze manually. Thus, there is a need for methods that can explain the molecular mechanisms underlying the information contained in the data.
Formulating models that integrate multilevel omics is nevertheless a complex task. This is largely due to a lack of comprehensive understanding of the interplay between the different omics levels. Additionally, due to complexity and evolution, tumorigenesis and the underlying causes are still only poorly characterized.
As a result, common approaches focus on unifying results on single omics levels: finding differentially expressed genes that overlap regions with frequent somatic mutations and copy number changes, or finding DNA regions with both altered methylation and copy number changes, or finding genes with altered expression and strong association to microRNAs that also show altered expression. For example, Koboldt et al. first analyzed DNA copy number arrays and identified somatic mutations in only three genes (the tumour repressor TP53, the oncogene PIK3CA and the GATA3 transcription factor) that occurred at $> 10\%$ incidence across all breast cancer; next they defined two novel subgroups based on protein expression and performed pathway enrichment for each subtype [80].

Other studies used integrative approaches with underlying supervised or unsupervised models: Kirk et al. designed a Bayesian method for the unsupervised integrative modeling of gene expression, chromatin immunoprecipitationchip, and protein-protein interactions data, to identify a set of protein complexes for which genes are co-regulated during the cell cycle [83]. Cho et al. developed a meta-model that summarizes the results of a large number of alternative models that use a given measure of phenotypic similarity between patients and a list of potential explanatory features, such as mutations, CNVs, microRNA levels, to return phenotypic similarities [84]. Hofree et al. introduced the network-based stratification (NBS) method that integrated somatic tumor genomes with gene networks to cluster patients with mutations in similar network regions [85].
Despite the diverse methods currently used for integrating multilevel of omics, the underlying models can only characterize specific aspects of tumorigenesis. Designing approaches that use patterns derived from multilevel omics allows us to comprehensively infer molecular mechanisms ranging from fundamental biological functions to disease-underlying events. Understanding cellular processes in an exhaustive manner can provide valuable knowledge for characterizing cancer, but also for other complex diseases that are still a global compelling public health burden.
In this thesis, I focused on uncovering and understanding distinct underlying processes of head and neck, liver, lung, and ovarian cancer by using approaches tailored to specific molecular aspects addressed in the next section. The results presented in this thesis contribute to understanding what models work and why do they work when integrating different levels of omics from cancer data for answering the following research questions.

1.4 Research questions

The main goal of this thesis is to better understand distinct mechanisms underlying tumorigenesis in head and neck, lung and liver cancer by exploiting the abundance of data made available by technological advances. For this purpose, I performed integrative analyses of multilevel omics and phenotypic data to reliably identify underlying molecular mechanisms that can improve our un-

derstanding of cancer biology.

Following, within this thesis I addressed the next main research questions:

A. How accurate are the commonly used CNA calling algorithms from SNP 6.0 array genotyping?

The accurate identification of CNAs from cancer tumor facilitates finding oncogenes or tumor suppressor genes. Hence, given the importance of reliably finding CNAs in cancer research, it is essential to assess the accuracy of CNA calling algorithms and which are the factors that affect it. To address this question, I performed a comparative study of CNA calling algorithms from single nucleotide polymorphism (SNP) array data in the presence of three cancer-specific confounding variables – tumor purity, CNA region length, and the amount of CNAs present in a tumor genome, both on synthetic data and real data comprising of TCGA head and neck cancer samples and HapMap samples.

B. How can the impact of viral infections on the protein-protein interactions network in cancer be confidently estimated?

OViruses are one of the main environmental factors that are known to cause cancer. Viral infections disrupt the cell replication mechanisms, affecting the normal cellular proteins such as cell cycle regulators of DNA repair proteins. The dysregulation of DNA repair proteins is followed by an increased rate of DNA changes. This mechanism coupled with the cell proliferation stimulated by viral replication can cause cancer [86, 32]. Consequently, it is essential to determine the impact of viral infections on the host protein-protein interactions in cancer. To address this question, I performed an integrative analysis that estimates the viral effect based on the mutational landscape of infected tumors – represented here by a TCGA liver cancer data set, and on the strength of physical interactions between viral and host proteins – in-house data.

C. How do non-coding RNAs contribute to disease, in particular to cancer?

Non-coding RNAs have been associated with many cancer types, independent of their length. However, the question of how do non-coding RNAs contribute to cancer remains to be answered. For this reason, part of this thesis focuses on understanding how do non-coding RNAs function and what conditions regulate their function. Specifically, I examine the functional roles of miRNAs (**Chapter**

6) and lncRNAs (**Chapter 7**). For this purpose, I assessed which are the genes and pathways affected by miRNAs in cancer. To address this question, I used miRlastic, introduced in Sass*, Pitea* et al., a method that identifies miRNA – mRNA interactions and functionally annotates target gene sets of miRNAs. With miRLastic, I was able to predict miRNA – mRNA regulatory networks in head and neck cancer and lung cancer TGGA data. Next, I analyzed the relationship between lncRNAs and mRNAs and how does tissue specificity affects this relationship when using data from GTEx, Roadmap, and TCGA. In the end, we provided a functional pipeline for a comprehensive and fully integrative study for inferring lncRNA functions while exploiting the wealth of newly available studies of larger sample sizes.

1.5 Overview

In this thesis, I addressed the research questions presented in the previous section. Figure 1.2 provides a brief overview of this thesis.
The first part of this thesis contains the introduction (**Chapter 1**), the biological and methodological background (**Chapter 2**) together with the materials used in the subsequent analyses (**Chapter 3**). Specifically, Chapter 2 introduces the technologies and the methods necessary for understanding the biological and technical background, which is relevant throughout the thesis. This includes a description of genomics and transcriptomics profiling and omics analysis together with an overview of the statistical theory used for model identification and inference techniques. Chapter 3 introduces the reader to the data sets that are used to investigate the molecular mechanisms of liver, lung, and head and neck cancer. This chapter includes results of statistical analyses for revealing differential expression of miRNAs and mRNAs induced by viral infections in liver and head and neck cancer, as well as differentially expressed miRNAs between different stages of lung cancer. Additionally, the chapter introduces the data measuring the strength of physical interactions between viral and human proteins in two liver cancer cell lines. Finally, this chapter introduces the Haplotype Map data set that was used for benchmarking CNA calling algorithms.

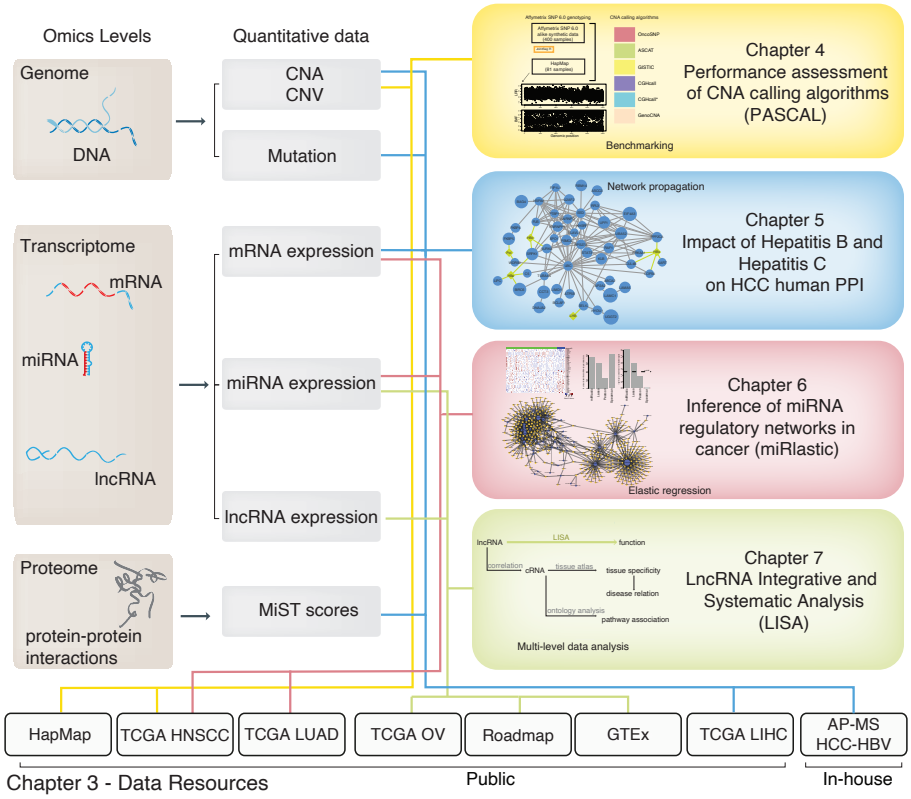

Figure 1.2: Multilevel omics from various data sources are combined in four main research studies: a benchmarking study on copy number calling algorithms (Chapter 4), a multi-level integrative analysis with the aim to identify the viral effect on human protein in liver cancer (Chapter 5), the application of miRlastic – an integrative framework for inference of miRNA-mRNA regulatory networks and functional characterization of specific miRNAs, on two TCGA cancer data sets: head and neck cancer and in lung cancer (Chapter 6) and an integrative framework at multiple molecular levels which aims to identify lncRNA functions (Chapter 7).

The next four chapters present the projects upon which this thesis was built.

Chapter 4 presents a benchmarking study aiming at finding the most suitable method for predicting DNA changes specific to cancer, given specific cancer confounding variables. Explicitly, Chapter 4 introduces and benchmarks several commonly used CNA calling algorithms on synthetic data given distinct tumor purities, CNA burdens, and altered DNA region lengths. Based on the results, I proposed and assessed an adjusted version of one of the algorithms. Next, I evaluated the performance of the algorithms on real data. Finally, I assessed the algorithms on a publicly available TCGA head and neck cancer data

set and I performed an explorative analysis of the results based on consensus results.

Chapter 5 is based on a collaboration project with the Ideker Lab (University of California San Diego) and the Krogan Lab (University of California San Francisco) that aimed to characterize viral-host protein-protein interactions in liver cancer patients infected with Hepatitis B. Subsequently, I present an integrative framework for confidently identifying the impact of viral infections on human protein interactions in cancer. The framework consists of a network propagation-based approach that integrates genomic and physical interaction measurements between viral and human proteins. The approach assessed the significance of interactions between human and viral proteins in liver cancer.

Chapter 6 aims to characterize miRNA – mRNA regulatory networks specific for specific conditions in head and neck cancer and lung cancer. In particular, Chapter 6 describes the application and the results of an inference tool that integrates protein coding expression and microRNA expression – miRlastic, on head and neck cancer and lung cancer data.

Chapter 7 introduces an integrative analysis of large scale multilevel data that aims to infer functions of long non-coding RNAs by investigating various omics levels in both normal and diseased tissues: LISA. LISA systematically explores lncRNA molecular mechanisms by exploiting genomics, transcriptomics and epigenomics together with functional and tissue annotations from four large sample size projects: Encyclopedia of DNA Elements - ENCODE [87], Roadmap Epigenomics Project [75], TCGA and GTEx [88].

The thesis concludes with the future perspective and the scientific contributions in the context of cancer genomics in Chapter 8.

Chapter 2

Background

This chapter introduces the theoretical concepts and the experimental techniques used in the projects presented in this thesis. These projects primarily derive from using statistics and machine learning concepts to formulate models that provide a plausible fit for experimental cancer data. Thus, we initially present the experimental techniques exploited for generating profiles from which scientists can estimate DNA copy number changes, transcriptomics levels, and protein-protein interaction strength. Next, several statistical concepts, like correlation, logistic, Bayesian, and elastic net regression, are briefly introduced for understanding the assumptions behind the underlying inference models. Lastly, we present the network propagation concept so the reader can grasp an idea of how we can combine and amplify signals from individual genes within a biological network.

2.1 Technologies that generate biological data

Several experimental techniques exist for generating data at multiple omics levels like DNA, RNA, proteins, and metabolites. The next section introduces the protocols used for generating profiles from which we can infer DNA copy number changes, transcriptomics expression, and the strength of protein-protein interactions.

Copy number profiling technologies - SNP arrays

High-throughput microarray-based assays and next-generation sequencing (NGS) have been used broadly to find disease-associated single SNP and CNV markers. NGS technologies include whole-exome and whole-genome sequencing (WES, WGS).

WGS provides a general view of the genome, thus improving the detection of shorter and novel CNAs. While slow and expensive initially, the current NGS techniques have improved by using parallelization and template generation via genome fragmentation [89, 90]. Parallelization allowed the scientific community to sequence up to billions of nucleotides concurrently, providing substantially more throughput. However, one resulting drawback of WGS is the size of the generated data - which requires intense computational power and large storage capacity. Moreover, scientists need to design appropriate pipelines for determining what is biological or medical relevant in the generated sequence.

Microarray-based assays include comparative genomic hybridization (CGH) arrays and SNP arrays. CGH arrays compare copy numbers between differentially labeled target and reference DNA by measuring the fluorescence ratio along each chromosome [91].

SNP arrays also enable finding CNVs, but unlike CGH arrays, they can determine the genotype of the SNP probes embedded in the chip [92]. Although SNP arrays provide less profiling information than WGS, they have the advantage of having been used already for over two decades in the lab [93]. SNP arrays are also considered more accessible, given that they require easier and milder sample preparation than NGS [94]. Despite the rapid NGS price decrease, SNP arrays are still cost-effective and can be used for genotype and copy number analysis [95]. SNP arrays also enabled the scientific community to characterize copy number changes and allelic imbalances of a sample [96, 97].

The first part of this thesis focused on benchmarking algorithms that predict CNAs from SNP arrays, specifically – Affymetrix SNP 6.0 arrays.

The current design of Affymetrix SNP arrays typically comprises approximately 1.8 million probes – of which 906,600 SNP probes and 946,000 CNV probes. The output consists of allele-specific signals at each marker of genetic variation [97]. These positions are known to vary within the population and allow us to explore

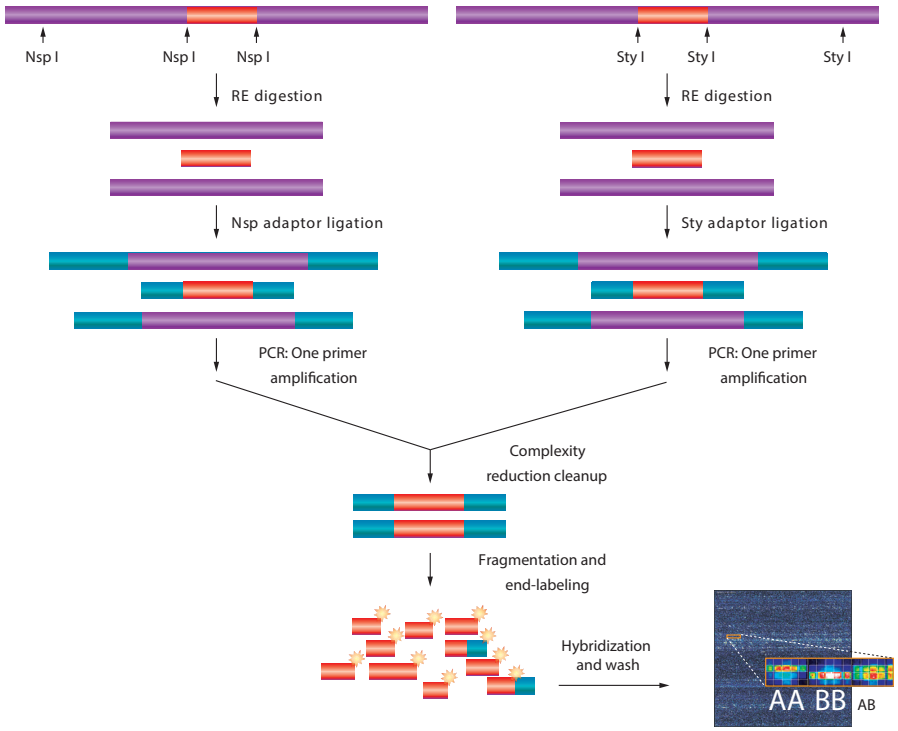

Figure 2.1: Overview of the Genome-Wide Human SNP Assay 5.0/6.0 pipeline. Figure taken from [98].

the variation in targeted genomic regions.

For genome-wide genotyping, the Affymetrix pipeline starts with two restriction enzymes (Nsp I and Sty I) digesting a total of 500 ng of genomic DNA [98]. After the two enzymes finish the digestion, all the resulting DNA fragments ligate to adaptors that recognize the cohesive four bp overhangs (Figure 2.1) [98]. Next, the technology includes generic primers that recognize the adaptor sequences and amplify the resulting adaptor-ligated DNA fragments with a preference for fragments of 200 to 1,100 bp length [98]. The products resulting after PCR amplification for each restriction enzyme digest are then combined and purified using polystyrene beads (Figure 2.1) [98]. The resulting amplified DNA is fragmented, labeled, and hybridized to the array [99, 98]. As described, Affymetrix SNP 6.0 array function based on the chemical attraction between DNA molecules: cytosine (C) attaches to guanine (G) and adenine (A) attaches to thymine (T). Each SNP probe set contains multiple oligonucleotide features that are identical copies of one of the two probes targeting the two possible

alleles (indicated as A and B in Figure 2.1). After hybridization, washing, and scanning, the technology produces a .CEL file for each sample. The .CEL file contains information regarding probe locations and signal intensities that are further used in downstream analyses like finding CNVs and SNPs associated with a specific condition.
This study focused on using the Affymetrix SNP 6.0 signals for benchmarking CNA calling algorithms (see Chapter 4).

Transcriptional profiling using NGS technologies

The transcriptomics field uses large-scale measurements of RNA molecules abundance to find changes at the molecular level associated with specific physiological or pathological conditions. Although in the beginning, microarrays remained the preferred choice for transcriptome profiling, with the improvement of NGS technologies, NGS-based RNA-Seq became the most used technology for transcriptional profiling.
RNA-seq allows scientists to carefully study the RNA abundance and sequences from a sample, enabling the analysis of varying RNA molecules.
By now, NGS has provided significant progress in terms of speed and resolution for transcriptome studies. More than that, NGS enabled finding and quantifying low-expressed genes that could not be revealed with microarrays [100].
NGS also enabled the analysis of known splice junctions together with the discovery of unknown splicing events [101]. Finally, NGS enabled analyzing allele-specific expression and finding fusion transcripts, which contribute to diseases like cancer [100, 102].

Modeling the transfer of information from DNA to protein depends on comprehending the transcription process. The reason for is that RNA-seq measurements indicate which genes are activated in a cell, what their abundance is, and what conditions influence their activation [103].
Although groundbreaking, the first sequencing RNA-seq technology – Sanger, was expensive, low-throughput, and imprecise [104]. However, recent advances provide highly parallel and high-throughput NGS sequencing methods.
As shown in Figure 2.2A, NGS RNA-seq methods start with extracting total

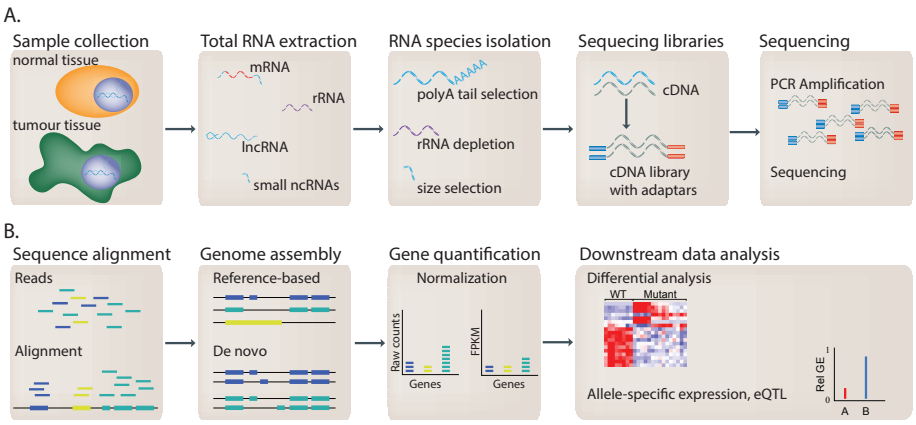

Figure 2.2: Next generation sequencing workflow.

RNA from collected samples. Next, RNA species are isolated and converted into complementary DNA fragments – cDNA libraries [105]. Adapters attach to each end of the fragments and enable sequencing (Figure 2.2A Sequencing libraries) [106].

With high-throughput sequencing technologies, scientists obtain from each cDNA, short RNAs that correspond to either one or both ends of the fragment [106]. One essential aspect of this step is the depth to which the library is sequenced, as detecting rare transcripts and variants requires more sequencing depth [106]. The sequencing returns millions of reads that are aligned to a reference genome and assembled in RNA sequences that span the transcriptome or form the transcriptome of a novel genome with no reference genome (Figure 2.2B) [107]. Once the sequence alignment is complete, we can count how many sequences map to each gene – this way, obtaining gene expression levels. However, raw read counts are affected by sequence length and the total number of reads. To make the gene expression levels comparable across a cohort, we need to normalize the read counts. Next, the normalized gene expression can be used for downstream analyses as differential gene expression (Figure 2.2B).

Dynamic measurements of protein-protein interactions

While SNP arrays and transcriptomics profiling techniques reveal omics relationships, affinity purification-mass spectrometry techniques (further referred

to as AP-MS) provide a dynamic view of the protein-protein interactions in healthy, diseased, and infected cells.

AP-MS emerged as a result of pairing methods that enrich biological material and perform chromatography (a technique that separates components of a mixture based on their differential interactions with two chemical or physical phases: a mobile phase and a stationary phase) with improved mass spectrometry [108]. Pairing affinity purification with mass spectrometry enables studying protein-protein interactions in protein complexes across different conditions, providing a more comprehensive and dynamic view of the physical interactome level. Specifically, AP-MS techniques use epitope tags on a single "bait" protein or molecule of interest, while probes of the interacting "prey" proteins are affinity captured for identification. Proteins that do not interact are separated and discarded. After purification, proteins are processed by MS.

Although AP-MS may reveal interacting proteins, it is constrained when distinguishing direct from indirect interactions and when proteins are highly co-participating in complex molecular processes [108].

2.2 Statistical methods

To comprehensively understand how biological systems function given a specific set of preexistent conditions, scientists use statistics regularly: from data mining to hypothesis testing, modeling, and prediction. We used statistics to identify patterns characterizing multiple molecular levels across different cancer types in the projects presented here. Ultimately, we aimed to enhance our knowledge about the initiation and molecular mechanisms of oncogenic processes.

To achieve our aim, we use both methods that predict *outcomes* (*responses*), based on several input variables – *features* (*predictors*) and methods that do not require an output measure. These methods are known as *supervised* and *unsupervised learning*, respectively.

An example of supervised learning included in this thesis is predicting DNA copy number states (the output) based on SNP array signal intensities, tumor purity levels, and tissues type (input variables) – see Chapter 4. Examples of unsupervised learning included in this thesis are hierarchical clustering of

miRNA – gene associations for intuitive analysis of miRNA function and the hierarchical clustering of lncRNA – pathway associations for intuitive analysis of lncRNA function.

This section introduces the statistical methods we used to learn about cancer-specific mechanisms based on multiple omics levels. In particular, we performed hypothesis testing to e.g., test for differential gene expression between healthy and cancer samples. We calculated correlations for finding associations between miRNAs and mRNAs in head and neck cancer samples. We designed, built, assessed, and selected prediction models - e.g., for robustly estimating DNA copy number changes in cancer. Finally, since cancer does not activate a single unit level, but rather a complex of molecules and interactions, we analyzed a molecular interaction network in the context of cancer. In detail, we used network propagation to identify how viral infections affect the human protein-protein interaction network in liver cancer patients. The following subsections intuitively describe the statistical concepts so that the reader can easily understand and follow why and how we used them.

Types of variables and data distributions

Given a sample space, scientists can already learn from the data using descriptive statistics (exploratory analysis). Summary statistics provide a numerical overview as a table or as *data frequency distribution* [109]. The frequency distribution is a parametrized mathematical function that indicates how frequently each sample value occurs [109]. The form of the distribution depends on the nature of the variables [109].

The first two classes of variables that one can distinguish are *numerical* and *categorical*. The numerical variables represent a *quantitative* measurement, while the categorical variables represent a *qualitative* measurement, a characteristic [109]. Numerical variables are either continuous — e.g., the gene expression levels, or discrete — e.g. survival days. Categorical variables can be nominal — e.g., indicating the ethnicity of a patient, or ordinal — indicating the grading of a tumor. A special type of categorical variables are the binary variables. An example of a binary variable is the infection status of a patient and it indicates

whether the patient is infected or not.
The nature of variables used in statistics and machine learning determines corresponding distributions types and influences the appropriate choice for model usage. For example, for normally distributed continuous data, it is meaningful is to calculate the mean and standard deviation, while for skewed continuous and categorical ordinal data, an informed choice is to calculate the median and the interquartile range [109].
Knowing the variable and distribution enables scientists to identify outliers and skewness in the sample space. Following, variable and distribution types aid in deciding what is the most appropriate statistical method to apply and how to interpret the results correctly.
Next, we will present what kind of tests and models we can use for data analysis.

Hypothesis testing

In research, we often aim at interpreting data for answering a specific scientific question. For example, we aim to identify if there are differences in gene expression between healthy and cancer samples. For this purpose, researchers often use inference methods such as hypothesis testing. A statistical hypothesis is a statement about the parameters of a population [110].
Generally, we state a hypothesis based on a scientific question we aim to answer based on sample data of a population. One example of a scientific question addressed in this thesis is: does the Hepatitis B viral infection affect the mutation rate of a given protein-coding gene in liver cancer patients?
To test whether a hypothesis is *statistically significant*, we need to perform the following steps:

- Formulate the null (H_0) and the alternative (H_1) hypotheses: H_0 states that the observations are the result of chance, while H_1 states that the observations are the result of a real effect. For our example, H_0 states that the mutation rate of a specific protein-coding gene is unaffected by the viral infection status in our liver cancer data, while H_1 states the opposite.

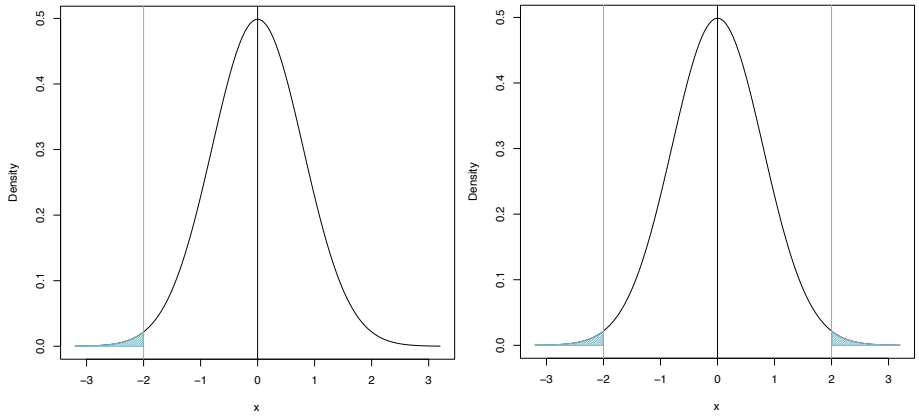

Figure 2.3: Statistical tests: A. One-tailed test - here a left-tailed test example: the shaded area represents the critical region limited by the test statistic of the sample data mean. The vertical line $x = 0$ represents the population mean, while the vertical line $x = -1$ represents the threshold for achieving statistical significance. B. Two-tailed test: the shaded regions represent the critical regions.

		Parametric	Nonparametric
One sample		t-test z test	Wilcoxon signed rank test Kolmogorov-Smirnov
Two samples	independent	two-group t-test Z test	Mann-Whitney Kolmogorov-Smirnov
	paired	paired t-test	Wilcoxon

Figure 2.4: An overview of statistical tests used in the next chapters.

- Select the most appropriate method and calculate the test statistic under the assumption that the null hypothesis is true.
- Determine if the test result is statistically significant.
- Interpret the statistical test results.

To establish if the result is statistically significant, one needs to calculate a *p-value*. A p-value represents the probability that a test statistic is at least as extreme as the one observed in the given data [109]. The null hypothesis will be rejected if the *p-value* is below the significance level α, commonly set to 5% or 0.05.

Depending on the research question, we can apply statistical tests such as *one-*

tailed or *two-tailed* tests.

For a left-tailed test, we reject H_0 if the test statistic is lower than the significance level. For a right-tailed test, the test statistic must be higher than the significance level (Figure 2.3A). In a two-tailed test, we reject H_0 for either lower or higher values of the test statistic (Figure 2.3B).

Depending on the variable type and data distribution, we distinguish between *parametric* and *non-parametric tests*.

While parametric tests require a specific distribution of the data, nonparametric tests can be applied for parameter-free distributions or when the parameters are unknown[111]. Furthermore, parametric tests use the mean as a measure of the shift, while nonparametric tests use the median [109].

Moreover, parametric tests apply to variables only, while nonparametric tests apply to both variables (dependent measurements) and attributes (independent measurements) [109].

As we can see in the overview of statistical tests presented in Table 2.4, we further distinguish between one-sample and two-sample tests.

A one-sample test determines if the mean of a sample $\bar{x}$ from a normal distribution is significantly different from a standard specific value c. To determine if the mean difference between two groups is zero, we use a *two-sample* test [109].

Parametric tests

To test if a sample mean is statistically different from a known population mean μ, we can use a *z-test* or *t-test*. Both z-tests and t-tests assume the sample follows a normal distribution. Given the sample size is sufficiently large or the population variance is known, we conduct a z-test and calculate the observed z-statistic as follow:

$$z_{obs} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1),$$

where $\bar{x}$ is the sample mean, μ is the population mean, σ is the population standard deviation, n is the sample size, N(0,1) is a normal distribution with $\mu = 0$ and $\sigma = 1$, and z_{obs} is the z-statistic. Next, given the symmetry of the normal distribution, we calculate the p-value by solving:

$$P(|z| \geq |z_{obs}|) = 2 \cdot \phi(-z_{obs}),$$

where ϕ is the cumulative distribution function of a standard normal $N(0,1)$. If we want to test if the p-value is lower than a predefined rejection threshold α, we reject the null hypothesis when $P(|z| > |z_{obs}|) < \alpha$.

For small sample sizes and unknown population variance, we use the t-test [109]. Since σ is unknown, the t-test approximates $sigma$ by the sample standard deviation s [109]. We then calculate the t-statistic as it follows:

$$t_{obs} = \frac{\bar{x} - \bar{\mu}}{\sqrt{\frac{s^2}{n}}},$$

where $\bar{x}$ is the sample mean, μ is the proposed value for the population mean, s is the sample standard deviation and n is the sample size. Therefore, for $t-tests$ the population standard deviation is replaced with the sample standard deviation.

However, this quantity does no longer follow a normal distribution, but a t-distribution [109]. The shape of the t distribution, also known as *Student's t* distribution, is defined by the sample size n and covers a group of curves ordered by the degrees of freedom [109]. The degrees of freedom represent the amount of independent values used for an estimate, for a specific number of samples [109]. We can calculate the p-value by solving:

$$P(|t_{df}| \geq |t_{obs}|) = 2 \cdot P(|t_{df}| \leq |t_{obs}|).$$

To determine if the average difference between two groups significantly differs from 0, one can use a two-sample t-test or a z-test. The independent two-samples t-test can be used only when the distribution of the sample in the two groups is normal and the variances within the two groups are equal. In this case the null hypothesis states that $\mu_1 = \mu_2$. The test statistic is calculated as follow:

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}},$$

where n_1, n_2 are the sample sizes, $\bar{x_1}, \bar{x_2}$ are the sample means, μ_1, μ_2 are the population means, and s_1, s_2 are the sample standard deviations.

We conduct the z-test to assess if the average difference between two groups significantly differs from 0, when n is large and the population variance is known

[112, 109]. The formula for calculating the z test statistic becomes:

$$z = \frac{\bar{x_1} - \bar{x_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}},$$

where σ_1, σ_2 represent the standard deviations of the two populations [112].

To compare the means of two groups with one-to-one relationship between their samples, we use the paired two-samples t-test. To perform a paired t-test, we first calculate the difference d between each pair of samples. Next, we calculate the mean $\bar{d}$, and the standard deviation of the differences d and compare the mean to 0. A mean that is far from 0, indicates a significant difference between the two pairs of samples. The t-test statistic value is calculated using as follow:

$$t = \frac{\bar{d}}{s/\sqrt{n}},$$

where $\bar{d}$ is the mean difference, s is the standard deviation of the distances and n is the sample size. If the p-value corresponding to the $|t|$ for the degrees of freedom $df = n - 1$ is smaller or equal to 0.05, the two paired samples are statistically significant.

Nonparametric tests

When no or few information about the distribution of the given data, we use nonparametric tests. For cases when the data is not assumed to be normally distributed, the alternative to the one-sample t-test is the *Wilcoxon signed-rank test.*

The Wilcoxon signed-rank test relies on ranks and, as a result, uses the median instead of the mean.

Given the set of observations $\{x_1, ..., x_i, ..., x_n\}$ and m a given value, we want to evaluate if the sample median is equal to m. For this, we assume that x_i is continuous and the probability function for x_i is symmetric around the median. Next, we calculate the differences between each observation and the given value: $d_i = x_i - m$. We assign to each rank r_i the sign of each d_i: $s_i = sign(d_i)r_i$ and calculate:

$$W = \sum s_i > 0.$$

The Wilcoxon signed-rank sum statistic is defined as:

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}},$$

where n is the number of samples. Finally, the probability of the test statistic z under the null hypothesis is indicated by the p-value.

The corresponding version of the Wilcoxon rank-sum test for comparing two independent samples with non-normal distribution, thus the substitute for the independent two-sample t-test, is also known as the Mann-Whitney U test [113]. Given two independent samples $x = \{x_1, ..., x_{n_1}\}$ and $y = \{y_1, ..., y_{n_2}\}$, the Mann-Whitney test compares every observation x_i with every observation y_j. The U statistic is defined as:

$$U_x = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum_{i=n_1+1}^{n_2} R_{xi}$$

and

$$U_y = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=n_2+1}^{n_2} R_{y_j},$$

where n_1 and n_2 are the sample sizes and R_i, R_j are the ranks. U_x indicates how many times the observations in x outrank the observations in y, while U_y indicates how many times the observations in y outrank the observations in x. To determine if the null hypothesis is rejected, we have to compare the U statistics to the statistical table corresponding to the Mann-Whitney U test for a significance level α. For large samples, U is asymptotically normally distributed and z is defined as:

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$$

with $z \approx N(0, 1)$.

A significant p-value indicates a significant difference between the medians of two groups. For paired samples, we use the Wilcoxon paired test.

To determine if two random variables X_1 and X_2 are drawn from the same

continuous distribution given two samples x_1 of length n and x_2 of length m, we use the Kolmogorov-Smirnov (KS) test.

Given the null hypothesis $H_0 : F_1(x) = F_2(x)$, where F_1 and F_2 are the empirical distribution functions of x_1 and x_2, the KS statistic D is calculated using the following formula:

$$D = \sup_z |F_1(x) - F_2(x)|.$$

The null hypothesis is rejected at significance level α if

$$c(\alpha) > D_{n,m}\sqrt{\frac{nm}{n+m}},$$

where

$$c(\alpha) = \sqrt{-\frac{1}{2}ln(\alpha)}$$

for large sample sizes. For small sample sizes the significance level α is drawn from the table of critical values.

To test if there is a non-random association between two nominal variables that result from classifying objects in two different ways, we use Fisher's exact test. Fisher's exact test is a distribution-free test.

Let c_1 and c_2 be the two nominal variables and n the number of observations in the sample population. The number of observations within class c_1 and c_2 can be arranged in a 2×2 contingency table (Table 2.1):

	c_1	$\bar{c}_1$	$\sum$
c_2	n_{11}	n_{12}	$n_{1\cdot}$
$\bar{c}_2$	n_{21}	n_{22}	$n_{2\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Table 2.1: A 2×2 contingency table.

The null hypothesis states that the observations of class c_1 and class c_2 are indepedent $H_0 : n_{11}/n_{\cdot 1} = n_{12}/n_{\cdot 2}$, where x is number of observations with property A and B, yielding $n_1 1 = x$. As seen in Table 2.1, the test takes row sums and column sums as given. We can then calculate $P(n_1 1 = x)$ using the hypergeometric distribution [114]:

$$P(n_{11} = x) = \frac{\binom{n_{1\cdot}}{x}\binom{n_{2\cdot}}{n_{21}}}{\binom{n}{n_{\cdot 1}}}$$

For a one-sided test, we can obtain a p-value for rejecting the null hypothesis by summing up the probabilities of observation frequencies and the probabilities of all other configurations that reflect a greater difference between conditions, i.e. higher values of n_{11}: $p = \sum_{i=x}^{\min(n_{1\cdot}, n_{\cdot 1})} P(n_{11} = i)$

Multiple testing correction

In the field of computational biology, we frequently deal with massive-scale data. Testing a broad set of hypotheses simultaneously increases the probability of falsely estimating random events as significant.
For example, in our research on the hepatitis B impact in liver carcinomas, we needed to determine which proteins were significantly affected by the virus. Therefore, we evaluated the cost associated with a false positive target and the advantage of revealing an unknown oncogenic molecular player.
Generally, this translates into associating a statistical confidence measure to each discovery. These measures may be p-values, false discovery rates, or q-values.

Bonferroni correction

A simple and conventional method to correct for multiple testing error is to apply the Bonferroni correction to the probability of a particular result occurring by chance: given a set of n hypotheses to be tested, with $p_i, i \in 1, ..., n$ as the corresponding p-values yielded by each test, and a significance threshold α, the Bonferroni correction considers a score significant only if $p_i \leq \frac{\alpha}{n}$ [115].
By applying a threshold α to a set of n significance scores, the Bonferroni correction controls the family-wise error rate. For example, for $n = 100$ and $\alpha = 0.01$, there is only a ~ 0.01 chance of observing at least one significant result, given all the test are not significant. Therefore, for most multiple testing corrections, minimizing the family-wise error rate is too strict and leads to not rejecting the null hypothesis when true effects exist.

Benjamini-Hochberg False Discovery Rate

When dealing with large-scale multiple testing, controlling the Benjamini and Hochberg (BH) false discovery rate (FDR) can be more relevant. The FDR is

defined as the percentage of false positives among all significant results.
For a significance level α, the BH procedure estimates a rejection region so that, on average, $FDR < \alpha$ [116]. The procedure follows the next steps:

- orders the unadjusted $p_i, i \in 1, ..., n$ values
- assign ranks to the p-values
- finds the test with the highest rank, p_r with $p_r \leq \alpha \frac{r}{n}$, where r is the rank of the p-value, n is the total number of tests and *alpha* is the proposed FDR [116].

The results with $p_i, i \in 1, ..., r$ and $p_1 \leq p_2 \leq ... \leq p_r$ are significant.

Storey's Empirical P-value based False Discovery Rate

Storey's correction, also known as the positive false discovery rate, adjusts the p-values by estimating the fraction of truly null tests π_0. Given that there are enough tests, we can robustly estimate π_0.
Furthermore, Storey defined the q-value as the minimum FDR attained at or above the significance level α. The q-value is then the expected percentage of false positives among all of the significant scores above α.
Choosing the appropriate multiple testing correction method depends on the number of tests and the threshold for false positives.

Hypothesis testing for multiple measurement variables.

Correlation analysis

One way to determine if and how strongly two random variables are associated is correlation analysis. The Pearson correlation coefficient is used for data that show a normal distribution, while the Spearman's rank correlation is used for data that follow a non-normal distribution or that have relevant outliers.
The Pearson coefficient is defined as:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

where X and Y represent the two random variables, $Cov(X,Y)$ is the covariance between X and Y and $\sigma_X \sigma_Y$ are the standard deviations of X and Y.

Given the paired data $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ and

$$Cov(X, Y) = E\left[(X - E\left[X\right])(Y - E\left[Y\right])\right]$$

and

$$\sigma_X = \sqrt{E(X^2) - (E(X))^2} \tag{2.1}$$

$$\sigma_Y = \sqrt{E(Y^2) - (E(Y))^2}, \tag{2.2}$$

The Pearson correlation coefficient can be calculated as:

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}},$$

where n is the sample size of x and y.

The correlation coefficient $\rho_{X,Y}$ ranges from -1 to 1, where -1 indicates a perfect inverse relationship between X and Y, 1 indicates a perfect correlation, and 0 indicates no correlation between the two variables.

A high correlation means that two or more variables show a strong association with each other, while a correlation close to 0 indicates that the variables are hardly related.

To test whether the correlation between two variables is statistically significant, we use a statistical test that builds on the *Fisher transformation* [117] of the correlation coefficient r as follows:

$$F(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right).$$

Given that X and Y are normally distributed, $F(r)$ is also normally distributed with mean $\mu = F(\rho)$ and standard error $\sigma = \frac{1}{\sqrt{n-3}}$. To determine, if the two variables are significantly correlated, we want to test the null hypothesis that $\rho = 0$. Hence, we can calculate a z-score as:

$$z = \frac{F(r) - \mu}{\sigma} = F(r)\sqrt{n-3}.$$

We then use the cumulative distribution function of the standard normal distribution ϕ to obtain a two-sided p-value indicating if the null hypothesis is rejected:

$$p = (1 - \phi(F(|r|)\sqrt{n-3})) * 2.$$

Spearman's rank correlation coefficient is used to determine non-linear relationships and is robust to outliers.
Given x and y two samples of size n drawn out from the random variables X and Y, and $R(x), R(y)$ the ranks of x and y, the Spearman's rank correlation coefficient can then be calculated as:

$$r_{x,y} = 1 - \frac{6\sum_{i=1}^{N}(R(x)_i - R(y)_i)^2}{N(N^2-1)}.$$

Since ordinal data can also be ranked, the Spearman's rank correlation is not limited to continuous variables. Through ranking, Spearman transforms a non-linear strictly monotonic relationship to a linear relationship and returns a coefficient that is also relatively robust against outliers.
To assess the statistical significance of Spearman's rank correlation coefficient, we can apply the Fisher transformation to $r_{x,y}$ similarly to Pearson's correlation. The standard error should be chosen as $\sigma = \sqrt{1.06/(n-3)}$ [118].

Linear regression analysis

Another way to analyze the relationship between two or more variables is regression. Regression comprises a set of statistical methods that estimate the relationship between a dependent variable – outcome $\mathbf{y_i}$ and m independent variables – predictors $\mathbf{x_i}$.
If the relationship between $\mathbf{y_i}$ and $\mathbf{x_i}$ is linear, we refer to the model as *linear regression*. If $m > 1$, we refer to the model as *multiple linear regression*.
A simple linear regression for modeling n data points has the following form:

$$\mathbf{y} \sim \beta_0 + \mathbf{x_1}\beta_1 + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma)$ represents the error term, β_1 represents the unknown parameter and β_0 represents the intercept. The equation describes a straight line.

For multiple linear regression, the model equation describes a hyperplane and has the following form:

$$\mathbf{y} \sim \beta_0 + \mathbf{x_1}\beta_1 + \mathbf{x_2}\beta_2 + \ldots + \boldsymbol{\epsilon},$$

To find a model optimal for predicting an outcome $\hat{\mathbf{y}}$, we estimate the corresponding coefficients $\hat{\boldsymbol{\beta}} = \{\hat{\beta_0}, \ldots, \hat{\beta_m}\}$. For this purpose, we need to calculate the residuals $e_i = y_i - \hat{y}_i$, i.e. the difference between the predicted and the actual value of the output.
One of the most common methods to find the optimal model is to find the optimal coefficient set that minimizes the residual sum of squares – also known as the least squares approach:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j)^2 = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2.$$

Given that all m predictors are linearly independent, the minimization problem has one unique solution:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Although this is a suitable solution for calculating the accuracy of $\hat{\beta}$, linear regression is sensitive to variance (the spread between the $\hat{\beta}$ parameters). Particularly, the accuracy of the model prediction is sensitive to correlated predictive features or a high number of predictive parameters.

Subsetting and regularization

Subsetting and regularization reduce the total error of prediction by scaling down the variance. However, this introduces a small bias. Regularization shrinks or removes the coefficients of the variables with weak effect on the response, thus, providing a subset of predicting variables with the strongest effects on the response.
One intuitive way to select a subset of $k \in \{1, \ldots, m\}$ variables for linear regression is to test all possible combinations and select the subset that minimizes the residual sum of squares – also known as *best subset regression* [119].
However, the residuals are smaller the more variables we introduce in the model.

Thus, the sum of squares cannot be used as a criterion for determining k. Moreover, this approach is not informative about the effect of the excluded variables on the response variable, becomes infeasible for large m, and is likely to lead to underfitting or overfitting.

Shrinkage methods are alternatives to variable selection for an optimal model. These methods regularize the $\hat{\beta}$ coefficients toward 0 by introducing a penalty on their size, and thus reducing the variance. The most commonly used shrinking methods are Ridge regression [120] and Lasso regression.

Ridge regression adds a penalty term – λ to the linear regression minimization problem:

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2 \right\}.$$

Note that when $\lambda = 0$ this is equivalent to the unpenalized regression. λ controls the magnitudes of $\hat{\boldsymbol{\beta}}$, and thus, the degree of shrinkage in the regression model. For large values of λ the $\hat{\boldsymbol{\beta}}$ coefficients are severely constrained and the degrees of freedom will descend, tending to 0 as $\lambda \to \infty$ [121]. The solution of the Ridge regression minimization problem can be solved in closed form as follows [122]:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mathbf{I}$ is the identity matrix.

Although it exploits the trade-off between variance and bias, Ridge regression includes all m predictors in the model regardless of the value of their $\hat{\boldsymbol{\beta}}$ coefficients. This becomes challenging for a model with an extensive number of predictors. This loss is overcome by another shrinkage method - *the least absolute shrinkage and selection operator (lasso)* [123].

The lasso estimator is defined as:

$$\hat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right\}.$$

Unlike Ridge regression, lasso regression uses the absolute value of the coefficients: $\lambda \sum_{j=1}^{m} |\beta_j|$. This penalty can be also expressed as $||\boldsymbol{\beta}||_1$ and is referred

to as the L_1 penalty. In contrast to the $L_2 = ||\boldsymbol{\beta}||_2$ penalty of ridge regression, the L_1 penalty forces some of the $\hat{\boldsymbol{\beta}}$ values to be exactly 0. Thus, lasso regression returns a sparse solution by removing variables that do not fit the model well.

Another important difference between Ridge and Lasso is how they handle the problem of multicollinearity between the predictors. Ridge regression returns similar values for the $\hat{\boldsymbol{\beta}}$ coefficients of correlated predictors, while LASSO selects and assigns the entire impact to one of them and removes the other ones.

As a result, Ridge regression is expected to perform better when most predictors have a true effect on the response, while lasso is expected to perform better when only a few predictors affect the response.

Ideally. we would like to be able to perform both feature selection and handle the correlated predictors. To account for the loss of information from lasso when there is a combined effect of the predictors, one needs to keep the correlated variables in the model. For this purpose, Zou et al. proposed the *elastic net* approach [124]:

$$\hat{\boldsymbol{\beta}}^{EN} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda P_\alpha(\boldsymbol{\beta}) \right\},$$

where the elastic net penalty is represented by $P_\alpha(\boldsymbol{\beta})$ and is defined as:

$$\begin{aligned} P_\alpha(\boldsymbol{\beta}) &= (1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1 \\ &= \sum_{j=1}^{p} \frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|. \end{aligned}$$

Note that P_α includes both, the lasso penalty (L_1) and the ridge penalty (L_2), and represents a trade-off between the two penalties. The trade-off is controlled by the α parameter: $\alpha = 0$ is equivalent to ridge regression, while $\alpha = 1$ is equivalent to lasso regression. When $0 < \alpha < 1$, both feature selection and appropriate handling of correlated variables are performed.

The elastic net is a good approach to achieve reduced model complexity, improved model prediction, but also to achieve model interpretability: with a subset of most powerful predictors, we can grasp relationships between the model variables.

The methods described so far are appropriate for continuous response variables. We now focus on regression models appropriate for binary response data.
Given a response variable that is binary and is e.g. defined as:

$$y_i = \begin{cases} 1, & \text{if y is mutated} \quad (2.3) \\ 0, & \text{if y is wild type,} \quad (2.4) \end{cases}$$

y_i represents a realization of a random variable Y_i, and p_i and $1 - p_i$ are the probabilities of y_i being mutated and wild type, respectively. Then, Y_i follows a *Bernoulli* distribution defined as:

$$Pr\{Y_i = y_i\} = {p_i}^{y}_{i}(1 - p_i)^{1-y_i}.$$

Given that under Bernoulli distribution, for $y_i = 1$, $Pr\{Y_i = y_i\} = p_i$, while for $y_i = 0$, we obtain $Pr\{Y_i = y_i\} = 1 - p_i$, the mean and variance of the distribution depend on p_i. This means that any feature that affects the probability affects both the mean but the variance of the response variables. Thus, a linear regression model under which the features influence the mean but not the variance is not suitable for binary data. Moreover, the predicted probabilities are not restricted to the $[0, 1]$, while probabilities must take values between 0 and 1. For binary response variables, we next introduce the concept of *binary logistic regression.*

Binary Logistic Regression

Applying a logit transformation to the model, helps us overcome the issues of using linear regression with binary data. Given the initial setup and supposing the logit is a linear function of the features, we can write:

$$logit(p_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

where $\mathbf{x}_i^T$ is the feature vector and $\boldsymbol{\beta}$ is the regression coefficients vector. Note that this model is a generalized linear regression model with a binomial output, where the $\boldsymbol{\beta}_i$ coefficient represents the variation of $logit(p_i)$. p_i is the probability associated with a unit change in the i-th feature while all the other features

are constant. When applying an exponential transformation to the previous equation, we obtain:

$$\frac{p_i}{1-p_i} = e^{\mathbf{x}_i^T \boldsymbol{\beta}}.$$

Note that if we now change the i-th feature by one unit while all the other features remain constant, we are simply multiplying the odds by $\exp \boldsymbol{\beta}_i$. To solve the equation, we can write:

$$p_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$$

Worth noting that while p_i represents a probability, the right side of the equation is a non-linear function of the features, thus it is difficult to estimate the change in probability introduced by a unit change of one of the features.
Following, we can calculate the logarithm of the odds ratio as follows:

$$\log \frac{Pr(y=1|\mathbf{X}, \boldsymbol{\beta})}{Pr(y=0|\mathbf{X}, \boldsymbol{\beta})} = \mathbf{X}\boldsymbol{\beta}$$

Under the assumption that the y outputs are generated independently given $\boldsymbol{\beta}, the Pr(y|\mathbf{X}, \boldsymbol{\beta})$ becomes:

$$Pr(y|\mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \sigma(\mathbf{x_i}^T \boldsymbol{\beta})_i^y (1 - \sigma(\mathbf{x_i}^T \boldsymbol{\beta}))^{1-y_i}$$

Note that the equation form will enforce that y_i is either 0 or 1, thus either $y_i \neq 0$, either $1 - y_i \neq 0$, allowing for the correct input to the likelihood.
The common method to solve a logistic regression problem is to maximize the likelihood as a function of the regression parameters $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} Pr(y|\mathbf{X}, \boldsymbol{\beta}).$$

Even though logistic regression does not have a closed-form solution for the maximum likelihood estimator, the negative log-likelihood function is convex. Thus, logistic regression has a unique solution given by the global minimum of the maximum likelihood estimator.

2.3 Network-based biological models

Network-based biological models are powerful resources for discovering genetic associations. Fundamental to network models is the concept that genes underlying the same phenotype tend to interact.

Given the overlap of phenotypic characteristics between a gene and its direct interactors, initial network models were built upon the "guilt-by-association" principle [125]. Subsequent methods used the concept of local network neighborhoods to find modules (Figure 2.5A). However, the "guilt-by-association" network models were proven to be cost-effective for gene functional annotation[126, 127].

A new class of network models relies on the concept of *network propagation* to prioritize phenotype-associated genes. Network propagation amplifies a biological signal by projecting prior gene-phenotype association knowledge (e.g. presence of disease-associated mutations) over the network nodes [127]. The method propagates the prior knowledge repetitively through the neighboring nodes of the biological network, until convergence is reached or for a given number of iterations (Figure 2.5B). Consequently, each node score is altered not only by the scores of its direct neighbors but also by the scores of the indirect neighbors.

Thus, the method assigns an association score for nodes with no prior knowledge, where their assigned score indicates the proximity to nodes with prior knowledge.

The straightforward approach of predicting all direct interactors of phenotype-associated genes can result in both high positive and negative rates. An alternative that corrects for this effect is to prioritize genes based on their distance to the initial gene list. However, since most distances are relatively small (biological networks are highly connected), this approach could also return a high false positive rate [127].

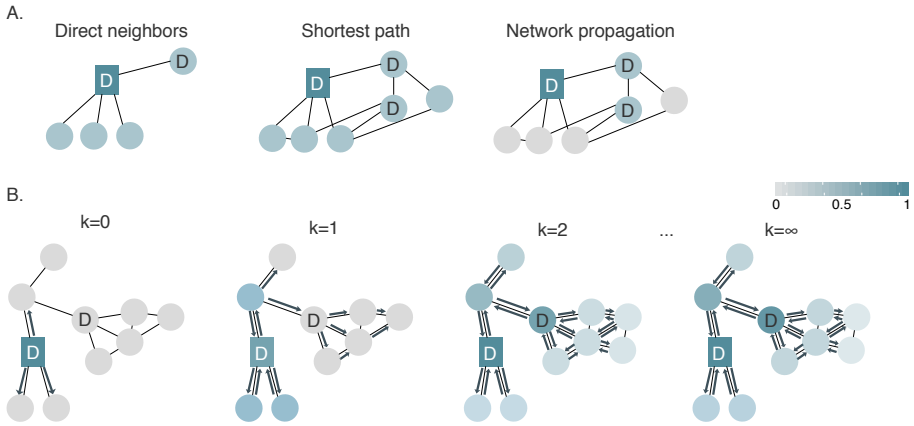

Figure 2.5: Network-based biological models. Nodes are colour-coded according to the scoring that they receive. A. Difference in node ranking between direct interactors, shortest path and network propagation. B. Example of a network propagation process process at different iterations. The process stops when convergence is reached. **D** indicates association with disease. Square node indicates prior knowledge, while circular node indicates predicted association.

Network propagation enables prioritizing genes associated with a specific phenotype, and that are more likely to interact with each other than with other genes by analyzing all possible paths concurrently. Explicitly, the method inhibits predicted node scores supported by only one interaction (path) and boosts predicted scores for nodes with no prior knowledge, but well connected to prior phenotype-associated nodes [127].

The propagation can be performed based on different formulations: from random walks on a graph to heat diffusion and computing minimum energy states in electric circuits. Given p_0 representing the prior knowledge, p_k the ranking vector at iteration k, and W the normalized version of the adjacency matrix A of the reference biological network, the propagation process can be described as follows:

$$p_k = W p_{k-1}, \quad \text{which} \quad \text{yields}$$
$$p_k = W^k p_0$$

When W is stochastic, the propagation is based on random walks, i.e. a transition from a current node is made to a random adjacent node with a probability given by W.

Using a random walk with a restart propagation process enables controlling the ratio between prior knowledge and network smoothing. The process can be described as follows:

$$p_k = \alpha p_0 + (1 - \alpha) W p_{k-1},$$

where α indicates the impact of prior knowledge and network smoothing.
As long as the convergence condition is satisfied, W can be defined in alternative ways based on A and the diagonal matrix D. Here, $W = AD^{-1}$.
The final ranking scores p_f - obtained when convergence is reached, can be written as function of the prior knowledge vector – $p_f = Sp_0$, where S can be viewed as a similarity matrix.
Biological network-based models continue to be successfully applied to rank disease-associated genes, find gene-gene similarities, and integrate multiple omics levels. They remain an effective resource for interpreting how diseases alter molecular processes.

Chapter 3

Materials

Throughout this thesis, we analyzed molecular mechanisms of head and neck squamous cell carcinoma (HNSCC) and lung adenocarcinoma (LUAD), together with the interactomes of hepatitis C and hepatitis B in liver hepatocellular carcinoma (LIHC) and functions of non-coding RNAs by integrating multilevel omics data. For this purpose, we used data provided by The Cancer Genome Atlas portal[1] (TCGA). Starting with 2005, TCGA has provided means for studying different aspects across a wide range of cancer types. Concerning the cancer types studied in this thesis, the TCGA consortium published a genomic characterization of HNSCC [32], a molecular profiling of LUAD [128], and an integrative genomic characterization of hepatocellular carcinomas [129]. Unlike TCGA, we used the HNSCC Affymetrix SNP 6.0 array data to perform a benchmark analysis for copy number calling methods including the standard method used by TCGA. Additionally, we used HNSCC miRNA and mRNA measurements to estimate miRNA-target networks in HNSCC. While TCGA focused on reporting a general molecular profiling of LUAD, we focused on finding miRNAs involved in metastasis of LUAD. Finally, we applied an integrative network approach on a LIHC subse to find viral-host protein-protein interactions in liver cancer.
In addition to the publicly available LIHC TCGA data, we analyzed experimental data that measure the interactions between hepatitis B proteins and human host proteins in two hepatocellular carcinomas cell lines. The experimental data was provided by Manon Eckhard, Ph.D., and John Gordan, MD Ph.D.,

[1] https://cancergenome.nih.gov

our collaboration partners from the Krogan Lab, University of California San Francisco.

Furthermore, we used a subset of the publicly available Haplotype map data (HapMap[2]) for benchmarking several commonly used copy number calling algorithms. Lastly, we used TCGA data on lung adenocarcinoma to investigate miRNA regulatory networks involved in metastasis.

3.1 HNSCC TCGA data

We used TCGA publicly available data sets comprising of mRNA and miRNA expression, DNA copy number data, as well as clinical data. The complete data sets consisted of 522 samples.

For our analyses, we obtained Level 1 Affymetrix Genome-Wide SNP 6.0 array data. We preprocessed the matched tumor and normal raw HNSCC .CEL files with the Aroma Affymetrix Power Tools (APT) package [130] and the PennCNV-Affy pipeline [131]. In this step, we performed quantile normalization and generated genotype calls from the .CEL files using the Birdseed algorithm [132]. Afterward, we extracted allele-specific signals, and calculated the canonical clustering parameters for each single nucleotide polymorphism (SNP) or copy number marker. We then calculated probe-wise logR ratios (LRR) and B allele frequencies (BAF) for each patient sample. For further downstream analysis, we split the signal file into individual files for each patient.

We used Level 3 IlluminaHiSeq RNASeqV2 mRNA and Illumina HiSeq 2000 miRNA data consisting of expression measurements that were generated following the protocol previously described by the TCGA consortium [81]. We used the *R* framework for statistical computing [133] to select mRNAs with non-zero count values in more than 80% of the patients, non-zero standard deviation and we applied a log2 transformation.

We selected miRNA precursors that accomplished the same expression criteria as the mRNAs. We overlapped the precursor entries with associated entries in TargetScan (version 6.2). By doing so, we considered only those miRNAs that

[2]https://www.broadinstitute.org/international-haplotype-map-project/haplotype-map

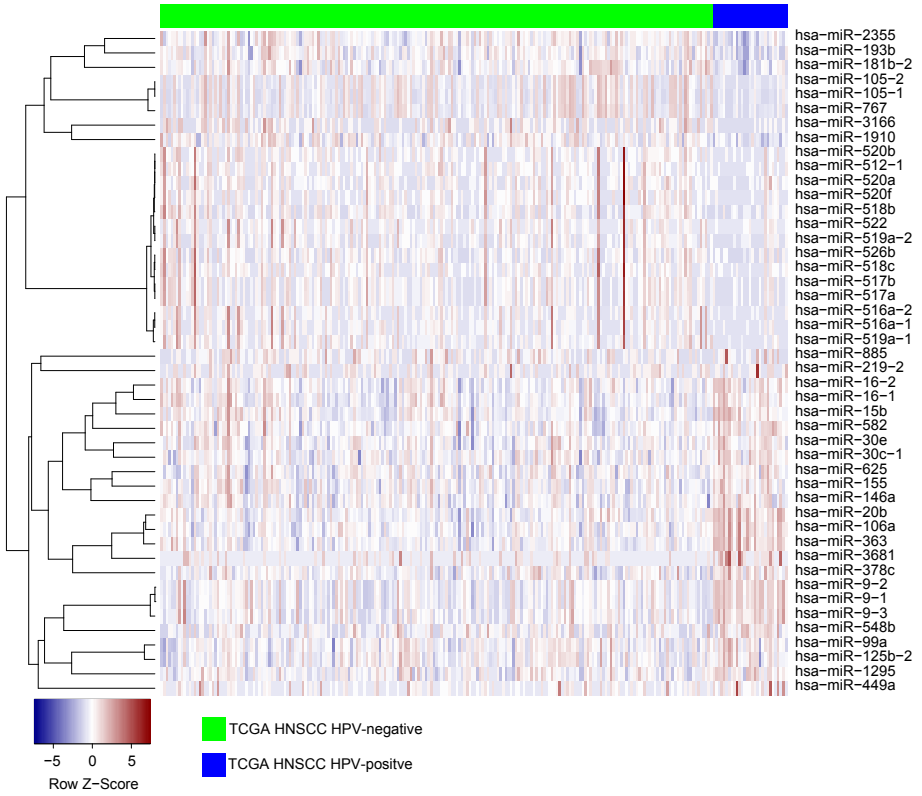

Figure 3.1: Heatmap of 44 miRNAs that show differential expression between HPV+ (blue) and HPV- (green) patients. The miRNA expression values were standardized row-wise. Low values are indicated in blue whereas high values are colored red.

are incorporated into the RNA-induced silencing complex (RISC) complex and, thus, not subject to degradation. Next, we merged the two sets of putative miRNA targets predicted by TargetScan on both forward and reverse strands for each miRNA precursor [134].

We selected a subcohort of 244 patients for which the human papillomavirus (HPV) status clinical parameter was provided [32] and tested for deregulated miRNAs between HPV+ and HPV- patients (Figure 3.1). When using the edgeR package [135], we additionally included age and gender as confounding variables. We controlled for a 5% false discovery rate (FDR) using Benjamini and Hochberg algorithm [116].

3.2 Haplotype Map (HapMap) data

We downloaded 98 Affymetrix 6.0 SNP array profiles of healthy patients from the publicly available HapMap repository [3][136, 97]. We preprocessed the HapMap .CEL files with the APT package [130] and the PennCNV-Affy pipeline [131] as described in the previous section. Next, we split the signal file into individual files for each patient. We then selected 81 patients that were further experimentally profiled by [136, 97].

3.3 LIHC TCGA data

For analyzing the hepatitis B interactomes in hepatocellular carcinomas (HCC), we used a liver hepatocellular carcinoma (LIHC) TCGA data subset consisting of 366 samples – for which both mutational and CNA data were available. We used the TCGA LIHC mutation data file and copy number calls on gene level provided by the Broad Institute TCGA Government Data Analytics Center (GDAC) on the Firehose portal.

We next used the Level 3 RNA-seq data to exclude genes with low expression from further analysis. We first normalized the expression levels: we divided the RNA-seq by Expectation Maximization values (RSEM [137]) by the 75th percentile of all genes RSEM values in the tumour samples and then we multiplied the resulting values 1000 times, as done by TCGA [4]. We selected genes with normalized RSEM values > 0.125 in over 50% of the samples, resulting in 16,895 expressed genes. These coverage criteria allowed us to remove genes with low-quality reads expressed in only a minority of the samples.
We obtained CNA regions from GISTIC2 on gene-level. This indicated the copy number state of the genomic region covering each gene, normalized to the average copy number state of the chromosome arm($+1$ = increased relative to the chromosome arm; 0 = same as the chromosome arm; -1 = decreased relative to the chromosome arm).
We downloaded the clinical data available on the Firehose portal. Based on the

[3] ftp://ftp.ncbi.nlm.nih.gov/hapmap/
[4] https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2

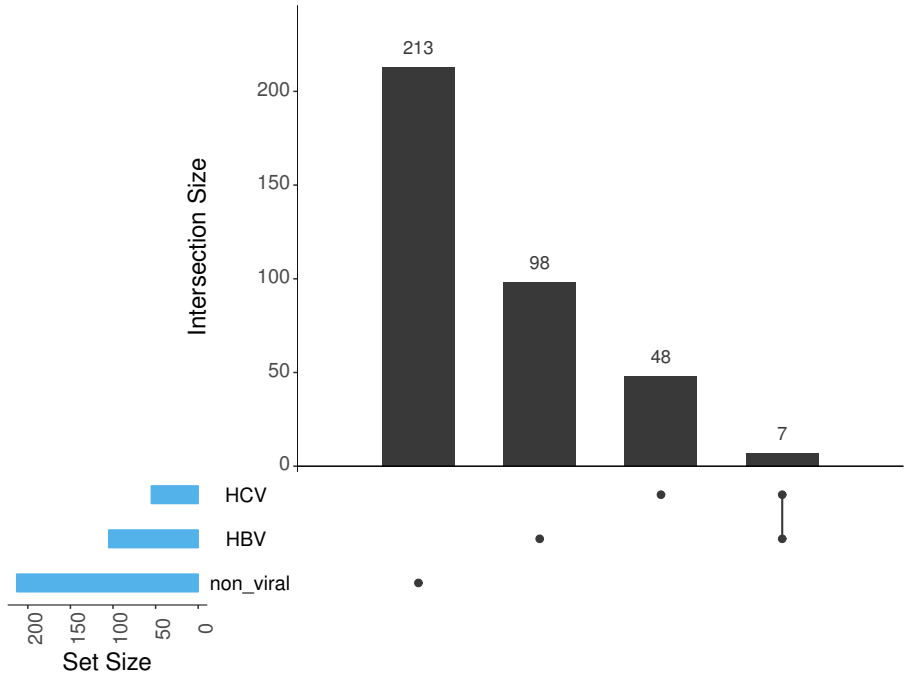

Figure 3.2: Distribution of hepatitis B and hepatitis C viral infection across 366 TCGA LIHC patients. Rows represent the sets of patients infected with Hepatitis C, Hepatitis B or non-infected, while and the columns represent the intersections of these sets.

viral status, the 366 selected samples were classified as follows: 213 patients were non-infected, 98 patients were infected with hepatitis B only, 48 patients were infected with hepatitis C only and 7 patients were infected with both hepatitis B and hepatitis C (Figure 3.2).

For a broader understanding of the hepatitis B interactomes in HCC, we used the Reactome Functional Interaction network, further referred as the ReactomeFI reference network [5]. ReactomeFI consisted of 229,300 experimentally validated and manually curated pathway relationships among 60 % of human proteins and included protein-protein interactions – interaction between protein A and protein B (low or high throughput experiments – e.g. yeast two-hybrid (Y2H), affinity purification coupled to mass spectrometry (AP-MS)), transcriptional interactions – e.g. protein A (encoded by gene A) is a transcription factor which regulates the expression of gene B, metabolic interactions – the product of enzyme A is the input of enzyme B, and many other types.

[5] https://reactome.org/

3.4 LIHC – HBV and LIHC – HCV PPI data

Our approach for learning how hepatitis B interacts with human proteins in HCC relied on combining differentially mutated protein data in HBV-associated LIHC with physical protein-protein interactions (PPIs) between viral infections and host in liver cancer cell lines.

The HBV-human protein-protein interaction network data were produced at the Krogan Lab, University of California San Francisco (UCSF). The strength of physical viral-host interactions in HCC was measured in two hepatocellular carcinoma cell lines through affinity purification - used to enhance the human proteins associated with viral proteins, followed by mass spectrometry - used to find interacting proteins (Figure **??**, joint work with Manon Eckhard, Ph.D.). To quantitatively measure each PPI, the MiST software was used [138, 139]. Hence, to every viral-host PPI, we assigned a confidence score which considered how abundant, how likely to reproduce and how specific it is the interaction between two co-purified proteins. In total 3,863 physical PPI strengths were measured. Each HBV protein physically interacted with host proteins (Figure 3.3c). The host interacting proteins enriched for several Reactome pathways and suggested disruption of essential molecular processes like transcription, endoplasmic reticulum-associated degradation (ERAD), and translation initiation (Figure 3.3)e.

We illustrated the resulting HBV-host interactome in Figure 3.3e and we could show that The AP-MS experiments were conducted by Manon Eckhard, Ph.D., and John Gordan, MD Ph.D., our collaboration partners from the Krogan Lab, UCSF.

3.5 LUAD TCGA data

For identifying miRNA–mRNA regulatory networks in LUAD, we used the TCGA publicly available data sets comprising of mRNA and miRNA expression levels. The complete data sets consist of Level 3 IlluminaHiSeq RNASeqV2 mRNA and Illumina HiSeq 2000 miRNA data across 181 samples.

We selected mRNAs with non-zero count values in more than 80% of the pa-

tients, non-zero standard deviation and we applied a log2 transformation to normalize the data. The miRNA precursor expression levels were preprocessed in the same manner. Following, 16,241 mRNAs were selected.
Since our collaboration partners – Margarita Gonzalez, Ph.D., and Kai Breuhahn, Prof. Ph. D, from the Heidelberg University were interested in distinguishing miRNAs that affect the metastasis process in lung cancer, they provided a set of preselected 24 miRNAs. The 24 miRNAs showed differential expression between NSCLC patients with and without lymph node metastasis (N1, N2 and N3 vs. N0) in a TCGA LUAD sub-cohort of n = 449 samples [140]. For each one of the 24 miRNAs, we applied the preprocessing pipeline used for the HNSCC TCGA miRNA set.

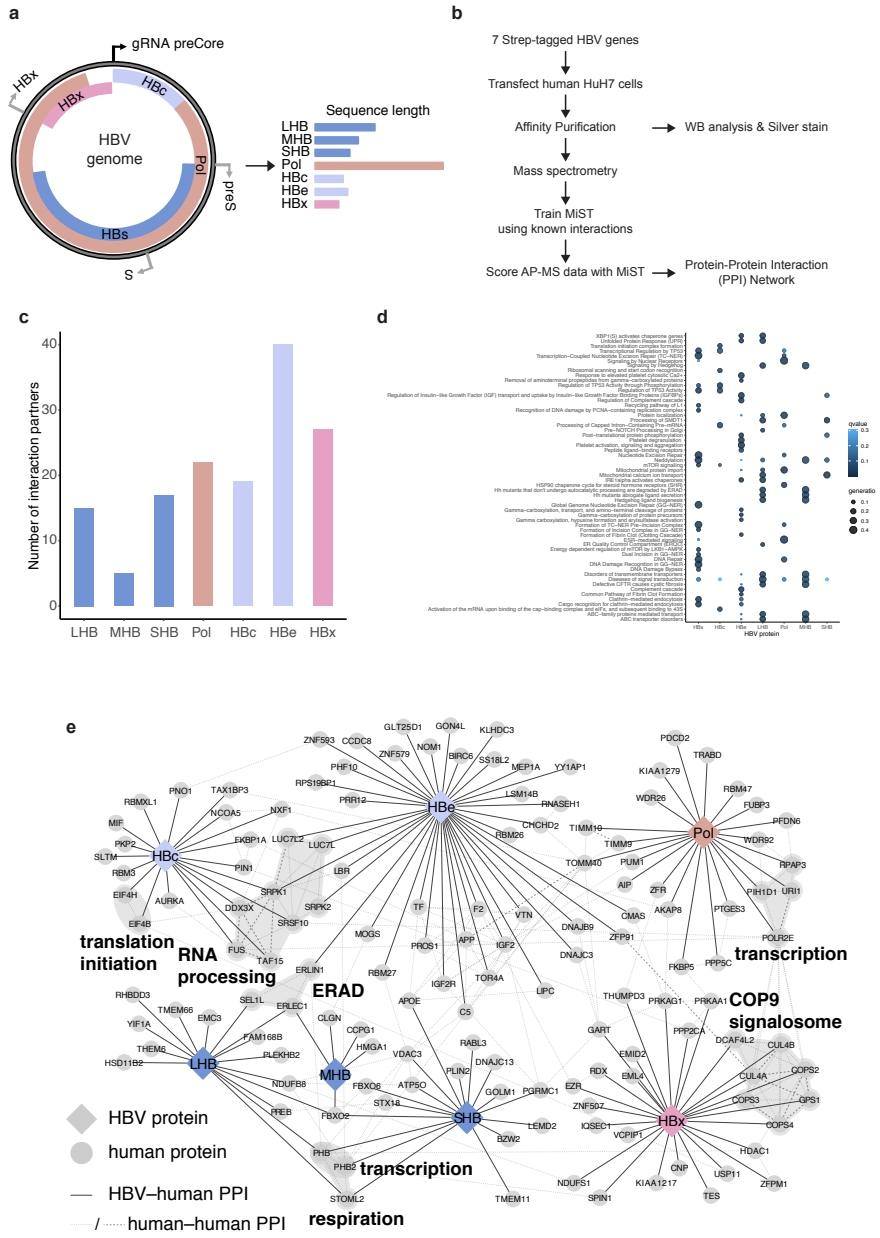

Figure 3.3: HBV-human protein-protein interaction map. (a) Representation of the HBV genome, as well as the individual viral proteins used in this study (to scale), colored accordingly to their canonical function during the viral replication cycle. (b) Summary of the experimental approach used to build the PPI map. (c) Distribution of human interactors (host proteins) for each individual HBV protein. (d) Reactome pathway enrichment of the host proteins interacting with HBV. (e) Network representation of the HBV-human PPI map in the HUH7 cell line. Diamond shaped nodes represent the seven individually expressed HBV proteins, while the circle shaped nodes represent the 140 high-confidence human interactors. The colors indicate the specific HBV proteins.

Chapter 4

Copy number aberrations detection - a benchmarking study

Predicting and characterizing oncogenic molecular phenotypes is decisive for personalized cancer medicine. A robust model for tackling these tasks requires the accurate identification of DNA copy number changes. A rigorous identification of changes in the tumor DNA enables distinguishing those CNAs that affect oncogenes or tumor suppressor genes. Following, one can provide the knowledge required for developing new targeted cancer therapies or patient stratification. Hence, CNAs play an essential role in cancer research and it is essential to assess the accuracy of CNA calling from tumor genomes.

To measure changes at the genomic level, technologies such as single nucleotide polymorphism (SNP) arrays, whole-genome sequencing (WGS), and array comparative genomic hybridization (aCGH) can be used. Of these technologies, SNP arrays come with the advantage that they can be used for both genotype and copy number analysis. Furthermore, this technology allows scientists to characterize both copy number changes and allelic imbalances of a sample. Estimating copy number changes and allelic imbalances require tailored methods that process and model the array signals [97].

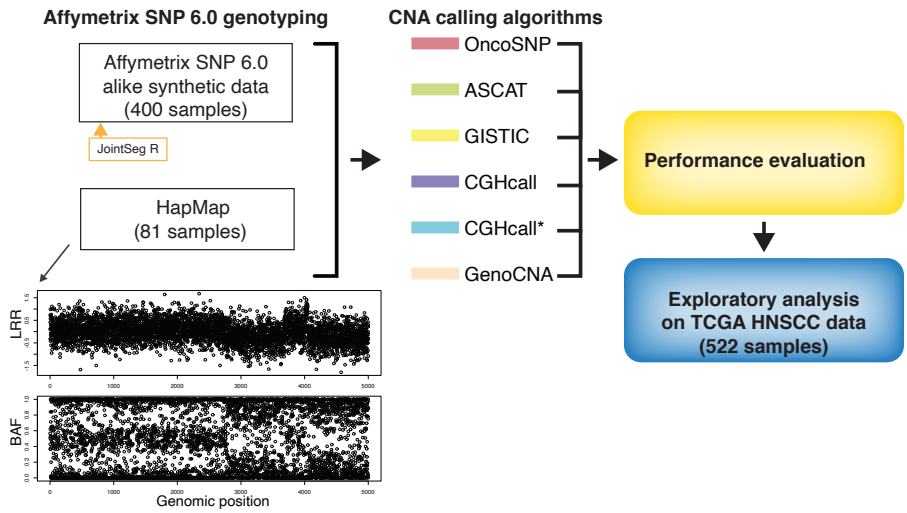

Figure 4.1: Copy number aberrations detection in cancer – a benchmarking study. The benchmarking pipeline first generates Affymetrix SNP 6.0 alike data (JointSeg R package). The performance of the algorithms was tested on both synthetic and real data (HapMap data). The benchmarking study concludes with an exploratory analysis on CNA calling from TCGA HNSCC data.

Despite the vast number of present methods, revealing cancer-related CNAs from SNP array data precisely is difficult to achieve [27, 37, 141]. One particular challenge in accurately estimating cancer-related CNAs is the presence of biological confounding variables specific to cancer, like tumor purity and length of an aberrated chromosomal segment [97].

This chapter introduces a comparative study designed to find the most suitable CNA calling algorithm from SNP array data, given the effect of cancer-specific biological confounding variables. Particularly, we test the performance of five algorithms commonly used for identifying CNAs from Affymetrix SNP 6.0 array data in the presence of three biological confounding variables.

Since the true copy number states for real cancer data are unknown and experimental validation on genome-wide level is not feasible (the human genome size is about 3.0×10^9bp and is affected by CNVs), a benchmarking study requires synthetic data mimicking Affymetrix SNP 6.0 array experiments. Synthetic data also allow us to tune confounding variables and observe how the algorithms are affected by them.

However, synthetic data are not sufficiently realistic due to simplifications of the molecular model. Realistic rendering of molecular phenotypes is challenging, as we have yet to completely uncover all the functions, connections, and interdependencies of the multiple molecular subsystems.
Since the real copy number state in tumor samples measurement is unknown, for evaluating the algorithms on real data, we used a HapMap cohort with subsequently experimentally validated CNAs genome regions.
To test the plausibility of CNA calling results in tumor samples, we examined the consistency between raw LRR signals from TCGA HNSCC samples and the predicted CNA calls overlapping the HNSCC consensus regions defined in [142].
The pipeline shows how the performance of commonly used CNA calling algorithms is altering in the presence of biological confounding variables and is available at https://github.com/adspit/PASCAL (PASCAL – Performance Assessment of CNA calling Algorithms, Figure 4.1).
This chapter is included in the following publication:

- **Adriana Pitea**, Ivan Kondofersky, Steffen Sass, Fabian J. Theis, Nikola S. Mueller and Kristian Unger. Copy number aberrations from Affymetrix SNP 6.0 genotyping data - how accurate are the commonly used prediction approaches? *Briefings in Bioinformatics*, 2018.

The study, text, and figures included in this publication represent entirely my work, with minor corrections from co-authors. The figures included in this chapter served as basis for the figures included in Pitea et al.

4.1 Biological confounding variables in CNA finding

A major difficulty in accurately identifying cancer-related CNAs is the effect of cancer-specific biological confounding variables [143, 144]. Among the variables specific to tumor samples, tumor purity, and the length of an aberrated chromosomal segmented are known to affect the prediction of copy number states.
The tumor purity represents the ratio between cancerous cells and all the cells present in a tumor sample – consisting of both cancerous and non-cancerous

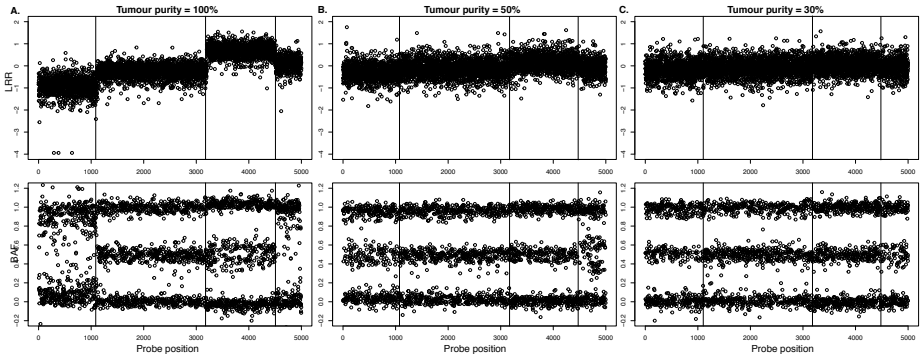

Figure 4.2: Effect of tumour purity on Affymetrix SNP 6.0 LRR and BAF signals. The left and middle panels depict the effect of contamination with 70 % and 50% non-cancerous cells on the LRR and BAF signals in regions with the same copy number states as in the right panel. The first row represents the LRR siganls, while the second represents the BAF. The x-axis represents the SNP probe position.

cells. The mixture of cancerous and non-cancerous cells affects the expected allelic fraction between germline and somatic variants: the intensity of the bulk measured signals is reduced by signals of the non-cancerous cells present in the sample (Figure 4.2). We demonstrate this behavior on three synthetic samples with the same copy number states and the same breakpoint positions. In panel B and C from figure 4.2 we observe that as the contamination of the sample increases, the LRR signals shrink towards 0. LRR values of 0 correspond to non-cancerous cells. In the 100% pure tumor sample, The BAF signals are scattered between 0 and 1 in aberrated regions indicating loss of heterozygosity. With 50% and 70% contamination in the sample, the BAF signals shift towards 0.5 – which indicates the heterozygous state.

If the algorithms require a certain signal intensity for classifying a genomic region as aberrated, the presence of non-cancerous cells can lead to missed CNA regions. In simple terms, the higher the non-tumor cell content within the assessed tissue sample, the lower the sensitivity of the copy number calling algorithm gets.

Another confounding variable shown to affect the sensitivity of CNA calling algorithms is the length of a CNA region, where a CNA region is a chromosomal segment for which the genetic markers present the same copy number state [97]. Longer CNA regions are easier to find [145, 146].

Finally, we observed another variable that influenced the performance of the CNA calling algorithms: the CNA burden, which represents the percentage of aberrated regions in a tumor sample. The CNA burden affects the number of genes mutated and thus the molecular phenotype of tumors.

4.2 Synthetic data

As the true states of genomic copy number data in cancer are unknown and experimental validation on genome-wide level is not feasible (the human genome is approximately 3.0×10^9bp and is affected by SNVs and CNVs), we generated synthetic data mimicking Affymetrix SNP 6.0 biological data.
We generated Affymetrix SNP 6.0 array-like synthetic tumor signals for 400 samples by using the jointseg R package [147, 97]. To make the samples as similar as possible to the Affymetrix SNP 6.0 array samples, we simulated data for 1.844.399 markers of genetic variation, comparable to the number of probes included in an Affymetrix SNP 6.0 array.

Jointseg was built to generate realistic synthetic DNA copy number profiles. The framework resamples signals corresponding to genomic regions with manually annotated copy number states from the publicly available lung cancer NCI-H1395 SNP microarray data [147, 148]. For analyzing the effect of tumor purity on the performance of the CNA calling algorithms, we generated 100 samples with each of the following tumor purity levels: 30%, 50%, 70%, and 100%. The tumor purity levels corresponded to the experimental settings of the Rasmussen et al study [148].

For each sample, the number of breakpoints ranged from 1 to 8, where a breakpoint represents a locus where one of the two parental copy number changes. The breakpoints were randomly placed in the simulated profiles. This setting allowed us to simulate samples with different ranges of CNA region lengths.

For the resulting regions we sampled the copy number states from a pre-defined set: (0,1), (0,2), (1,1), (1,2), (1,3), (2,2), (3,2), where (0,1) represented the loss

of a single copy, (0,2) and (1,1) represented normal, and (1,2), (1,3), (2,2) and (3,2) represented the gain of one, two or three copies.

4.3 Genomic copy number calling algorithms

We selected five commonly used copy number calling algorithms for comparison: CGHCall, OncoSNP (version 2.1), ASCAT (version 2.4), GenoCNA, and GISTIC (version 2.0). Additionally, we developed and included CGHcall* – an adjusted version of CGHcall that prevents the shift of profiles

OncoSNP

OncoSNP labels SNP array signals from cancer genomes based on 21 states dictionary that includes multiple arrangements of allele losses and amplifications [97]. The model accounts for the effects of tumor purity, polyploidy, and intratumor heterogeneity [149]. We applied OncoSNP on the synthetic data with the arguments specific for Affymetrix SNP array, together with the predefined number of training states and tumor states. We used the intratumor mode and set the tumor purity parameter to 30%, 50%, 70%, and 100% .

ASCAT

ASCAT was designed to analyze allele-specific copy number in tumor samples. The algorithm corrects for the effects of tumor purity and tumor aneuploidy and infers copy number classes, loss of heterozygosity, and homozygous deletions. However, the algorithm requires a threshold for the segmenting of the SNP profile into regions that have the same copy number states. ASCAT estimates the number of copies for both alleles at all SNP marker positions [150]. For our study, we preprocessed the synthetic data and generated the ASCAT–format input tumor LRR and BAF files. Next, we used the ascat.predictGermlineGenotypes R function with the platform parameters set to "AffySNP6" to generate germline genotype profiles [97]. Finally, we segmented the data with the ASPCF segmentation algorithm (default parameters) and applied the ASCAT copy number calling function [97].

GenoCNA

GenoCN performs simultaneous searches for CNAs and CNVs while correcting for tumor purity [97]. However, the framework does not account for a chromosomal background that is non-diploid [151].
For our benchmark, we performed a GenoCNA search given the synthetic data sample files and the human genome assembly hg18 genetic marker information [97]. We chose the output format 2 that included the most probable copy number and genotype state for all the genetic markers.

GISTIC

GISTIC represents the standard CNA calling algorithm used to estimate copy number changes from Affymetrix SNP 6.0 arrays in TCGA studies [152]. GISTIC was designed to find genomic regions that are significantly amplified or deleted across a set of samples, and not on a single patient level. Following, GISTIC eliminates common chromosome arm-level events that are unspecific to cancer and retains the focal events based on a significance measure. The significance measure relies on the amplitude of the CNA, on how frequently the CNA occurs across samples, and a user-defined threshold for the discovery rate. GISTIC required as input a segmentation file and a reference genome file.
One disadvantage of GISTIC is the fact that the algorithm does not correct for the effect of the biological confounding variables.
In our study, we first segmented the samples by using the segment function of the CGHcall R package, and then we applied the GISTIC algorithm.

CGHcall

Just like GISTIC, CGHcall uses breakpoint information from circular binary segmentation [153]. However, CGHcall processes raw log2-ratios between reference and tumor DNA and estimates their belonging to one of the next five copy number states: double loss – homozygous (biallelic) deletion, loss – hemizygous deletion (loss of one of the alleles), normal – two copies, gain – three to four copies, and amplification – more than four copies [154, 97].
For analyzing the synthetic data, we log-transformed the total copy numbers and we applied the CGHcall pipeline on the resulting signals with adjustment

for the corresponding tumor purity. The HNSCC TCGA data set comprised of normal-tumor matched patient samples. We calculated the $\log_2$-ratios between the tumor and the normal matched patient samples by using a Python script [97]. As the HapMap data consist of samples collected from healthy patients, we calculated log2-ratios between each LRR signal and the mean LRR signals of the 81 selected samples [97].

4.4 Performance analysis of genomic copy number calling algorithms

In addition to realistic synthetic data, benchmarking studies require a suitable performance metric for copy number calling algorithms. In general, to show how prediction algorithms perform, receiver operating characteristics (ROC) curves are used [155]. However, when the distribution of the classes is imbalanced, as in our case (Figure 4.3), ROC curves can present an over-optimistic view on how an algorithm performs, while the recall and the precision have been shown to give a more informative view [156, 157]. Since the F-score represents the balance between the precision and the recall of an algorithm, we selected it as a suitable metric to evaluate the performance of the copy number algorithms for each class.

We were interested whether the algorithms can classify correctly the log R ratio (LRR) and the B allele frequency (BAF) signals into three states: loss, normal, and gain. Therefore, we split the multiclass classification problem into three binary classification problems and we collapsed the resulting calls of each of the algorithms to loss, normal and gain. For CGHcall and GISTIC, we unified losses and double losses into one loss class and the gains and amplifications into one gain class [97]. For OncoSNP, the homozygous and the hemizygous deletion states were collapsed to loss, and all the states that were defined by more than two copies were considered gain [97]. For ASCAT and GenoCNA, probes with a number of copies lower represented a loss, while the probes a number of copies higher than two copies represented a gain [97].

We calculated the sample-wise confusion matrix, precision, recall and balanced

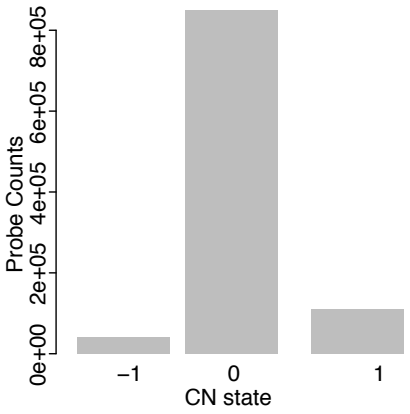

Figure 4.3: Example of the class imbalance present in one sample. On the y axis we observe the number of probes from a simulated Affymetrix SNP 6.0 array-alike sample covered by each of the three CN states: -1 (loss), 0 (normal) and 1 (gain).

F-score [158] as follows:

$$\text{precision}_c = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.1}$$

$$\text{recall}_c = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.2}$$

$$F_c = 2 \cdot \frac{\text{precision}_c \cdot \text{recall}_c}{\text{precision}_c + \text{recall}_c}, \tag{4.3}$$

where c represented the class: loss, normal, or gain. True positives (TP) represent the number of probes that were classified correctly for each class c, while false positives (FP) are the probes classified incorrectly as class c. The false negatives (FN) comprised of the total of probes that initially were classified as belonging to group c but were classified as belonging to another group. To test for statistically significant shifts between F-score distributions of the algorithms, we performed non-parametric pairwise comparison Wilcoxon tests [159]. We adjusted the resulting p-values for multiple testing error through Bonferroni correction [160].

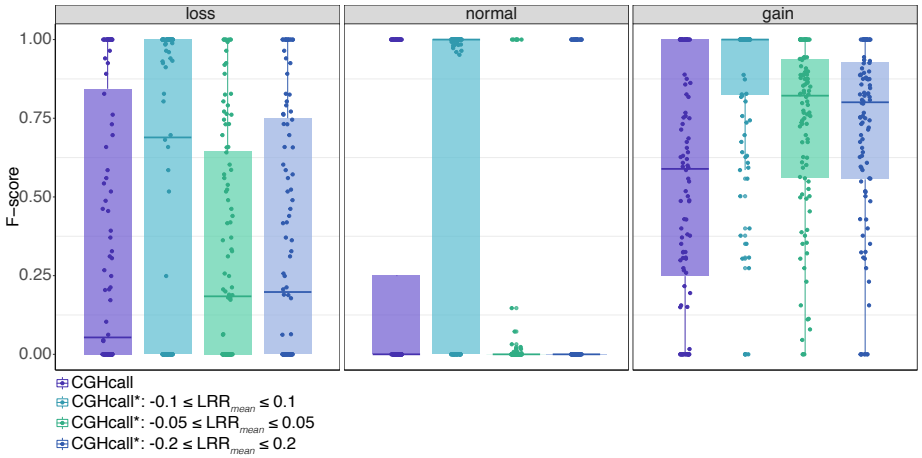

Figure 4.4: F-scores for different filtering criteria for the LRR signals considered for the mean value of the CGHcall normalizationa and postsegmentation normalization. On the y axis we observe the F-scores. Each panel represents a CN state and includes the corresponding scores for CGHcall and the three adjusted versions.

4.5 An improved algorithm for CNA calling from Affymetrix SNP 6.0 data: CGHcall*

During a manual inspection of the CGHcall pipeline, we observed that all the normalized log2 signals before and after segmentation in the synthetic samples with more 50% non-normal states covering the profiles, were incorrectly shifted (either to -1, either to 1). The high ratio of aberrated regions caused the algorithm to assign a faulty baseline level. In consequence, CGHcall returned inaccurate estimates for copy numbers states [97]. Since cases in which more than half of the genotyped probes are in a non-normal state have been reported in a pan-cancer study on somatic genomic CNAs [37], we developed a model to adjust for the CNA burden effect [97].

The problem arose from the LRR levels being normalized to the median level over a sample. If more than half of the genomic profiles were lost or gained, CGHcall was unable to correctly estimate the baseline level and assigned the 0 level to loss or gain. We observed the same behavior when we applied the post-segmentation normalization – which assigns the baseline segment to a segment that is either lost or gained.

To correct for this effect, we selected three different intervals as constrains for

the LRR signals: [-0.1, 0.1], [-0.05, 0.05], and [-0.2, 0.2] and analyzed how the performance of the algorithms changed. To eliminate the effect of tumor purity, we performed this analysis on samples with 100% tumor purity. The resulting F-scores suggested that the LRR signals within the [-0.1, 0.1] interval provided the optimal mean for normalization and post-segmentation normalization (Figure 4.4). Accordingly, we developed a model that normalized the LRR signals using as mean the LRR signals included in the [-0.1, 0.1] interval. We applied the same strategy for the post-segmentation normalization step.

4.6 Tumor purity strongly influenced the performance of CNA calling algorithms

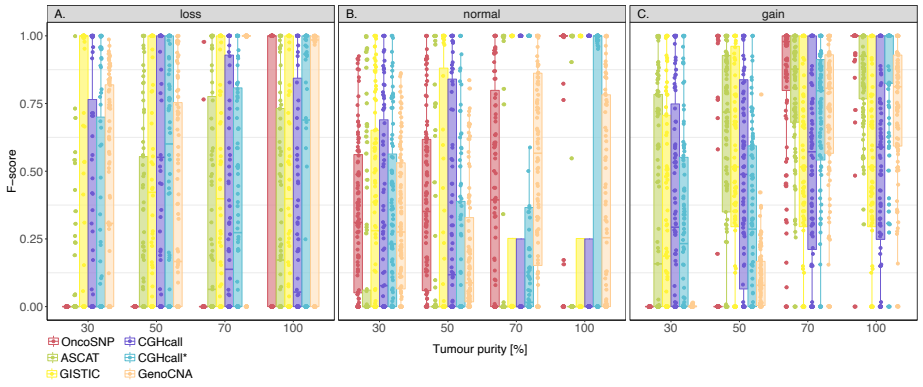

Figure 4.5: **Performance of calling algorithms on synthetic data across different tumor purities.** We assessed the performance of the following algorithms: OncoSNP - coral red, ASCAT - light green, GISTIC - yellow, CGHcall - purple, CGHcall* - cyan and GenoCNA - pale pink. The y-axis indicated the F-score, while the x-axis indicated the tumor purity level in %. The three different classes were depicted in the three different panels: A. loss, B. normal and C. gain. Each boxplot comprised of the F-scores corresponding to the 100 synthetic samples. This figure served as component for Figure 1. in the Pitea et al. publication [97].

During the first phase of our benchmarking, we evaluated how the algorithms performed on synthetic data based on their F-score distributions (Figure 4.5). Figure 4.5A. depicted how accurately the six algorithms estimated losses at tumor purity levels (depicted on the x-axis) varying from 30% to 100%. OncoSNP

could not predict losses in impure tumor samples (mean F-score = 0.03). Similarly, ASCAT, GISTIC, and CGHcall performed poorly when predicting losses regardless of the tumor purity level (mean ASCAT F-score = 0.26, mean GISTIC F-score = 0.34, mean CGHcall F-score = 0.39). Figure 4.5)A. also suggested that CGHcall* and GenoCNA performed competently in samples with 100% tumor purity. GenoCNA returned admissible predictions for losses in samples with tumor purities > 50% (mean F-score = 0.68).

OncoSNP showed increasing performance for calling normal states as the tumor purity level increased (Figure 4.5B). We hypothesized the presence of normal DNA drove the log2 ratios towards the 0 baseline and influenced the performance of OncoSNP. Besides, the normal state represented the majority class. Thus, the results indicated that OncoSNP can not handle the class imbalance problem. Except for CGHcall*, the algorithms performed rather poorly when predicting normal states (mean GISTIC F-score = 0.28, mean CGHcall F-score = 0.29, mean GenoCNA F-score = 0.35, mean ASCAT F-score = 0.36). CGHCall* performed well overall but especially in samples with tumor purity 100% (mean F-score = 0.70, 4.5B).

Finally, we examined how well the algorithms predicted gains. We observed that OncoSNP performed well for samples tumor purity was > 50% (4.2C), while ASCAT correctly predicted gains in samples with tumor purities > 30% (mean F-score = 0.76). All algorithms improved their performance for calling gains as tumor purity increased.

Our aggregated results showed that CGHcall* strengthened the prediction of all copy number states for all tumor purities relative to CGHcall [97]. Moreover, our results also indicated that GISTIC and CGHcall performed similarly [97]. One aspect that can explain the similarity in performance is that both methods use CBS segmentation results and do not exploit the information contained by the B allele frequency [97]. The benchmark hinted that CGHcall* and OncoSNP are the better performing algorithms in samples with high tumor purities regardless of the class of the DNA change [97]. The key message of this analysis is that tumor purity strongly influences the results of the CNA calling algorithms [97]. Our finding provides valuable information for the scientific community, especially for cohort including samples with tumor purities markedly below 50%

[97].

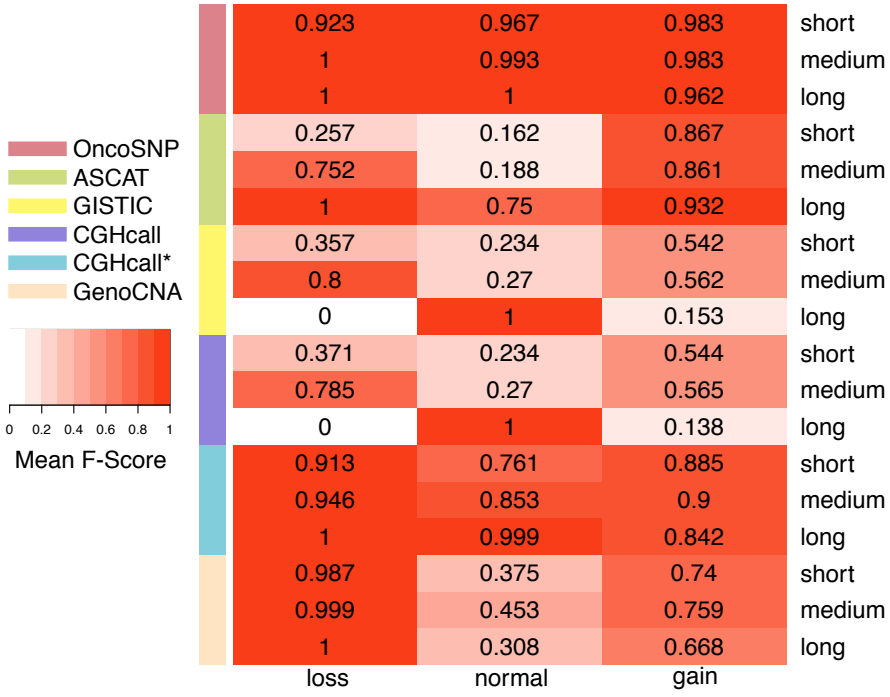

Figure 4.6: **Performance of benchmarked copy number calling algorithms in synthetic data across different lengths of genomic regions.** The columns indicated the three copy number states: loss, normal and gain, while the rows indicated the different ranges of genomic region lengths.

4.7 The effect of copy number region length

Next, we aimed to comprehend how the algorithm performed relative to the effect of the length of a copy number region. Therefore, we assessed the predictions of the algorithms for different region lengths: $\leq 10^5$probes (short), between 10^5 and 10^6 probes (medium), and $> 10^6$ (long) (Figure 4.6). To remove a combined effect with tumor purity, we examined only samples with 100% tumor purity. We defined the region length was as the number of genetic markers with the same copy number state overlapping a chromosomal segment. A chromosomal segment overlapped between 3 kilobase pairs (kbp) and 1.8 million base pairs (Mbp).

The results shown in Figure 4.6 indicated that OncoSNP, GenoCNA, and CGH-

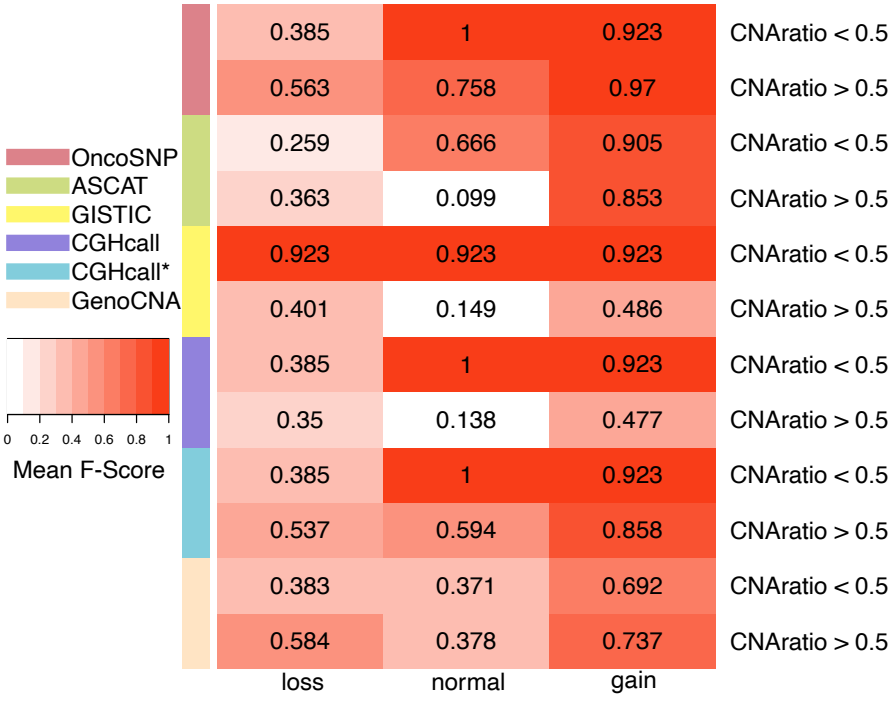

Figure 4.7: **Performance of the benchmarked copy number calling algorithms for samples with a CNA burden smaller or higher than 0.5.** The columns indicated the three copy number states: loss, normal and gain, while the rows indicated the CNA burden ratio.

call* were barely sensitive to region length. The heat map in Figure 4.6 showed that CGHcall* and OncoSNP performed well regardless of the class, while GenoCNA floundered when predicting normal genomic regions. For ASCAT, we experienced a decline in performance for short and medium-length CNA regions. GISTIC missed predicting losses or gains, regardless of the length of the region. CGHcall displayed a comparable behavior. Once again, the aspect that may be affecting CGHcall and GISTIC is the CBS algorithm.

CGHcall* and OncoSNP were again our top performers regardless of the region length and the copy number state.

4.8 The effect of CNA burden

Given that the CNA burden influenced the CGHcall normalization of the log2 ratios, we tested whether it also affected the prediction of the other copy number calling algorithms.

Accordingly, we examined the mean F-scores for samples with a CNA burden $>$ 50% and samples with CNA burden $<$ 50% for each copy number class (Figure 4.7). CGHcall and GISTIC performed poorly for samples with a CNA burden $>$ 50%. However, GISTIC improved its performance in samples with a CNA burden $<$ 50%. ASCAT returned low performance for the normal state and samples with a high CNA burden.CGHcall* outperformed CGHcall in samples with CNA burden $>$ 50% and verified our revision of the initial pipeline, notably for normal and gain classes.

Generally, OncoSNP and CGHcall* remained the best performing algorithms assessed in this study, independent of the CNA burden.

4.9 Performance of the copy number calling algorithms on SNP 6.0 array profiles of healthy patients (HapMap)

Assessing how OncoSNP, ASCAT, CGHcall, CGHcall*, GenoCNA, and GISTIC perform on real data required knowing the true copy number states of the genome positions – i.e., having a gold standard. Due to the human genome size – 3.0×10^9 bp, the scientific community has yet to provide an Affymetrix SNP 6.0 array gold standard. The HapMap project comprehensively experimentally validated DNA copy number changes estimated from Affymetrix SNP 6.0 arrays in a cohort of healthy patients. Based on that, we designed a benchmarking approach in which we used the copy number profiles annotated by Redon et al. as our golden standard [136]. In total, Redon et al. provided experimental validation for 14,500 genomic regions.

The distribution of F-scores from Figure 4.8 showed that OncoSNP, ASCAT, CGHcall, CGHcall*, and GenoCNA performed a fairly accurate prediction for the normal class (mean OncoSNP F-score = 0.91, mean ASCAT F-score = 0.85,

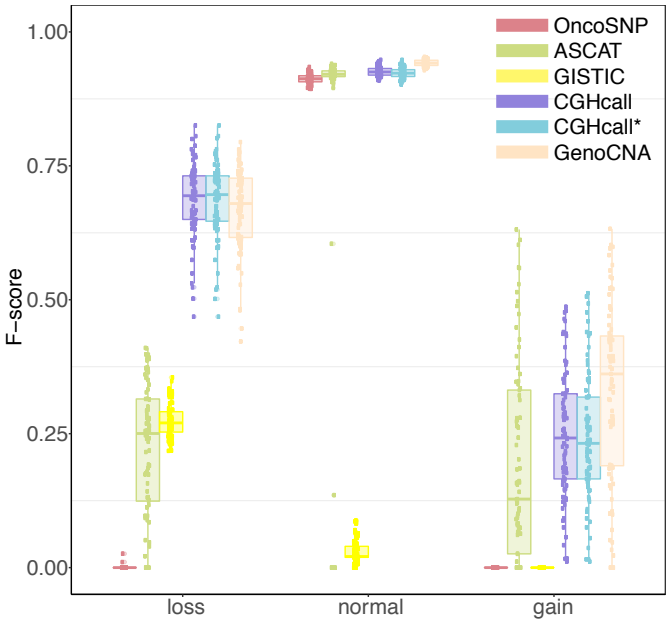

Figure 4.8: Distribution of F-scores for OncoSNP, ASCAT, CGHcall, CGHcall*, GenoCNA and GISTIC in 81 healthy HapMap subjects.

mean CGHcall F-score = 0.92, mean CGHcall* F-score = 0.92, mean GenoCNA F-score = 0.94). However, GISTIC only made poor predictions for 381 regions overlapping our ground truth (mean F-score = 0.10).

None of the algorithms returned reasonable predictions for the gain class. However, CGHcall, GenoCNA, and CGHcall* performed reasonably well for the loss class (mean F-score $\approx$ 0.69).

Although our results seem pessimistic, one should consider that ASCAT, OncoSNP, and GISTIC aim to find somatic CNAs in tumor data, and HapMap provides profiles of healthy human blood samples. However, we theorized that germline DNA changes would be easier to identify.

In conclusion, the HapMap project allowed us to assess the CNA calling algorithm performance on real data and determine how applying them on non-tumor data affects them.

4.10 CNAs in HNSCC patients

Lastly, we investigated two other aspects: the agreement between raw LRR signals and predictions and the consistency of predictions for consensus CNA regions of HNSCC samples. We used the consensus CNA regions characterized by Gollin et al. and selected the overlapping genes, SNPs, and CNVs [142]. Next, we examined the raw LRR signals together with the CNA predictions in two genes known to be involved in HNSCC: CCND1 and CDKN2A (Fig. 4.9, Fig. 4.10). Both Figure 4.9 and 4.10 showed a consensus between high raw LRRs and the predicted gains. Analogously, the benchmarked algorithms estimated that the genomic regions with low raw LRR values overlapping the two genes represented losses.

Additionally, the frequencies of CCND1 gains predicted by the algorithms were comparable to the frequencies of CCND1 gains reported by Golin et al. from CGH data – 32%: CGHcall - 26.5%, CGHcall* - 24.9%, OncoSNP - 44%, and GISTIC - 43% [142]. CGHcall, CGHcall*, OncoSNP, and GISTIC predicted similar frequencies of CDKN2A losses: CGHcall - 39.8%, CGHcall* - 35.4%, and GISTIC - 59%. Generally, tumor purity was $> 60\%$, but the data set also included samples with tumor purity as low as 27.9%.

Our results suggested that given realistic tumor purity levels, CGHcall* and OncoSNP remained consistent and performed similarly in the TCGA HNSCC data.

4.11 Conclusions

This chapter described a benchmarking study on five CNA calling algorithms from Affymetrix SNP 6.0 array data: OncoSNP, ASCAT, GISTIC, CGHcall and GenoCNA. We chose to benchmark these algorithms because the scientific community often uses them to estimate copy number states in tumor samples. Except for GISTIC, all the algorithms employed adjustment for tumor purity, intra-tumor heterogeneity, and even tumor cell ploidy (ASCAT and OncoSNP).(ASCAT and OncoSNP).

Another reason for conducting this study was the lack of cancer-specificity in

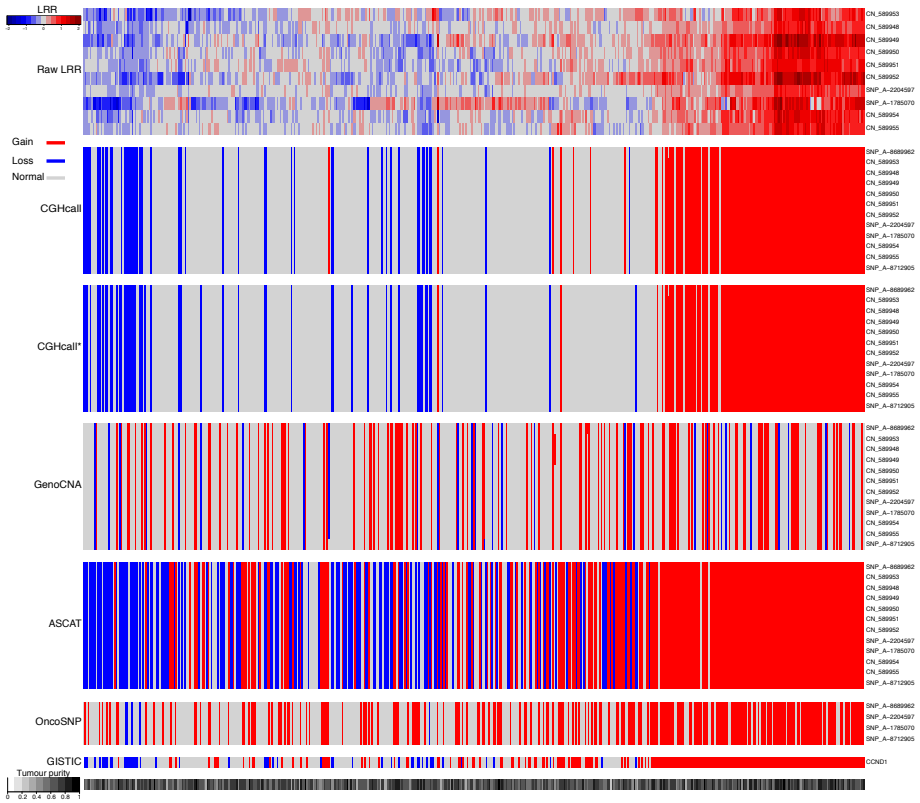

Figure 4.9: **Consensus of raw data and algorithm predictions in TCGA HNSCC CCND1.** The columns indicated the patients clustered by raw LRR signals in the probes overlapping the CCND1 genomic region. The rows indicated the Affymetrix SNP 6.0 probes that overlapped the CCND1 region. The heat map bar indicated the tumor purity of each sample.

previous studies [146, 161], or the complexity and feasibility of the model that generated synthetic data [162, 147].

Our benchmark introduced a pipeline for CNA calling algorithms that used realistic synthetic data, which accounted for cancer-specific confounding variables. Our overall benchmark results indicated that tumor purity and CNA burden significantly influence CNA calling algorithm results. Furthermore, our study allowed us to recognize a weakness of CGHcall and provide an adjusted version – CGHcall*, that corrects for the CNA burden effect.and to validate the consensus between predicted CNA regions in the TCGA HNSCC cohort and the previous finding of Redon et al.

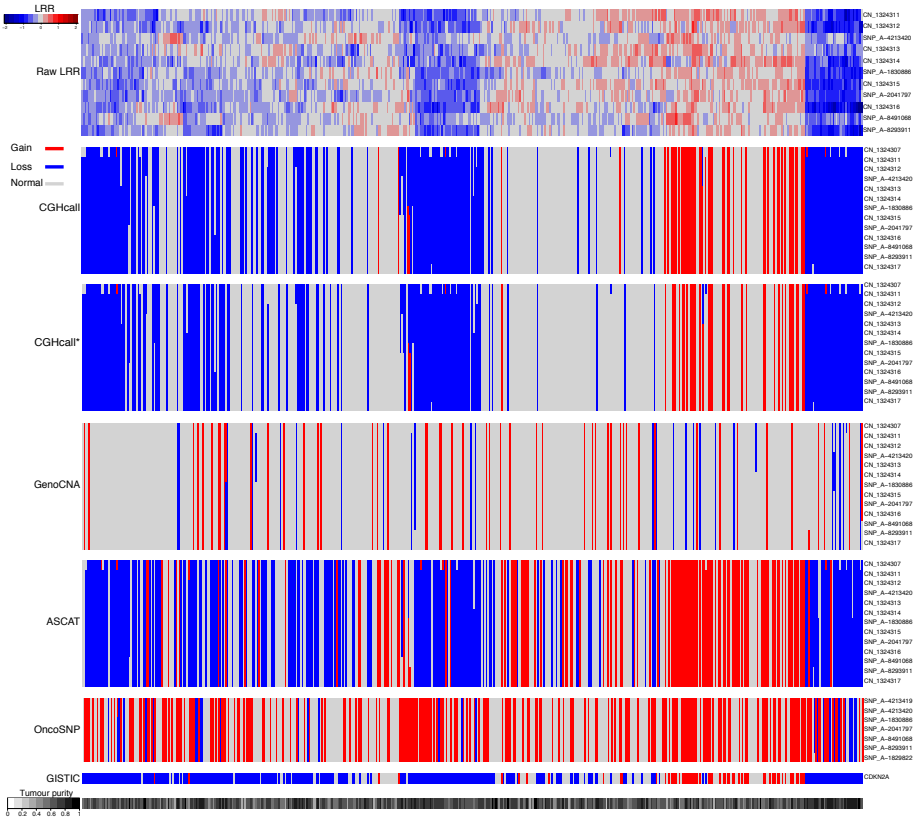

Figure 4.10: **Consensus of raw data and algorithm predictions in TCGA HNSCC CDKN2A.** The columns indicated the patients clustered by raw LRR signals in the probes overlapping the CDKN2A genomic region. The rows indicated the Affymetrix SNP 6.0 probes that overlapped the CDKN2A region. The heat map bar indicated the tumor purity of each sample.

Chapter 5

Hepatitis B alters host protein-protein interaction networks of liver cancer

Although some human diseases, such as Huntington's disease, cystic fibrosis, and fragile X syndrome, are caused by mutations in a single gene, the majority of disorders, including cancer, result from interactions between multiple gene products and environmental factors. Viruses are one of the leading environmental factors that contribute to cancer. Specifically, more than 15% of human cancers are attributed to viral infections [163].

Viruses are small infectious entities that consist of a core of DNA or RNA surrounded by a protein coat [164]. Viruses can lead to cancer directly – the cell machinery is disrupted by viral protein expression, or indirectly – when the virus integrates itself in the DNA of the host cell [165]. They target host genes involved in immunocompetence pathways, triggering chronic infection and inflammation [165]. Subsequently, viruses induce carcinogenic mutations and transform the host cells [165]. We addressed this aspect by using a network propagation approach that allowed us to estimate the strength of the viral hit within neighborhoods of the human protein-protein interaction network.

Here, we developed a network-based integrative strategy combining viral-human

physical interactions with genomic alterations of tumors to characterize the oncogenic viral impact on host proteins. The computational framework consists of two integrative analyses:

- an integrative analysis of transcriptomics, genomics, and clinical data for determining the effect of a viral infection on the mutational landscape of tumors

- a network propagation-based approach using genomic and physical interactions between viral and human proteins to assess the significance of oncogenic interactions between virus and host.

Our approach made use of the network propagation concept (Chapter 2) that allowed us to discover not only the direct interactions between virus and host but also the interactions that were not experimentally assessed or observed in the mutation rates. Finding these high-confidence interactions allowed us to discover relevant players involved in viral-mediated oncogenic pathways.

This project aimed to reveal how hepatitis B (HBV) proteins interact with host proteins in hepatocellular carcinomas (HCC). The integrative analysis of transcriptomics, genomics, and clinical data provided an estimated effect of viral infections on the HCC mutational landscape. Explicitly, we determined genes in which genetic alterations are dependent on the HBV status in HCC tumors. The integrative analysis of genomic and physical interactions between viral and human proteins revealed genes differentially mutated in viral cancers. Following, we identified both known and novel oncogenic interactions associated with the viral infections.

The project aim was defined together with Wei Zhang, PhD. from the Ideker Lab (UCSD) and John Gordan, MD PhD., and Manon Eckhardt, PhD., from the Krogan Lab (UCSF).

The analysis and modeling of the data represent my work entirely. The design of the approach represents joint work with Dr. Wei Zhang, Ideker Lab. My contribution also comprised implementing the method, conducting the statistical analysis, and visualizing the data and the results of the computational analysis. The figures included in this chapter represent my work optimized after receiving feedback from my collaborators and my UCSD visit supervisor, Trey Ideker,

Ph.D.
The interpretation of the computational results represents my work, while the interpretation of the biological results represents join work with John Gordan. The analyses and the materials presented in this chapter will be fundamental for the manuscript in preparation.

5.1 Viral mutational landscapes in HCC

Hepatocellular carcinoma represents the second leading cause of cancer death worldwide [166], with increasing incidence [167]. Despite the comprehensive genomic profiling of HCC in the LIHC data set, few actionable molecular targets have emerged [168]. HCC typically arises in the context of co-morbid hepatitis due to HBV or Hepatitis C (HCV) infections or non-alcoholic fatty liver disease. HBV and HCV represent the primary causes of HCC worldwide [86]. Given all these reasons, we investigated the oncogenic effects of HBV in HCC by integrating transcriptomics, genomics, and clinical data from the liver hepatocellular carcinoma TCGA data set (further referred to as LIHC).

5.1.1 Mutated genes in the TCGA liver cancer data

To determine altered protein-coding genes from the LIHC data set, we used the corresponding mutation files, and copy number calls on gene-level as provided by the Broad Institute TCGA GDAC (Chapter 2). The mutation annotation file comprised 53,777 missense mutations in 14,901 RNAs and 373 patients, as determined by Mutation Assessor [169]. We classified genes as altered (Mut) or wild type (WT) as follows. Since we are interested in non-silent mutations, we removed silent mutations, i.e., we removed variants classified as 'Silent', 'IGR', '5'UTR', '3'UTR', '5'Flank', '3'Flank', 'RNA', 'Intron'}. Following, we selected 41,263 non-silent mutations across 13,675 RNAs. Additionally, we considered mutated genes overlapping amplifications or deletions as determined by GISTIC.
The interactions between viral and human proteins can lead to cancer. Thus, we further focused on DNA changes that affect protein-coding genes: we inter-

sected the RNAs impacted either by somatic mutations either by CNAs with the protein-coding genes included in the ReactomeFI PPI reference network (https://reactome.org/).

As a result, we determined $m = 8,765$ protein-coding genes altered by mutations, amplifications, or deletions, in a set of $n = 366$ patients. For the following analysis, we binarized this information for each gene: {0=WT, 1=Mut}.

5.1.2 Differential mutation analysis revealed significant HBV impact on 46 genes

To determine the effect of the viral infections on the mutational status of each gene in cancer, we assessed the differential mutation rates at gene-level between:

A. HCV(+) and HCV(-) HCC cases.

B. HBV(+) and HBV(-) HCC cases.

For this purpose, we set up to map the inputs $x_{g_{hepC}}, x_{g_{hepB}}$ to the output y_g, where $g \in \{g_1, ..., g_m\}$. The output y_g is a one dimensional (1d) vector of length n representing the mutational status across the n HCC patients for each mutated gene g (0 = wild type; 1 = altered). The features $x_{g_{hepC}}$ and $x_{g_{hepB}}$ are 1d binary vectors of length n representing the viral infection status for HCV (1 = HCV(+); 0 = HCV(-)) – $x_{g_{hepC}}$, and the viral infection status for HBV (1 = HBV(+); 0 = HBV(-)) – $x_{g_{hepB}}$.

Given the binary nature of the response variable y_g and our aim to learn the dependence between mutation status and viral infections, the first choice to formalize the problem was logistic regression. However, logistic regression returned perfect separation of the response, which is a common problem in imbalanced small sample size studies. Unstable regression coefficients accompany perfect separation. Since we aimed to estimate the risk of a mutation happening due to viral infection and not solve a binary classification problem, we used the solution proposed by Gelman et al. [170] to obtain stable regression coefficients.

Following, we formally defined three Bayesian logistic regression models conditioned by independent Student-t prior distributions on the coefficients for each

g out of the m mutated protein-coding genes:

$$\pi_{g_{complete}} : p(y_g|x_{g_{hepC}}, x_{g_{hepB}}) \tag{5.1}$$

$$\pi_{g_{null_1}} : p(y_g|x_{g_{hepC}}) \tag{5.2}$$

$$\pi_{g_{null_2}} : p(y_g|x_{g_{hepB}}), \tag{5.3}$$

where $\pi_{g_{complete}}$ is defined by the probability mass function of the output y_g given $x_{g_{hepC}}$ and $x_{g_{hepB}}$, $\pi_{g_{null_1}}$ is defined by the probability mass function of the output y_g given $x_{g_{hepC}}$, and $\pi_{g_{null_2}}$ is defined by the probability mass function of the output y_g given $x_{g_{hepB}}$ only. We used the default Cauchy distribution with mean 0 and prior scale 2.5 – in the simplest scenario, "a longer-tailed version of the distribution attained by assuming one-half additional success and one-half additional failure in logistic regression" [170].

To examine the effect of a specific viral infection on the mutational status of HCC tumors, we compared the likelihood of the $\pi_{g_{complete}}$ model to the likelihood of each of the two alternative models. Hence, we calculated the deviances between the complete model and each of the two alternative models:

$$D_{g_{hepB}} = -2\ln\left(\frac{\mathcal{L}_{g_{null_1}}}{\mathcal{L}_{g_{complete}}}\right) \tag{5.4}$$

$$D_{g_{hepC}} = -2\ln\left(\frac{\mathcal{L}_{g_{null_2}}}{\mathcal{L}_{g_{complete}}}\right) \tag{5.5}$$

with $\mathcal{L}$ being the maximum likelihood, i.e., the probability of the data given the inputs x_g and the parameter vector θ that maximizes $p(y_g|\theta, x_g) = \prod_{i=1}^{n} p(y_{g_i}|\theta, x_{g_i})$, with n representing the number of samples.

We selected protein-coding genes that accomplished a mutation threshold of 10 samples and corrected their p-values resulting from the likelihood test for multiple testing (FDR$<$ 20%). We identified 46 protein-coding genes with mutational status dependent on the viral infection status (Figure 5.1). Of these, BAP1, MAP2K4, and TP53 are previously established tumor suppressors and GNAQ is an oncogene. Alterations of these genes have been shown to impact DNA repair, p53-apoptosis [171], signal transduction or transcription (MAP2K4), and oncogenesis through the G-protein signaling pathway [172].

The results revealed a negative effect on the mutation rate driven by the HBV in-

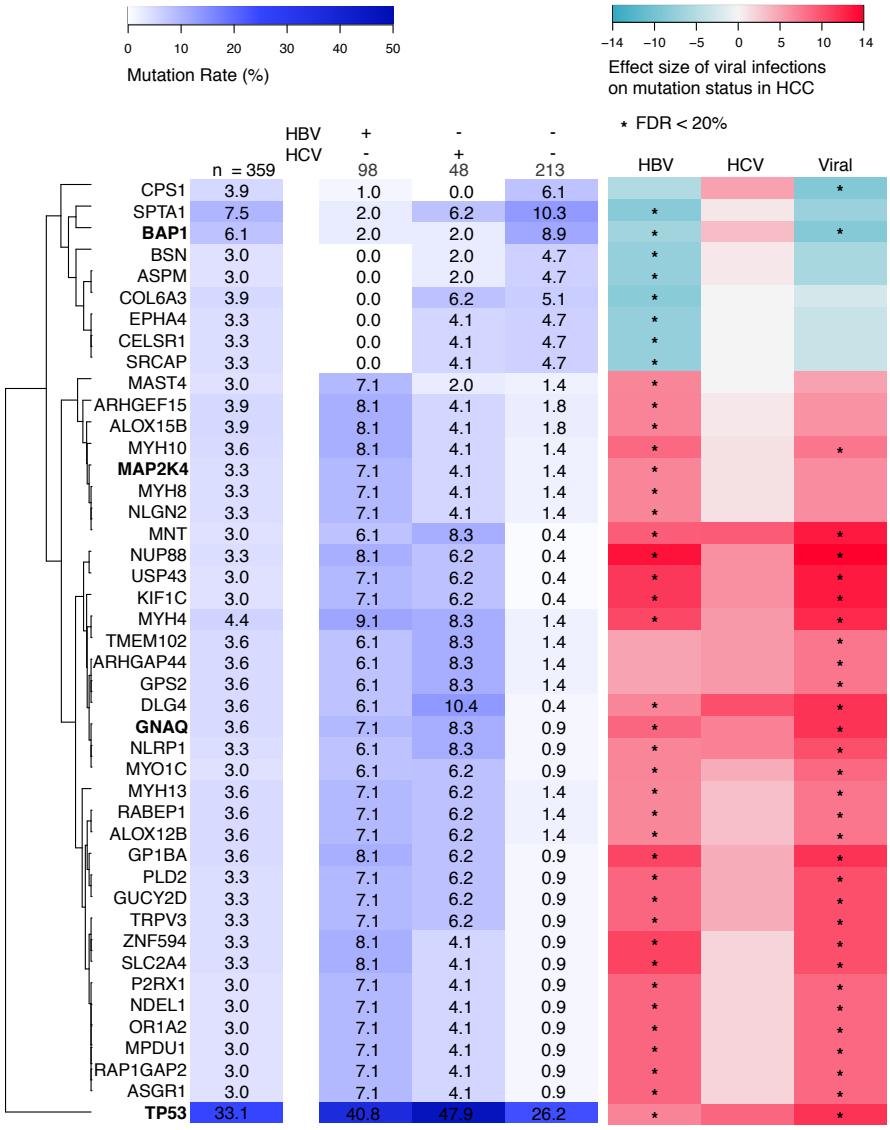

Figure 5.1: Differences in mutation rates between HBV, HCV and non-infected patients. The left panel of the heat map represents the mutation rates in the different patient groups. The right panel represents the effect size of the viral infection on the mutation status in the LIHC data set.

fection (Figure 5.1, right panel) for the following genes: CPS1, BAP1, SPTA1, ASPM, COL6A3, EPHA4, CELSR1, and SRCAP. The role of CPS1 in cell growth, metabolism, and prognosis in LKB1-Inactivated lung adenocarcinoma [173], together with its annotation as a prognosis marker in chronic HBV infection [174], suggested involvement in viral-host pathways activation. The differ-

ential mutation rate observed for BAP1 confirmed its involvement in hepatic-induced tumors [175].
Motivated by these findings, we aimed to increase the confidence of the results by using another data level: physical viral-host interactions.

5.2 Viral-host interactions with oncogenic effect

5.2.1 Network propagation estimated genomic and physical HBV impact on human PPI interactions

Our next research goal was to find viral-host interactions with a strong involvement in the underlying development of tumors. The oncogenic effect of viral-host interactions may be reflected in proteins that serve as both viral targets and cancer drivers. Hence, we combined the physical HCV-human and HBV-human interactomes with the mutational landscapes of HCV and HBV in HCC.
Given the two different means of measuring the strength of viral-host interactions in HCC – the deviances resulting from the differential mutation analysis and the strength of physical host-viral interactions, we aimed to understand the broad viral effect on human pathways in HCC. For this purpose, we used the ReactomeFI network (further referred to as the reference network), consisting of 229,300 manually curated pathway-based protein functional interaction network and the network propagation framework [85]. By applying network propagation within the reference network, we spread the influence of each differential mutation and the influence of each viral physical interaction over their network neighborhoods.
We first propagated the HBV deviances – $D_{g_{hepB}}$, through the reference network. We retained the propagated deviance scores in S_{d_g}. Conceptually, S_{d_g} indicated how likely it is that gene g is affected by proteins with differential viral-associated mutations. We obtained estimates of the viral effect on human proteins within the reference network by scoring the proximity of protein in the reference network to the HBV-interacting proteins at the genomic level. Next, we propagated the HBV MiST scores through the reference network. S_{p_g} denoted the propagated MiST scores. S_{p_g} represented the likelihood of gene g

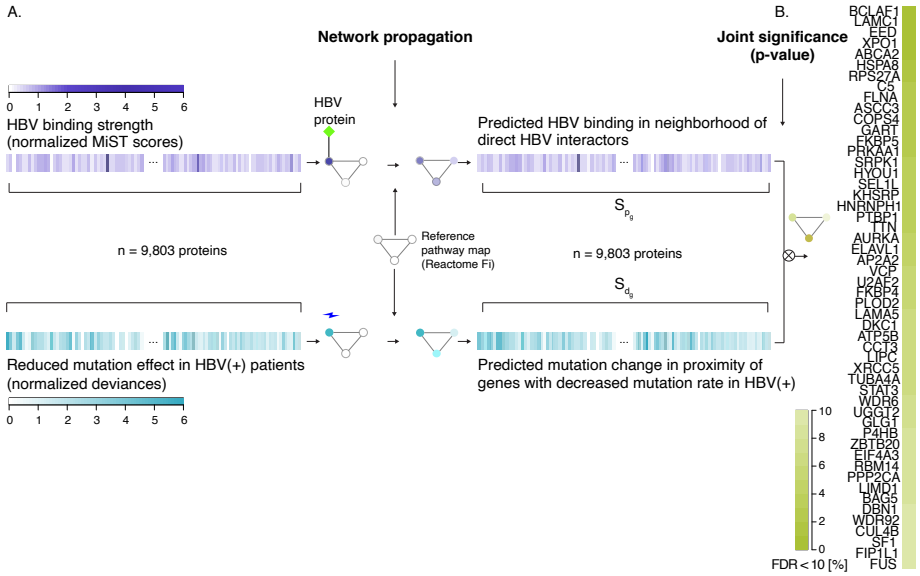

Figure 5.2: Network propagation identified Reactome Fi neighborhoods enriched in both HBV interactors and genes with a decreased mutation rate in HBV(+) HCC tumors. The initial and predicted scores for each gene are indicated by color intensity: purple for HBV physical interaction scores (MiST), turquoise for differential mutation scores, and green for combined significance (FDR value).

being affected by proteins that were physically interacting with viral proteins. Thus, we estimated the viral effect on proteins within the reference network by scoring their proximity to the HBV-interacting proteins at the physical level. An overview of the network propagation can be seen in Figure 5.2A. The change introduced by network propagation is indicated by color intensity.

We used the human protein-coding genes present in the reference network for network propagation, that were also expressed in the LIHC data set (p = 9,803 proteins). Given the topology of the reference network, certain nodes (e.g., hubs) will be 'hot' regardless of the initial scores represented by either deviances or MiST scores. To estimate the expected background of S_p scores given the network topology, we performed 10,000 permutations in which we randomly reassigned the deviances $D_{g_{hepB}}$ and the HBV MiST scores. To calculate the significance of the propagation score of a specific gene, we ran the network propagation algorithm separately with the permutated deviances and MiST scores as input scores. Next, we calculated empirical p-values. The p-values indicated how many times the propagated scores after permutation are greater than the

real scores.

For each of the 9,803 proteins, we obtained two confidence scores. The two scores indicated the likelihood of a given protein being altered at the genomic level due to HBV infection and the likelihood of the same protein physically interacting with HBV proteins.

Both times, network propagation allowed us to learn from different data modalities and revealed:

- neighborhoods within the reference network enriched with genes with a decreased mutation rate in HBV(+) HCC tumors
- neighborhoods within the reference network enriched with genes physically interacting with the virus.

5.2.2 Measure of joint significance reveals

We used the gene-wise propagated MiST and deviances scores to calculate a measure of joint significance for each protein-coding gene (5.2B). Given that normalization brought the two types of measurement, we defined the joint significance score as:

$$S_{c_g} = S_{d_g} \cdot S_{p_g} \tag{5.6}$$

To obtain the null hypothesis distribution of the joint score given the network topology, we performed 10,000 permutations through which we randomly reassigned MiST scores and deviances. We applied the network propagation algorithm and calculated the product of the two propagated scores. We then calculated empirical p-values corresponding to the joint score. The p-values indicated which genes had network neighborhoods significantly enriched for both viral interactors and genes with a decreased mutation rate in HBV(+) HCC tumors. We calculated the false discovery rate using the Benjamini-Hochberg method [116]. The values represented the probabilities of incorrectly finding genes with neighborhoods significantly enriched for both viral interactors and genes with differential mutation rates.

We identified 54 proteins that showed significant proximity to both proteins with

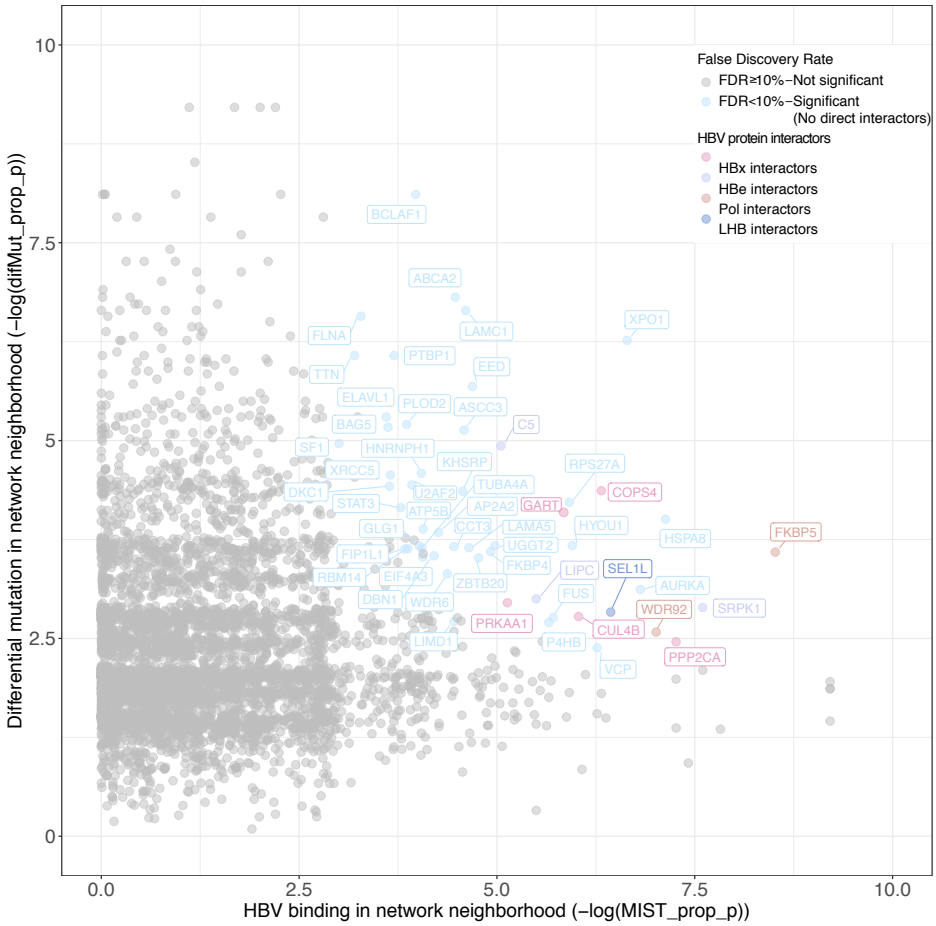

Figure 5.3: Empirical p-values of propagated MiST (x-axis) and differential mutation scores (y-axis). The empirical p-values were derived from 10,000 permutations. Neighborhoods of proteins shown in color are significantly enriched for both proteins physically interacting with viral proteins and proteins with decreased mutation rate in HBV(+) HCC tumors (FDR $<$ 10%). The proteins in significantly enriched protein that are directly interacting with HBV protein are further color-coded according to their HBV protein interactor: pink (HBx), lavender-blue (HBe), Jordy blue (LHB), rose (Pol), and light sky blue (non-direct interactors)

a decreased mutation rate in HBV(+) HCC tumors and proteins with strong physical binding to HBV proteins (Figure 5.3, FDR $<$ 10%).

Next, we superimposed the interactions in the reference network between proteins with joint significance with the viral-host physical interactions to build an integrated interactome of HBV in HCC (Figure 5.4).

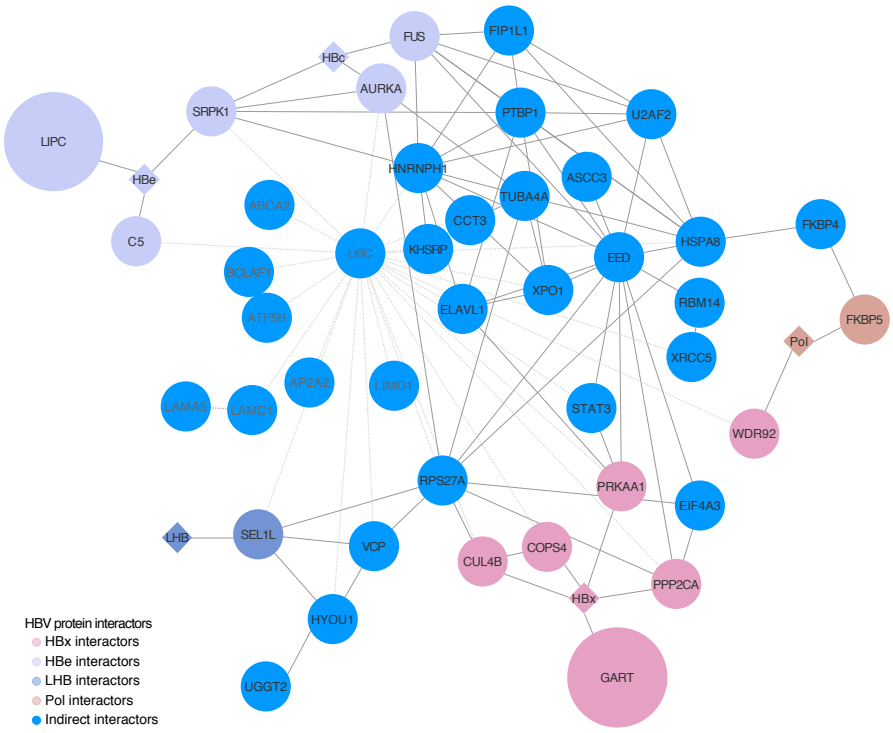

Figure 5.4: HBV interactome consisting of reference network pathways of combined significance after network propagation (FDR<10%) and HBV-host interactions (MiST scores > 0.75).

Finally, we identified neighborhoods enriched for both physical viral-host interactors and genes with differential mutation rates in infected patients.

5.3 Ubiquitylation and phosphorylation affected by both HBV-related mutation and viral physical interaction

The results of our approach confirmed known viral-host interactions but also revealed previously unknown interactions. In particular, we observed that CUL4A, CUL4B [176], CDKN2A [177], and TP53 [178, 179] – known cancer drivers in HCC, showed differential mutation status in HBV infected HCC patients and were bound physically to the HBV proteins in the HCC cell lines data (Figure 5.4).

Additionally, the integrative analysis of HBV data revealed strong effects in several previously unknown protein genes, among which COPS3 and PPP2CA. COPS3 is part of the COP9 signalosome complex (CSN) – an essential regulator of ubiquitination, while PPP2CA is the main phosphatase for microtubule-associated proteins (MAPs) – a negative regulator of cell growth and division. The identified interactions indicated ubiquitination and phosphorylation include both physical and genetic interactions with HBV. Given these findings, we further focused on examining their involvement in viral two pathways in particular. The experiments examining the remodeling of the COP9 and PP2A complexes upon binding of HBx (the core HBV protein) confirmed the involvement of HBV infection on global ubiquitylation and phosphorylation. These results are included in the manuscript describing this study (to be submitted).

5.4 Conclusions and outlook

This chapter introduced a multilevel omics data integration approach for identifying protein interactions involved in viral oncogenesis. To determine the impact of viral infections on oncogenesis, we combined two layers of information. The first layer included multilevel omics data – gene expression, somatic mutations, and CNA copy number aberrations from tumor samples. The second layer consisted of experimental data measuring physical interactions between viral and human proteins observed in cancer cell lines.

We showed that integrating different data levels constitutes a powerful learning approach for finding viral interactors that are also involved in cancer. Additionally, we showed that the results of our integrative approach enable us to broaden findings from the level of proteins to pathways. Our analysis revealed knowledge related to HBV remodeling the PP2A and COP9 protein complex and altering their function in ubiquitination and phosphorylation.

The approach described in this chapter can be used for any viral infection and cancer, considering the availability of the required data. We focused on HBV infection in 366 TCGA LIHC samples and physical interactions between HBV and human proteins in HCC cell lines. The first part of this study revealed hu-

man protein genes with mutational status dependent on viral infections. In the second part, we used network propagation to identify human pathways affected by both viral-host physical interactions and tumor mutation status.

One can easily adjust the model estimating the effect of viral infection at the genomic level to correct for confounder variables of interest. For example, since it is known that HBV infection is more frequent in males [180], the Bayesian logistic model can be broadened to correct for gender. Furthermore, considering the highly variant liver cancer incidence rates across world regions [177], ethnicity can be included in the model as a covariate.

Another way to adjust the method is to measure the viral expression in the tumor samples and replace the binary infection status with continuous viral expression values. Although computational intensive (due to the mapping of the viral sequences to the whole genome sequencing of the liver cancer samples), this additional analysis may enable the accuracy of the viral status provided by the clinicians. Furthermore, we can test whether there is an effect of the viral expression abundance on the mutation rate.

Lastly, although our study focused on the known link between HBV and liver cancer, the method can also be applied to tumors that have not yet been associated with viral infections - to discover whether the viral infection is a risk factor.

Altogether, this work shows that integrating physical protein-protein interactions with multilevel omics data represents a suitable framework that will assist the progress of many other cancers with viral risk factors.

Chapter 6

Regulatory network inference in head and neck and lung cancers reveal miRNAs involved in oncogenic pathways

Dysregulation of miRNAs that act as oncogenes or tumor suppressor genes was associated with many cancer types [181]: Medina et al. were the first to show in an *in vivo* model that overexpression of miR-21 initiated a pre-B malignant lymphoid-like phenotype and proved that miR-21 acted as an oncogene [182]. Later, Ma et al. showed that miR-21 enhanced cellular necrosis by negatively regulating tumor suppressor genes associated with the death-receptor-mediated intrinsic apoptosis pathway [183]. In mouse models, miRNA dysregulation was sufficient for driving oncogenesis, while, in humans, changes on the genetic and epigenetic levels of the miRNA biogenesis were associated with cancer initiation [184]. We explored the potential role of miRNAs as prognostic markers in HNSCC tumors treated with radiotherapy and showed that changes in the

abundance of circulating miRNAs during radiochemotherapy affect the therapy response of primary HNSCC cells after an in vitro treatment [185, 186]. Given the miRNA involvement in cancer, it is essential to understand how miRNAs act in a multilevel omics framework.

For this purpose, we developed a novel method, miRlastic, consisting of two successive steps: identification of miRNA–mRNA interactions and functional annotation of miRNA target gene sets (Sass & Pitea et al.). Our method infers miRNA–mRNA interactions using transcriptomic data and prior knowledge and performs functional annotation of target genes by exploiting the local structure of the inferred network [134]. Moreover, miRlastic comes with the advantage that it can be used for any specific biological condition.

As discussed in Sass & Pitea et al., inferring miRNA–mRNA interactions to further reveal miRNA functions can uncover how miRNAs impact molecular pathways. In particular, inferring miRNA–mRNA interactions and functionally annotating miRNA target in cancer can show how miRNAs regulate underlying cellular mechanisms and contribute to oncogenic pathways.

One of the main parts of this chapter introduces our application of miRlastic to infer miRNA–mRNA regulatory networks linked to human papillomavirus (HPV)-associated miRNAs in HNSCC. We also investigated the miRNA impact of HPV-associated dysregulation. Another central part of this chapter includes the application of miRlastic to study how dysregulated miRNAs affect NSCLC metastasis.

The approach and the applications described in this chapter are part of the following publications:

- Steffen Sass*, **Adriana Pitea***, Kristian Unger, Julia Hess, Nikola S. Mueller and Fabian J. Theis.

  MicroRNA-Target Network Inference and local Network Enrichment Analysis Identify Two microRNA Clusters with Distinct Functions in Head and Neck Squamous Cell Carcinoma. *Int J Mol Sci*, 16(12): 30204-30222, 2015.

- Margarita Gonzalez-Vallinas, Manuel Rodriguez-Paredes, Marco Albrecht,

Carsten Sticht, Damian Stichel, Julian Gutekunst, **Adriana Pitea**, Steffen Sass et al.

Epigenetically Regulated Chromosome 14q32 miRNA Cluster Induces Metastasis and Predicts Poor Prognosis in Lung Adenocarcinoma Patients. *Mol Cancer Res.*, 10.1158/1541-7786.MCR-17-0334, 2018.

The previous analyses on the potential role of miRNAs as prognostic markers in HNSCC tumors treated with radiotherapy are included in the following publications:

- Isolde Summerer, Julia Hess, **Adriana Pitea**, Kristian Unger, Ludwig Hieber, Martin Selmansberger, Kirsten Lauber and Horst Zitzelsberger

  Integrative analysis of the microRNA-mRNA response to radiochemotherapy in primary head and neck squamous cell carcinoma cells. *BMC Genomics*, 16:654, 2015.

- Isolde Summerer, Maximilian Niyazi, Kristian Unger, **Adriana Pitea**, Verena Zangen, Julia Hess, Michael J Atkinson, Claus Belka, Simone Moertl, Horst Zitzelsberger.

  Changes in circulating microRNAs after radiochemotherapy in head and neck cancer patients. *Radiat Oncol.*, 8:296, 2013.

The work presented in this chapter focuses on an integrative approach to study the role of miRNAs in miRNA–mRNA–pathway interactions specific to cancer. The data analyses and figures presented here represent my work entirely. The miRlastic R package, Figure 6.3, together with the new scoring procedure for the local enrichment analysis, represents joint work with my former colleague, Steffen Sass, Ph.D. My contribution to the lung cancer collaboration project consisted of inferring the miRNA–mRNA regulatory network for 23 candidate pre-miRNAs. The 23 candidate pre-miRNAs were selected and provided by my collaboration partner, Margarita Gonzalez, Ph.D.

6.1 miRNA–mRNA–pathway interactions

MiRNAs can bind to the complementary 3'-untranslated region (3' UTR) of messenger RNA (mRNA) sequences to post-transcriptionally fine-tune target

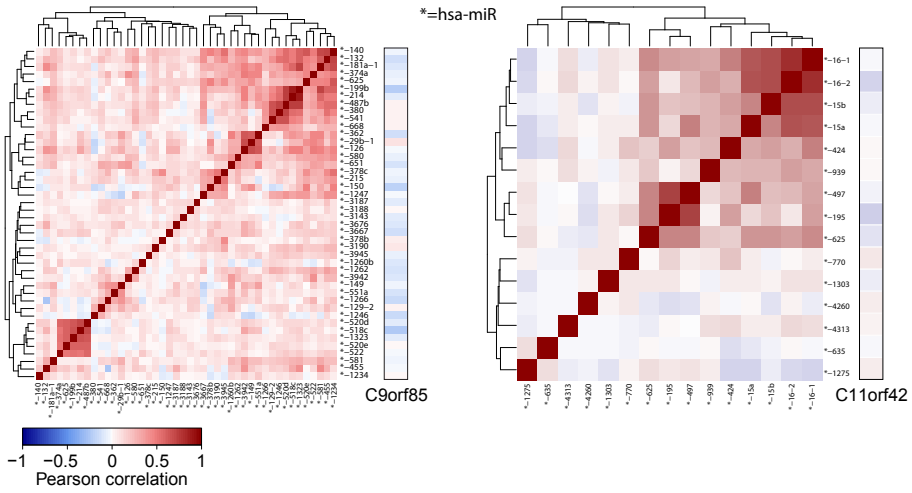

Figure 6.1: **Joint effect of co-expressed miRNAs with a common predicted target.** Pairwise correlation of predicted expressed miRNA regulators together with their corresponding target genes predicted by TargetScan [187] C9orf85 and C11orf42 (obtained from the HNSCC dataset). The miRNAs are themselves clustered into several co-expressed groups.

mRNA expression [188]. MiRNA biogenesis was proven to be under tight spatio-temporal control, and targeting relationships were shown to be cell type or tissue-specific [189]. Additionally, the scientific community theorized that miRNAs act in a combinatorial manner [190, 191]. To obtain condition-specific miRNA–mRNA interactions, we developed a statistical inference method that:

- used matched miRNA and mRNA expression data of the underlying conditions
- integrated prior knowledge of sequence-based predictions
- integrated the joint regulation of miRNAs that target the same mRNA.

Group correlation of miRNAs with a common target

Since our method relied on the concept of joint regulation of miRNAs that target the same mRNA, we evaluated the correlation between miRNAs with a common target. In particular, we calculated the pairwise Pearson correlation coefficients among all miRNA expression profiles from the HNSCC data that TargetScan predicted to have a common target. Figure 6.1 showed the correlation for two genes predicted to be targeted by multiple miRNAs: C9orf85 and

C11orf42 (Figure 6.1). We observed subgroups of high correlation, which confirmed the predicted co-expression of miRNAs that were functionally related or resided nearby the chromosome.

To systematically analyze whether miRNA expression profiles typically corre-

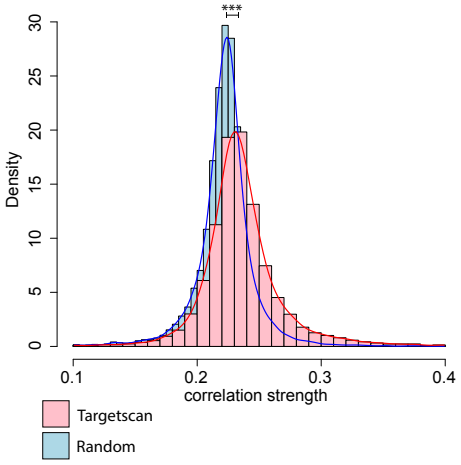

Figure 6.2: The distribution of correlation strengths $c(\boldsymbol{X})$ of miRNA sets, which are predicted to target a common gene, (red curve and histogram) is higher than for randomly re-sampled miRNA-mRNA associations (blue, Wilcoxon rank sum test has $p < 1 \times 10^{-80}$) in the HNSCC miRNA expression dataset.

lated when predicted to target the same mRNA, we assessed the Pearson (anti-)correlation strength across all expressed miRNAs in the HNSCC data with a common target. Our analysis revealed a stronger correlation between such miRNAs relative to correlation in randomly sampled sets of miRNAs ($p < 1 \times 10^{-80}$ Wilcoxon rank sum test, Figure 6.2).

Following, we applied a multiple linear regression model with an elastic net penalty, which represented a tradeoff between the lasso and ridge regression and accounted for both joint effects of several miRNAs on a common target and co-expression between miRNAs [124]. We imposed a negativity-constraint on the regression coefficients to choose only down-regulation effects. To functionally annotate miRNAs based on the inferred miRNA–mRNA network, we introduced a local enrichment analysis that scored miRNAs based on the underlying network structure and the functional annotations of their target genes.

miRNA–pathway scoring

We evaluated whether node arrangements were assigned to a specific term, describing, e.g., a molecular function or biological process, occurred by chance or not. To characterize the importance of miRNAs in the inferred network, we define the following score:

$$S_{miR}(v_i) = \left(\frac{1}{|\mathcal{V}_i|} \sum_{v_j \in \mathcal{V}_i} S(v_j) \right) \cdot \sqrt{|\mathcal{V}_i|}, \tag{6.1}$$

where $v_i \in V^{miR}$ represented every miRNA node in the inferred network, $\mathcal{V}_i$ indicated the set of predicted targets, $|\mathcal{V}_i|$ indicated the number of inferred targets and $S(v_j)$ represented the enrichment of a term for a given functional group around gene j. $|\sqrt{\mathcal{V}_i}|$ represented a weight that corrected for the number of the corresponding targets – $|\mathcal{V}_i|$, for each miRNA. Specifically, $|\sqrt{\mathcal{V}_i}|$ ascertained that miRNA nodes with a significantly reduced number of predicted targets were not highly ranked. The resulting score indicated how strongly miRNA disrupted a particular pathway.

We extensively evaluated the miRlastic network inference module and further applied miRlastic on an HNSCC TCGA subset (Chapter 3).

6.2 miRNA–mRNA regulatory networks in HNSCC

6.2.1 Human papilloma viral impact on HNSCC

HPV infection presents specific molecular characteristics that include gene mutations, CNAs, changes in DNA methylation, mRNA, and miRNA expression patterns [32, 192]. Lajer *et al.* identified a set of core miRNAs implicated in HPV pathogenesis [193, 192] related to viral-human pathways of HPV induced malignant pathogenesis. As studies showed, HPV disrupted cellular differentiation in HNSCC [32, 194]. We then chose to focus on the 244 patient samples with reported HPV status. The involvement of miRNAs in HPV-related induced malignant pathogenesis motivated us to further explore miRNAs in this

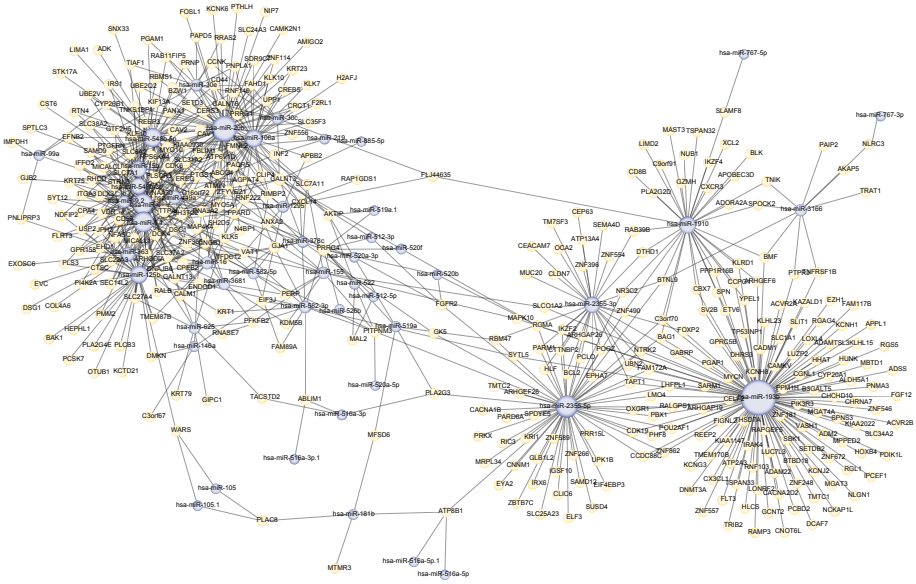

Figure 6.3: miRNA–mRNA regulatory network generated by miRlastic. The network consists of 766 interactions between 44 miRNAs (light blue) and 16,617 genes (light yellow). The edges represent miRNA–mRNA relationships within the TCGA HNSCC sub-cohort.

sub-cohort.

For this, we performed a differential analysis between miRNA expression of HPV+ and HPV- HNSCC samples and identified 44 deregulated miRNAs between HPV+ and HPV- patients (Figure 3.1).

The set of differentially expressed miRNAs included the miR-9 family, miR-363, miR-20b, confirming the reported association between the HPV status and miRNA expression in several independent studies [195, 196, 193].

6.2.2 miRNA–mRNA interactions in HNSCC

To understand the HPV-associated miRNA-mediated gene regulation of HNSCC tumors, we performed a miRlastic inference using 135,391 targets predicted by TargetScan in combination with the respective miRNA and mRNA expression values. Our method inferred 766 miRNA–mRNA interactions (Fig. 6.3). The underlying miRNA–mRNA predictions were extracted from the TargetScan database (version 6.2) [197] and included only conserved target sites for conserved miRNAs families. Figure 6.3 depicts the HPV-associated miRNA nodes

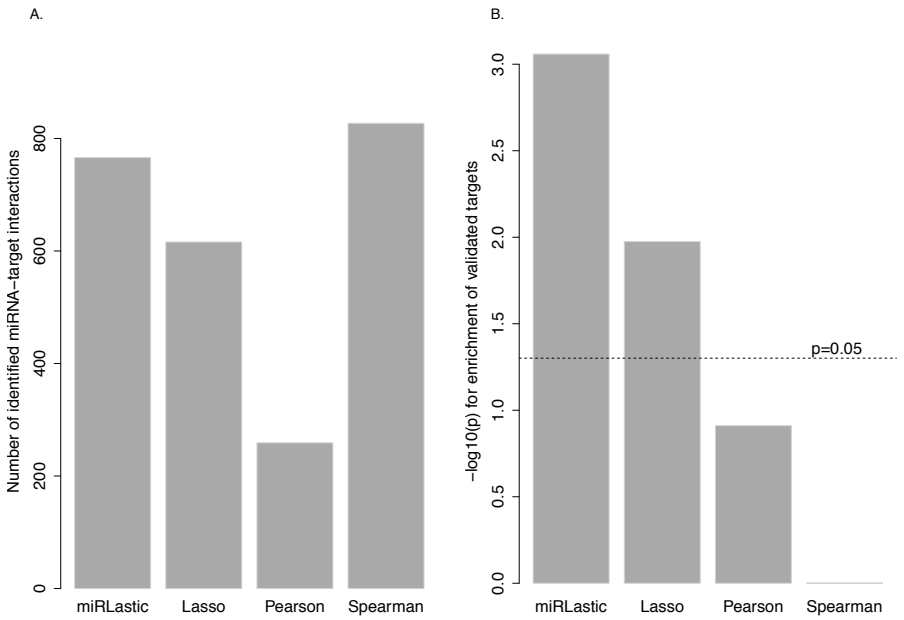

Figure 6.4: Performance evaluation of the miRlastic on HNSCC data. Panel A. indicates the number of miRNA – mRNA interactions detected by miRlastic in comparison to the number interactions detected by lasso, Pearson and Spearman. Panel B shows the $-$log10-transformed p-values when testing for enrichment of experimentally validated targets within the results of each method.

and their inferred targets. The network itself already provided insights into the functional roles of miRNAs. We presented an interactive representation of this network at: http://icb.helmholtz-muenchen.de/mirlastic/hnscc.

Next, we evaluated miRlastic on the HNSCC sub-cohort. Specifically, we tested whether the regulatory network predicted by miRlastic included significantly more experimentally validated miRNA–mRNA interactions relative to commonly used methods - TargetScan, Spearman and Pearson correlation, lasso regression. Following, we collect experimentally validated interactions from starBase v2.0 obtained using HITS-CLIP or PAR-CLIP (high stringency) [198]. Of 766 interactions predicted by miRlastic, 87 miRNA–mRNA were experimentally validated. To determine whether the fraction of inferred and validated interactions was higher than expected from the prior target network (TargetScan) relative to the number of inferred interactions, we applied Fisher's exact test. For miRlastic, the test yielded a highly significant p-value of $p =$

8.736821×10^{-4}(Figure 6.4B).

To compare miRlastic to other methods, we applied three further commonly used miRNA-mRNA network inference methods [186, 185] on the same data: Pearson correlation, Spearman correlation, and lasso. We obtained a p-value of $p = 1.059271 \times 10^{-2}$ for lasso, $p = 1.228527 \times 10^{-1}$ for Pearson correlation and $p = 9.978787 \times 10^{-1}$ for Spearman (Figure 6.4). The results indicated that miRlastic identified a higher fraction of validated target predictions than the other methods. Lasso also performed well predicting a significant fraction of experimentally determined interactions, but lower relative to miRlastic (Figure 6.4). Pearson and Spearman correlations did not show any significant difference.

In conclusion,, miRlastic inference outperformed the other three methods in over-representing experimentally validated miRNA–mRNA interactions.

6.2.3 LEA identifies two miRNA clusters associated with tumorigenesis regulating processes: apoptosis, immune response and proliferation

In the previous section, we provided a network characterizing miRNA–mRNA interactions in TCGA HNSCC samples with known HPV status. To finally reveal how miRNAs impact the underlying cellular mechanisms in HNSCC concerning HPV infection, we assessed the local enrichment in the miRNA–mRNA genes, given the HNSCC network.

For this purpose, we downloaded 108 pathways from KEGG for gene annotations [199]. To identify closely-connected functions within the previously inferred network, we used LEA to evaluate whether node arrangements assigned to a specific term, describing, e.g., a molecular function or a biological process, occurred by chance or not. We obtained nine significantly locally enriched pathways (Fig. 6.5).

Next, we clustered the miRNAs according to their functional score. Our analysis revealed two clusters with a similar pattern of miRNA scores across pathways (Figure 6.5).

The clusters associated with the MAPK- and Neurotrophin-signaling pathways. MAPK- and Neurotrophin-signaling pathways comprised common elements re-

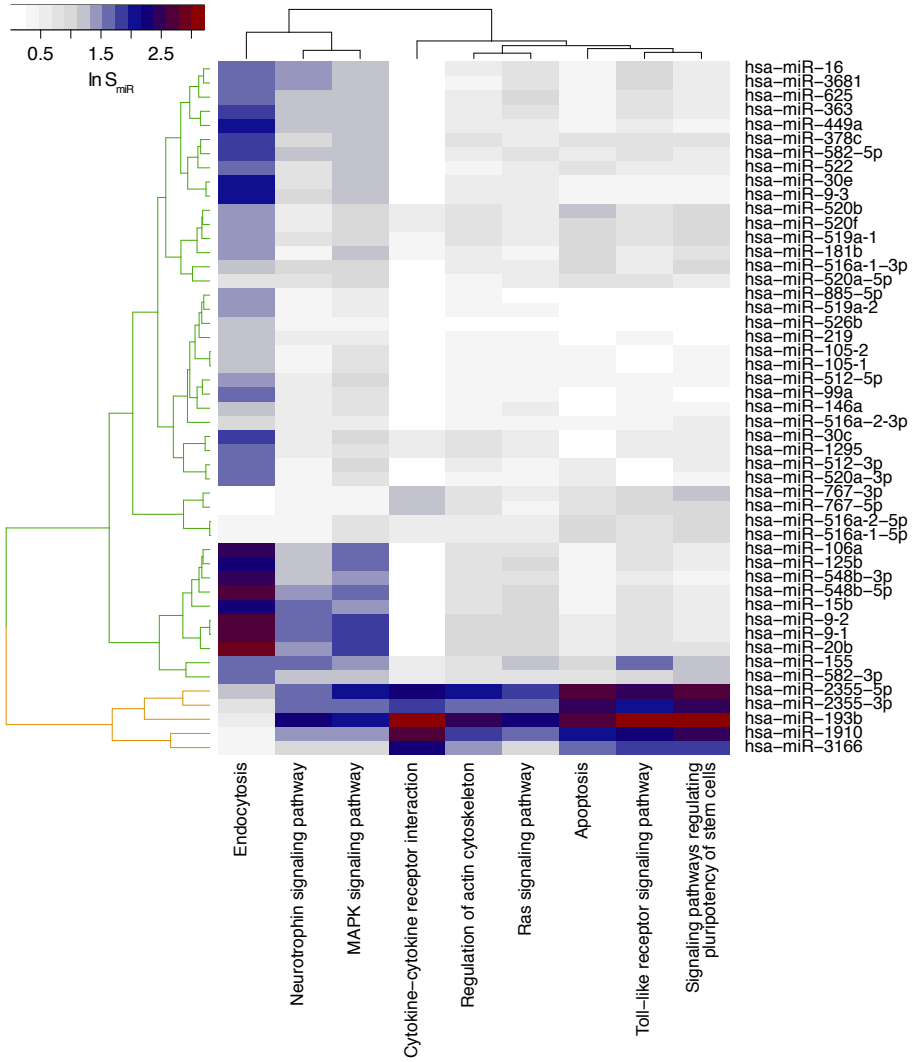

Figure 6.5: **Functional characterization of the HPV-associated miRNA–mRNA network in HNSCC samples.** Heatmap of miRNA scores $S_{miR}(v)$ for each miRNA v in the network indicating the functional role in the significantly locally enriched KEGG pathways.

lated to tumorigenesis-regulating processes such as immune response, apoptosis, proliferation, and angiogenesis [200, 201]. Among the clustered miRNAs, we distinguished hsa-miR-193b, which was shown to affect the the MAPK signaling pathway and enhance tumor progression in HNSCC [202, 203].
In a more detailed analysis, we observed that one of the miRNA clusters comprised links between hsa-miR-106a, -125b, -548b-3p/5p, -15b, -9-2, -9-1, -20b, -155, and -582-3p and the Endocytosis pathway – prone to deregulation in cancer cells [134, 204].
The other miRNA cluster comprised a broader range of pathways and included links between hsa-miR-2355-5p/3p, -193b, -1910, and 3166 with pathways involved in apoptosis, regulation of stem cell pluripotency and metastasis [5, 205, 134].
The scientific community linked cancer stem cells (CSC) to therapy resistance of HNSCC and HPV infection: HPV+ HNSCCs presented smaller CSC proportions relative to HPV- HNSCCs [206]. Our functional analysis provided a reason for HPV+ HNSCC patients presenting a better prognosis relative to HPV-, and linked the "Signaling pathways regulating pluripotency of stem cells" to HPV-infection [134].
Overall, our results suggested that only specific miRNAs mediate gene dysregulation primarily through pathways regulating stem cell pluripotency [134].

6.3 miRNA–mRNA–pathway interactions in NSLSC

In the previous section, we investigated the role of HPV-associated miRNAs in HNSCC by using miRlastic. Within this section, we continued to explore the impact of dysregulated miRNAs on cancer pathways. Specifically, we applied miRlastic to distinguish the relevant miRNA–mRNA network driving non-small cell lung cancer metastasis (further referred to as lung cancer). For performing the miRlastic inference, we used 135,391 targets predicted by TargetScan combined with expression levels of 23 pre-miRNAs and 16,241 mRNAs from lung cancer. The 23 pre-miRNAs showed differential expression between lung cancer samples from patients with and without lymph node metastasis (N1, N2, and N3 vs. N0) in a TCGA sub-cohort (n=449).

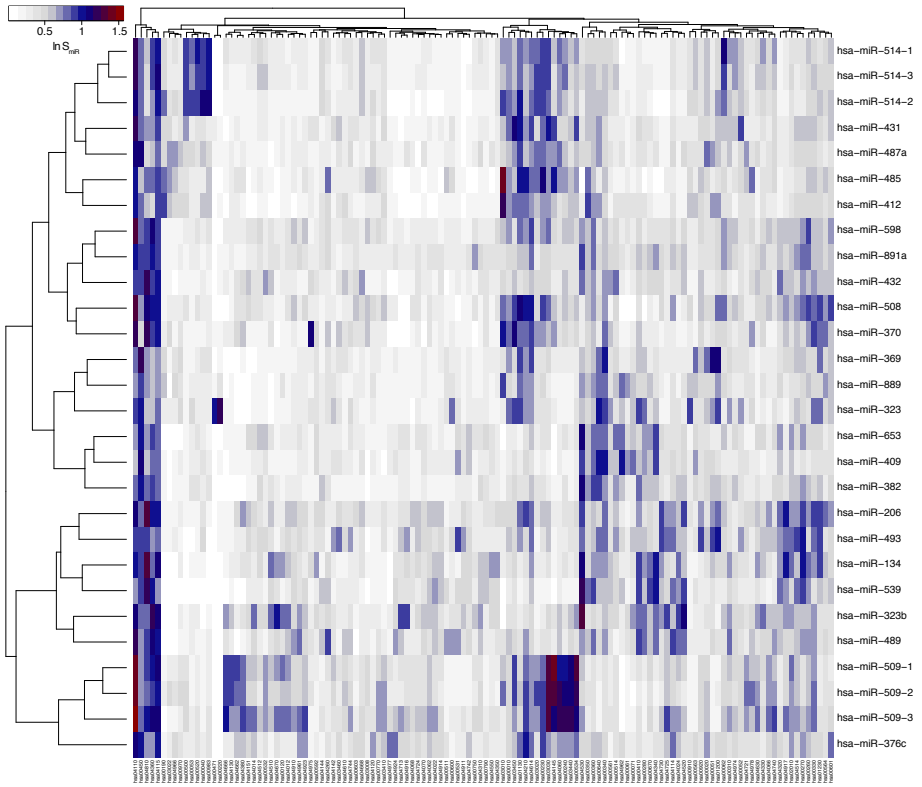

Figure 6.6: **Functional characterization of the metastasis-associated miRNA–mRNA network in TCGA lung cancer samples.** Heatmap of miRNA scores for each miRNA in the network indicating the functional role in the significantly locally enriched KEGG pathways.

Our method inferred 8310 miRNA–mRNA interactions. Next, we applied LEA on the resulting network with the aim of finding functional groups enriched in a specific neighborhood of the network. Out of the 226 KEGG pathway annotations used when running LEA, 126 showed significant enrichment in local communities of our network (Figure 6.6).

Notably, the results suggested that miR-509 had a significant impact on cell cycle (hsa04110), cytokine-cytokine receptor interaction (hsa04060), homologous recombination (hsa03440), p53 signaling pathway (hsa04115). Our finding agreed with previous results showing that miR-509 regulated cancer cell growth by affecting the p53 signaling pathway and, subsequently, cell cycle [207]. Another relevant pathway that was locally enriched in our network and was involved in metastasis was the regulation of actin cytoskeleton (hsa04810) [208, 209, 210].

Since we were interested in metastasis, we explored whether the inferred network locally enriched for genes involved in the pathways related to this process: cell migration and cell invasion. Thus, we extracted the terms related to these two molecular processes from the Gene Ontology (GO) database together with their child nodes: cell adhesion, cell migration, epithelial to mesenchymal transition, negative and positive regulation of cell migration, negative and positive regulation of cell adhesion, and negative and positive regulation of epithelial to mesenchymal transition. Based on these terms, we built a metastasis-specific set of genes involved in invasion and migration pathways. We tested whether these two molecular processes were locally enriched in our inferred miRNA–mRNA regulatory network. Out of the selected pathways, only three were significantly locally enriched (Figure 6.7).

We observed comparable scores for miR-514, miR-323b, miR-489, and miR-509 for local enrichment for targets involved in cell adhesion, cell migration, and epithelial to mesenchymal transition. Our results confirmed the association between these miRNAs and metastasis across different cancer types [211, 212].

6.4 Discussion and Conclusion

In this chapter, we described two studies on miRNA regulatory networks within cancer. In particular, we aimed to:

- distinguish miRNA–mRNA regulatory networks associated with HPV infection in HNSCC
- distinguish miRNA–mRNA regulatory networks specific for metastasis in lung cancer.

For this purpose, we used miRlastic – a multiple regression approach with an elastic net penalty that accounted for the joint effect of miRNAs with a common target. By using miRlastic, we identified all miRNA–mRNA interactions that were specific to their expression levels. Thus, we removed those miRNA–mRNA predicted interactions from TargetScan that were not related to the investigated biological condition.

Since a gene can be regulated by multiple miRNAs and miRlastic was designed

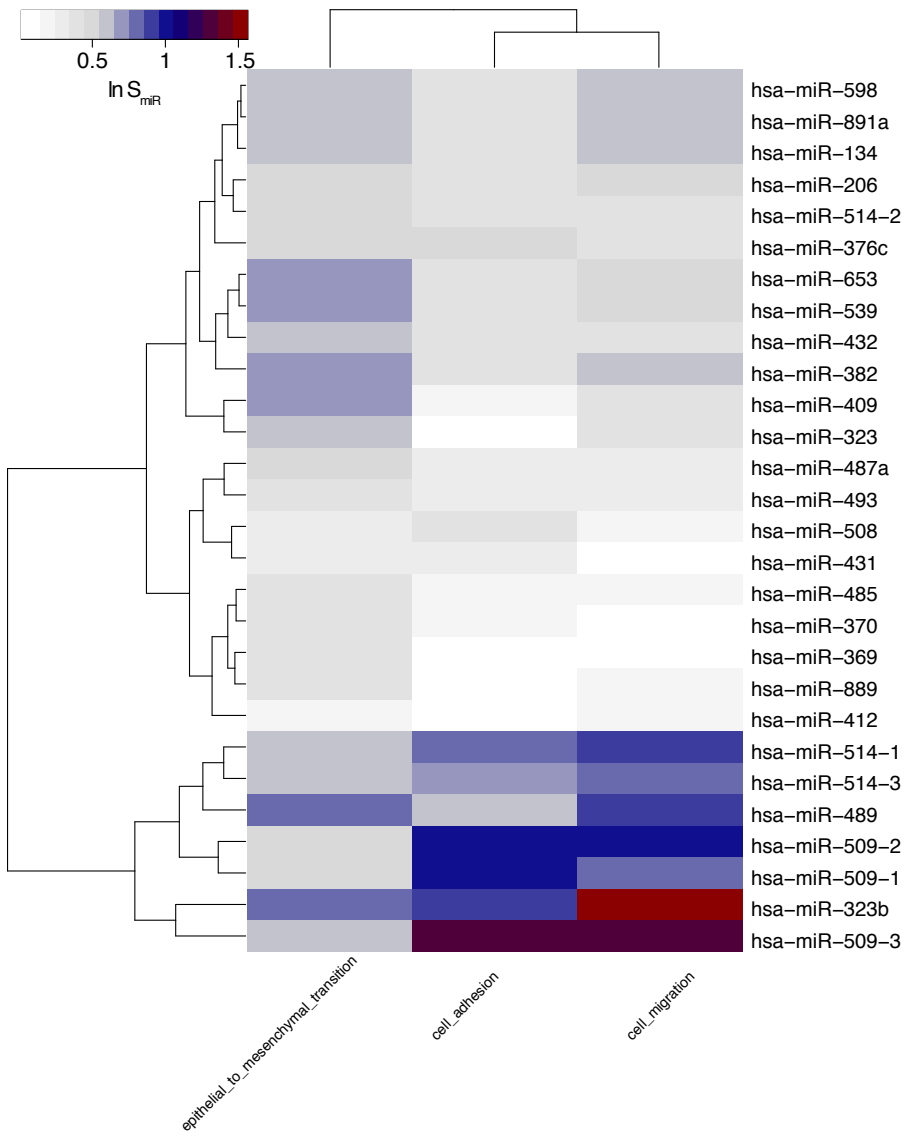

Figure 6.7: **Specific functional characterization of the metastasis-associated miRNA–mRNA network in TCGA lung cancer samples.** Heatmap of significance scores for each miRNA in the network indicating the functional role in the significantly locally enriched metastasis-related KEGG pathways.

to consider the putative joint effect of miRNAs targeting a common gene, we first explored this effect in HNSCC expression data. Following, we inferred the HPV-associated miRNA-target in HNSCC by applying miRlastic. For the HPV-associated miRNAs, the network consisted of 766 miRNA–mRNA interactions.

The subsequent step of miRlastic identified two functionally distinct miRNA clusters predicted to mediate HPV-associated dysregulation in HNSCC.
Overall, miRlastic revealed dysregulation of pathways known to be associated with HNSCC tumorigenesis and HPV infection in HNSCC. Our analysis also showed that miRlastic provided a miRNA–mRNA network significantly enriched for experimentally validated miRNA-target interactions compared to networks resulting from Pearson's correlation, Spearman's correlation, and lasso regression.
Our study on dysregulated miRNAs involved in tumor migration or tumor invasion in non-small cell lung cancer revealed a cluster of miRNAs with significant impact on pathways related to metastasis: cell cycle regulation, regulation of actin cytoskeleton, cytokine-cytokine receptor interaction and p53 signalling pathways. These results provided our collaboration partners a subset of miRNAs with potential involvement in metastasis for further experimental validation.
Lastly, although we focused on only two specific types of cancer, we showed that we could successfully use miRlastic to identify miRNA–mRNA relationships playing essential roles in the context of cancer.

Chapter 7

Functional characterization of long non-coding RNAs through multi-level data integration

Although several studies have annotated and investigated long non coding RNAs (further referred to as lncRNAs) [213, 214, 215], the understanding of lncRNA functional mechanisms is still limited. So far lncRNAs were shown to impact a largely heterogenous class of biological processes like: the guidance of protein complexes to their correct genomic locations [216], the activation of contiguous genes [217], chromatin changes [216, 218] and gene regulation [219, 216]. Recently, lncRNAs overlapping trait-associated SNPs were shown to be highly expressed in cell types relevant to the traits, and thus suggested involvement of lncRNAs in multiple diseases [213]. Additionally, lncRNAs were shown to be involved in diverse cancer types. For example lncRNAs DNM3OS, MEG3 and MIAT were overexpressed in ovarian cancer epithelial-to-mesenchymal transition, a pathway related to tumour migration [220], while the long intergenic non-coding RNA 152 (LINC00152) was shown to promote cell proliferation, metastasis and resistance to treatment in colorectal cancer [221].

To better understand lncRNAs, several resources were developed. However, each is limited by small sample size or by individual functional levels. The catalog of human lncRNAs GENCODE represents the most complete lncRNA annotation resource, including a database of 9,277 manually curated genes [222]. GENCODE quantified the co-expression between lncRNAs and protein coding RNAs (pcRNAs) to identify subclasses and analyze lncRNA biogenesis [222]. However, GENCODE did not provide further functional annotation of individual lncRNAs. LncRNA2function performed enrichment analysis with correlated pcRNAs using only 19 human tissue samples [223]. The lncRNAtor portal, which aimed to offer a comprehensive resource for functional investigation of lncRNAs, was also limited by small sample sizes [224]. Cabili et al focused only on tissue-specificity of intergenic lncRNAs [225], while Gong et al constrained their study to the relationships between lncRNA functions and SNPs [226].
So far, a comprehensive and fully integrative study to infer lncRNA functions exploiting the wealth of newly available studies of larger sample sizes remains non-existent. Given the next-generation sequencing technology and the current availability of numerous public data sets, we now have the possibility to study the heterogeneous, yet poorly characterized, pool of lncRNAs.
Hence, in Pitea & Krause et al, we introduced an integrative analysis of broad scale multilevel data that aimed to explore the molecular functions of lncRNAs: LncRNA Integrative and Systematic Analysis – LISA. To meet the drawbacks of present methods, LISA systematically explored lncRNA molecular mechanisms by integrating genomics, transcriptomics and epigenomics together with functional and tissue annotations from four large sample size projects: Encyclopedia of DNA Elements - ENCODE [87], Roadmap Epigenomics Project [75], the Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) data portal [88]. With LISA we analyzed lncRNA-to-protein coding RNAs (pcRNAs) and lncRNA-to-epigenetic marks associations. Furthermore, we performed functional annotation and tissue-specificity analysis of lncRNAs. Ultimately, we used LISA to identify lncRNA-to-human disease associations.
LISA was discussed in the following publication (in review):

- **Adriana Pitea***, Linda Krause*, Gökcen Eraslan, Steffen Sass, Janine Arloth, Martin Presse, Christoph Ogris and Nikola S. Mueller. Holistic

multilevel analysis of long non-coding RNAs using flexible graph database *Nucleic Acids Research*, (in review).

The lncRNA anotation, the lncRNA expression data analysis and the lncRNA – pcRNA association analysis together with the genomic proximity analysis represent entirely my own work. The tissue specificity analysis and the functional analysis are joint work with my colleagues Dr. Steffen Sass and Gökcen Eraslan. Figure 7.2 represents joint work with Dr. Steffen Sass. The joint work sections and figures were included in this chapter for completeness.

lncRNA annotation

Throughout the LISA approach we used the high quality human reference lncRNA annotation produced by the GENCODE Release 23 [227]. The list consisted of 15,931 common lncRNA IDs annotated within the Ensembl project [228] and experimental validated by the HAVANA group. According to GENCODE, we categorized the lncRNAs into eight distinct biotypes: 3prime overlapping ncRNA ($n = 6,621$), antisense ($n = 323,370$), long intervening noncoding RNA (lincRNA,$n = 159,679$), processed transcript (n=497), asense intronic ($n = 917$), sense overlapping ($n = 194$), to be experimentally confirmed (TEC, $n = 1,050$) and macro lncRNA ($n = 1$).

7.1 Distinct correlation patterns between lncRNA and pcRNA expression across different tissues

One natural way to investigate functions of lncRNAs is to evaluate the co-expression of lncRNAs and pcRNAs based on their corresponding pairwise correlations. To test for lncRNA – pcRNA interactions, we used expression measurements from three projects: Roadmap, GTEx and TCGA.

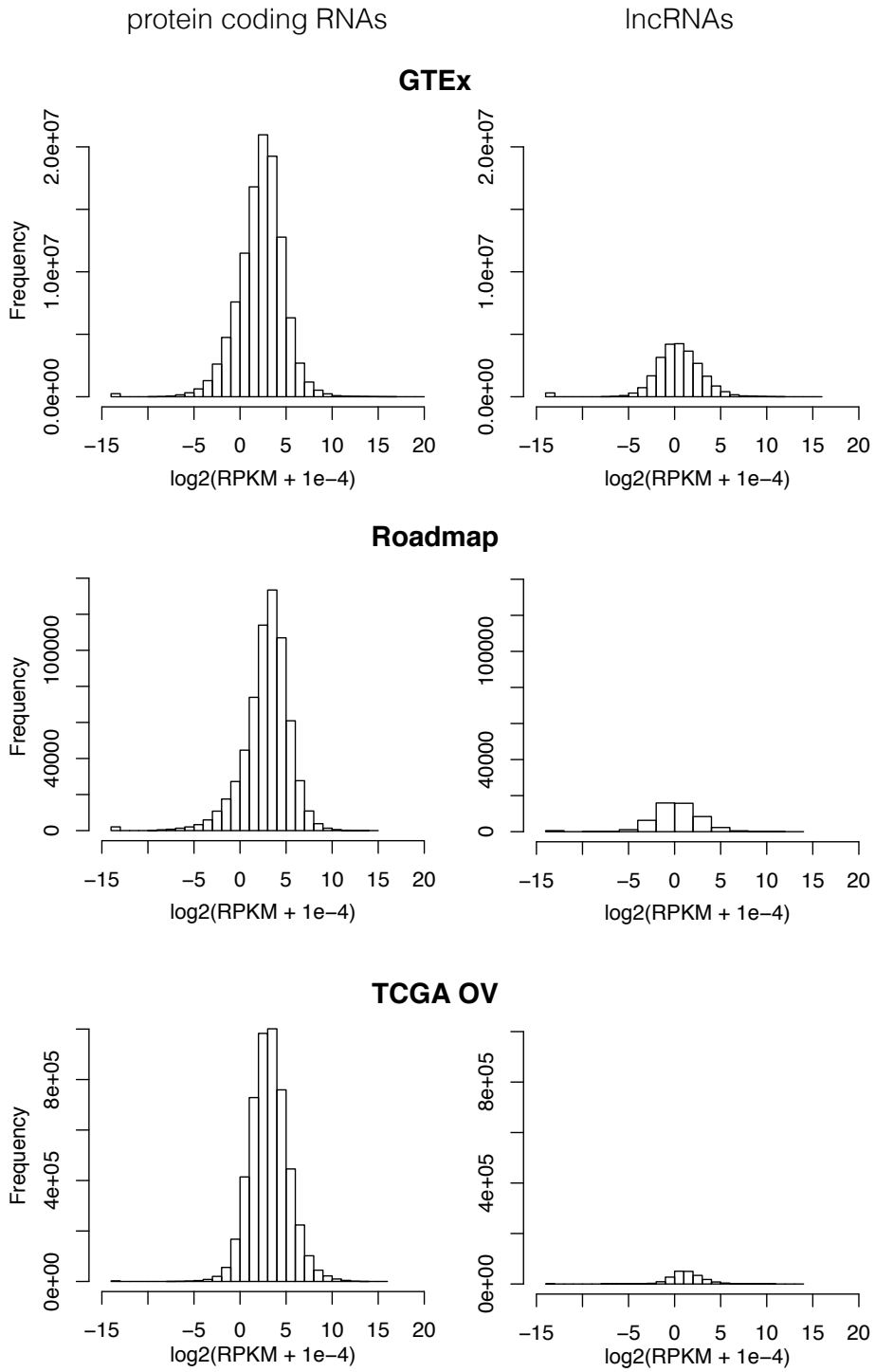

Figure 7.1: Distribution of the log2-normalized expression levels of pcRNAs and lncRNAs within the GTEx, Roadmap and TCGA OV data sets. To avoid non-defined values we added 10^4 to the initial values.

7.1.1 LncRNAs expressed in GTEx and Roadmap data

We exploited the RNA-seq expression data from 53 human reference epigenomes available on the Roadmap Epigenomics Project portal, 8,555 samples from the GTEx portal and 407 samples from ovarian cancer patients (OV TCGA).
Both, lncRNAs and pcRNAs were considered expressed if their normalized expression values are > 0.1 reads per kilo-base per million mapped (RPKM $>$ 0.1) in at least 80% of the samples. The chosen coverage criteria was shown to correspond to about 5 reads in most genes [229] and it was used in several previous studies [229, 88].
We identified 964 expressed lncRNAs and 12,263 pcRNAs across the 53 samples from the Roadmap data. The analysis further identified 2,863 lncRNAs and 12,865 pcRNAs expressed in GTEx data, and 511 lncRNAs and 12,279 pcRNAs expressed in the OV TCGA data set (Figure 7.1). All three data sets were complete, e.g. there were no missing values.

7.1.2 lncRNA - pcRNA correlation

Given the filtered expression profiles, we next calculated pairwise Pearson correlation coefficients for all lncRNA – pcRNA expression profiles. To test whether the association between a lncRNA and a pcRNA was significant, we applied the Fisher transformation on the previously calculated correlation coefficients. Note that p-values depended on the sample number. Thus, due to the large sample size, lncRNA – pcRNA associations tested as significant already for moderate absolute correlation coefficients. Accordingly, we additionally constrained associations to absolute values above Pearson correlation of 0.4. If a lncRNA – pcRNA correlation meat the criteria, the pair was further considered and referred to as associated.
Among the associated lncRNA – pcRNA pairs, we distinguished significant correlation between HOTAIRM1 and the HOX genes – HOXA4, HOXB2, HOXB3 and HOXB4. The identified correlation agreed with the previously established correlation between HOTAIRM1 and HOX genes [230, 231].

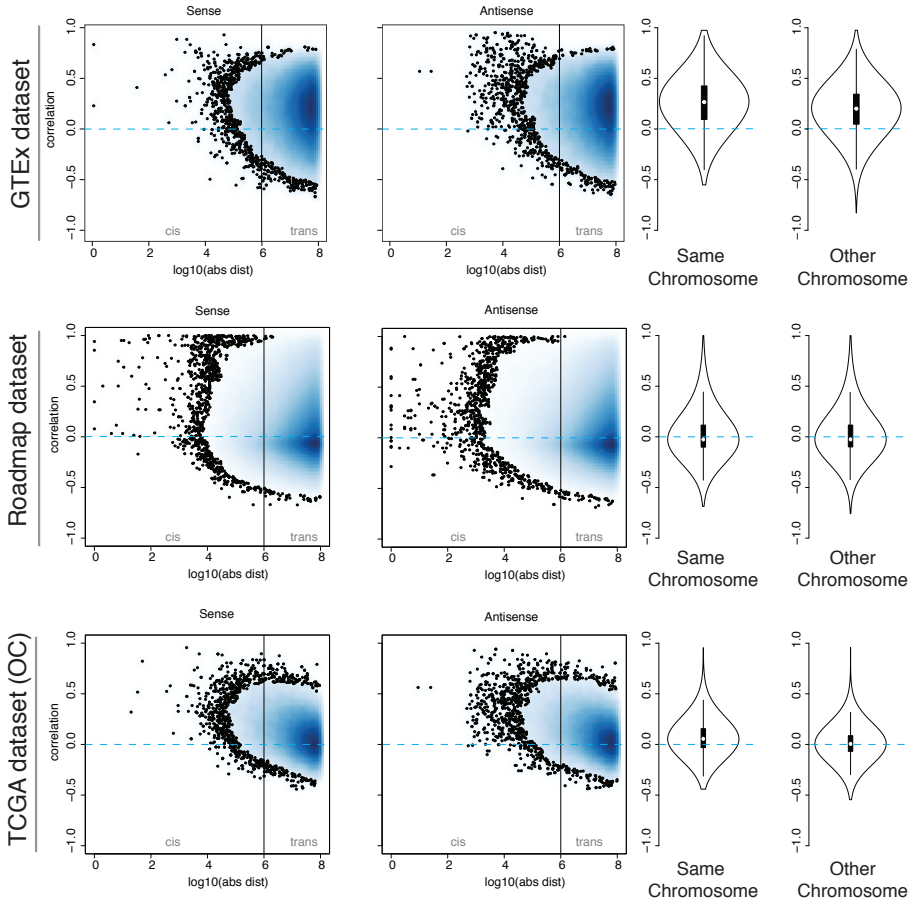

Figure 7.2: **Relationships between lncRNA – pcRNA correlation strength (y-axis) and distances between the lncRNA and the pcRNA (x-axis) across GTEx, Roadmap and TCGA OV.** *Trans* region indicated associations between lncRNAs and pcRNAs situated at a distance >1Mb or on different chromosomes. The violin plots summarized the distribution of correlation of lncRNA – pcRNA pairs located within *cis* (regions on the same chromosome) or *trans* (regions on different chromosomes).

7.1.3 Genomic proximities

Since lncRNAs were shown to be co-expressed with cis neighboring genes [225] and to interact with genes located on the same chromosome [232], we analyzed the interdependence between lncRNA – pcRNA correlations and their corresponding genomic proximities. For this purpose, we merged the lncRNA and the pcRNA comprehensive genome annotation lists available in the GENCODE Release 23, by chromosome. Next, we calculated the absolute distance between the transcription start site of lncRNA and pcRNAs (genome assembly hg19 and ENSEMBL Gene identifier).

We observed a trend towards positive correlations between lncRNA and pcRNAs across all data sets, irrespective of their genomic proximity (Figure 7.2). The orientation of the pairs did not influence the correlation direction (sense and antisense). Notably, in Roadmap and GTEx samples, lncRNAs and pcRNAs showed strong positive correlation 0.99, while in the TCGA OV data we observed a decrease in correlation. This suggested changes in the interactions between lncRNAs and pcRNAs in ovarian tumour tissue.

7.1.4 Similarity of lncRNA – pcRNA correlation across different data sets

To determine whether lncRNAs were associated with the same pcRNAs across different data sets, we compared sets of correlated pcRNA sets for each expressed lncRNA across Roadmap, GTEx and TCGA OV (Figure 7.3). To test for similarity across all three data sets, we selected only lncRNAs and pcRNAs expressed across all data sets and we computed the corresponding Jaccard coefficients (Figure 7.3B). The Jaccard coefficient has been established as a statistic used for comparing the similarity of two data sets [233]. Figure 7.3 suggested a higher overlap between lncRNA – pcRNA correlated pairs in GTEx and Roadmap (the highest Jaccard coefficient is 0.3 for Roadmap – TCGA OV and Roadmap – GTEx, and 0.5 in GTEx – TCGA OV). To test if the overlap between the GTEx and Roadmap data was significantly higher than the overlap between GTEx and TCGA OV and Roadmap and TCGA OV, we applied a

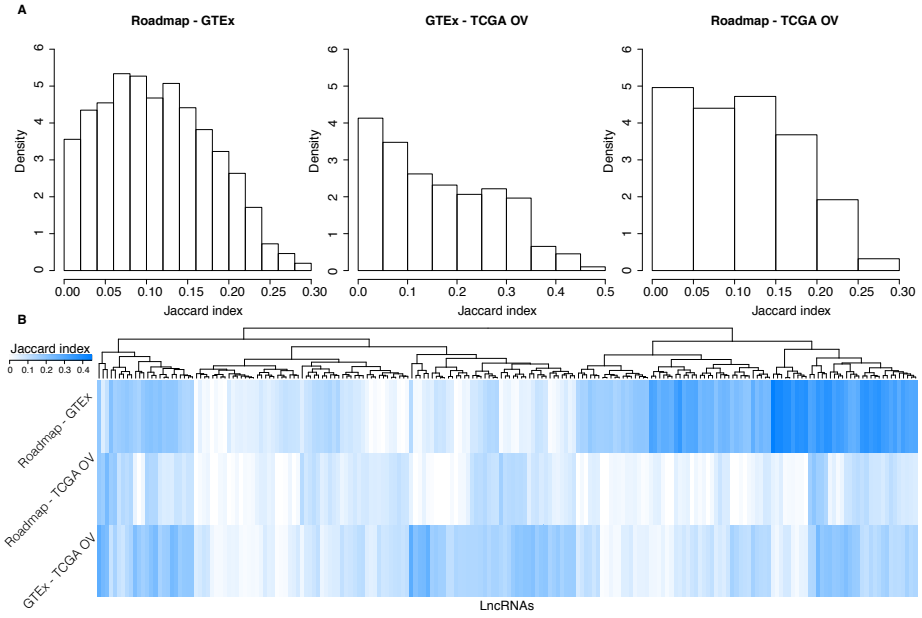

Figure 7.3: **Similarity between lncRNA-correlated pcRNA sets across Roadmap, GTEx and TCGA OV** A. Distribution of Jaccard coefficients for lncRNA-correlated pcRNA sets across the three data sets for each combination of two data sets. B. Overlap of lncRNA-correlated pcRNA sets expressed in all data sets. The columns represented the commonly expressed lncRNAs.

Wilcoxon Signed-Rank test. We found that indeed the distribution of Jaccard coefficient in Roadmap - GTEx was significantly shifted when compared to the other two distributions (p.value $< 10^{-6}$).

The differential correlation between lncRNAs and pcRNAs in the TCGA OV data and the correlation between lncRNAs and pcRNAs in the other data sets suggested tissue-specific expression regulation.

7.2 Tissue enrichment analysis revealed lncRNAs specific to blood-related and liver tissues

To assess whether the lncRNA expression was tissue specific, we explored data from the EBI Expression Atlas, Roadmap and GTEx. Since TCGA OV represented data from one tissue, it was not used in the analysis. We used the EBI Expression Atlas resource to map pcRNAs expression to human tissue

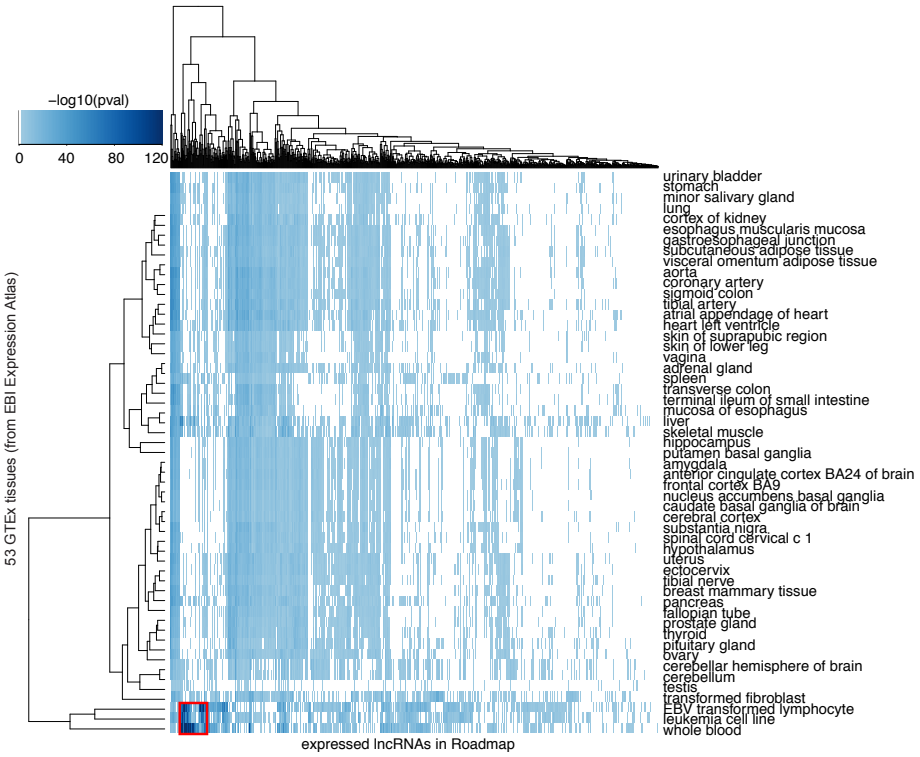

Figure 7.4: **Tissue specificity of lncRNAs across Roadmap data.** Highlighted cluster of lncRNAs (red) indicated high association with blood-related tissues. The color coding of the heat map denoted FDR-adjusted p-values. Significant associations (FDR $< 1\%$) were color-coded in blue, whereas non-significant ones were represented in white.

[234]. The expression profiles of pcRNAs from the EBI portal were obtained from eight projects: ENCODE, GENETECH, FANTOM5, GTEX, ILLUMINA BODY MAP, NCI60 CANCER, MAMALIAN KAESSMANN and UHKENS LAB. Within the EBI data, we considered pcRNAs as being expressed if the fragments of transcript per million mapped reads was higher than 0.5 (FPKM>0.5) – the default threshold used by EBI.

To determine the significance of the tissue specificity, we applied a Fischer's exact test. In particular, for each lncRNA expressed in the Roadmap/GTEx data set, we tested whether the set of correlated genes was over-represented in the tissue-specific genes within each EBI project. We defined as background set the overlap between all the pcRNAs expressed in Roadmap/GTEx and pcRNAs

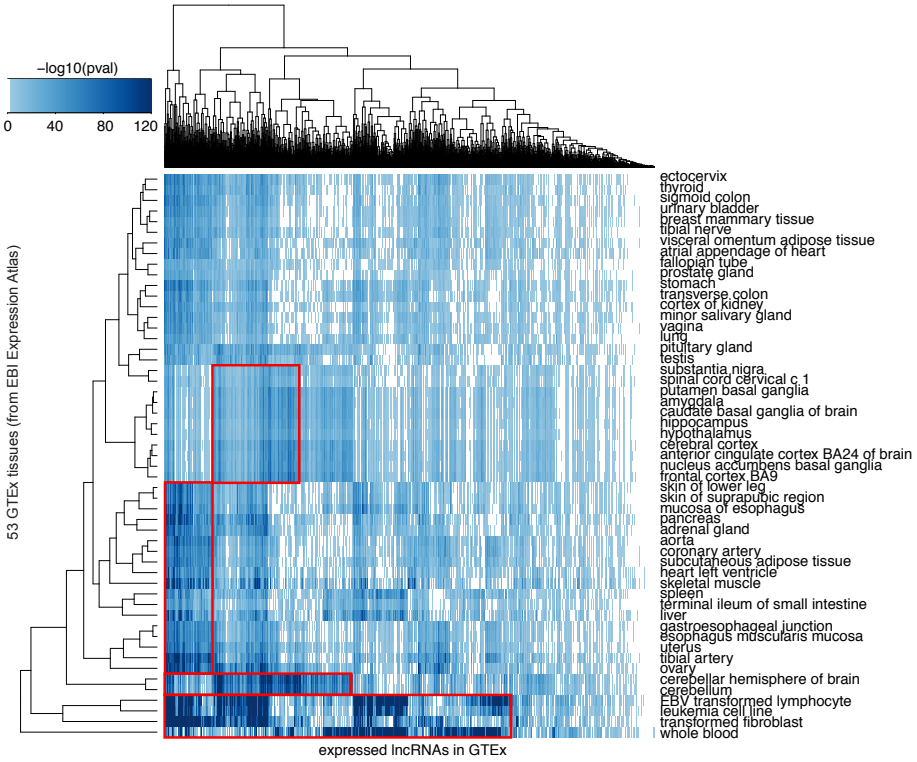

Figure 7.5: **Tissue specificity of lncRNAs across GTEx data.** Highlighted clusters of lncRNAs (red) showed high association with blood-related tissues. The color coding of the heat map denotes FDR-adjusted p-values. Significant associations (FDR $< 1\%$) were indicated with blue, whereas non-significant ones were indicated with white.

expressed within each EBI project. To correct for multiple testing, we then applied an FDR-based correction to the resulting p-values [116].

The Roadmap results revealed a cluster of lncRNAs that significantly associated with EBV transformed lymphocyte, whole blood and leukemia cell line tissues (Figure 7.4). Specifically, the cluster included 13 lncRNAs: RP4-639F20.1, LINC01123, LINC01106, LINC00883, CTD-2015H6.3, LINC00472, BAALC-AS1, RP11-475I24.3, RP11-111F5.4, RP11-211N8.2, LINC01503, RP11-539I5.1 and RP1-122P22.2. Of these, LINC01123 was shown to be part of a plasma lncRNA signature that distinguished aggressive/malignant intraductal papillary mucinous neoplasms [235]. The published results confirmed the specificity of the clustered lncRNAs to blood-related tissue.

The lncRNA specificity pattern to whole blood, leukemia cell line and liver tissues remained consistent in the GTEx samples (Figure 7.5). Additionally, we distinguished several lncRNA clusters specific to brain, heart, liver and skin tissues. The remaining tissue expression data sets form the EBI Expression Atlas also showed comparable results.
In summary, blood-related and liver tissues showed a reproducible and characteristic expression profile for lncRNA-associated genes.

7.3 LncRNA functional analysis suggested involvement in translational regulation

In order to identify the functional roles of lncRNAs, we examined the association between the pcRNAs associated with lncRNA expression, and the biological pathways from WikiPathways. For this purpose, we used the model-based ontology analysis (MGSA) [236], which was built to cope effectively with redundancies in ontologies.
For each lncRNA, we applied MGSA on the set of associated pcRNAs and calculated term posterior probabilities of gene sets retrieved from WikiPathways. The term probabilities denoted the enrichment of lncRNA-correlated genes in each pathway, e.g. the strength of the association between a lncRNA and a biological pathway. Given the underlying pcRNAs set, pathways were considered active if the pathway probability was above 50%.

Functional annotation of lncRNAs in Roadmap identified three pathways, which were associated with clusters of lncRNAs, namely proteasome degradation, T-cell receptor (TCR) receptor singling and mRNA processing (Figure 7.6). These results suggested involvement of lncRNAs in translational regulation. Figure 7.7 revealed clusters of functionally highly similar lncRNAs in the GTEx data set. In accordance with the blood tissue-specific expression results, we distinguished a cluster of lncRNAs which were also functionally associated with immune system related signaling pathways. Involvement of lncRNA in gene regulation was again supported by association with active DNA replication and energy active pathways. Additionally, we identified a group of lncRNAs that were associated with metabolism-related pathways.

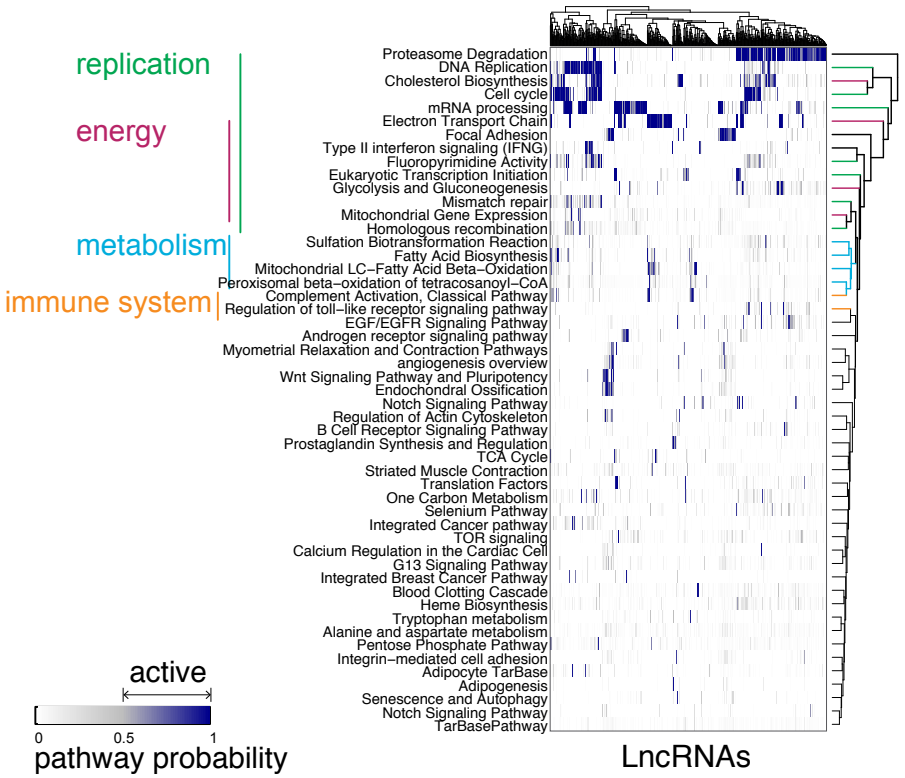

Figure 7.6: **Guilty-by-association: MGSA functional annotation of pcRNAs that were significantly correlated with lncRNAs within the Roadmap data.** WikiPathways with probabilities ranging from 0(white) to 1(dark blue). All pathways with a probability > 0.5 were considered active.

To validate the functional associations across Roadmap, we further analyzed the GTEx data. Since GTEx included 7,049,286 associated lncRNA – pcRNA pairs (while Roadmap includes only 919,121 associated lncRNA – pcRNA pairs), we expected finding a higher variety of functional pathways active. Indeed, we distinguished a higher number of active pathways, among which we again identified the replication-related pathways, the immune system signaling, metabolic and energy-related pathways. Additionally, we identified clusters of lncRNAs associated with early development, hormone-response and breast cancer pathways.

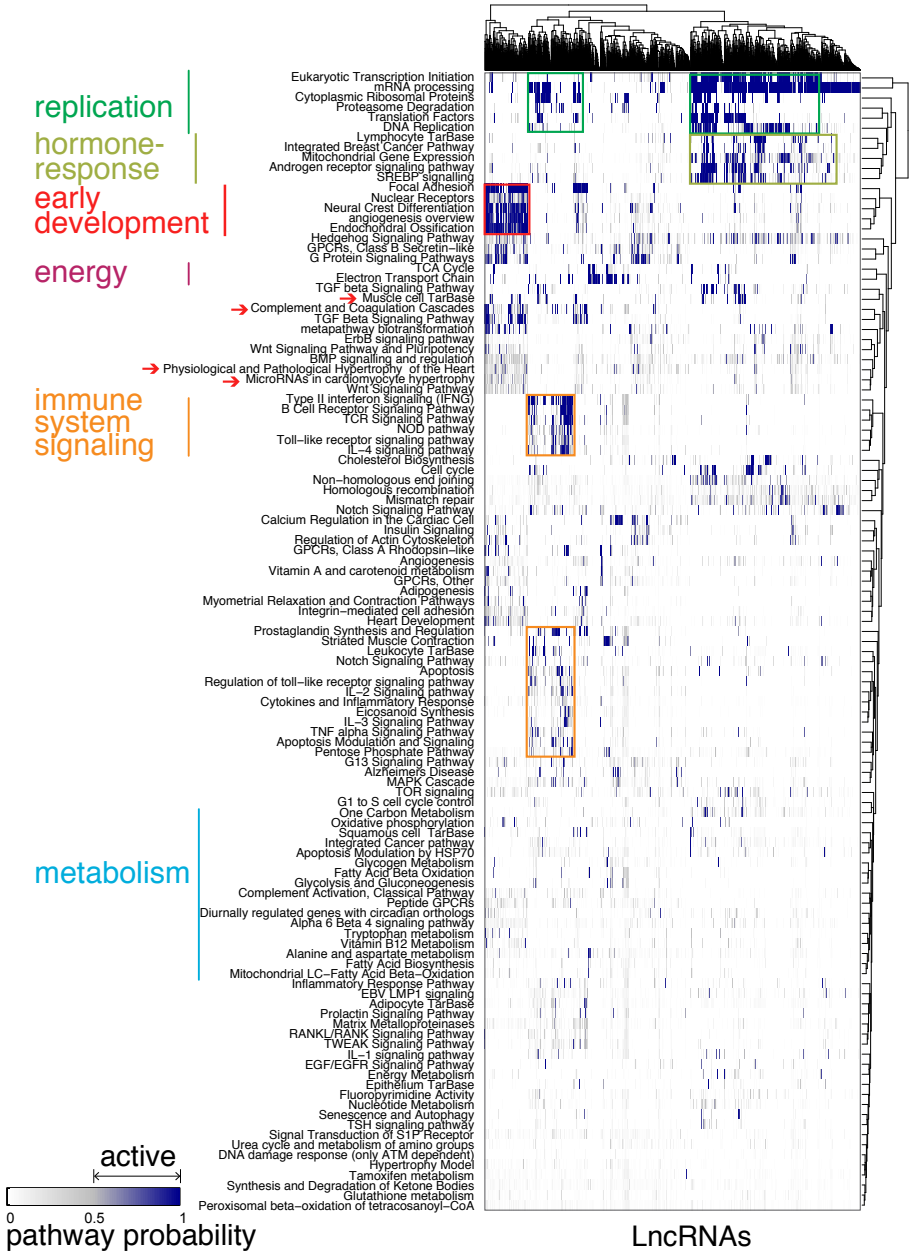

Figure 7.7: **Guilty-by-association: MGSA functional annotation for pcRNAs that were significantly correlated with lncRNAs within the GTEx data.** WikiPathways with probabilities ranging from 0(white) to 1(dark blue). All pathways with a probability > 0.5 were considered active.

7.4 Discussion and Conclusion

In this chapter, we introduced a multilevel data analysis framework for exploring lncRNA molecular functions – LISA. With this approach we aimed to identify lncRNA function in cellular processes through integrating multilevel data across multiple large-scale next-generation projects. The different modules of LISA consisted of lncRNA and protein coding gene expression analysis, similarity analysis across different data sets, tissue-specificity analysis and functional analysis of lncRNAs. The essential idea of LISA was to use guilt-by-association assumptions, which linked lncRNAs to pcRNAs. As a result, for each lncRNA expressed in a data set, one could define a set of pcRNAs that were associated with the respective lncRNA. The defined sets could then be used to find associations with functional pathways or to test for tissue-specificity.

This study focused on evaluating the tissue-specificity in two data sets: Roadmap and GTEx. The results revealed groups of lncRNAs with specific expression in blood, liver, brain, heart and skin related tissues. Specifically, our tissue enrichment analysis distinguised a lncRNA cluster including LINC01123 – which was shown to be part of a plasma lncRNA signature that distinguished aggressive/malignant intraductal papillary mucinous neoplasms [235]. Furthermore, the MGSA functional analysis within the same data sets confirmed the role that lncRNAs play in translation regulation, but also revealed associations between groups of lncRNAs and cellular maintenance and immune system signaling pathways.

Even though the results of the individual analysis provided valuable knowledge for a better understanding of the molecular mechanisms of lncRNAs, the most comprehensive and consistent results could only be obtained by the integrative approach of multiple molecular levels.

Chapter 8

Conclusion

The molecular mechanisms underlying cancer are highly complex and, as seen throughout this thesis, involve molecules from multilevel omics: DNA, RNA, proteins. Genomics, transcriptomics and proteomics profiles are frequently used for both oncotarget discovery and precision medicine, but also to decode cancer-specific mechanisms for fundamental research. Understanding the functional relationships between different omics levels in tumorigenesis, remains however a challenging task.

While technological innovations have enabled tumor omics profiling at unprecedented scale, depth, and speed, most studies are still assessing and learning from single-level omics. For example, while projects as TCGA made available large-scale manifold cancer data, multiple studies still focused on knowledge learnt from transcriptomics only: Wu et. al used transcription levels to find breast cancer risk genes [237], while Uhlen et. al predicted a pathology atlas for human cancer transcriptome [238].

It is critical to acknowledge that different omics levels are not isolated and need to be analyzed and interpreted in the context of complex and dynamic molecular processes through effective integrative models. However, formulating such models to mimic the initiation and evolution of tumorigenesis remains one of the most complex unsolved tasks in the scientific research field.

In this thesis we examined multiple cancer data and provided integrative models of different data modalities. We aimed to distinguish cancer-specific regulatory

interactions between different molecular levels, so that we can better understand the interplay between multilevel omics that characterizes malignant cell proliferation.

Another essential aspect of integrative methods often overlooked is the feasibility of integrative approaches on multiple levels and how it affects the method demand. We discussed and addressed this aspect together with the level-wise variety of data types across four main studies that analyzed and integrated DNA measurements, mRNA, miRNA and lncRNA expression, as well as measurements of physical interactions between human and viral proteins. We particularly examined data from either one of the next four cancer types: head and neck, liver, lung or ovarian tumours.

The following section summarizes the scientific contributions developed in this thesis, and present possible extensions and future directions.

Scientific contributions

Benchmarking study of commonly used CNA calling algorithms reveals significant effect of tumour purity and CNA burden on performance

One important aspect of cancer cells is the presence of DNA changes. Accurately predicting DNA changes that are involved in tumor development represents a challenging task. We showed tumor purity significantly influenced the performance of commonly used CNA calling algorithms (Chapter 4). Additionally, our study revealed another variable that influenced the performance of the algorithms: the CNA burden. Considering the identified CNA burden effect on CGHcall, we developed an adjusted version of the algorithm that corrects for the confounding effect. Next, we showed how integrating data as DNA changes with several other data levels can reveal putative biomarkers in viral liver cancer.

Joint significance of genomic and physical viral-human interactions in liver cancer

In our next study, we introduced a multilevel integrative analysis to examine the impact of ongoing Hepatitis B viral infection in HCC cases (Chapter 5). We estimated the viral impact on mutation status with a Bayesian logistic regression model. In the next step, we examined another data layer that reflected the strength of viral-human physical interactions at the protein level. We estimated the impact at both genomic and physical levels within the ReactomeFi PPI network by using network propagation. Since we were interested in protein genes that were affected at both genomic and physical level, we proposed and calculated a joint significance representing a confidence score for the given interaction.

Our approach allowed us to identify those viral-human interacting protein genes that were underrepresented in individual data analysis. Based on the confidence scores, we were able to formulate, test, and validate new hypotheses related to the impact of Hepatitis B on phosphorylation and ubiquitylation.

Most importantly, we showed that our multilevel integrative approach revealed relevant pathways for oncogenesis in liver cancer with ongoing Hepatitis B infection.

Multilevel integrative approaches for distinguishing functional roles of non-coding RNAs in cancer: miRNAs and lncRNAs

While the importance of protein-coding RNAs has been thoroughly investigated, the functional roles of non-coding RNAs are yet to be fully characterized.

In the projects introduced in Chapters 6 and 7, we aimed to reveal properties of miRNA and lncRNA in regulation.

We used an elastic-net based multiple regression model to infer a miRNA-mRNA regulatory network for HPV-associated miRNAs in HNSCC. Subsequently, we performed a local enrichment analysis that identified two functional clusters of miRNAs that were predicted to mediate HPV-associated dysregulation in HNSCC (Chapter 6).

Using miRlastic, we were able to include prior information together with condition-

based transcriptomics for inferring the miRNA-mRNA interactions specific to HNSCC and LUAD. Furthermore, miRlastic allowed us to infer how miRNAs contribute to the disruption of molecular pathways in HNSCC and LUAD. Thus, our approach allows scientists to comprehend the changes in expressions related to a specific condition, and to reveal miRNA functions that are activated by such a specific condition.

Given the lack of a comprehensive study on lncRNA functional roles and the availability of large public data sets, we next introduced an integrative analysis of broad-scale multilevel data that aimed to explore the molecular functions of lncRNAs. We first explored the correlation patterns between lncRNAs and protein-coding genes across different data sets. Next, we performed a tissue enrichment analysis of lncRNAs in multiple data sets. The results revealed a consistent specificity of a plasma lncRNA signature to blood-related tissues both in GTEx and Roadmap. Ultimately, our pipeline used MGSA to examine the association between protein-coding genes correlated with lncRNAs and the molecular pathways defined in WikiPathways. We distinguished clusters of lncRNAs that correlated with pathways involved in metabolism, immune system, and DNA replication.

The resulting pipeline offers scientists an approach that addresses multiple aspects of lncRNA involvement in molecular processes. Using LISA, scientists can assimilate the advancement offered by public repositories like GTEx, Encode, and TCGA, and use it to predict lncRNA functions in epigenetics and transcriptomics across various tissues and genomic regions.

Our methods for examining the roles of non-coding RNAs served as a fundament for biomedical researchers to formulate and evaluate alternative hypothesis – e.g. our collaboration with the Breuhahn group [140].

Extensions and future directions

The success of developing and optimizing integrative models that learn from multiple biological data modalities heavily relies not only the mathematical prerequisites and understanding of the given data, but also on the biological interpretation of the input, output, and performance of a model.

Following, every model and pipeline presented within this thesis can be extended as follows when considering these two aspects:

- From a methodological point of view, our benchmarking study can be extended to evaluate the performance of more algorithms, different data technologies (whole genome sequencing) and different data resolution (single-cell sequencing).

  Going beyond that, an interdisciplinary analysis can reveal elements involved in generating DNA copy number changes in carcinogenesis. For example, the extent of DNA damage in cancer may depend on clinical parameters such as tumour stage or tumour tissue type [239]. Subsequent extensions of our pipeline can include evaluating the effect of such parameters. Gathering experimental validation data sets that cover an increasing area of the DNA could also contribute to an improved evaluation.

- Our integrative pipeline that aims to identify more compelling evidence for interactions between human and viral proteins relies on a Bayesian logistic model and network propagation. Although, our approach provides a joint significance for protein interactions from two data modalities, there are options for a simultaneously learning model. For example, we can consider using the mutation status and physical interactions scores as attributes for protein nodes and develop a graphical neural network model for estimating the joint effect of these attributes on the reference network.

  Another way to improve the sensitivity of the model is to use the viral expression instead of a binary vector. This would indicate how active is the virus in the sample and how strong is the association between viral expression and mutation status.

  Given the modular aspect of our design, researchers can easily adapt our approach to other biological conditions.

- When examining the functional roles of non-coding RNAs we could also integrate additional molecular levels that can describe their activity: overlapping SNPs, CNAs, or point mutations. Adding imaging data of protein localization (available on TCGA) could also pinpoint which are the cellular elements involving mechanisms of action of non-coding RNAs.

Final statement

Complex diseases like cancer function on multiple molecular levels that can be quantitatively measured with different high-throughput omics experimental technologies. In this thesis, we addressed the challenge of formulating befitting models to explain the interactions between the different omics layers in cancer. Our approaches revealed known cancer-specific molecular interactions, but also provided novel insights for further experimental validation. Moreover, our methods and pipelines allowed us to develop collaborations that catalyzed the progress of translational biology and to demonstrate unknown cancer-specific interactions at various data levels, elucidating key concepts in cancer biology. We thereby provided integrative approaches that can provide valuable insights on cancer-specific multilevel omics interactions.

Bibliography

[1] C. Fitzmaurice, C. Allen, R. M. Barber, L. Barregard, Z. A. Bhutta, H. Brenner, D. J. Dicker, O. Chimed-Orchir, R. Dandona, L. Dandona, T. Fleming, M. H. Forouzanfar, J. Hancock, R. J. Hay, R. Hunter-Merrill, C. Huynh, H. D. Hosgood, C. O. Johnson, J. B. Jonas, J. Khubchandani, G. A. Kumar, M. Kutz, Q. Lan, H. J. Larson, X. Liang, S. S. Lim, A. D. Lopez, M. F. MacIntyre, L. Marczak, N. Marquez, A. H. Mokdad, C. Pinho, F. Pourmalek, J. A. Salomon, J. R. Sanabria, L. Sandar, B. Sartorius, S. M. Schwartz, K. A. Shackelford, K. Shibuya, J. Stanaway, C. Steiner, J. Sun, K. Takahashi, S. E. Vollset, T. Vos, J. A. Wagner, H. Wang, R. Westerman, H. Zeeb, L. Zoeckler, F. Abd-Allah, M. B. Ahmed, S. Alabed, N. K. Alam, S. F. Aldhahri, G. Alem, M. A. Alemayohu, R. Ali, R. Al-Raddadi, A. Amare, Y. Amoako, A. Artaman, H. Asayesh, N. Atnafu, A. Awasthi, H. B. Saleem, A. Barac, N. Bedi, I. Bensenor, A. Berhane, E. Bernabé, B. Betsu, A. Binagwaho, D. Boneya, I. Campos-Nonato, C. Castañeda-Orjuela, F. Catalá-López, P. Chiang, C. Chibueze, A. Chitheer, J.-Y. Choi, B. Cowie, S. Damtew, J. das Neves, S. Dey, S. Dharmaratne, P. Dhillon, E. Ding, T. Driscoll, D. Ekwueme, A. Y. Endries, M. Farvid, F. Farzadfar, J. Fernandes, F. Fischer, T. T. G/hiwot, A. Gebru, S. Gopalani, A. Hailu, M. Horino, N. Horita, A. Husseini, I. Huybrechts, M. Inoue, F. Islami, M. Jakovljevic, S. James, M. Javanbakht, S. H. Jee, A. Kasaeian, M. S. Kedir, Y. S. Khader, Y.-H. Khang, D. Kim, J. Leigh, S. Linn, R. Lunevicius, H. M. A. E. Razek, R. Malekzadeh, D. C. Malta, W. Marcenes, D. Markos, Y. A. Melaku, K. G. Meles, W. Mendoza, D. T. Mengiste, T. J. Meretoja, T. R. Miller, K. A. Mohammad, A. Mohammadi, S. Mohammed, M. Moradi-Lakeh,

G. Nagel, D. Nand, Q. L. Nguyen, S. Nolte, F. A. Ogbo, K. E. Oladimeji, E. Oren, M. Pa, E.-K. Park, D. M. Pereira, D. Plass, M. Qorbani, A. Radfar, A. Rafay, M. Rahman, S. M. Rana, K. Søreide, M. Satpathy, M. Sawhney, S. G. Sepanlou, M. A. Shaikh, J. She, I. Shiue, H. R. Shore, M. G. Shrime, S. So, S. Soneji, V. Stathopoulou, K. Stroumpoulis, M. B. Sufiyan, B. L. Sykes, R. Tabarés-Seisdedos, F. Tadese, B. A. Tedla, G. A. Tessema, J. S. Thakur, B. X. Tran, K. N. Ukwaja, B. S. C. Uzochukwu, V. V. Vlassov, E. Weiderpass, M. W. Terefe, H. G. Yebyo, H. H. Yimam, N. Yonemoto, M. Z. Younis, C. Yu, Z. Zaidi, M. E. S. Zaki, Z. M. Zenebe, C. J. L. Murray, and M. Naghavi, "Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015," *JAMA Oncology*, vol. 3, p. 524, apr 2017.

[2] B. Alberts, D. Bray, K. Hopkin, A. D. Johnson, K. Roberts, and J. Lewis, *Essential Cell Biology*. Butterworth-Heinemann, 4th ed., 2013.

[3] G. A. Colditz, T. A. Sellers, and E. Trapido, "Epidemiology — identifying the causes and preventability of cancer?," *Nature Reviews Cancer*, vol. 6, pp. 75–83, dec 2005.

[4] F. Biemar and M. Foti, "Global progress against cancer-challenges and opportunities.," *Cancer biology & medicine*, vol. 10, pp. 183–186, Dec. 2013.

[5] D. Hanahan and R. A. Weinberg, "Hallmarks of Cancer: The Next Generation," *Cell*, vol. 144, pp. 646–674, Mar. 2011.

[6] C. Willyard, "Cancer therapy: an evolved approach," *Nature*, vol. 532, pp. 166–168, apr 2016.

[7] G. R. Bignell, C. D. Greenman, H. Davies, A. P. Butler, S. Edkins, J. M. Andrews, G. Buck, L. Chen, D. Beare, C. Latimer, S. Widaa, J. Hinton, C. Fahey, B. Fu, S. Swamy, G. L. Dalgliesh, B. T. Teh, P. Deloukas, F. Yang, P. J. Campbell, P. A. Futreal, and M. R. Stratton, "Signatures of mutation and selection in the cancer genome," *Nature*, vol. 463, pp. 893–898, feb 2010.

[8] X.-J. Ma, R. Salunga, J. T. Tuggle, J. Gaudet, E. Enright, P. McQuary, T. Payette, M. Pistone, K. Stecker, B. M. Zhang, Y.-X. Zhou, H. Varnholt, B. Smith, M. Gadd, E. Chatfield, J. Kessler, T. M. Baer, M. G. Erlander, and D. C. Sgroi, "Gene expression profiles of human breast cancer progression," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 5974–5979, apr 2003.

[9] G. A. Calin, M. Ferracin, A. Cimmino, G. D. Leva, M. Shimizu, S. E. Wojcik, M. V. Iorio, R. Visone, N. I. Sever, M. Fabbri, R. Iuliano, T. Palumbo, F. Pichiorri, C. Roldo, R. Garzon, C. Sevignani, L. Rassenti, H. Alder, S. Volinia, C. gong Liu, T. J. Kipps, M. Negrini, and C. M. Croce, "A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia," *New England Journal of Medicine*, vol. 353, pp. 1793–1801, oct 2005.

[10] E. A. Vucic, K. L. Thu, K. Robison, L. A. Rybaczyk, R. Chari, C. E. Alvarez, and W. L. Lam, "Translating cancer 'omics' to improved outcomes," *Genome Research*, vol. 22, pp. 188–195, feb 2012.

[11] D. Branzei and M. Foiani, "Regulation of DNA repair throughout the cell cycle," *Nature Reviews Molecular Cell Biology*, vol. 9, pp. 297–308, feb 2008.

[12] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nature Genetics*, vol. 36, pp. 949–951, aug 2004.

[13] D. F. Conrad, , D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles, "Origins and functional impact of copy number variation in the human genome," *Nature*, vol. 464, pp. 704–712, oct 2009.

[14] A. M. Rice and A. McLysaght, "Dosage sensitivity is a major determinant of human copy number variant pathogenicity," *Nature Communications*, vol. 8, p. 14366, feb 2017.

[15] E. R. Gamazon and B. E. Stranger, "The impact of human copy number variation on gene expression," *Briefings in Functional Genomics*, vol. 14, pp. 352–357, apr 2015.

[16] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis, "Relative impact of nucleotide and copy number variation on gene expression phenotypes," *Science*, vol. 315, pp. 848–853, feb 2007.

[17] J. R. Lupski, "Genome structural variation and sporadic disease traits," *Nature Genetics*, vol. 38, pp. 974–976, sep 2006.

[18] L. A. Pérez Jurado, R. Peoples, P. Kaplan, B. C. Hamel, and U. Francke, "Molecular definition of the chromosome 7 deletion in williams syndrome and parent-of-origin effects on growth.," *American journal of human genetics*, vol. 59, pp. 781–792, Oct. 1996.

[19] S. A. McCarroll, A. Huett, P. Kuballa, S. D. Chilewski, A. Landry, P. Goyette, M. C. Zody, J. L. Hall, S. R. Brant, J. H. Cho, R. H. Duerr, M. S. Silverberg, K. D. Taylor, J. D. Rioux, D. Altshuler, M. J. Daly, and R. J. Xavier, "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and crohn's disease," *Nature Genetics*, vol. 40, pp. 1107–1112, aug 2008.

[20] R. G. Walters, S. Jacquemont, A. Valsesia, A. J. de Smith, D. Martinet, J. Andersson, M. Falchi, F. Chen, J. Andrieux, S. Lobbens, B. Delobel, F. Stutzmann, J. S. E.-S. Moustafa, J.-C. Chèvre, C. Lecoeur, V. Vatin, S. Bouquillon, J. L. Buxton, O. Boute, M. Holder-Espinasse, J.-M. Cuisset, M.-P. Lemaitre, A.-E. Ambresin, A. Brioschi, M. Gaillard, V. Giusti, F. Fellmann, A. Ferrarini, N. Hadjikhani, D. Campion, A. Guilmatre, A. Goldenberg, N. Calmels, J.-L. Mandel, C. L. Caignec, A. David, B. Isidor, M.-P. Cordier, S. Dupuis-Girod, A. Labalme, D. Sanlaville,

M. Béri-Dexheimer, P. Jonveaux, B. Leheup, K. Õunap, E. G. Bochukova, E. Henning, J. Keogh, R. J. Ellis, K. D. MacDermot, M. M. van Haelst, C. Vincent-Delorme, G. Plessis, R. Touraine, A. Philippe, V. Malan, M. Mathieu-Dramard, J. Chiesa, B. Blaumeiser, R. F. Kooy, R. Caiazzo, M. Pigeyre, B. Balkau, R. Sladek, S. Bergmann, V. Mooser, D. Waterworth, A. Reymond, P. Vollenweider, G. Waeber, A. Kurg, P. Palta, T. Esko, A. Metspalu, M. Nelis, P. Elliott, A.-L. Hartikainen, M. I. McCarthy, L. Peltonen, L. Carlsson, P. Jacobson, L. Sjöström, N. Huang, M. E. Hurles, S. O'Rahilly, I. S. Farooqi, K. Männik, M.-R. Jarvelin, F. Pattou, D. Meyre, A. J. Walley, L. J. M. Coin, A. I. F. Blakemore, P. Froguel, and J. S. Beckmann, "A new highly penetrant form of obesity due to deletions on chromosome 16p11.2," *Nature*, vol. 463, pp. 671–675, feb 2010.

[21] S. B. Cassidy and D. J. Driscoll, "Prader–willi syndrome," *European Journal of Human Genetics*, vol. 17, pp. 3–13, sep 2008.

[22] L. Kalsner and S. J. Chamberlain, "Prader-willi, angelman, and 15q11-q13 duplication syndromes," *Pediatric Clinics of North America*, vol. 62, pp. 587–606, jun 2015.

[23] M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, K. Stephens, and A. Amemiya, *GeneReviews. 15q Duplication Syndrome and Related Disorders.* University of Washington, Seattle, 1993.

[24] A. Dagli, K. Buiting, and C. A. Williams, "Molecular and clinical aspects of angelman syndrome.," *Molecular syndromology*, vol. 2, pp. 100–112, Apr. 2012.

[25] C. DiStefano, A. Gulsrud, S. Huberty, C. Kasari, E. Cook, L. T. Reiter, R. Thibert, and S. S. Jeste, "Identification of a distinct developmental and behavioral profile in children with dup15q syndrome," *Journal of Neurodevelopmental Disorders*, vol. 8, may 2016.

[26] E. D. Pleasance, R. K. Cheetham, P. J. Stephens, D. J. McBride, S. J. Humphray, C. D. Greenman, I. Varela, M.-L. Lin, G. R. Ordóñez, G. R. Bignell, K. Ye, J. Alipaz, M. J. Bauer, D. Beare, A. Butler, R. J. Carter,

L. Chen, A. J. Cox, S. Edkins, P. I. Kokko-Gonzales, N. A. Gormley, R. J. Grocock, C. D. Haudenschild, M. M. Hims, T. James, M. Jia, Z. Kingsbury, C. Leroy, J. Marshall, A. Menzies, L. J. Mudie, Z. Ning, T. Royce, O. B. Schulz-Trieglaff, A. Spiridou, L. A. Stebbings, L. Szajkowski, J. Teague, D. Williamson, L. Chin, M. T. Ross, P. J. Campbell, D. R. Bentley, P. A. Futreal, and M. R. Stratton, "A comprehensive catalogue of somatic mutations from a human cancer genome," *Nature*, vol. 463, pp. 191–196, Jan. 2010.

[27] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. M. Henry, R. M. Pinchback, A. H. Ligon, Y.-J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M.-S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, and M. Meyerson, "The landscape of somatic copy-number alteration across human cancers," *Nature*, vol. 463, pp. 899–905, feb 2010.

[28] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, pp. 719–724, apr 2009.

[29] N. Bardeesy, K.-h. Cheng, J. H. Berger, G. C. Chu, J. Pahler, P. Olson, A. F. Hezel, J. Horner, G. Y. Lauwers, D. Hanahan, and R. A. DePinho, "Smad4 is dispensable for normal pancreas development yet critical in progression and tumor biology of pancreas cancer," *Genes Dev.*, vol. 20, pp. 3130–3146, Nov. 2006.

[30] A. K. Witkiewicz, E. A. McMillan, U. Balaji, G. Baek, W.-C. Lin, J. Mansour, M. Mollaee, K.-U. Wagner, P. Koduru, A. Yopp, M. A. Choti, C. J. Yeo, P. McCue, M. A. White, and E. S. Knudsen, "Whole-exome sequenc-

ing of pancreatic cancer defines genetic diversity and therapeutic targets," *Nature Communications*, vol. 6, p. 6744, Apr. 2015.

[31] E. Leucci, R. Vendramin, M. Spinazzi, P. Laurette, M. Fiers, J. Wouters, E. Radaelli, S. Eyckerman, C. Leonelli, K. Vanderheyden, A. Rogiers, E. Hermans, P. Baatsen, S. Aerts, F. Amant, S. Van Aelst, J. van den Oord, B. de Strooper, I. Davidson, D. L. J. Lafontaine, K. Gevaert, J. Vandesompele, P. Mestdagh, and J.-C. Marine, "Melanoma addiction to the long non-coding RNA SAMMSON," *Nature*, vol. 531, pp. 518–522, Mar. 2016.

[32] Cancer Genome Atlas Network, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature*, vol. 517, pp. 576–582, Jan. 2015.

[33] V. L. Bauer, H. Braselmann, M. Henke, D. Mattern, A. Walch, K. Unger, M. Baudis, S. Lassmann, R. Huber, J. Wienberg, M. Werner, and H. F. Zitzelsberger, "Chromosomal changes characterize head and neck cancer with poor prognosis," *J Mol Med*, vol. 86, pp. 1353–1365, Dec. 2008.

[34] J. Hess, K. Unger, M. Orth, U. Schötz, L. Schüttrumpf, V. Zangen, I. Gimenez-Aznar, A. Michna, L. Schneider, R. Stamp, M. Selmansberger, H. Braselmann, L. Hieber, G. A. Drexler, S. Kuger, D. Klein, V. Jendrossek, A. A. Friedl, C. Belka, H. Zitzelsberger, and K. Lauber, "Genomic amplification of fanconi anemia complementation group a (FancA) in head and neck squamous cell carcinoma (HNSCC): Cellular mechanisms of radioresistance and clinical relevance," *Cancer Letters*, vol. 386, pp. 87–99, feb 2017.

[35] A. Rossi, Z. Kontarakis, C. Gerri, H. Nolte, S. Hölper, M. Krüger, and D. Y. R. Stainier, "Genetic compensation induced by deleterious mutations but not gene knockdowns," *Nature*, vol. 524, pp. 230–233, jul 2015.

[36] J. Ding, M. K. McConechy, H. M. Horlings, G. Ha, F. C. Chan, T. Funnell, S. C. Mullaly, J. Reimand, A. Bashashati, G. D. Bader, D. Huntsman, S. Aparicio, A. Condon, and S. P. Shah, "Systematic analysis of somatic

mutations impacting gene expression in 12 tumour types," *Nature Communications*, vol. 6, p. 8554, Oct. 2015.

[37] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhang, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, and R. Beroukhim, "Pan-cancer patterns of somatic copy number alteration," *Nat Genet*, vol. 45, pp. 1134–1140, Oct. 2013.

[38] C. M. Croce, "Oncogenes and cancer," *New England Journal of Medicine*, vol. 358, pp. 502–511, jan 2008.

[39] B. Alaei-Mahabadi, J. Bhadury, J. W. Karlsson, J. A. Nilsson, and E. Larsson, "Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 13768–13773, Nov. 2016.

[40] L. A. Pray, "Dna replication and causes of mutation," *Nature Education*, vol. 1, no. 1, p. 214, 2008.

[41] I. D'Souza, P. Poorkaj, M. Hong, D. Nochlin, V. M.-Y. Lee, T. D. Bird, and G. D. Schellenberg, "Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements," *Proceedings of the National Academy of Sciences*, vol. 96, pp. 5598–5603, may 1999.

[42] L. Fedele, J. Newcombe, M. Topf, A. Gibb, R. J. Harvey, and T. G. Smart, "Disease-associated missense mutations in GluN2b subunit alter NMDA receptor ligand binding and ion channel properties," *Nature Communications*, vol. 9, mar 2018.

[43] K. S. Amy Ralston, Laura Hoopes, "Environmental factors like viral infections play a role in the onset of complex diseases,"

[44] S. L. Fishman, S. H. Factor, C. Balestrieri, X. Fan, A. M. DiBisceglie, S. M. Desai, G. Benson, and A. D. Branch, "Mutations in the hepatitis

c virus core gene are associated with advanced liver disease and hepatocellular carcinoma," *Clinical Cancer Research*, vol. 15, pp. 3205–3213, apr 2009.

[45] "Beyond the genome," *Nature*, vol. 518, pp. 273–273, feb 2015.

[46] R. Nativio, G. Donahue, A. Berson, Y. Lan, A. Amlie-Wolf, F. Tuzer, J. B. Toledo, S. J. Gosai, B. D. Gregory, C. Torres, J. Q. Trojanowski, L.-S. Wang, F. B. Johnson, N. M. Bonini, and S. L. Berger, "Dysregulation of the epigenetic landscape of normal aging in alzheimer's disease," *Nature Neuroscience*, vol. 21, pp. 497–505, mar 2018.

[47] H. Wu, Y. Deng, Y. Feng, D. Long, K. Ma, X. Wang, M. Zhao, L. Lu, and Q. Lu, "Epigenetic regulation in b-cell maturation and its dysregulation in autoimmunity," *Cellular & Molecular Immunology*, vol. 15, pp. 676–684, jan 2018.

[48] T. Fleischer, , X. Tekpli, A. Mathelier, S. Wang, D. Nebdal, H. P. Dhakal, K. K. Sahlberg, E. Schlichting, A.-L. Børresen-Dale, E. Borgen, B. Naume, R. Eskeland, A. Frigessi, J. Tost, A. Hurtado, and V. N. Kristensen, "DNA methylation at enhancers identifies distinct breast cancer lineages," *Nature Communications*, vol. 8, nov 2017.

[49] J. Ju, A. Chen, Y. Deng, M. Liu, Y. Wang, Y. Wang, M. Nie, C. Wang, H. Ding, B. Yao, T. Gui, X. Li, Z. Xu, C. Ma, Y. Song, M. Kvansakul, K. Zen, C.-Y. Zhang, C. Luo, M. Fang, D. C. S. Huang, C. D. Allis, R. Tan, C. K. Zeng, J. Wei, and Q. Zhao, "NatD promotes lung cancer progression by preventing histone h4 serine phosphorylation to activate slug expression," *Nature Communications*, vol. 8, oct 2017.

[50] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley,

J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis,

R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, and M. J. Morgan, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, feb 2001.

[51] K. V. Desai, N. Xiao, W. Wang, L. Gangi, J. Greene, J. I. Powell, R. Dickson, P. Furth, K. Hunter, R. Kucherlapati, R. Simon, E. T. Liu, and J. E. Green, "Initiating oncogenic event determines gene-expression patterns of human breast cancer models," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 6967–6972, may 2002.

[52] Z. Cheng, Q. Zhang, A. Yin, M. Feng, H. Li, H. Liu, Y. Li, and L. Qian, "The long non-coding RNA uc.4 influences cell differentiation through the TGF-beta signaling pathway," *Experimental & Molecular Medicine*, vol. 50, p. e447, feb 2018.

[53] R. Cloney, "Deciphering the rules of microRNA targeting," *Nature Reviews Genetics*, vol. 17, pp. 718–718, oct 2016.

[54] G. D. Leva, M. Garofalo, and C. M. Croce, "MicroRNAs in cancer," *Annual Review of Pathology: Mechanisms of Disease*, vol. 9, pp. 287–314, jan 2014.

[55] N. Bartonicek, J. L. V. Maag, and M. E. Dinger, "Long noncoding RNAs in cancer: mechanisms of action and technological advancements," *Molecular Cancer*, vol. 15, may 2016.

[56] R. R. Pandey and C. Kanduri, "Transcriptional and posttranscriptional programming by long noncoding RNAs," in *Long Non-Coding RNAs*, pp. 1–27, Springer Berlin Heidelberg, nov 2010.

[57] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding rnas: insights into functions.," *Nature reviews. Genetics*, vol. 10, pp. 155–159, Mar. 2009.

[58] A. M. Schmitt and H. Y. Chang, "Long noncoding RNAs in cancer pathways," *Cancer Cell*, vol. 29, pp. 452–463, apr 2016.

[59] A. Schwarzer, S. Emmrich, F. Schmidt, D. Beck, M. Ng, C. Reimer, F. F. Adams, S. Grasedieck, D. Witte, S. Käbler, J. W. H. Wong, A. Shah, Y. Huang, R. Jammal, A. Maroz, M. Jongen-Lavrencic, A. Schambach, F. Kuchenbauer, J. E. Pimanda, D. Reinhardt, D. Heckl, and J.-H. Klusmann, "The non-coding RNA landscape of human hematopoiesis and leukemia," *Nature Communications*, vol. 8, aug 2017.

[60] E. Anastasiadou, L. S. Jacob, and F. J. Slack, "Non-coding RNA networks in cancer," *Nature Reviews Cancer*, vol. 18, pp. 5–18, nov 2017.

[61] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology. 4th edition.* W. H. Freeman and Company, 2000.

[62] J. D. Lapek, P. Greninger, R. Morris, A. Amzallag, I. Pruteanu-Malinici, C. H. Benes, and W. Haas, "Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities," *Nature Biotechnology*, vol. 35, pp. 983–989, sep 2017.

[63] S. Ge, X. Xia, C. Ding, B. Zhen, Q. Zhou, J. Feng, J. Yuan, R. Chen, Y. Li, Z. Ge, J. Ji, L. Zhang, J. Wang, Z. Li, Y. Lai, Y. Hu, Y. Li, Y. Li, J. Gao, L. Chen, J. Xu, C. Zhang, S. Y. Jung, J. M. Choi, A. Jain, M. Liu, L. Song, W. Liu, G. Guo, T. Gong, Y. Huang, Y. Qiu, W. Huang, T. Shi, W. Zhu, Y. Wang, F. He, L. Shen, and J. Qin, "A proteomic landscape of diffuse-type gastric cancer," *Nature Communications*, vol. 9, mar 2018.

[64] M. Eckhardt, W. Zhang, A. M. Gross, J. V. Dollen, J. R. Johnson, K. E. Franks-Skiba, D. L. Swaney, T. L. Johnson, G. M. Jang, P. S. Shah, T. M. Brand, J. Archambault, J. F. Kreisberg, J. R. Grandis, T. Ideker, and N. J. Krogan, "Multiple routes to oncogenesis are promoted by the human papillomavirus-host protein network," *Cancer Discovery*, pp. CD–17–1018, sep 2018.

[65] P. S. Ward and C. B. Thompson, "Metabolic reprogramming: A cancer hallmark even warburg did not anticipate," *Cancer Cell*, vol. 21, pp. 297–308, mar 2012.

[66] L. B. Sullivan, D. Y. Gui, and M. G. V. Heiden, "Altered metabolite levels in cancer: implications for tumour biology and cancer therapy," *Nature Reviews Cancer*, vol. 16, pp. 680–693, sep 2016.

[67] A. Chiarugi, C. Dölle, R. Felici, and M. Ziegler, "The NAD metabolome — a key determinant of cancer cell biology," *Nature Reviews Cancer*, vol. 12, pp. 741–752, sep 2012.

[68] A. B. Hall, A. C. Tolonen, and R. J. Xavier, "Human genetic variation and the gut microbiome in disease," *Nature Reviews Genetics*, vol. 18, pp. 690–699, aug 2017.

[69] X. Zhang, S. A. Deeke, Z. Ning, A. E. Starr, J. Butcher, J. Li, J. Mayne, K. Cheng, B. Liao, L. Li, R. Singleton, D. Mack, A. Stintzi, and D. Figeys, "Metaproteomics reveals associations between microbiome and intestinal extracellular vesicle proteins in pediatric inflammatory bowel disease," *Nature Communications*, vol. 9, jul 2018.

[70] S. Jangi, R. Gandhi, L. M. Cox, N. Li, F. von Glehn, R. Yan, B. Patel, M. A. Mazzola, S. Liu, B. L. Glanz, S. Cook, S. Tankou, F. Stuart, K. Melo, P. Nejad, K. Smith, B. D. Topçuolu, J. Holden, P. Kivisäkk, T. Chitnis, P. L. D. Jager, F. J. Quintana, G. K. Gerber, L. Bry, and H. L. Weiner, "Alterations of the human gut microbiome in multiple sclerosis," *Nature Communications*, vol. 7, p. 12015, jun 2016.

[71] L. Zitvogel, R. Daillère, M. P. Roberti, B. Routy, and G. Kroemer, "Anticancer effects of the microbiome and its products," *Nature Reviews Microbiology*, vol. 15, pp. 465–478, may 2017.

[72] V. Matson, J. Fessler, R. Bao, T. Chongsuwat, Y. Zha, M.-L. Alegre, J. J. Luke, and T. F. Gajewski, "The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients," *Science*, vol. 359, pp. 104–108, jan 2018.

[73] V. Gopalakrishnan, C. N. Spencer, L. Nezi, A. Reuben, M. C. Andrews, T. V. Karpinets, P. A. Prieto, D. Vicente, K. Hoffman, S. C. Wei, A. P. Cogdill, L. Zhao, C. W. Hudgens, D. S. Hutchinson, T. Manzo, M. P. de Macedo, T. Cotechini, T. Kumar, W. S. Chen, S. M. Reddy, R. S. Sloane, J. Galloway-Pena, H. Jiang, P. L. Chen, E. J. Shpall, K. Rezvani, A. M. Alousi, R. F. Chemaly, S. Shelburne, L. M. Vence, P. C. Okhuysen, V. B. Jensen, A. G. Swennes, F. McAllister, E. M. R. Sanchez, Y. Zhang, E. L. Chatelier, L. Zitvogel, N. Pons, J. L. Austin-Breneman, L. E. Haydu, E. M. Burton, J. M. Gardner, E. Sirmans, J. Hu, A. J. Lazar, T. Tsujikawa, A. Diab, H. Tawbi, I. C. Glitza, W. J. Hwu, S. P. Patel, S. E. Woodman, R. N. Amaria, M. A. Davies, J. E. Gershenwald, P. Hwu, J. E. Lee, J. Zhang, L. M. Coussens, Z. A. Cooper, P. A. Futreal, C. R. Daniel, N. J. Ajami, J. F. Petrosino, M. T. Tetzlaff, P. Sharma, J. P. Allison, R. R. Jenq, and J. A. Wargo, "Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients," *Science*, vol. 359, pp. 97–103, nov 2017.

[74] L. Zitvogel, M. Ayyoub, B. Routy, and G. Kroemer, "Microbiome and anticancer immunosurveillance," *Cell*, vol. 165, pp. 276–287, apr 2016.

[75] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. e. a. Ziller, and M. Kellis, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, Feb 2015.

[76] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca,

D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore, "The genotype-tissue expression (GTEx) project," *Nature Genetics*, vol. 45, pp. 580–585, jun 2013.

[77] R. McLendon, A. Friedman, D. Bigner, E. G. V. Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, K. Aldape, W. K. A. Yung, O. Bogler, S. VandenBerg, M. Berger, M. Prados, D. Muzny, M. Morgan, S. Scherer, A. Sabo, L. Nazareth, L. Lewis, O. Hall, Y. Zhu, Y. Ren, O. Alvi, J. Yao, A. Hawes, S. Jhangiani, G. Fowler, A. S. Lucas, C. Kovar, A. Cree, H. Dinh, J. Santibanez, V. Joshi, M. L. Gonzalez-Garay, C. A. Miller, A. Milosavljevic, L. Donehower, D. A. Wheeler, R. A. Gibbs, K. Cibulskis, C. Sougnez, T. Fennell, S. Mahan, J. Wilkinson, L. Ziaugra, R. Onofrio, T. Bloom, R. Nicol, K. Ardlie, J. Baldwin, S. Gabriel, E. S. Lander, L. Ding, R. S. Fulton, M. D. McLellan, J. Wallis, D. E. Larson, X. Shi, R. Abbott, L. Fulton, K. Chen, D. C. Koboldt, M. C. Wendl, R. Meyer, Y. Tang, L. Lin, J. R. Osborne, B. H. Dunford-Shore, T. L. Miner, K. Delehaunty, C. Markovic, G. Swift, W. Courtney, C. Pohl, S. Abbott, A. Hawkins, S. Leong, C. Haipek, H. Schmidt, M. Wiechert, T. Vickery, S. Scott, D. J. Dooling, A. Chinwalla, G. M. Weinstock, E. R. Mardis, R. K. Wilson, G. Getz, W. Winckler, R. G. W. Verhaak, M. S. Lawrence, M. O'Kelly, J. Robinson, G. Alexe, R. Beroukhim, S. Carter, D. Chiang, J. Gould, S. Gupta, J. Korn, C. Mermel, J. Mesirov, S. Monti, H. Nguyen, M. Parkin, M. Reich, N. Stransky, B. A. Weir, L. Garraway, T. Golub, M. Meyerson, L. Chin, A. Protopopov,

J. Zhang, I. Perna, S. Aronson, N. Sathiamoorthy, G. Ren, J. Yao, W. R. Wiedemeyer, H. Kim, S. W. Kong, Y. Xiao, I. S. Kohane, J. Seidman, P. J. Park, R. Kucherlapati, P. W. Laird, L. Cope, J. G. Herman, D. J. Weisenberger, F. Pan, D. V. D. Berg, L. V. Neste, J. M. Yi, K. E. Schuebel, S. B. Baylin, D. M. Absher, J. Z. Li, A. Southwick, S. Brady, A. Aggarwal, T. Chung, G. Sherlock, J. D. Brooks, R. M. Myers, P. T. Spellman, E. Purdom, L. R. Jakkula, A. V. Lapuk, H. Marr, S. Dorton, Y. G. Choi, J. Han, A. Ray, V. Wang, S. Durinck, M. Robinson, N. J. Wang, K. Vranizan, V. Peng, E. V. Name, G. V. Fontenay, J. Ngai, J. G. Conboy, B. Parvin, H. S. Feiler, T. P. Speed, J. W. Gray, C. Brennan, N. D. Socci, A. Olshen, B. S. Taylor, A. Lash, N. Schultz, B. Reva, Y. Antipin, A. Stukalov, B. Gross, E. Cerami, W. Q. Wang, L.-X. Qin, V. E. Seshan, L. Villafania, M. Cavatore, L. Borsu, A. Viale, W. Gerald, C. Sander, M. Ladanyi, C. M. Perou, D. N. Hayes, M. D. Topal, K. A. Hoadley, Y. Qi, S. Balu, Y. Shi, J. Wu, R. Penny, M. Bittner, T. Shelton, E. Lenkiewicz, S. Morris, D. Beasley, S. Sanders, A. Kahn, R. Sfeir, J. Chen, D. Nassau, L. Feng, E. Hickey, J. Zhang, J. N. Weinstein, A. Barker, D. S. Gerhard, J. Vockley, C. Compton, J. Vaught, P. Fielding, M. L. Ferguson, C. Schaefer, S. Madhavan, K. H. Buetow, F. Collins, P. Good, M. Guyer, B. Ozenberger, J. Peterson, and E. Thomson, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, pp. 1061–1068, sep 2008.

[78] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, G. Alexe, M. Lawrence, M. O'Kelly, P. Tamayo, B. A. Weir, S. Gabriel, W. Winckler, S. Gupta, L. Jakkula, H. S. Feiler, J. G. Hodgson, C. D. James, J. N. Sarkaria, C. Brennan, A. Kahn, P. T. Spellman, R. K. Wilson, T. P. Speed, J. W. Gray, M. Meyerson, G. Getz, C. M. Perou, and D. N. Hayes, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, pp. 98–110, jan 2010.

[79] H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, K. Pujara, B. P. Berman, F. Pan, C. E. Pelloski, E. P. Sulman, K. P. Bhat, R. G. Ver-

haak, K. A. Hoadley, D. N. Hayes, C. M. Perou, H. K. Schmidt, L. Ding, R. K. Wilson, D. V. D. Berg, H. Shen, H. Bengtsson, P. Neuvial, L. M. Cope, J. Buckley, J. G. Herman, S. B. Baylin, P. W. Laird, and K. Aldape, "Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma," *Cancer Cell*, vol. 17, pp. 510–522, may 2010.

[80] D. C. Koboldt, R. S. Fulton, M. D. McLellan, H. Schmidt, J. Kalicki-Veizer, J. F. McMichael, L. L. Fulton, D. J. Dooling, L. Ding, E. R. Mardis, R. K. Wilson, A. Ally, M. Balasundaram, Y. S. N. Butterfield, R. Carlsen, C. Carter, A. Chu, E. Chuah, H.-J. E. Chun, R. J. N. Coope, N. Dhalla, R. Guin, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, A. J. Mungall, E. Pleasance, A. G. Robertson, J. E. Schein, A. Shafiei, P. Sipahimalani, J. R. Slobodan, D. Stoll, A. Tam, N. Thiessen, R. J. Varhol, N. Wye, T. Zeng, Y. Zhao, I. Birol, S. J. M. Jones, M. A. Marra, A. D. Cherniack, G. Saksena, R. C. Onofrio, N. H. Pho, S. L. Carter, S. E. Schumacher, B. Tabak, B. Hernandez, J. Gentry, H. Nguyen, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, G. Getz, S. B. Gabriel, M. Meyerson, L. Chin, P. J. Park, R. Kucherlapati, K. A. Hoadley, J. T. Auman, C. Fan, Y. J. Turman, Y. Shi, L. Li, M. D. Topal, X. He, H.-H. Chao, A. Prat, G. O. Silva, M. D. Iglesia, W. Zhao, J. Usary, J. S. Berg, M. Adams, J. Booker, J. Wu, A. Gulabani, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, J. S. Parker, D. N. Hayes, C. M. Perou, S. Malik, S. Mahurkar, H. Shen, D. J. Weisenberger, T. T. Jr, P. H. Lai, M. S. Bootwalla, D. T. Maglinte, B. P. Berman, D. J. V. D. Berg, S. B. Baylin, P. W. Laird, C. J. Creighton, L. A. Donehower, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C.-J. Wu, S. Y. Liu, M. S. Lawrence, L. Zou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, J. Cho, R. Sinha, R. W. Park, M.-D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, S. Reynolds, R. B. Kreisberg, B. Bernard, R. Bressler, T. Erkkila, J. Lin, V. Thorsson, W. Zhang, I. Shmulevich, G. Ciriello, N. Weinhold, N. Schultz, J. Gao, E. Cerami, B. Gross, A. Jacobsen, R. Sinha, B. A. Aksoy, Y. Antipin, B. Reva, R. Shen, B. S. Taylor, M. Ladanyi, C. Sander, P. Anur, P. T. Spellman, Y. Lu, W. Liu,

R. R. G. Verhaak, G. B. Mills, R. Akbani, N. Zhang, B. M. Broom, T. D. Casasent, C. Wakefield, A. K. Unruh, K. Baggerly, K. Coombes, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Zhu, C. C. Szeto, G. K. Scott, C. Yau, E. O. Paull, D. Carlin, C. Wong, A. Sokolov, J. Thusberg, S. Mooney, S. Ng, T. C. Goldstein, K. Ellrott, M. Grifford, C. Wilks, S. Ma, B. Craft, C. Yan, Y. Hu, D. Meerzaman, J. M. Gastier-Foster, J. Bowen, N. C. Ramirez, A. D. Black, R. E. X. E. unknown variable "tname"., P. White, E. J. Zmuda, J. Frick, T. M. Lichtenberg, R. Brookens, M. M. George, M. A. Gerken, H. A. Harper, K. M. Leraas, L. J. Wise, T. R. Tabler, C. McAllister, T. Barr, M. Hart-Kothari, K. Tarvin, C. Saller, G. Sandusky, C. Mitchell, M. V. Iacocca, J. Brown, B. Rabeno, C. Czerwinski, N. Petrelli, O. Dolzhansky, M. Abramov, O. Voronina, O. Potapova, J. R. Marks, W. M. Suchorska, D. Murawa, W. Kycler, M. Ibbs, K. Korski, A. Spychała, P. Murawa, J. J. Brzeziński, H. Perz, R. Łaźniak, M. Teresiak, H. Tatka, E. Leporowska, M. Bogusz-Czerniewicz, J. Malicki, A. Mackiewicz, M. Wiznerowicz, X. V. Le, B. Kohl, N. V. Tien, R. Thorp, N. V. Bang, H. Sussman, B. D. Phu, R. Hajek, N. P. Hung, T. V. T. Phuong, H. Q. Thang, K. Z. Khan, R. Penny, D. Mallery, E. Curley, C. Shelton, P. Yena, J. N. Ingle, F. J. Couch, W. L. Lingle, T. A. King, A. M. Gonzalez-Angulo, G. B. Mills, M. D. Dyer, S. Liu, X. Meng, M. Patangan, F. Waldman, H. Stöppler, W. K. Rathmell, L. Thorne, M. Huang, L. Boice, A. Hill, C. Morrison, C. Gaudioso, W. Bshara, K. Daily, S. C. Egea, M. D. Pegram, C. Gomez-Fernandez, R. Dhir, R. Bhargava, A. Brufsky, C. D. Shriver, J. A. Hooke, J. L. Campbell, R. J. Mural, H. Hu, S. Somiari, C. Larson, B. Deyarmin, L. Kvecher, A. J. Kovatich, M. J. Ellis, T. A. King, H. Hu, F. J. Couch, R. J. Mural, T. Stricker, K. White, O. Olopade, J. N. Ingle, C. Luo, Y. Chen, J. R. Marks, F. Waldman, M. Wiznerowicz, R. Bose, L.-W. Chang, A. H. Beck, A. M. Gonzalez-Angulo, T. Pihl, M. Jensen, R. Sfeir, A. Kahn, A. Chu, P. Kothiyal, Z. Wang, E. Snyder, J. Pontius, B. Ayala, M. Backus, J. Walton, J. Baboud, D. Berton, M. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. Kigonya, S. Alonso, R. Sanbhadti, S. Barletta, D. Pot, M. Sheth, J. A. Demchok, K. R. M. Shaw, L. Yang, G. Eley,

M. L. Ferguson, R. W. Tarnuzzer, J. Zhang, L. A. L. Dillon, K. Buetow, P. Fielding, B. A. Ozenberger, M. S. Guyer, H. J. Sofia, and J. D. Palchik, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, sep 2012.

[81] The Cancer Genome Atlas Research Network, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, pp. 519–525, Sept. 2012.

[82] Cancer Genome Atlas Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, pp. 330–337, jul 2012.

[83] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild, "Bayesian correlated clustering to integrate multiple datasets," *Bioinformatics*, vol. 28, pp. 3290–3297, oct 2012.

[84] D.-Y. Cho and T. M. Przytycka, "Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model," *Nucleic Acids Research*, vol. 41, pp. 8011–8020, jul 2013.

[85] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker, "Network-based stratification of tumor mutations," *Nature Methods*, vol. 10, pp. 1108–1115, sep 2013.

[86] Y. A. Ghouri, I. Mian, and J. H. Rowe, "Review of hepatocellular carcinoma: Epidemiology, etiology, and carcinogenesis.," *Journal of carcinogenesis*, vol. 16, p. 1, 2017.

[87] E. P. Consortium, "An integrated encyclopedia of dna elements in the human genome.," *Nature*, vol. 489, pp. 57–74, Sept. 2012.

[88] G. Consortium, "Human genomics. the genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans.," *Science (New York, N.Y.)*, vol. 348, pp. 648–660, May 2015.

[89] E. R. Mardis, "A decade's perspective on DNA sequencing technology," *Nature*, vol. 470, pp. 198–203, feb 2011.

[90] J. J. Salk, M. W. Schmitt, and L. A. Loeb, "Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations," *Nature Reviews Genetics*, vol. 19, pp. 269–285, mar 2018.

[91] W. Li and M. Olivier, "Current analysis platforms and methods for detecting copy number variation," *Physiological Genomics*, vol. 45, pp. 1–16, jan 2013.

[92] B. Keren, "The advantages of SNP arrays over CGH arrays," *Molecular Cytogenetics*, vol. 7, no. Suppl 1, p. I31, 2014.

[93] F. Robert and J. Pelletier, "Exploring the impact of single-nucleotide polymorphisms on translation," *Frontiers in Genetics*, vol. 9, oct 2018.

[94] Z. Peng, Z. Zhao, J. P. Clevenger, Y. Chu, D. Paudel, P. Ozias-Akins, and J. Wang, "Comparison of SNP calling pipelines and NGS platforms to predict the genomic regions harboring candidate genes for nodulation in cultivated peanut," *Frontiers in Genetics*, vol. 11, mar 2020.

[95] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances," *Nucleic Acids Research*, vol. 37, pp. 4181–4193, jul 2009.

[96] J. Staaf, D. Lindgren, J. Vallon-Christersson, A. Isaksson, H. Goransson, G. Juliusson, R. Rosenquist, M. Hoglund, A. Borg, and M. Ringner, "Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays," *Genome Biology*, vol. 9, no. 9, p. R136, 2008.

[97] A. Pitea, I. Kondofersky, S. Sass, F. J. Theis, N. S. Mueller, and K. Unger, "Copy number aberrations from affymetrix SNP 6.0 genotyping data—how accurate are commonly used prediction approaches?," *Briefings in Bioinformatics*, oct 2018.

[98] Affymetrix, "Genome-wide human snp array 6.0," tech. rep., Affymetrix, 2009.

[99] N. Rabbee and T. P. Speed, "A genotype calling algorithm for affymetrix SNP arrays," *Bioinformatics*, vol. 22, pp. 7–12, nov 2005.

[100] A. Casamassimi, A. Federico, M. Rienzo, S. Esposito, and A. Ciccodicola, "Transcriptome profiling in human diseases: New advances and perspectives," *International Journal of Molecular Sciences*, vol. 18, p. 1652, jul 2017.

[101] M. Scarpato, A. Federico, A. Ciccodicola, and V. Costa, "Novel transcription factor variants through RNA-sequencing: The importance of being "alternative"," *International Journal of Molecular Sciences*, vol. 16, pp. 1755–1771, jan 2015.

[102] N. N. Vellichirammal, A. Albahrani, J. K. Banwait, N. K. Mishra, Y. Li, S. Roychoudhury, M. J. Kling, S. Mirza, K. K. Bhakat, V. Band, S. S. Joshi, and C. Guda, "Pan-cancer analysis reveals the diverse landscape of novel sense and antisense fusion transcripts," *Molecular Therapy - Nucleic Acids*, vol. 19, pp. 1379–1398, mar 2020.

[103] F. Ozsolak and P. M. Milos, "RNA sequencing: advances, challenges and opportunities," *Nature Reviews Genetics*, vol. 12, pp. 87–98, dec 2010.

[104] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-throughput sequencing technologies," *Molecular Cell*, vol. 58, pp. 586–597, may 2015.

[105] S. Gilbert, *Developmental biology*. Sunderland, Mass: Sinauer Associates, 2000.

[106] Z. Wang, M. Gerstein, and M. Snyder, "RNA-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57–63, jan 2009.

[107] Illumina, "Hiseq x ten series of sequencing systems," *Illumina Documents*, 2014.

[108] J. H. Morris, G. M. Knudsen, E. Verschueren, J. R. Johnson, P. Cimermancic, A. L. Greninger, and A. R. Pico, "Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions," *Nature Protocols*, vol. 9, pp. 2539–2554, oct 2014.

[109] B. H. University and H. M. S. Rosner, *Fundamentals of Biostatistics*. Cengage Learning, Inc, 2015.

[110] S. M. Ross, "Testing statistical hypotheses," in *Introductory Statistics*, pp. 381–432, Elsevier, 2017.

[111] W. J. Conover, *Practical Nonparametric Statistics*. John Wiley & Sons, 1998.

[112] D. Voelker, *Cliffsnotes statistics : quickreview*. Hoboken, NJ: Wiley, 2011.

[113] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, pp. 50–60, Mar 1947.

[114] R. A. Fisher, "On the interpretation of χ^2 from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, vol. 85, p. 87, Jan 1922.

[115] C. E. Bonferroni, *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato, 1935.

[116] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[117] R. A. Fisher *et al.*, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, vol. 10, no. 4, pp. 507–521, 1915.

[118] J. C. Caruso and N. Cliff, "Empirical size, coverage, and power of confidence intervals for spearman's rho," *Educational and Psychological Measurement*, vol. 57, pp. 637–654, Aug 1997.

[119] G. M. Furnival and R. W. Wilson, "Regressions by leaps and bounds," *Technometrics*, vol. 16, pp. 499–511, Nov 1974.

[120] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[121] S. Zeger, "Regression shrinkage methods."

[122] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, 2011.

[123] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[124] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 301–320, apr 2005.

[125] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein–protein interactions in yeast," *Nature Biotechnology*, vol. 18, pp. 1257–1261, dec 2000.

[126] J. Song and M. Singh, "How and when should interactome-derived clusters be used to predict functional modules and protein function?," *Bioinformatics*, vol. 25, pp. 3143–3150, sep 2009.

[127] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, pp. 551–562, jun 2017.

[128] Cancer Genome Atlas Network, "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, vol. 511, July 2014.

[129] A. Ally, M. Balasundaram, R. Carlsen, E. Chuah, A. Clarke, N. Dhalla, R. A. Holt, S. J. Jones, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, D. Cheung, T. Wong, D. Brooks, A. G. Robertson, R. Bowlby, K. Mungall, S. Sadeghi, L. Xi, K. Covington, E. Shinbrot, D. A. Wheeler, R. A. Gibbs, L. A. Donehower, L. Wang, J. Bowen, J. M. Gastier-Foster, M. Gerken, C. Helsel, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, S. B. Gabriel, M. Meyerson, C. Cibulskis, B. A. Murray, J. Shih, R. Beroukhim, A. D. Cherniack, S. E. Schumacher, G. Saksena, C. S. Pedamallu, L. Chin, G. Getz, M. Noble, H. Zhang, D. Heiman,

J. Cho, N. Gehlenborg, G. Saksena, D. Voet, P. Lin, S. Frazer, T. Defreitas, S. Meier, M. Lawrence, J. Kim, C. J. Creighton, D. Muzny, H. Doddapaneni, J. Hu, M. Wang, D. Morton, V. Korchina, Y. Han, H. Dinh, L. Lewis, M. Bellair, X. Liu, J. Santibanez, R. Glenn, S. Lee, W. Hale, J. S. Parker, M. D. Wilkerson, D. N. Hayes, S. M. Reynolds, I. Shmulevich, W. Zhang, Y. Liu, L. Iype, H. Makhlouf, M. S. Torbenson, S. Kakar, M. M. Yeh, D. Jain, D. E. Kleiner, D. Jain, R. Dhanasekaran, H. B. El-Serag, S. Y. Yim, J. N. Weinstein, L. Mishra, J. Zhang, R. Akbani, S. Ling, Z. Ju, X. Su, A. M. Hegde, G. B. Mills, Y. Lu, J. Chen, J.-S. Lee, B. H. Sohn, J. J. Shim, P. Tong, H. Aburatani, S. Yamamoto, K. Tatsuno, W. Li, Z. Xia, N. Stransky, E. Seiser, F. Innocenti, J. Gao, R. Kundra, H. Zhang, Z. Heins, A. Ochoa, C. Sander, M. Ladanyi, R. Shen, A. Arora, F. Sanchez-Vega, N. Schultz, K. Kasaian, A. Radenbaugh, K.-D. Bissig, D. D. Moore, Y. Totoki, H. Nakamura, T. Shibata, C. Yau, K. Graim, J. Stuart, D. Haussler, B. L. Slagle, A. I. Ojesina, P. Katsonis, A. Koire, O. Lichtarge, T.-K. Hsu, M. L. Ferguson, J. A. Demchok, I. Felau, M. Sheth, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. Zhang, C. M. Hutter, H. J. Sofia, R. G. Verhaak, S. Zheng, F. Lang, S. Chudamani, J. Liu, L. Lolla, Y. Wu, R. Naresh, T. Pihl, C. Sun, Y. Wan, C. Benz, A. H. Perou, L. B. Thorne, L. Boice, M. Huang, W. K. Rathmell, H. Noushmehr, F. P. Saggioro, D. P. da Cunha Tirapelli, C. G. C. Junior, E. D. Mente, O. de Castro Silva, F. A. Trevisan, K. J. Kang, K. S. Ahn, N. H. Giama, C. D. Moser, T. J. Giordano, M. Vinco, T. H. Welling, D. Crain, E. Curley, J. Gardner, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, R. Kelley, J.-W. Park, V. S. Chandan, L. R. Roberts, O. F. Bathe, C. H. Hagedorn, J. T. Auman, D. R. O'Brien, J.-P. A. Kocher, C. D. Jones, P. A. Mieczkowski, C. M. Perou, T. Skelly, D. Tan, U. Veluvolu, S. Balu, T. Bodenheimer, A. P. Hoyle, S. R. Jefferys, S. Meng, L. E. Mose, Y. Shi, J. V. Simons, M. G. Soloway, J. Roach, K. A. Hoadley, S. B. Baylin, H. Shen, T. Hinoue, M. S. Bootwalla, D. J. V. D. Berg, D. J. Weisenberger, P. H. Lai, A. Holbrook, M. Berrios, and P. W. Laird, "Comprehensive and integrative genomic characterization of hepatocellular carcinoma," *Cell*, vol. 169, pp. 1327–1341.e23, jun 2017.

[130] H. E. Lockstone, "Exon array data analysis using Affymetrix power tools and R statistical software," *Brief Bioinform*, vol. 12, pp. 634–644, Nov. 2011.

[131] K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. A. Grant, H. Hakonarson, and M. Bucan, "PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data," *Genome Res.*, vol. 17, pp. 1665–1674, Nov. 2007.

[132] J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler, "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs," *Nat Genet*, vol. 40, pp. 1253–1260, Oct. 2008.

[133] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[134] S. Sass, A. Pitea, K. Unger, J. Hess, N. Mueller, and F. Theis, "MicroRNA-target network inference and local network enrichment analysis identify two microRNA clusters with distinct functions in head and neck squamous cell carcinoma," *International Journal of Molecular Sciences*, vol. 16, pp. 30204–30222, dec 2015.

[135] M. Robinson, D. McCarthy, and G. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, pp. 139–140, 2009.

[136] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Abu-

ratani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles, "Global variation in copy number in the human genome," *Nature*, vol. 444, pp. 444–454, Nov. 2006.

[137] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[138] S. Jäger, P. Cimermancic, N. Gulbahce, J. R. Johnson, K. E. McGovern, S. C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, H. Hernandez, G. M. Jang, S. L. Roth, E. Akiva, J. Marlett, M. Stephens, I. D'Orso, J. Fernandes, M. Fahey, C. Mahon, A. J. O'Donoghue, A. Todorovic, J. H. Morris, D. A. Maltby, T. Alber, G. Cagney, F. D. Bushman, J. A. Young, S. K. Chanda, W. I. Sundquist, T. Kortemme, R. D. Hernandez, C. S. Craik, A. Burlingame, A. Sali, A. D. Frankel, and N. J. Krogan, "Global landscape of HIV–human protein complexes," *Nature*, vol. 481, pp. 365–370, dec 2011.

[139] E. Verschueren, J. Von Dollen, P. Cimermancic, N. Gulbahce, A. Sali, and N. J. Krogan, "Scoring large-scale affinity purification mass spectrometry datasets with mist.," *Current protocols in bioinformatics*, vol. 49, pp. 8.19.1–8.1916, Mar. 2015.

[140] M. González-Vallinas, M. Rodríguez-Paredes, M. Albrecht, C. Sticht, D. Stichel, J. Gutekunst, A. Pitea, S. Sass, F. J. Sánchez-Rivera, J. Lorenzo-Bermejo, J. Schmitt, C. D. L. Torre, A. Warth, F. J. Theis, N. S. Müller, N. Gretz, T. Muley, M. Meister, D. F. Tschaharganeh, P. Schirmacher, F. Matthäus, and K. Breuhahn, "Epigenetically regulated chromosome 14q32 miRNA cluster induces metastasis and predicts poor prognosis in lung adenocarcinoma patients," *Molecular Cancer Research*, vol. 16, pp. 390–402, jan 2018.

[141] W. Zhou, Z. Zhao, R. Wang, Y. Han, C. Wang, F. Yang, Y. Han, H. Liang, L. Qi, C. Wang, Z. Guo, and Y. Gu, "Identification of driver copy number alterations in diverse cancer types and application in drug repositioning," *Mol Oncol*, vol. 11, pp. 1459–1474, Oct. 2017.

[142] S. M. Gollin, "Cytogenetic alterations and their molecular genetic correlates in head and neck squamous cell carcinoma: a next generation window to the biology of disease," *Genes Chromosomes Cancer*, vol. 53, pp. 972–990, Dec. 2014.

[143] S. L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P. W. Laird, R. C. Onofrio, W. Winckler, B. A. Weir, R. Beroukhim, D. Pellman, D. A. Levine, E. S. Lander, M. Meyerson, and G. Getz, "Absolute quantification of somatic DNA alterations in human cancer," *Nat Biotech*, vol. 30, pp. 413–421, May 2012.

[144] T. Tony Cai, X. Jessie Jeng, and H. Li, "Robust detection and identification of sparse segments in ultrahigh dimensional data analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, pp. 773–797, Nov. 2012.

[145] Y. Chen, L. Zhao, Y. Wang, M. Cao, V. Gelowani, M. Xu, S. A. Agrawal, Y. Li, S. P. Daiger, R. Gibbs, F. Wang, and R. Chen, "SeqCNV: a novel method for identification of copy number variations in targeted next-generation sequencing data," *BMC Bioinformatics*, vol. 18, p. 147, Mar. 2017.

[146] X. Zhang, R. Du, S. Li, F. Zhang, L. Jin, and H. Wang, "Evaluation of copy number variation detection for a SNP array platform," *BMC Bioinformatics*, vol. 15, p. 50, Feb. 2014.

[147] M. Pierre-Jean, G. Rigaill, and P. Neuvial, "Performance evaluation of DNA copy number segmentation methods," *Brief Bioinform*, vol. 16, pp. 600–615, July 2015.

[148] M. Rasmussen, M. Sundström, H. G. Kultima, J. Botling, P. Micke, H. Birgisson, B. Glimelius, and A. Isaksson, "Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity," *Genome Biology*, vol. 12, p. R108, Oct. 2011.

[149] C. Yau, D. Mouradov, R. N. Jorissen, S. Colella, G. Mirza, G. Steers, A. Harris, J. Ragoussis, O. Sieber, and C. C. Holmes, "A statistical approach for detecting genomic aberrations in heterogeneous tumor samples

from single nucleotide polymorphism genotyping data," *Genome Biology*, vol. 11, p. R92, 2010.

[150] P. V. Loo, S. H. Nordgard, O. C. Lingjærde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Børresen-Dale, and V. N. Kristensen, "Allele-specific copy number analysis of tumors," *PNAS*, vol. 107, pp. 16910–16915, Sept. 2010.

[151] W. Sun, F. A. Wright, Z. Tang, S. H. Nordgard, P. V. Loo, T. Yu, V. N. Kristensen, and C. M. Perou, "Integrated study of copy number states and genotype calls using high-density SNP arrays," *Nucleic Acids Res*, vol. 37, pp. 5365–5377, Sept. 2009.

[152] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome Biol*, vol. 12, no. 4, p. R41, 2011.

[153] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostat*, vol. 5, pp. 557–572, Oct. 2004.

[154] M. A. van de Wiel, K. I. Kim, S. J. Vosse, W. N. van Wieringen, S. M. Wilting, and B. Ylstra, "CGHcall: calling aberrations for array CGH tumor profiles," *Bioinformatics*, vol. 23, pp. 892–894, Apr. 2007.

[155] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recogn. Lett.*, vol. 27, pp. 861–874, June 2006.

[156] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, (New York, NY, USA), pp. 233–240, ACM, 2006.

[157] J. Lever, M. Krzywinski, and N. Altman, "Points of Significance: Classification evaluation," *Nat Meth*, vol. 13, pp. 603–604, Aug. 2016.

[158] C. J. V. Rijsbergen, *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, 2nd ed., 1979.

[159] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[160] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilità," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3–62, 1936.

[161] J. Metzger, U. Philipp, M. S. Lopes, A. da Camara Machado, M. Felicetti, M. Silvestrelli, and O. Distl, "Analysis of copy number variants by three detection algorithms and their association with body size in horses," *BMC Genomics*, vol. 14, p. 487, July 2013.

[162] D. Mosén-Ansorena, A. M. Aransay, and N. Rodriguez-Ezpeleta, "Comparison of methods to detect copy number alterations in cancer using simulated and real genotyping data," *BMC Bioinformatics*, vol. 13, p. 192, 2012.

[163] M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi, "Global burden of cancers attributable to infections in 2012: a synthetic analysis," *The Lancet Global Health*, vol. 4, pp. e609–e616, sep 2016.

[164] H. zur Hausen, "Oncogenic DNA viruses," *Oncogene*, vol. 20, pp. 7820–7823, nov 2001.

[165] P. S. Moore and Y. Chang, "Why do viruses cause cancer? highlights of the first century of human tumour virology," *Nature Reviews Cancer*, vol. 10, pp. 878–889, nov 2010.

[166] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, pp. 87–108, feb 2015.

[167] M. C. S. Wong, J. Y. Jiang, W. B. Goggins, M. Liang, Y. Fang, F. D. H. Fung, C. Leung, H. H. X. Wang, G. L. H. Wong, V. W. Wong, and H. L. Y. Chan, "International incidence and mortality trends of liver cancer: a global profile," *Scientific Reports*, vol. 7, p. 45846, mar 2017.

[168] Y. Totoki, K. Tatsuno, K. R. Covington, H. Ueda, C. J. Creighton, M. Kato, S. Tsuji, L. A. Donehower, B. L. Slagle, H. Nakamura, S. Ya-

mamoto, E. Shinbrot, N. Hama, M. Lehmkuhl, F. Hosoda, Y. Arai, K. Walker, M. Dahdouli, K. Gotoh, G. Nagae, M.-C. Gingras, D. M. Muzny, H. Ojima, K. Shimada, Y. Midorikawa, J. A. Goss, R. Cotton, A. Hayashi, J. Shibahara, S. Ishikawa, J. Guiteau, M. Tanaka, T. Urushidate, S. Ohashi, N. Okada, H. Doddapaneni, M. Wang, Y. Zhu, H. Dinh, T. Okusaka, N. Kokudo, T. Kosuge, T. Takayama, M. Fukayama, R. A. Gibbs, D. A. Wheeler, H. Aburatani, and T. Shibata, "Trans-ancestry mutational landscape of hepatocellular carcinoma genomes," *Nature Genetics*, vol. 46, pp. 1267–1273, nov 2014.

[169] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic Acids Research*, vol. 39, pp. e118–e118, jul 2011.

[170] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, vol. 2, pp. 1360–1383, dec 2008.

[171] S. P. Hussain, J. Schwank, F. Staib, X. W. Wang, and C. C. Harris, "TP53 mutations and hepatocellular carcinoma: insights into the etiology and pathogenesis of liver cancer," *Oncogene*, vol. 26, pp. 2166–2176, apr 2007.

[172] X.-Q. He, Y.-F. Zhang, J.-J. Yu, Y.-Y. Gan, N.-N. Han, M.-X. Zhang, W. Ge, J.-J. Deng, Y.-F. Zheng, and X.-M. Xu, "High expression of g-protein signaling modulator 2 in hepatocellular carcinoma facilitates tumor growth and metastasis by activating the PI3k/AKT signaling pathway," *Tumor Biology*, vol. 39, p. 101042831769597, mar 2017.

[173] M. Çeliktaş, I. Tanaka, S. C. Tripathi, J. F. Fahrmann, C. Aguilar-Bonavides, P. Villalobos, O. Delgado, D. Dhillon, J. B. Dennison, E. J. Ostrin, H. Wang, C. Behrens, K.-A. Do, A. F. Gazdar, S. M. Hanash, and A. Taguchi, "Role of CPS1 in cell growth, metabolism, and prognosis in LKB1-inactivated lung adenocarcinoma," *Journal of the National Cancer Institute*, vol. 109, p. djw231, dec 2016.

[174] R. M. El-Sheikh, S. S. Mansy, I. G. Nessim, H. N. Hosni, A. E. Hindawi, M. H. Hassanein, and A. S. AbdelFattah, "Carbamoyl phosphate synthetase 1 (CPS1) as a prognostic marker in chronic hepatitis c infection," *APMIS*, vol. 127, pp. 93–105, jan 2019.

[175] X.-X. Chen, Y. Yin, J.-W. Cheng, A. Huang, B. Hu, X. Zhang, Y.-F. Sun, J. Wang, Y.-P. Wang, Y. Ji, S.-J. Qiu, J. Fan, J. Zhou, and X.-R. Yang, "BAP1 acts as a tumor suppressor in intrahepatic cholangiocarcinoma by modulating the ERK1/2 and JNK/c-jun pathways," *Cell Death & Disease*, vol. 9, oct 2018.

[176] M. Minor and B. Slagle, "Hepatitis b virus HBx protein interactions with the ubiquitin proteasome system," *Viruses*, vol. 6, pp. 4683–4702, nov 2014.

[177] J. M. Llovet, J. Zucman-Rossi, E. Pikarsky, B. Sangro, M. Schwartz, M. Sherman, and G. Gores, "Hepatocellular carcinoma," *Nature Reviews Disease Primers*, vol. 2, p. 16018, apr 2016.

[178] L. Belloni, T. Pollicino, F. D. Nicola, F. Guerrieri, G. Raffa, M. Fanciulli, G. Raimondo, and M. Levrero, "Nuclear HBx binds the HBV minichromosome and modifies the epigenetic regulation of cccDNA function," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 19975–19979, nov 2009.

[179] L.-N. Qi, T. Bai, Z.-S. Chen, F.-X. Wu, Y.-Y. Chen, B.-D. Xiang, T. Peng, Z.-G. Han, and L.-Q. Li, "The p53 mutation spectrum in hepatocellular carcinoma from guangxi, china : role of chronic hepatitis b virus infection and aflatoxin b1 exposure," *Liver International*, vol. 35, pp. 999–1009, jan 2014.

[180] L.-H. Zhao, X. Liu, H.-X. Yan, W.-Y. Li, X. Zeng, Y. Yang, J. Zhao, S.-P. Liu, X.-H. Zhuang, C. Lin, C.-J. Qin, Y. Zhao, Z.-Y. Pan, G. Huang, H. Liu, J. Zhang, R.-Y. Wang, Y. Yang, W. Wen, G.-S. Lv, H.-L. Zhang, H. Wu, S. Huang, M.-D. Wang, L. Tang, H.-Z. Cao, L. Wang, T. Lee, H. Jiang, Y.-X. Tan, S.-X. Yuan, G.-J. Hou, Q.-F. Tao, Q.-G. Xu, X.-Q. Zhang, M.-C. Wu, X. Xu, J. Wang, H.-M. Yang, W.-P. Zhou, and

H.-Y. Wang, "Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma," *Nature Communications*, vol. 7, p. 12992, oct 2016.

[181] A. Esquela-Kerscher and F. J. Slack, "Oncomirs - microrna with a role in cancer.," *Nature reviews. Cancer*, vol. 6, pp. 259–269, Apr. 2006.

[182] P. P. Medina, M. Nolde, and F. J. Slack, "Oncomir addiction in an in vivo model of microrna-21-induced pre-b-cell lymphoma.," *Nature*, vol. 467, pp. 86–90, Sept. 2010.

[183] X. Ma, D. J. Conklin, F. Li, Z. Dai, X. Hua, Y. Li, Z. Y. Xu-Monette, K. H. Young, W. Xiong, M. Wysoczynski, S. D. Sithu, S. Srivastava, A. Bhatnagar, and Y. Li, "The oncogenic microrna mir-21 promotes regulated necrosis in mice.," *Nature communications*, vol. 6, p. 7151, May 2015.

[184] S. Lin and R. I. Gregory, "Microrna biogenesis pathways in cancer.," *Nature reviews. Cancer*, vol. 15, pp. 321–333, June 2015.

[185] I. Summerer, J. Hess, A. Pitea, K. Unger, L. Hieber, M. Selmansberger, K. Lauber, and H. Zitzelsberger, "Integrative analysis of the microrna-mrna response to radiochemotherapy in primary head and neck squamous cell carcinoma cells.," *BMC genomics*, vol. 16, p. 654, Sept. 2015.

[186] I. Summerer, M. Niyazi, K. Unger, A. Pitea, V. Zangen, J. Hess, M. J. Atkinson, C. Belka, S. Moertl, and H. Zitzelsberger, "Changes in circulating microrna after radiochemotherapy in head and neck cancer patients.," *Radiation oncology (London, England)*, vol. 8, p. 296, Dec. 2013.

[187] V. Agarwal, G. W. Bell, J.-W. Nam, and D. P. Bartel, "Predicting effective microRNA target sites in mammalian mRNAs," *eLife*, vol. 4, aug 2015.

[188] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge, "Prediction of mammalian microrna targets.," *Cell*, vol. 115, pp. 787–798, Dec. 2003.

[189] S. M. Cohen, "Use of microRNA sponges to explore tissue-specific microRNA functions in vivo," *Nature Methods*, vol. 6, pp. 873–874, dec 2009.

[190] A. E. Pasquinelli, "MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship," *Nature Reviews Genetics*, vol. 13, pp. 271–282, mar 2012.

[191] S. Sass, S. Dietmann, U. Burk, S. Brabletz, D. Lutter, A. Kowarsch, K. F. Mayer, T. Brabletz, A. Ruepp, F. Theis, and Y. Wang, "MicroRNAs coordinately regulate protein complexes," *BMC Systems Biology*, vol. 5, no. 1, p. 136, 2011.

[192] K. John, J. Wu, B.-W. Lee, and C. S. Farah, "Micrornas in head and neck cancer.," *International journal of dentistry*, vol. 2013, p. 650218, 2013.

[193] C. B. Lajer, E. Garnæs, L. Friis-Hansen, B. Norrild, M. H. Therkildsen, M. Glud, M. Rossing, H. Lajer, D. Svane, L. Skotte, L. Specht, C. Buchwald, and F. C. Nielsen, "The role of mirnas in human papilloma virus (hpv)-associated cancers: bridging between hpv-related head and neck cancer and cervical cancer.," *British journal of cancer*, vol. 106, pp. 1526–1534, Apr. 2012.

[194] A. M. Gross, R. K. Orosco, J. P. Shen, A. M. Egloff, H. Carter, M. Hofree, M. Choueiri, C. S. Coffey, S. M. Lippman, D. N. Hayes, E. E. Cohen, J. R. Grandis, Q. T. Nguyen, and T. Ideker, "Multi-tiered genomic analysis of head and neck cancer ties tp53 mutation to 3p loss.," *Nature genetics*, vol. 46, pp. 939–943, Sept. 2014.

[195] A. B. Y. Hui, A. Lin, W. Xu, L. Waldron, B. Perez-Ordonez, I. Weinreb, W. Shi, J. Bruce, S. H. Huang, B. O'Sullivan, J. Waldron, P. Gullane, J. C. Irish, K. Chan, and F.-F. Liu, "Potentially prognostic mirnas in hpv-associated oropharyngeal carcinoma.," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 19, pp. 2154–2162, Apr. 2013.

[196] G. Gao, H. A. Gay, R. D. Chernock, T. R. Zhang, J. Luo, W. L. Thorstad, J. S. Lewis, and X. Wang, "A microrna expression signature for the prognosis of oropharyngeal squamous cell carcinoma.," *Cancer*, vol. 119, pp. 72–80, Jan. 2013.

[197] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets," *Cell*, vol. 120, pp. 15–20, jan 2005.

[198] J.-H. Li, S. Liu, H. Zhou, L.-H. Qu, and J.-H. Yang, "starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-seq data," *Nucleic Acids Research*, vol. 42, pp. D92–D97, dec 2013.

[199] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "Kegg for integration and interpretation of large-scale molecular data sets.," *Nucleic acids research*, vol. 40, pp. D109–D114, Jan. 2012.

[200] T. D. Gilmore, "Introduction to NF-κb: players, pathways, perspectives," *Oncogene*, vol. 25, pp. 6680–6684, oct 2006.

[201] C. C. Bancroft, Z. Chen, J. Yeh, J. B. Sunwoo, N. T. Yeh, S. Jackson, C. Jackson, and C. Van Waes, "Effects of pharmacologic antagonists of epidermal growth factor receptor, pi3k and mek signal kinases on nf-kappab and ap-1 activation and il-8 and vegf expression in human head and neck squamous cell carcinoma lines.," *International journal of cancer*, vol. 99, pp. 538–548, June 2002.

[202] Y. Ikeda, E. Tanji, N. Makino, S. Kawata, and T. Furukawa, "MicroRNAs associated with mitogen-activated protein kinase in human pancreatic cancer," *Molecular Cancer Research*, vol. 10, pp. 259–269, dec 2011.

[203] M. Lenarduzzi, A. B. Y. Hui, N. M. Alajez, W. Shi, J. Williams, S. Yue, B. O'Sullivan, and F.-F. Liu, "MicroRNA-193b enhances tumor progression via down regulation of neurofibromin 1," *PLoS ONE*, vol. 8, p. e53765, jan 2013.

[204] I. Mellman and Y. Yarden, "Endocytosis and cancer.," *Cold Spring Harbor perspectives in biology*, vol. 5, p. a016949, Dec. 2013.

[205] L. C. Kelley, S. Shahab, and S. A. Weed, "Actin cytoskeletal mediators of motility and invasion amplified and overexpressed in head and neck cancer.," *Clinical & experimental metastasis*, vol. 25, pp. 289–304, 2008.

[206] M. M. Rietbergen, S. R. Martens-de Kemp, E. Bloemena, B. I. Witte, A. Brink, R. J. Baatenburg de Jong, C. R. Leemans, B. J. M. Braakhuis, and R. H. Brakenhoff, "Cancer stem cell enrichment marker cd98: a prognostic factor for survival in patients with human papillomavirus-positive oropharyngeal cancer.," *European journal of cancer (Oxford, England : 1990)*, vol. 50, pp. 765–773, Mar. 2014.

[207] Z.-J. Ren, X.-Y. Nong, Y.-R. Lv, H.-H. Sun, P. p An, F. Wang, X. Li, M. Liu, and H. Tang, "Mir-509-5p joins the mdm2/p53 feedback loop and regulates cancer cell growth," *Cell Death & Disease*, vol. 5, pp. e1387–e1387, aug 2014.

[208] H. Yamaguchi and J. Condeelis, "Regulation of the actin cytoskeleton in cancer cell migration and invasion," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1773, pp. 642–652, may 2007.

[209] H. T. Morris and L. M. Machesky, "Actin cytoskeletal control during epithelial to mesenchymal transition: focus on the pancreas and intestinal tract," *British Journal of Cancer*, vol. 112, pp. 613–620, jan 2015.

[210] J.-H. Hong, Y. Kwak, Y. Woo, C. Park, S.-A. Lee, H. Lee, S. J. Park, Y. Suh, B. K. Suh, B. S. Goo, D. J. Mun, K. Sanada, M. D. Nguyen, and S. K. Park, "Regulation of the actin cytoskeleton by the ndel1-tara complex is critical for cell migration," *Scientific Reports*, vol. 6, aug 2016.

[211] P. Yuan, X.-H. He, Y.-F. Rong, J. Cao, Y. Li, Y.-P. Hu, Y. Liu, D. Li, W. Lou, and M.-F. Liu, "KRAS/NF-κb/YY1/miR-489 signaling axis controls pancreatic cancer metastasis," *Cancer Research*, vol. 77, pp. 100–111, oct 2016.

[212] Y. Tao, T. Han, T. Zhang, C. Ma, and C. Sun, "LncRNA CHRF-induced miR-489 loss promotes metastasis of colorectal cancer via TWIST1/EMT signaling pathway," *Oncotarget*, vol. 8, apr 2017.

[213] C.-C. Hon, J. A. Ramilowski, J. Harshbarger, N. Bertin, O. J. L. Rackham, J. Gough, E. Denisenko, S. Schmeier, T. M. Poulsen, J. Severin, M. Lizio, H. Kawaji, T. Kasukawa, M. Itoh, A. M. Burroughs, S. Noma,

S. Djebali, T. Alam, Y. A. Medvedeva, A. C. Testa, L. Lipovich, C.-W. Yip, I. Abugessaisa, M. Mendez, A. Hasegawa, D. Tang, T. Lassmann, P. Heutink, M. Babina, C. A. Wells, S. Kojima, Y. Nakamura, H. Suzuki, C. O. Daub, M. J. L. de Hoon, E. Arner, Y. Hayashizaki, P. Carninci, and A. R. R. Forrest, "An atlas of human long non-coding RNAs with accurate 5′ ends," *Nature*, vol. 543, pp. 199–204, mar 2017.

[214] J. E. Wilusz, H. Sunwoo, and D. L. Spector, "Long noncoding RNAs: functional surprises from the RNA world," *Genes & Development*, vol. 23, pp. 1494–1504, jul 2009.

[215] J. S. Mattick and J. L. Rinn, "Discovery and annotation of long noncoding RNAs," *Nature Structural & Molecular Biology*, vol. 22, pp. 5–7, jan 2015.

[216] K. C. Wang and H. Y. Chang, "Molecular mechanisms of long noncoding rnas.," *Molecular cell*, vol. 43, pp. 904–914, Sept. 2011.

[217] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander, "Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals.," *Nature*, vol. 458, pp. 223–227, Mar. 2009.

[218] S. Geisler and J. Coller, "Rna in unexpected places: long non-coding rna functions in diverse cellular contexts.," *Nature reviews. Molecular cell biology*, vol. 14, pp. 699–712, Nov. 2013.

[219] A. M. Khalil, M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander, and J. L. Rinn, "Many human large intergenic noncoding rnas associate with chromatin-modifying complexes and affect gene expression.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 11667–11672, July 2009.

[220] R. Mitra, X. Chen, E. J. Greenawalt, U. Maulik, W. Jiang, Z. Zhao, and C. M. Eischen, "Decoding critical long non-coding RNA in ovarian can-

cer epithelial-to-mesenchymal transition," *Nature Communications*, vol. 8, nov 2017.

[221] Z. Bian, J. Zhang, M. Li, Y. Feng, S. Yao, M. Song, X. Qi, B. Fei, Y. Yin, D. Hua, and Z. Huang, "Long non-coding RNA LINC00152 promotes cell proliferation, metastasis, and confers 5-FU resistance in colorectal cancer by inhibiting miR-139-5p," *Oncogenesis*, vol. 6, nov 2017.

[222] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, and R. Guigo, "The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression," *Genome Research*, vol. 22, pp. 1775–1789, sep 2012.

[223] Q. Jiang, R. Ma, J. Wang, X. Wu, S. Jin, J. Peng, R. Tan, T. Zhang, Y. Li, and Y. Wang, "Lncrna2function: a comprehensive resource for functional investigation of human lncrnas based on rna-seq data.," *BMC genomics*, vol. 16 Suppl 3, p. S2, 2015.

[224] C. Park, N. Yu, I. Choi, W. Kim, and S. Lee, "lncrnator: a comprehensive resource for functional investigation of long non-coding rnas.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 2480–2485, Sept. 2014.

[225] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn, "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses," *Genes & Development*, vol. 25, pp. 1915–1927, sep 2011.

[226] J. Gong, W. Liu, J. Zhang, X. Miao, and A.-Y. Guo, "lncrnasnp: a database of snps in lncrnas and their potential functions in human and mouse.," *Nucleic acids research*, vol. 43, pp. D181–D186, Jan. 2015.

[227] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes,

A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo, and T. J. Hubbard, "GENCODE: The reference human genome annotation for the ENCODE project," *Genome Research*, vol. 22, pp. 1760–1774, sep 2012.

[228] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. Searle, G. Spudich, S. J. Trevanion, A. Yates, D. R. Zerbino, and P. Flicek, "Ensembl 2015," *Nucleic Acids Research*, vol. 43, pp. D662–D669, oct 2014.

[229] M. Melé, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segrè, S. Djebali, A. Niarchou, G. Consortium, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, and R. Guigó, "Human genomics. the human transcriptome across tissues and individuals.," *Science (New York, N.Y.)*, vol. 348, pp. 660–665, May 2015.

[230] B. D. Kumar and R. Krumlauf, "HOXs and lincRNAs: Two sides of the same coin," *Science Advances*, vol. 2, pp. e1501402–e1501402, jan 2016.

[231] J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang, "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs," *Cell*, vol. 129, pp. 1311–1323, jun 2007.

[232] M. Baker, "Long noncoding rnas: the search for function," *Nat Meth*, vol. 8, no. 5, pp. 379–383, May 2011.

[233] P. Jaccard, "THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1," *New Phytologist*, vol. 11, pp. 37–50, feb 1912.

[234] R. Petryszak, T. Burdett, B. Fiorelli, N. A. Fonseca, M. Gonzalez-Porta, E. Hastings, W. Huber, S. Jupp, M. Keays, and N. e. a. Kryvych, "Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments," *Nucleic Acids Research*, vol. 42, no. Database issue, pp. D926–932, Jan 2014.

[235] J. B. Permuth, D.-T. Chen, S. J. Yoder, J. Li, A. T. Smith, J. W. Choi, J. Kim, Y. Balagurunathan, K. Jiang, D. Coppola, B. A. Centeno, J. Klapman, P. Hodul, F. A. Karreth, J. G. Trevino, N. Merchant, A. Magliocco, M. P. Malafa, and R. Gillies, "Linc-ing circulating long non-coding RNAs to the diagnosis and malignant prediction of intraductal papillary mucinous neoplasms of the pancreas," *Scientific Reports*, vol. 7, sep 2017.

[236] S. Bauer, J. Gagneur, and P. N. Robinson, "GOing Bayesian: model-based gene set analysis of genome-scale data," *Nucleic Acids Research*, vol. 38, no. 11, pp. 3523–3532, Jun 2010.

[237] L. Wu, , W. Shi, J. Long, X. Guo, K. Michailidou, J. Beesley, M. K. Bolla, X.-O. Shu, Y. Lu, Q. Cai, F. Al-Ejeh, E. Rozali, Q. Wang, J. Dennis, B. Li, C. Zeng, H. Feng, A. Gusev, R. T. Barfield, I. L. Andrulis, H. Anton-Culver, V. Arndt, K. J. Aronson, P. L. Auer, M. Barrdahl, C. Baynes, M. W. Beckmann, J. Benitez, M. Bermisheva, C. Blomqvist, N. V. Bogdanova, S. E. Bojesen, H. Brauch, H. Brenner, L. Brinton, P. Broberg, S. Y. Brucker, B. Burwinkel, T. Caldés, F. Canzian, B. D. Carter, J. E. Castelao, J. Chang-Claude, X. Chen, T.-Y. D. Cheng, H. Christiansen, C. L. Clarke, M. Collée, S. Cornelissen, F. J. Couch, D. Cox, A. Cox, S. S. Cross, J. M. Cunningham, K. Czene, M. B. Daly, P. Devilee, K. F. Doheny, T. Dörk, I. dos Santos-Silva, M. Dumont, M. Dwek, D. M. Eccles, U. Eilber, A. H. Eliassen, C. Engel, M. Eriksson, L. Fachal, P. A. Fasching, J. Figueroa, D. Flesch-Janys, O. Fletcher, H. Flyger, L. Fritschi,

M. Gabrielson, M. Gago-Dominguez, S. M. Gapstur, M. García-Closas, M. M. Gaudet, M. Ghoussaini, G. G. Giles, M. S. Goldberg, D. E. Goldgar, A. González-Neira, P. Guénel, E. Hahnen, C. A. Haiman, N. Håkansson, P. Hall, E. Hallberg, U. Hamann, P. Harrington, A. Hein, B. Hicks, P. Hillemanns, A. Hollestelle, R. N. Hoover, J. L. Hopper, G. Huang, K. Humphreys, D. J. Hunter, A. Jakubowska, W. Janni, E. M. John, N. Johnson, K. Jones, M. E. Jones, A. Jung, R. Kaaks, M. J. Kerin, E. Khusnutdinova, V.-M. Kosma, V. N. Kristensen, D. Lambrechts, L. L. Marchand, J. Li, S. Lindström, J. Lissowska, W.-Y. Lo, S. Loibl, J. Lubinski, C. Luccarini, M. P. Lux, R. J. MacInnis, T. Maishman, I. M. Kostovska, A. Mannermaa, J. E. Manson, S. Margolin, D. Mavroudis, H. Meijers-Heijboer, A. Meindl, U. Menon, J. Meyer, A. M. Mulligan, S. L. Neuhausen, H. Nevanlinna, P. Neven, S. F. Nielsen, B. G. Nordestgaard, O. I. Olopade, J. E. Olson, H. Olsson, P. Peterlongo, J. Peto, D. Plaseska-Karanfilska, R. Prentice, N. Presneau, K. Pylkäs, B. Rack, P. Radice, N. Rahman, G. Rennert, H. S. Rennert, V. Rhenius, A. Romero, J. Romm, A. Rudolph, E. Saloustros, D. P. Sandler, E. J. Sawyer, M. K. Schmidt, R. K. Schmutzler, A. Schneeweiss, R. J. Scott, C. G. Scott, S. Seal, M. Shah, M. J. Shrubsole, A. Smeets, M. C. Southey, J. J. Spinelli, J. Stone, H. Surowy, A. J. Swerdlow, R. M. Tamimi, W. Tapper, J. A. Taylor, M. B. Terry, D. C. Tessier, A. Thomas, K. Thöne, R. A. E. M. Tollenaar, D. Torres, T. Truong, M. Untch, C. Vachon, D. V. D. Berg, D. Vincent, Q. Waisfisz, C. R. Weinberg, C. Wendt, A. S. Whittemore, H. Wildiers, W. C. Willett, R. Winqvist, A. Wolk, L. Xia, X. R. Yang, A. Ziogas, E. Ziv, A. M. Dunning, P. D. P. Pharoah, J. Simard, R. L. Milne, S. L. Edwards, P. Kraft, D. F. Easton, G. Chenevix-Trench, and W. Z. and, "A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer," *Nature Genetics*, vol. 50, pp. 968–978, jun 2018.

[238] M. Uhlen, C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors, K. Sanli, K. von Feilitzen, P. Oksvold, E. Lundberg, S. Hober, P. Nilsson, J. Mattsson, J. M. Schwenk, H. Brunnström, B. Glimelius, T. Sjöblom, P.-H. Edqvist,

D. Djureinovic, P. Micke, C. Lindskog, A. Mardinoglu, and F. Ponten, "A pathology atlas of the human cancer transcriptome," *Science*, vol. 357, p. eaan2507, aug 2017.

[239] S. Mamlouk, L. H. Childs, D. Aust, D. Heim, F. Melching, C. Oliveira, T. Wolf, P. Durek, D. Schumacher, H. Bläker, M. von Winterfeld, B. Gastl, K. Möhr, A. Menne, S. Zeugner, T. Redmer, D. Lenze, S. Tierling, M. Möbs, W. Weichert, G. Folprecht, E. Blanc, D. Beule, R. Schäfer, M. Morkel, F. Klauschen, U. Leser, and C. Sers, "DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer," *Nature Communications*, vol. 8, jan 2017.