TUM SCHOOL OF LIFE SCIENCES

TECHNISCHE UNIVERSITÄT MÜNCHEN

Dissertation

Towards precision medicine: computational approaches for patient stratification and biomarker identification in oncology

Fabio Boniolo

TUM SCHOOL OF LIFE SCIENCES

TECHNISCHE UNIVERSITÄT MÜNCHEN

Towards precision medicine: computational approaches for patient stratification and biomarker identification in oncology

Fabio Boniolo

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines **Doktors der Naturwissenschaften (Dr. rer. nat.)** genehmigten Dissertation.

Vorsitz: Prof. Dr. Mathias Wilhelm
Prüfer der Dissertation:
1. Prof. Dr. Jan Baumbach
2. Prof. Dr. Dieter Saur

Die Dissertation wurde am 10.05.2022 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 16.09.2022 angenommen.

Acknowledgements

First and foremost, I would like to thank my mentor and main PI Prof. Dr. Dieter Saur, for giving me the chance to join his group and to start my doctorate under his supervision. You gave me the chance to learn and thrive in a highly interdisciplinary setting and I am extremely thankful for this. Additionally, I would like to thank Prof. Dr. Jan Baumbach, who allowed me to join his Chair and welcomed me in the fantastic environment he had created in Fresing at the Chair of Experimental Bioinformatics. Your mentorship and support have been invaluable. I would also like to express my gratitude to Prof. Dr. Bernhard Küster for agreeing to become my main supervisor when needed, and for his comments and feedbacks. Last but not least, I would like to thank my mentor, Dr. Markus List. Your support and motivation over the years have been critical and your mentorship has been impeccable.

Furthermore, I would like to thank all the colleagues, collaborators, and scientists I had the privilege to work with these past few years. Thank you, Ananth, Daniele, Markus, Stefanie, Sebastian, Stella, Christian, and Markus. I am extremely proud and honored to have met you and wish you all the best for your journey. A special thank you goes to Chiara, Valentina, and Melissa. I have no doubt you will all achieve everything you desire.

The biggest thank you of all goes to my family, papà, mamma, and Irene. You have been the best example of perseverance and sacrifice, while showing me constant love and support. I owe everything I am to you and I realize more every day how blessed I am to have you around me.

Finally, I would like to express my gratitude to Belle. You have been the unexpected element in an otherwise carefully designed plan. Your enthusiasm, motivation, and support push me every day and I cannot wait for the many adventures we have ahead of us.

Abstract

Cancer is a class of diseases characterized by the accumulation of mutations in healthy cells and the progressive aberrations of physiological mechanisms that lead to abnormal growth, proliferation, and ultimately the invasion of neighboring and distant tissues. In the past decades, improvements in cancer prevention and treatment and an increased understanding of the basic mechanisms of cancer genesis, progression, and maintenance have led to significant improvements in outcomes for many cancer types. The rise of new medical approaches, such as precision medicine, and powerful technologies, like artificial intelligence, is destined to give further impulse to this trend and result in more efficient diagnostic, prognostic, and therapeutic strategies. In this work, I show how computational approaches can pave the way for precision medicine approaches in oncology and present two studies where I exploit machine learning techniques to analyze large molecular datasets to stratify observations and identify mechanistic biomarkers. In the first study, I present the design of a new tool for the inference of patient- or sample-specific post-transcriptional regulatory subnetworks. The identified subnetworks, or modules, summarise the contributions of miRNAs and competing endogenous RNAs, also known as microRNA sponges, in the regulation of RNAs with shared microRNA binding sites and allow for the identification of important biomarkers. I showcase the designed method by applying it to a breast cancer subtype classification example. In the second study, I introduce an innovative pharmacogenomic pipeline designed to predict drug response values resulting from high-throughput drug screens from the transcriptional profiles of 251 murine pancreatic ductal adenocarcinoma cell cultures. I show how the integration of *a priori* knowledge, in the form of gene sets, and overall general levels of drug sensitivity across the screened cohort substantially increases the performance of the prediction models and leads to the identification of response biomarkers that can be further validated with functional

assays. This work lays the foundation for the implementation of advanced computational methods for precision medicine-based approaches such as patient stratification and biomarker identification in pre-clinical and clinical datasets.

Kurzfassung

Krebs umfasst eine Klasse von Krankheitsbildern, die durch die Anhäufung von Mutationen in gesunden Zellen und die fortschreitende Abweichung von physiologischen Mechanismen gekennzeichnet sind. Diese führen zu abnormalem Wachstum, Proliferation und schließlich zur Invasion von benachbartem und entferntem Gewebe. In den letzten Jahrzehnten haben Fortschritte in der Krebsvorbeugung und -behandlung sowie ein besseres Verständnis der grundlegenden Mechanismen der Krebsentstehung, -progression und -erhaltungstherapie bei vielen Krebsarten zu deutlich verbesserten Behandlungsergebnissen geführt. Das Aufkommen neuer medizinischer Ansätze wie der Präzisionsmedizin sowie die Entwicklung leistungsfähiger Technologien wie der künstlichen Intelligenz werden diesen Trend weiter vorantreiben und zu effizienteren diagnostischen, prognostischen und therapeutischen Strategien führen. In dieser Arbeit zeige ich, wie computergestützte Ansätze den Weg für präzisionsmedizinische Konzepte in der Onkologie ebnen können. Daher stelle ich zwei Studien vor, in denen ich Machine-Learning-Technologien zur Analyse großer molekularer Datensätze nutze, um Beobachtungen zu stratifizieren und mechanistische Biomarker zu identifizieren. In der ersten Studie zeige ich das Design eines neuen Tools, das die Einflussnahme von patienten- oder probenspezifische post-transkriptionellen regulatorischen Subnetzwerken zeigt. Die identifizierten Subnetzwerke oder Module fassen die Beteiligungen von miRNAs und konkurrierenden endogenen RNAs, auch bekannt als microRNA-Sponges, bei der Regulierung von RNAs mit gemeinsamen microRNA-Bindungsstellen zusammen und ermöglichen die Identifizierung wichtiger Biomarker. Ich demonstriere die Nutzung der entwickelten Methode anhand von Beispielen zur Klassifizierung von Brustkrebs-Subtypen. In der zweiten Studie stelle ich eine innovative pharmakogenomische Pipeline vor, die darauf ausgelegt ist, aus den Transkriptionsprofilen von 251 Zellkulturen des duktalen Adenokarzinoms der Bauchspe-

icheldrüse von Mäusen Werte für das Ansprechen auf Arzneimitteltherapie vorauszusagen, die sich aus high-throughput Wirkstoffscreens ergeben. Ich zeige, wie die Integration von A-priori-Wissen in Form von Gene-sets und allgemeiner Arzneimittelempfindlichkeit in der untersuchten Kohorte die Leistung der Vorhersagemodelle erheblich steigert. Dies führt des Weiteren zur Identifizierung von Response-Biomarkern, die mit funktionellen Assays weiter validiert werden können. Diese Arbeit legt den Grundstein für die Entwicklung fortschrittlicher Berechnungsmethoden für Präzisions-Ontologie basierte Ansätze wie Patientenstratifizierung und Biomarker-Identifizierung in präklinischen und klinischen Datensätzen.

Contents

1 Introduction

Cancer is the second-leading cause of death globally, accounting for nearly 1 out of 6 deaths in 2020 [1], and is projected to become the leading cause of premature deaths worldwide by the end of this century [2].

The term cancer is used to define a broad class of diseases sharing the production of abnormal cells that rapidly grow, divide, spread, and eventually invade neighboring tissues of the host organisms in a process called metastasis. The transformation of healthy normal cells into tumor ones is extremely complex and multifaceted. It takes place as a sequence of steps that drives the degeneration of healthy tissues into malignant ones. Various causes have been linked to cancer, such as internal genetic factors and exposure to external agents such as radiation and chemical or biological carcinogens. At the same time, multiple factors have been associated with increased susceptibility to these diseases, such as the use of tobacco or alcohol, unhealthy diets, air pollution, or infections (e.g., from the human papillomavirus) [3, 4, 5]. Moreover, socio-economical disparities have been shown to contribute to differences in cancer incidence and mortality numbers, mainly due to limited access to healthcare services in disadvantaged or isolated communities [6].

Despite the 19.3 million new cases and almost 10 million deaths in 2020 [7], death rates from many cancer types have been steadily falling in the last decade [8, 9], with improvements for 11 of the 19 and 14 of the 20 most common cancers in men and women respectively [10]. While being partly due to broadened access to basic cancer care services and improvements in preventive, diagnostic, and prognostic technologies, these improvements can be directly linked to an increased understanding of the biological mechanisms driving cancer formation, progression, and maintenance. These discoveries have been successfully translated to the

clinical setting and fueled new approaches in oncology, such as the use of patient-specific information to drive clinical decisions, i.e., precision oncology.

In parallel to these advancements, biology and medicine have witnessed the unprecedented production and accumulation of large quantities of different types of data, offering the chance for exploration, mining, and hypothesis validation while exploiting the power of emerging technologies such as artificial intelligence (AI) and machine learning (ML). These tools are poised to deliver further impulses to cancer research along the translational pipeline and impact the way medicine is perceived and performed [11, 12]. AI applications in oncology have found most of their success in imagining applications, where seminal works have shown the potential of these agents for image-based diagnosis and prognosis [13]. Moreover, they have demonstrated the potential to tackle tasks such as prediction of treatment response, design of novel therapies, and clinical decision-making [14].

In addition to clinical applications, machine learning tools are becoming integral tools of the scientific process. They can be designed and trained to run *in silico* experiments and interrogated to study critical biological mechanisms [15, 16, 17]. This is partly due to the widespread accumulation of molecular information, made possible by technological advancements that now allow the collection of thousands of measurements from multiple patients simultaneously.

In this work, I investigate the potential of computational techniques for biological discoveries in cancer biology. In particular, I focus on the importance these tools have towards the realization of precision oncology approaches. In Chapter 1, I will introduce the main biological mechanisms regulating cellular homeostasis (i.e., equilibrium) that are relevant for this thesis, emphasizing how they are altered in cancer. In Chapter 2, I follow with an overview of precision medicine and artificial intelligence. I describe the potential of the integration of the two fields and describe a few applications where computational approaches are already leading to new insights and discoveries. Chapter 3 and Chapter 4 contain the methodological contributions of my work and showcase two potential applications of computational techniques for the analysis of large and multi-dimensional datasets, namely the inference of post-transcriptional regulatory networks and the pharmacogenomic analysis of

high-throughput drug screens. I conclude this work with a personal take on the potential of Artificial Intelligence in biomedicine, with a particular focus on promising future research directions.

2 A primer to cancer biology

The human body is estimated to consist, on average, of 4×1013 cells [18] that constantly work together to give rise to a considerable diversity of structures. Such organization is managed through a complex network of interactions and signals that set whether every single cell should rest, divide, differentiate or die. Cell-cell interactions determine the possibility for cells to cooperate and, ultimately, allow the preservation and maintenance of tissues and organs throughout the lifespan of an organism. In physiological conditions, cellular behavior is tightly controlled to maintain homeostasis and the stability of the whole system [19]. In particular, basic functions such as cell duplication, differentiation, or apoptosis (i.e., programmed cell death) must be aligned, given that improper control of any of these mechanisms may disrupt the equilibrium and lead to abnormal behaviors, such as uncontrolled proliferation, which can, in turn, lead to cancer.

Proteins are one of the main molecules playing a role in controlling intracellular and intercellular communications. The process that drives the synthesis of proteins can be schematically described following the central dogma of biology, which defines the directionality of the process that leads from DNA to RNA and finally to proteins. The copying of a DNA sequence, and more specifically of the DNA functional units (i.e., genes), into an RNA one is called transcription and is strictly controlled during the life cycle of a cell. Transcribed genes are considered to be expressed, while those not actively taking part in the transcription process are deemed to be repressed. Once transcribed, RNA molecules are further translated into sequences of amino acids that ultimately form proteins. While the long-standing problem of the definition of the unique 3D structure of proteins based on the amino acid sequence has received a great impulse from computational technologies and artificial intelligence [20, 21], the link between genotypes, i.e., the genetic makeup of organisms defined by the sequence

of DNA bases, and phenotypes, i.e., observable or measurable traits that can range from complex behaviors to morphology, is still obscure in many complex diseases such as cancer, where different alterations of physiological mechanisms in different patients might lead to the same observable phenotype [22].

All the information necessary to maintain equilibrium, i.e., homeostasis, is contained in the DNA sequence, structured as a sequence of four nucleotides, adenine (A), cytosine (C), guanine (G), and thymine (T). Variations in the DNA sequence, structure, and organization can take many forms and involve portions of varying lengths, ranging from a single nucleotide to megabases. The size of the DNA sequence affected by the aberration defines the type of variations. Changes in a single nucleotide (e.g., substitutions, insertions, or deletions) are typically classified as Single Nucleotide Variants (SNVs) or Single Nucleotide Polymorphisms (SNPs), depending on whether their population frequencies are, respectively, below or above 1%. Longer aberrations may consist of insertion or deletion of a few nucleotides (indels) or whole segments, Copy Number Variations (CNVs), and are typically grouped under the umbrella of structural variants together with more significant aberrations that modify chromosome structure, such as translocations or inversions [23]. Generally, these aberrations are the results of mistakes occurring during DNA replication and can accumulate during the lifespan of an individual.

Genetic variants may affect both genomic regions that serve as templates for the production of proteins, called coding regions, or regions not directly associated with any protein, non-coding regions, and thus result in the production of aberrant proteins or the alteration of key mechanisms such as gene regulation. Genetic variants can be of interest if associated with specific diseases and have drawn a lot of interest as genetic markers of disease [23]. Interestingly, multiple studies have found that genetic variants can often be associated with non-coding regions, offering the chance to investigate how aberrations impact gene regulation and, ultimately, the progression and maintenance of a disease [24, 25].

Cells that accumulate aberrant modifications may escape homeostasis control mechanisms and give rise to cancer. Typically, cancers are classified according to the tissue and cell types of origin. Most frequently diagnosed tumors arise from epithelial cells, i.e., cells that form layers

covering channels and ducts [26]. Cancers arising from these cell types are referred to as carcinomas. They can be further characterized as squamous cell carcinomas if they originated from epithelial cells forming protective layers or as adenocarcinomas if the epithelial cells differentiated to secrete substances into ducts. Other cell types, such as those constituting connective tissues or muscles (mesenchymal cells or fibroblasts), give rise to sarcomas. Finally, cancers may arise from cell types present in the blood that are part of the immune system and are classified based on specific cell types, e.g., leukemia.

Multiple single and independent mutational events must happen for the carcinogenesis process to start. Tumor progression is often a slow process defined by the accumulation of mutations in multiple genes of the cancer cells, driving their behavior from an initial status of disorder to a malignant one. Such progression is a cycle in which cells descended from a single mutant ancestor evolve to more aggressive stages by successive steps of mutation and selection. At each step, new mutations are introduced to overcome the complexity and interconnection of cellular systems. Mutations conferring further selective advantages to tumor cells are called 'driver' mutations, as opposed to neutral, or 'passenger', mutations that do not directly impact tumor cells' fitness but may confound the search for causal mechanisms driving tumorigenesis. This cycle is aided by specific characteristics shared by tumor cells, like: i) resisting cell death, ii) sustained proliferative signaling, iii) evasion of growth suppressor, iv) enablement of replicative immortality, v) induction of angiogenesis, and vi) activation of invasion and metastasis (as reported in [27, 28, 29, 26]). To understand the process underlying cancer initiation, progression, and maintenance and ultimately define efficient treatment strategies, it is pivotal to understand the molecular mechanisms that give rise to malignant cells to identify which genes harbor the relevant mutations and how they cooperate.

Cancer-associated genes tend to be classified into oncogenes and tumor suppressor genes [30]. Such classes operate in opposite ways, increasing cancer cells' proliferation and survival. Typically, oncogenes act in a dominant manner, where gain-of-function mutations (i.e., mutations in a copy of the gene lead to overactivity) in specific regions of proto-oncogenes drive a cell towards cancer. A typical example of an oncogene is KRAS, which leads to uncontrolled

cell division and survival when mutated [29]. On the other hand, tumor suppressor genes act in a recessive manner, where a loss-of-function mutation allows cancer cells to overcome barriers to proliferation and division. The first example of a tumor suppressor was the RB gene, a major cell cycle regulator [31].

2.1 Transcription and gene expression regulation

Over 21000 coding genes have been cataloged in the human genome, working and being activated in different configurations to give rise to the extensively observed heterogeneity in functions, structures, and cell types at the basis of living tissues and organs. Phenotypic heterogeneity results from changes in the expression of genes, driving the synthesis of sets of molecules without altering the DNA sequence. Such changes happen in response to specific stimuli and may differ from cell type to cell type.

RNA synthesis, a direct result of gene expression, is a complex process. The initial RNA molecule can be as long as the parent gene it derives from. In the first steps, segments of different lengths called introns are cut out of the pre-mRNA. The remaining sequences, called exons, are flanked together in a process that takes the name of splicing and results in a molecule called messenger RNA (mRNA). Diversity in spliced regions leads to variety in downstream proteins starting from a single gene. Alternative splicing, i.e., the different combinations of exons that form mature mRNA strands, has been shown to play a role in offering evolutionary advantage [32], differentiation [33], and development [34] and to be regulated at the tissue level so that tissue-specific variants can cooperate at in protein-protein interactions [35]. Being such an important step in the synthesis of RNA, it is no surprise that alternative splicing and its deregulation play a role in the biology of cancer [36]. After splicing, the mature molecule of mRNA is exported to the cytoplasm and to the ribosomes, where it serves as a template for the synthesis of proteins.

Different steps during transcription and translation can be regulated (Figure 2.1). These are briefly described here, but the reader is encouraged to read more at [26], given that in this work, I will focus only on one of these mechanisms. Any given cell can adjust protein

synthesis by controlling i) the transcription rate of a gene (transcriptional control), ii) RNA splicing, iii) the transport of RNA transcripts outside of the nucleus and to specific areas in the cytosol, iv) selection of which mRNA to translate via ribosome, and v) degradation of specific mRNA molecules in the cytoplasm.

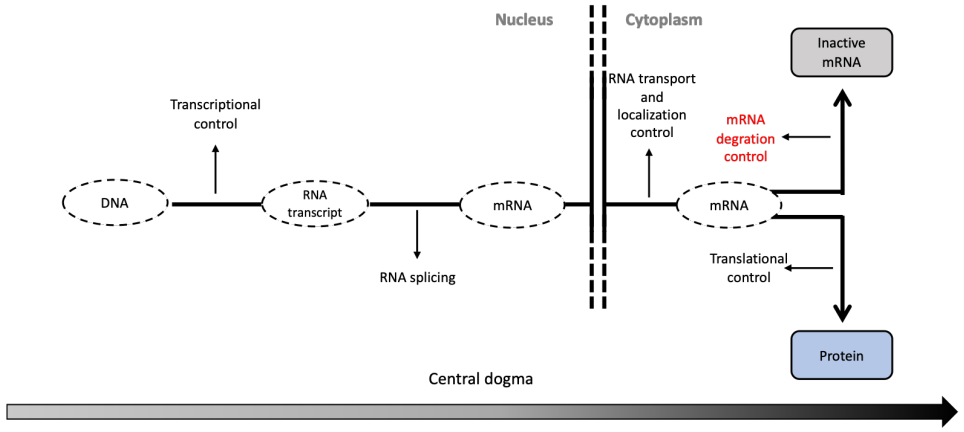

Figure 2.1: Schematic representation of the process leading to the synthesis of new proteins. In red, the control mechanism further investigated in this work.

2.1.1 Transcription factors

Complex phenotypes derive from the coordinated transcription of groups of genes and simultaneous repression of others. Such coordination is achieved through transcription factors (TFs), specialized proteins designed to target and bind specific regions of a gene, such as the promoter or upstream enhancer regions, to control its transcription [26]. Their binding is facilitated by the recognition of sequences in the promoter region of a gene, called motif. Every single TF has the potential to recognize multiple motifs in different genes, thus having the ability to control the transcription of many different elements simultaneously. In pathological conditions, e.g., in cancer, malfunctioning of one of these proteins may lead to downstream activation/repression of multiple genes that might contribute to the observed aggressive phenotypes.

Gene activation depends upon the action of several of such molecules that must be recognized by the enhancer sequences of the target genes to activate expression. 1600 coding

genes in the human genome are identified as transcription factor genes [37], making the analysis of the combined effect of this multitude of interactions often intractable. Given their importance, the dysfunction of TFs can lead to aberrant cell behaviors and assume a leading role in tumorigenesis, tumor progression, and maintenance. Indeed, around 20% of typically identified oncogenes are TFs [38] supporting key processes in cancer cells. Similarly, loss-of-function of tumor-suppressor TFs may lead to uncontrollable proliferation, as shown by loss-of-function events in the *TP53* gene that has been measured in 50% of cancers [39], or an increase in metastatic potential, as happens for example with mutations of the *KLF4* gene, known to maintain E-cadherin expression while reducing SLUG expression to control metastasis [40, 41][40,41]. Other TFs of interest, whose activity is often mentioned in cancer studies, are the ones belonging to the Myc family (containing three proteins, c-MYC, N-MYC, and L-MYC), known to be involved in cell growth, proliferation, and differentiation and to be often dysfunctional in cancer [42]. Given their pivotal role, the study of TFs in cancer regulation is a key topic, both to elucidate tumor initiation and progression mechanisms and to identify new potential druggable targets. Given the complexity of the interactions involved, computational methods able to capture patterns and identify important connections have assumed a key role in this effort. In particular, algorithms inferring transcriptional regulatory networks based on different data sources and existing biological knowledge and designed to highlight the role of TFs in regulating groups of genes (collectively referred to as regulon) [43] that have been shown to play a role in the biology of cancers such as breast adenocarcinoma [44, 45].

2.1.2 Chromatin accessibility and methylation

TFs and the RNA transcription machinery (e.g., RNA polymerase II, a molecule driving transcription) directly interact with DNA. In the absence of transcription, the DNA is packed together with multiple proteins into a tight structure called chromatin. The way these proteins allow the interaction between TFs and the DNA chain is a major determinant of gene expression.

The functional units of chromatin are called nucleosomes, which are made up of four

histone proteins (H2A, H2B, H3, and H4) that behave as spools around which small portions of DNA (147 bp ca.) are wrapped [46, 47]. Tails of the core histones are exposed from the nucleosome and subject to modifications that alter chromatin structure. Different families of proteins, containing domains such as the bromodomain, target and bind these units and lead to structural changes that open genomic regions to interact with transcriptional regulators such as TFs [48].

Given its role in regulating gene expression, many cancer genomes are characterized by mutations in chromatin-related structures [49] and histone modifications [50] directly linked to tumor development. For example, inactivation of the SWI/SNF complex, responsible for chromatin remodeling, resulted in the direct silencing of the *CDKN2* gene, a widely acknowledged tumor-suppressor controlling cell proliferation [51, 52]. Likewise, alterations in the coding portions of histone H3 have been identified as of important in cancers such as pediatric glioma [53].

Changes happening without modification of the DNA sequence are typically grouped under the umbrella of epigenetic changes, such as modifications of histones (e.g. acetylation), or DNA methylation. DNA methylation can be loosely defined as the addition of a methyl group (-CH3) to the DNA sequence. CH3 addition often happens to a cytosine ring found next to a guanine base, giving rise to the so-called CpG sites. While the majority of CpG sites in the genome are methylated, those found in gene start sites are often protected from such modification. In homeostatic conditions, methylation patterns are believed to preserve DNA packaging and control unwanted transcription and gene expression [54]. On the contrary, it has been found that cancer genomes are often characterized by global losses of methylation patterns (i.e., hypomethylation) [55], with frequent modifications of CpG sites at the start sites of genes typically involved in key pathways regulating cell growth, cell cycle, proliferation and differentiation [56]. Hypermethylation events in cancer are commonly observed in tumor suppressor genes regulating cell cycle, as in the case of *CDKN2* on chromosome 9p21, frequently a target of methylation in breast and non-small cell lung cancer [57, 58]. Unlike genetic changes, epigenetic aberrations are reversible and offer an appealing target for the development of new targeted inhibitors [50].

2.1.3 Small and non-coding RNAs

Only 3% of the genome is constituted by protein-coding genes [59]. The remaining portion has been historically referred to as non-coding or "junk DNA", given the absence of indications that these DNA regions had a clear biological purpose. On the opposite, projects like the Encyclopedia of DNA Elements (ENCODE) [60] have revealed that at least 75% of the DNA is transcribed into RNAs, opening new research avenues for the understanding of non-coding RNAs [61] In particular, a growing body of evidence suggests that non-coding genes play an important role in gene regulation [62]. Here, I focus on two classes of non-coding RNA, microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), and on their relationship in the framework of a recently introduced layer of post-transcriptional regulation, i.e. competing endogenous RNA (ceRNA) networks.

MicroRNAs

Towards the end of the 20th century, a new class of RNA molecules, alongside mRNAs and other RNA molecules such as ribosomal RNAs, transfer RNAs, and small nuclear RNAs, has emerged as involved in controlling mRNA levels and translation. These molecules, called microRNAs (miRNAs), are 21-25bp long (when mature) and able to bind with different mRNA targets to drive their post-transcriptional activities. MiRNAs are typically encoded in intronic regions, with 54% of them originating from non-coding transcripts [63], and often occupy neighboring genomic regions, called clusters, that are collectively transcribed. Transcription of these loci results in the production of primary miRNAs (pri-miRNAs) that are processed by the microprocessor complex into precursor miRNAs (pre-miRNAs), 70bp long molecules containing a terminal stem-loop [64]. Pre-miRNAs are then exported to the cytoplasm, where the loop is clipped by DICER to produce mature miRNA strands (Figure 2.2.a) [65]. Overall, miRNA biogenesis has not been fully understood yet, limiting the investigation of the mechanisms behind miRNA regulation. For example, the location of promoter regions regulating miRNA transcription is still a matter of debate [66]. Moreover, details about the subcellular localization and transport of miRNAs are still lacking [65].

Once mature, miRNA strands are loaded into the 4 AGO proteins encoded by the mam-

malian genome (argonaute proteins) [67] to form the RNA-induced silencing complex (RISC) [68]. Once loaded onto the complex, miRNAs pair with their regulatory targets by matching their seed region, located at the 5' end, to a specific mRNA binding site, called microRNA response element (MRE). While typical binding regions have been found at the 3' ends of target miRNAs (canonical targeting), the action of other locations outside of the seed has been shown to contribute to the recognition of the target (non-canonical targeting) and the downstream effects of the pairing [69, 70]. Indeed, perfect matching of the seed region to the MRE results in the cleavage of the target by the AGO proteins, while incomplete matching leads to the recruitment of additional proteins that can mediate silencing through a combination of various mechanisms [67].

miRNAs are traditionally grouped in families, whose members are defined based on the sharing of the same seed sequences and/or similar pre-miRNA structures [71]. miRNA families have assumed an important role in the study of these molecules, in that miRNAs belonging to the same family are believed to share specific biological functions [71]. Furthermore, genomic studies highlighted that miRNAs part of the same families tend to be localized around key genes involved in crucial cellular processes such as signal transduction, proliferation, and inflammation [72].

While the exact number of existing miRNAs and their targets is unclear, it is estimated that they regulate at least half of the genes in our genome [73]. As in the case of TFs described above, a single miRNA can regulate multiple mRNAs and can thus exercise its regulatory control on multiple cellular processes and pathways. Given their function, it has been demonstrated that miRNAs could play a key role in health and disease. In cancer, miRNAs have been shown to act either as tumor suppressors, taking the name of tumor-suppressing miRs (TSmiRs), or as oncogenes (oncomiRs). For example, members of the miR34 family have been found to be dysregulated in many cancers and to be directly associated with the activity of *TP53*. The miR-34 family is known to inhibit tumorigenesis and has therefore been suggested as a potential therapeutic target of interest [74]. Differently from the miR-34 family, members of the miR-99 family may have both tumor-suppressing and oncogenic roles based on the cellular context and tumor type [75], confirming the importance of these molecules in

health and disease.

Long non-coding RNAs

Together with miRNAs, another class of RNAs has captured the focus of researchers in the past few years, given their role in a growing number of cellular processes [76]. Long non-coding RNAs (lncRNA) are a class of molecules 200bp long that do not contain protein-coding sequences. Compared to mRNAs, lncRNAs are thought to undergo different transcription and regulation processes that are closely linked to their functions. In addition, they are commonly localized in the cellular nucleus at lower expression levels than their coding counterparts [77, 78, 79].

lncRNAs are believed to be involved in multiple gene regulation levels (Figure 2.2.b) (see [76] for a detailed overview of the role of lncRNAs). They have been shown to be associated with chromatin changes via the interaction with chromatin modifiers that they recruit to activate or suppress the expression of target genes, both in genomic sites distant from the genomic locus of origin of the lncRNA (trans-activity) [80] or based on the loci from which they were transcribed (cis-activity) [81]. For example, lncRNAs mediate the activity of Polycomb Repressive Complexes (PRCs) [82], multiprotein complexes that modify histones when gene silencing is required [83]. LncRNA ANRIL mediates PRC1, acting on histone H2A, and PRC2, acting on histone H3 and recruits them to the promoter region of nearby genes *CDKN2A* and *CDKN2B* thus influencing cell senescence [84] (cis-activity). The same gene has also been studied for its trans-activity in association with Alu motifs across the genome [80]. Another important lncRNA mediating gene silencing via PRC2 recruitment is HOTAIR, whose overexpression has been measured in different tumor types and demonstrated to contribute to their metastatic behaviors [85]. lncRNAs can also promote gene activation by recruiting chromatin modifiers, as in the case of lncRNA HOTTIP regulating the HOXA gene cluster [86], or by working as a decoy, as the TP53-regulated lncRNA PRESS1 does by binding to the pluripotency repressing *SIRT6* [86].

Further evidence of lncRNA activity has been found at the transcriptional level, where these molecules have been observed to interfere with the transcriptional machinery of a

cell and thus result in gene silencing via altered recruitment of transcription factors [87] or modification of histones [88] and chromatin accessibility [89]. Finally, lncRNAs can work as post-transcriptional regulators by interacting with proteins or nucleic acids to amper further processing of mRNAs. Notably, they have been proven to bind with RNA-binding proteins to form complexes that result in alterations of RNA splicing mechanism [90], mRNA stability [91], and even in the modulation of signaling pathways [78, 76]. Importantly for this work, lncRNAs have been observed to often harbor various MREs and have thus been hypothesized to constitute a layer of post-transcriptional regulation in the shape of competing endogenous RNAs, as discussed below.

Competing endogenous RNA networks

In light of the discovery of the role of miRNAs in many biological processes, miRNAs have been suggested as key components for the regulation and control of gene expression at the RNA level (i.e. post-transcriptional modification). As described earlier, miRNAs can recognize target sites on molecules belonging to different classes of RNAs such as such as circular RNAs (circRNAs), long non-coding RNAs (lncRNAs), and messenger RNAs (mRNAs) [92, 93, 94, 95]. Importantly for this thesis, non-coding RNAs have been observed to carry many miRNA target sequences and have thus been identified as important putative targets for miRNA binding [96].

Given their potential affinity with multiple RNA classes, miRNAs are seen as regulatory molecules that can mediate the communication between RNAs sharing the same MREs, which end up indirectly regulating their respective expression levels by binding to the miRNAs first thus sequestering them from the cellular environment. This type of mutual regulatory relationship can be extended to the full transcriptome, resulting in (indirect) post-transcriptional networks of regulatory interactions, typically referred to as competing endogenous RNA (ceRNA) networks (Figure 2.2.c) [94]. The name derives from the fact that RNAs must "compete" for a limited pool of miRNAs (2600 mature miRNAs are estimated to be encoded in the human genome, against more than 200 000 transcripts [97]). The hypothesis offers a way to provide mechanisms behind unexpected changes in expression [96]. For

example, *ZEB2*, a master regulator of the epithelial to mesenchymal transition [98], has been shown to modulate *PTEN*, an important tumor suppressor [99], in a miRNA-mediated and protein-coding-independent way [100, 95]. Following the ceRNA hypothesis, low expression of RNAs harboring miRNA targets would lead to the release of many miRNA transcripts that would then be free to target and silence other molecules. On the opposite, high expression of the target would end in a lower amount of miRNAs and thus in the decrease of their regulatory activity on the other target RNAs.

Examples of ceRNA interactions have been observed both in health and disease. These regulatory relationships have been shown to play a role in brain tissue development [101] and liver regeneration [102], and to be involved in fundamental cellular processes such as reprogramming [103] and differentiation [104]. Moreover, investigation of dysregulation of these post-translational mechanisms has assumed importance in the framework of complex diseases such as cancer, where alteration of gene expression regulation is known to play a fundamental role in the appearance of malignant phenotypes. It is in this setting that non-coding RNAs have become an object of study, given the possibility of defining the biological role of these previously poorly understood molecules [96]. For example, lncRNA HOTAIR, already mentioned in the previous chapter, is broadly known for its role in tumor development and is often used as a prognostic biomarker in different cancer types, e.g., nasopharyngeal carcinoma [105]. In addition to its regulatory role in association with PRC2, HOTAIR acts as ceRNA by competing for binding with miRNA130a in gallbladder cancer [106] and with mir-331-3p in gastric cancer, where it regulates *HER2* [107]. Finally, HOTAIR is known for its relationship with tumor suppressor *PTEN* [95].

Understanding ceRNA regulatory mechanisms in cancer has proven to be a valuable task, given the importance that uncovering altered or aberrant relationships might have in elucidating the biology of cancer. This has not been always possible, given the multiplicity of potential miRNA-target pairs and the size and complexity of the regulatory interactions involved in ceRNA networks. Computational models have quickly become an efficient way to infer biologically plausible ceRNA networks and further investigate them. These methods typically rely on two different approaches that dictate network inference. On the one hand,

they exploit the fact that ceRNAs should be positively correlated with each other, while simultaneously being negatively correlated with miRNAs they are regulated by and use miRNA and ceRNA expression data to estimate these associations [108, 109]. On the other, miRNA-ceRNA interactions are predicted by matching of the seed regions of the miRNAs with the target region of potential transcripts of interest [112,113]. More recent methods tried to combine the two steps, to reduce the number of false-positive interactions and define robust ceRNA networks [110].

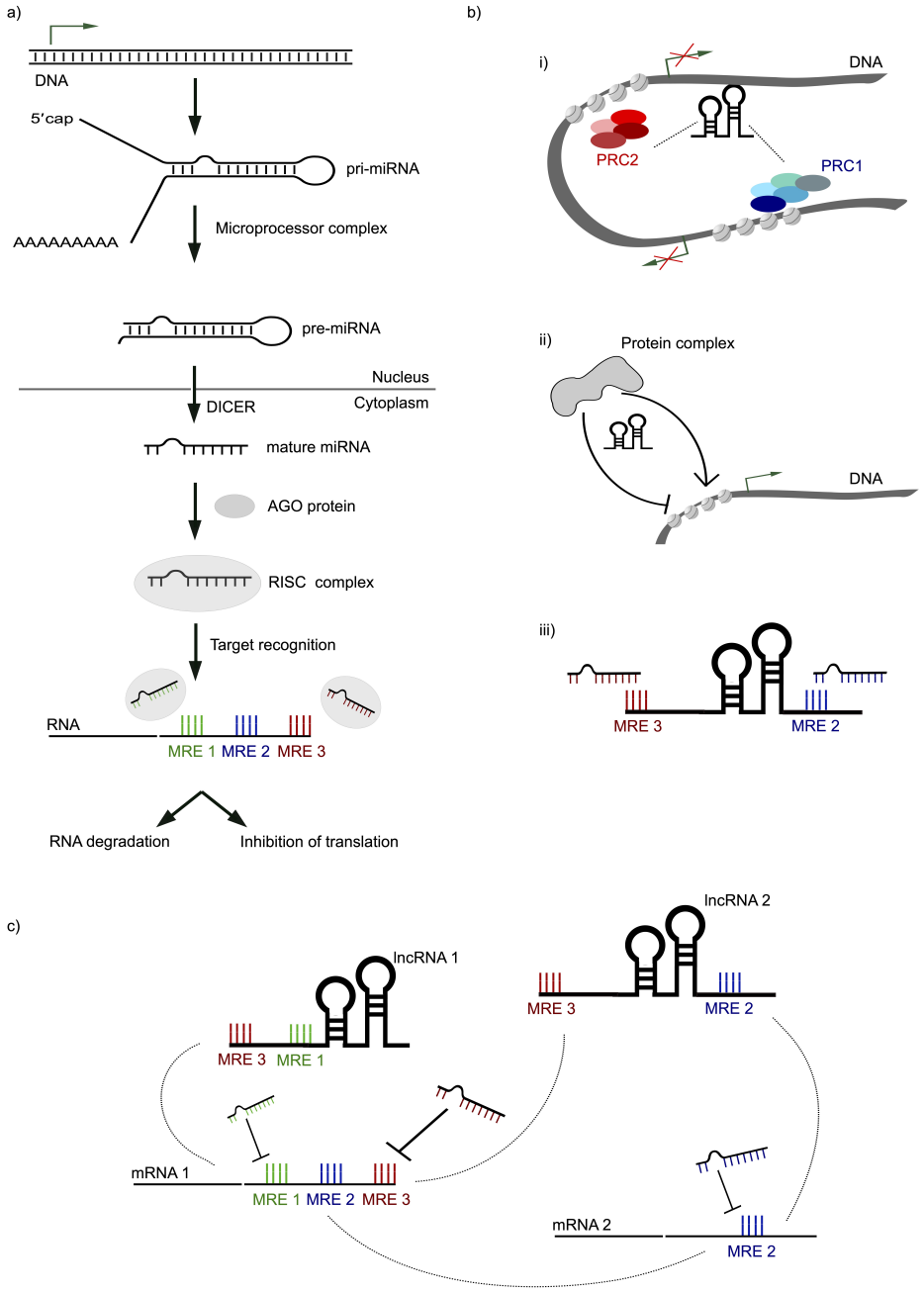

Figure 2.2: Overview of the regulatory role of non-coding RNAs in health and disease. a) Schematic representation of miRNA biogenesis. From the top, pri-miRNAs are transcribed from intronic regions and processed by the microprocessor complex into pre-miRNAs. pre-miRNAs are transported to the cytoplasm and cut by DICER. Mature miRNAs bind with AGO proteins and bind to target mRNAs. b) Overview of the regulatory tasks of lncRNAs described in this thesis. i) lncRNAs recruit chromatin modifier complexes to induce chromatin changes and inhibit transcription, ii) lncRNAs can work as promoters or decoys of transcription, iii) lncRNAs can influence post-transcriptional regulation. c) Depiction of a ceRNA network comprising 4 RNAs in total. Potential matching is indicated by the same coloring. Binding of the target genes with the miRNAs establishes a cross-talk between genes. Once extended to the whole genome, these cross-talks can be seen as a regulatory layer of interactions, i.e. a ceRNA network.

2.2 Intercellular signaling

As previously mentioned, cellular homeostasis relies on a complex and precise communication network between cells to control growth, division, and proliferation. These mechanisms are modified to allow higher proliferation rates and faster growth in cancer. Intercellular signaling mechanisms can be broadly grouped into three steps, (signal) reception, transduction, and response [111]. Much of the communication between cells happens through growth factors, small proteins that allow cell-cell communication. Growth factors are sensed by receptor proteins extruding from the cell membrane. Once phosphorylated, these proteins are functionally altered and proceed to alter further downstream cells to propagate the external stimulus. Epidermal growth factors (EGFs) were the first family of growth factors to be discovered. EGFs and their receptors, normally involved in early embryonic development and stem cell renewal in healthy tissues such as liver and skin, have surged as an important player in tumorigenesis and progression of different cancer types [112]. EGFs are recognized by cells via surface proteins identified as EGF receptors (EGFRs), belonging to the class of receptor tyrosine kinases (RTKs) and some of the most common types of observed receptors. RTKs in the inactive state, i.e., in the absence of a ligand, present themselves as a pair of unconnected monomers. Upon ligand binding, the monomers dimerize, leading to phosphorylation of the tyrosine domain part of the intracellular monomer and to the subsequent activation of the receptor. Activation then drives the recruitment of new proteins and their phosphorylation to propagate the signal further.

Kinases are the most frequently mutated proteins in cancer [113], and phosphorylation is one of the main post-translational mechanisms of signal transduction, making it suitable as a therapeutic target [114]. Kinases act by removing a high-energy phosphate group from a GTP molecule and transferring it to other available proteins. Tyrosine kinases are a particular class of kinases, so defined given the fact that they phosphorylate tyrosine, as opposed to the action of serine/threonine kinases that phosphorylate serine and threonine [115], as in the case of CKD proteins known to regulate the cell cycle [116]. Many ligand-receptor pairs have been identified since the discovery of EGFs-EGFRs, such as platelet-derived growth factors and receptors (PDGFs and PDGFRs), vascular endothelial growth factors and receptors (VEGFs

and VEGFRs), or fibroblast growth factors and receptors (FGFs and FGFRs).

Once receptors bind with their respective signal, signaling cascades are propagated to achieve the desired target response, either directed toward the cell nucleus, e.g., to induce changes in gene expression, or towards the cytoplasm, e.g., to reorganize the cytoskeleton structure. While the pairing is necessary for healthy cells to start the signaling cascade, cancer cells can become independent of the availability of growth factors in the extracellular space to grow and proliferate constantly. For example, mutations in genes encoding growth factor receptors may drive activation of the signaling cascade independently of the presence of a ligand. Alternatively, tumor cell surface receptors' overexpression might increase their signaling output [112].

Ligand-receptor binding is followed by various downstream signal-transducing cascades to the nucleus. While a wide range of pathways is known to be altered in cancer [117], such as the TGF- pathway [118], the PI3K-AKT pathway [119], or the JAK-STAT pathway [120], I here focus on the RAS-RAF-MEK-ERK signaling pathway, that plays a role in many cancer types and particular in pancreatic cancer, which is important for this thesis (Figure 2.3).

The RAS-RAF-MEK- signaling pathway is activated upon ligand binding to tyrosine kinase receptors and subsequent recruitment of adaptor proteins such as GRB2 and SOS. Signals are further transmitted via activation of GTPases (enzymes able to bind GTP and hydrolyze it to GDP), e.g., belonging to the RAS family. GTP activated RAS activates downstream RAF isoforms such as ARAF, BRAF, and CRAF, all able to activate MEK1/2 and its downstream effector ERK1/2. The simplicity of the RAS-RAF-MEK-ERK cascade is opposed to the complexity of the negative feedback mechanisms that developed to maintain ERK activation (for more see [121, 122]). ERK activity is directly related to cellular proliferation, differentiation, and apoptosis via multiple substrates, for example, via the downstream pathway ERK-MSK-CREB which leads to the expression of cyclin D, required for CDK proteins activity to control cell cycle arrest, making this pathway a critical component for cancer cells to drive enhanced proliferation and other mechanisms [123]. Other important substrates of ERK are RSK, an inhibitor of tumor suppressor p27 [124] and activator of the PI3K-AKT pathway [125], and MYC, an important transcription factor known to enhance DNA transcription [126, 127]

and to be over-activated in many cancer types [128]. ERK is recognized as an important part of many cancer hallmarks, such as cell proliferation or avoidance of cell death [129]. Its over-activation can be achieved in multiple ways, either via overexpression of tyrosine receptors (e.g., ERBB family amplifications [130]) or amplification or mutational activation of the downstream molecules, e.g., kinases such as RAS and BRAF [131, 132]

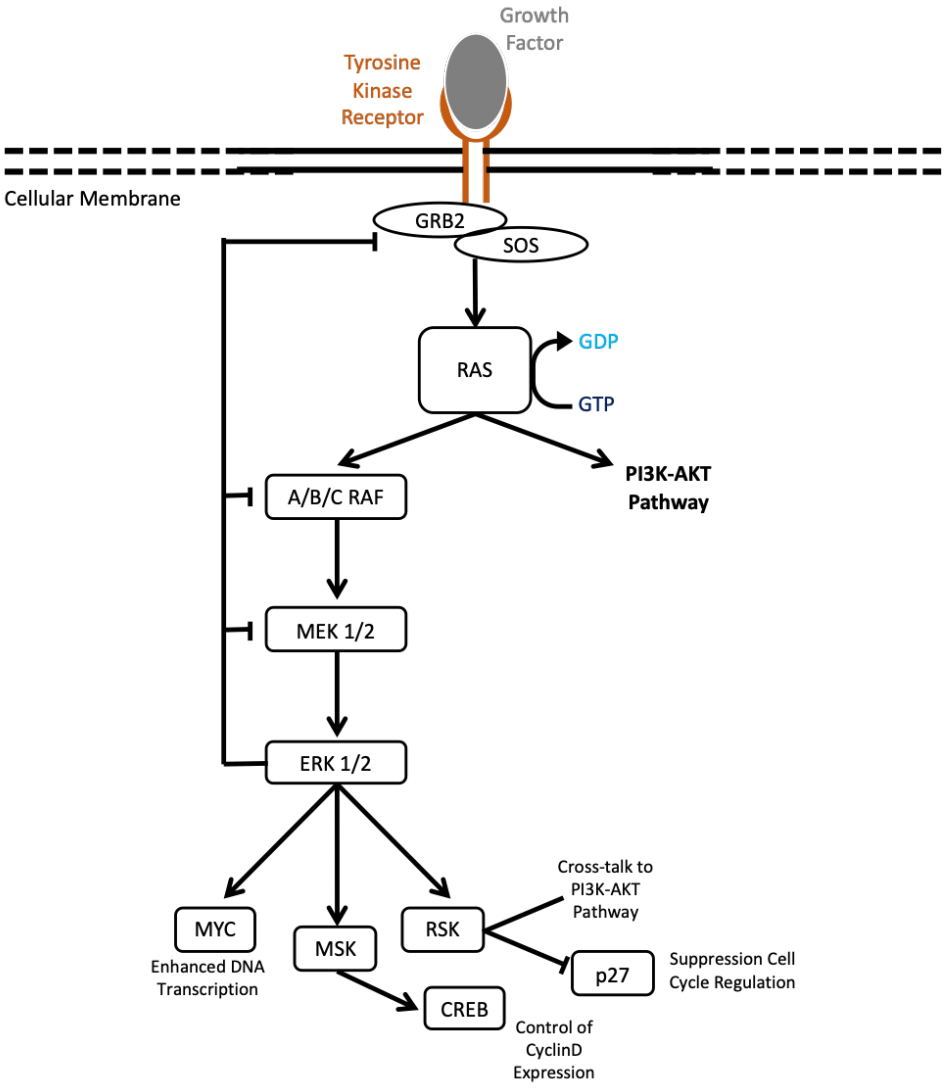

Figure 2.3: Schematic illustration of the main molecules involved in the ERK signaling pathway.

Intracellular signaling is a complex and dynamic process characterized by high redundancy in routes activating the same pathways. Cancer cells take advantage of these mechanisms

by promoting the rewiring of existing pathways, negative feedback signals, and pathway cross-talks. For example, the multiple negative feedback loops between ERK and its upstream molecules described above are known to grant robustness to the cascade [121], as shown in experiments where multiple members of the pathway were targeted in melanoma samples and led to improved therapeutic response and prognosis [133]. Cross-activation has been observed between the PI3K-AKT and the RAF-MEK-ERK pathway, with observed resistance to the inhibition of the PI3K-pathway in murine lung cancer samples harboring a KRAS mutation [134].

2.3 Breast adenocarcinoma

Breast cancer is the most frequent cancer diagnosed in the world (11.7% of newly diagnosed cases) and the main cause of cancer-related deaths in female patients [7]. 10% of breast cancers have been linked to hereditary factors and genetic predisposition with the most common germline mutations being observed in the *BRCA1* and *BRCA2* genes [135]. The advent of next-generation sequencing and large availability of genomic datasets brought to the surface additional genes related to the disease, such as *ATM*, *CHEK2*, *PALB2* (stabilized by BRCA2), and *TP53* [136]. In addition to genetic aberrations, other elements associated with a higher risk of breast cancer are genetic syndromes (e.g., the Lynch syndrome), pregnancy-derived events, obesity and unhealthy lifestyles, and hormonal therapies (e.g., Menopausal hormone therapy). Early screening has shown to be beneficial to decrease breast cancer-related mortality, thanks to improvements in techniques such as mammography, ultrasonography, and MRI. Moreover, preventive care options such as treatment with tamoxifen, raloxifene, or mastectomies can reduce breast cancer development or recurrence [137].

Breast cancer is a complex disease that is characterised by the alteration of physiological mechanisms at multiple levels [138]). This results in an extremely diverse set of diseases that may present drastically diverse clinical phenotypes. Multiple stratification efforts identified five breast cancer subtypes (Luminal A, Luminal B, Basal, HER2-positive, and Normal-like) [139] that have been linked to specific oncogenes and tumor suppressors and that present

clear differences in aggressiveness and metastatic potential [136]. Clinical decisions are typically driven by the stratification of patients into subgroups defined by expression of estrogen receptors (ER), progesterone receptors (PR), and human epidermal growth factor receptors 2 (HER2). Despite the connection between histopathological subtypes and intrinsic cancer subtypes, highlighted in the 2013 St. Gallen Consensus Recommendations [140], misalignments between protein-based and gene expression-based subtypes have been reported [141] and highlighted the importance of identifying robust biomarkers for patient stratification beyond established molecular signatures such as PAM50 and similar [142, 143, 144].

2.4 Pancreatic ductal adenocarcinoma (PDAC)

Pancreatic ductal adenocarcinoma (PDAC) is the most frequent form of pancreatic neoplasms, accounting for 90 to 95% of all pancreatic neoplasms [145], characterized by an overall 5-year survival rate of 9% (in the United States) [9]. It is thought to develop via pancreatic intraepithelial neoplasia (PanIN) and from cystic lesions (e.g., intraductal papillary mucinous neoplasm (IPMN), intraductal tubulopapillary neoplasm (ITPN), and mucinous cystic neoplasm (MCN)). PanIN lesions are known to be the most common precursor [146]. Such lesions give rise to cancer through the gradual accumulation of genetic alterations that lead to phenotypic changes and, ultimately, to the progression of invasive pancreatic cancer, or rapidly progress through catastrophic events, such as chromothripsis to invasive PDAC. The earliest lesion is defined as Acinar-to-ductal metaplasia (ADM), during which mutations in key oncogenes such as *KRAS*, mutated in >90% of PDACs, initiate the differentiation of pancreatic acinar cells to ductal-like ones [147, 148] and whose impairment has been shown to impact further degeneration in later steps of tumorigenesis [149].

KRAS oncogenic activation has been observed in 80-90% of all early-stage lesions [150], and is involved in the dysregulation of cell differentiation and inhibition of tissue repair mechanisms. *KRAS* point mutations, with the most frequent ones being G12D, G12V, and G12R [151], result in the activation of downstream pathways (see paragraph above) such as the MAPK and PI3K pathways. *CDKN2A* loss-of-function is present in more than 80%

of PDACs [152] via, e.g., loss of both alleles or hypermethylation in the promoter region [153]. *CDKN2A* is known to encode two important tumor suppressor proteins, p14 and p16, controlling checkpoints of the G1/S transition during the cell cycle by binding to CDK proteins such as CDK4 and CDK6 [154].

Moreover, further aberrations in later steps, such as in *TP53* and *SMAD4*, are often found in pancreatic cancer patients. 62.5% of PDAC patients harbor loss-of-function modification in the *TP53* gene by homozygous deletion and/or intragenic mutations [155], driving genomic instability in PDAC [156]. Similarly, *SMAD4* intragenic and hemizygous deletions are observed in 50% of PDAC patients, leading to alterations of the TGF- pathway and correlated with metastasis and poor prognosis [157]. The development of next-generation sequencing technologies and the advent of large international consortia dedicated to the collection and analysis of molecular data for different cancer types (see next chapter for more details) have given the opportunity to analyze the molecular pathology of pancreatic cancer more in detail and highlighted key aspects that could have deep implications for the advancement of new therapeutic strategies.

Massive sequencing efforts highlighted the heterogeneity of pancreatic cancer beyond a few key frequent mutations [158]. They accentuated alterations in germline DNA damage repair genes such as *BRCA1*, *BRCA2*, or *ATM/ATR*, leading to genomic instability [159, 160, 161]. Further works reported complex chromosomal rearrangements as a feature of PDAC [162, 163]. Whole-exome sequencing efforts revealed frequent aberrations (30% of pancreatic cancer) in chromosome arms such as deletion of 8p, 9p, 18p, and 18q and amplification of 1q [164, 165, 166]. Other recurrent events are amplifications of *GATA6*, *KRAS*, and *MYC* and deletions of *CDKN2A*, *SMAD4*, *ARID1A*, and *PTEN* [166]. In parallel to genomic studies, the use of gene expression (both from sequencing and array-based based technologies) for PDAC subtyping (see next chapter for a description of tumor subtyping approaches) identified two main broad and general subtypes of pancreatic cancer, a more differentiated, less aggressive classical subtype and an undifferentiated, aggressive and more therapy-resistant mesenchymal subtypes, one [167, 168, 162].

2.5 Summary

Cancer is an extremely heterogeneous class of diseases characterized by the accumulation of genetic aberrations that alter cellular homeostasis. Gene expression, a process that allows the synthesis of proteins and is key to preserving cellular equilibrium in healthy tissues, is particularly susceptible to modifications induced by mutations and structural variants and must be carefully characterized and analyzed to understand how its dysregulation gives rise to malignant phenotypes.

Gene expression is regulated at different levels and by different mechanisms. Transcription factors are a class of proteins primarily designed to modulate gene expression. The action of transcription factors depends on the availability of open genomic regions, which depends on the structure of the chromatin, the protein-bound DNA. Chromatin structure can be modified via epigenetic changes such as changes in its structural components, e.g., histones, or via changes in the DNA methylation patterns. Finally, transcriptional products, such as messenger RNAs can be further processed by small RNAs, e.g., miRNAs, and other classes of non-coding RNAs, e.g., long non-coding RNAs (lncRNAs). These two classes of molecules have drawn a lot of attention in the past few years in cancer research, thanks to their potential as diagnostic biomarkers and therapeutic targets. Importantly, new hypotheses regarding their activity have surfaced. For example, the competing endogenous RNA (ceRNA) hypothesis suggested that different RNA molecules, e.g., lncRNAs and mRNAs, compete for the binding with miRNAs, shaping complex post-transcriptional regulatory networks that can be exploited to study cancer mechanisms.

Intercellular signaling is another important process governing cellular functions such as growth, proliferation, and death throughout the cell cycle. It works as a cascade of signals that are transduced from the cell membrane to the nucleus through the work of signaling proteins. Kinases are one of the main classes of signaling proteins and constitute the core of some of the most important signaling pathways in a cell, such as the RAF-MEK-ERK signaling pathway. Kinases have also acquired importance in the framework of cancer research, being the most frequently mutated class of proteins in tumors. In particular, they have become one of the main targets for therapeutic strategies aiming at inhibiting signaling pathways to stop

tumor cells' growth and proliferation.

3 Precision oncology in the age of Artificial Intelligence

In recent years, the traditional paradigm "one symptom-one target-one drug" [169] has shown its limitations, with the ten highest-grossing drugs in the United States resulting in improved conditions only for a small proportion of patients [170]. Moreover, increased access to healthcare services has highlighted differences in performances between ethnic groups [171]. Precision medicine has emerged as a possible alternative by offering the chance to tailor medical decisions to patients' clinical and molecular profiles. In particular, stratified medicine [172] allows the identification and prediction of clinically relevant strata that share molecular disease mechanisms, offering the chance for the development of mechanism-based diagnostic and therapeutic strategies [169].

Oncology has pioneered the transition toward this new paradigm, recognizing that similar clinical phenotypes may be the result of different tumor development routes that impact treatment response [173]. This effort has been aided by the systematic collection of genomic alteration information [174, 175, 176] that resulted in the identification of a wide range of cancer-specific alterations. Current precision medicine approaches rely on these to investigate "first-order" relationships (i.e., linking of patients' mutation and copy number profiles with specific clinical strategies) [177] and mechanisms of "oncogenic addiction" (i.e., the dependency of tumor cells on a specific oncogene) to identify biomarkers able to stratify patients between potential responders (sensitivity biomarkers) and non-responders (resistance biomarkers) [178]. Targeted therapies exploiting these characteristics are the foundation of modern cancer treatment, with multiple compounds available in the clinics exploiting genomic biomarkers such as Crizotinib (targeting *ALK* rearrangements) [179], Nilotinib (*BRC-ABL*

fusion) [180], and Dabrafenib ($BRAF^{V600E}$ and $BRAF^{V600K}$) [181].

Recently, innovative cancer treatment options have surfaced, leveraging and targeting different aspects of cancer biology. For example, increased understanding of the mechanisms driving alternative splicing and their relevance in cancer has led to the development of treatment options designed to correct or modulate alternative splicing events [182]. Different approaches try to exploit the patients' own immune systems to fight cancer progression. For example, immune checkpoint blockades, such as anti-PD-1, anti-PD-L1, and anti-CTL-4 therapies [183], control the inhibition of tumor-infiltrating T cells. The use of chimeric antigen receptor T (CAR-T) cells for adoptive T-cell transfer therapies has a similar goal and has been shown to be effective in non-solid tumors such as B-cell lymphoblastic leukaemia [184]. Finally, recent advancements raised high hopes for the development of effective cancer vaccines able to identify antigen peptides and boost a patient's immune system [185]. Despite the lack of concrete examples of successful use of cancer vaccines in clinical practice, they have shown promising results in different cancer types [186, 187, 188, 189, 190].

All these approaches are set to benefit from the technical and methodological advancements that have brought to the surface the complexity of cancer genomes and that have highlighted the need for a more comprehensive molecular profiling of cancer patients, going past "first-order" relationships. Molecular characterization efforts allowed the creation of complete datasets encompassing different molecular layers, such as the genome [191], transcriptome [192], epigenome [193], and proteome [194], typically collected under the term "omics" technologies. The accumulation of multiple information layers has made clear the need for the use of advanced computational techniques to analyze and interpret biological data [195], creating a fertile ground for the use of computational techniques such as machine learning (ML) and artificial intelligence (AI) in precision medicine [196]. Successful integration of these disciplines will pave the way to a new data-driven age for medicine and biology [197] and will enable the leveraging of the whole molecular landscape of patients to drive treatment decisions and potentially design new therapeutic strategies.

3.1 Artificial intelligence

Artificial intelligence-based technologies have established themselves as a disruptive force taking different fields and industries by storm [198, 199, 200]. Particularly important for this work, they are assuming an increasingly important role in biology, medicine, and healthcare [201], in particular in the medical imaging field and in sectors such as radiology and pathology, where automated classification agents have achieved excellent performances in diagnosis, risk prediction, and as decision-support systems for selecting treatment across different diseases and applications [202, 203, 204]. Similar technologies have been successfully employed beyond imaging tasks, where AI technologies have shown their potential as signal-processing tools for medical signals such as electrocardiograms (ECGs) [205] or electroencephalograms (EEGs) [206], together with the great advances in the field of biochemistry and structural biology [207].

Artificial intelligence is an umbrella term that refers to all the techniques based on the simulation of human intelligence by machines, such as natural language processing, robotics, computer vision, machine learning (ML), and deep learning. In this work, I will mainly focus on machine learning, an AI research area focused on the design of agents able to learn general rules and patterns from predefined example datasets [208]. ML applications can be categorized into three broad frameworks. Supervised ML learning methods are designed to identify and approximate the relations between input features and outcomes of interest. Supervised approaches can be categorized based on the type of outcome variable of interest, with regression approaches analyzing numerical or continuous variables and classification ones where the outcome variable is categorical. On the other hand, unsupervised machine learning methods try to define hidden patterns in the features of interest. Finally, reinforcement learning has grown alongside these two more traditional methods, establishing a framework where agents take actions in predefined environments while maximizing user-defined and task-specific reward functions [209].

Different types of machine learning methods have been introduced, differing in the complexity of the patterns they can learn and identify, in the type of data they can be applied to, and in the degree to which they can offer explanations of their inner functioning (interpretability).

For instance, linear models have been extensively used in statistical literature because of their straightforward implementation and inherent interpretability [210]. On the opposite hand, neural network-based strategies such as deep neural networks can automatically identify complex patterns while offering limited room for interpretation [211].

Artificial intelligence is poised to revolutionize the way precision medicine and the broader medical industry are defined. However, broad translation of new technologies and tools to daily clinical use is still lacking, mainly due to important technical, ethical, and regulatory challenges [201]. The main technical challenges are related to building models that are trustworthy, reliable, easy to use and understand, and easily integrable into existing clinical frameworks [212]. Explainability is another characteristic often mentioned as one of the key obstacles to the widespread deployment of these technologies. Despite recent advancements in this direction, current strategies are very limited [210] and require further research.

Regulatory challenges are mainly related to the accuracy, robustness, and fairness of AI models across different settings, e.g., hospitals and patient populations. Furthermore, it is necessary to define the relationship between humans and automated agents and how the two systems interact and exchange information [213]. Finally, the introduction of such technologies might imply shifts in responsibility accountability and might lead to new sets of rights and duties for all the stakeholders in the healthcare field [214, 215]. Such issues would require AI technologies to be explainable and justifiable, i.e., they should provide reasons for their decisions in the framework of rights, laws, and norms in our society [216].

Significant problems are then related to the ethical use of patient data, which these technologies are intrinsically dependent on. These must be protected from potentially malicious agents interested in such highly sensitive data. Approaches such as federated learning might ease decentralization while making the calibration of AI models using data from different centers/locations easier [217]. Problems related to the exacerbation of existing inequalities based on biases hidden in the data are a known problem for ML models that need to be tackled to assure fairness in healthcare [218, 219, 220] with specific actions at every step of an ML pipeline [221].

Despite the appearance of the first approaches able to successfully exploit algorithms in the field of precision oncology [177], the true potential of ML in this field remains untapped. In particular, the potential of these technologies to identify complex disease biomarker signatures across multiple omics layers offer the chance to advance precision medicine [222]. In this chapter, I describe a handful of potential applications of ML in precision oncology. In particular, I focus on disease subtyping (Figure 3.1.a), drug response prediction (Figure 3.1.b), drug repurposing tasks (Figure 3.1.c), and design of drug combinations (Figure 3.1.d). The intent here is not to discuss these topics exhaustively but rather to draw a general overview of the main models and methods used in these applications. Drug response prediction will be discussed more in detail as it is of particular relevance for this work and will be the main topic of Chapter 4.

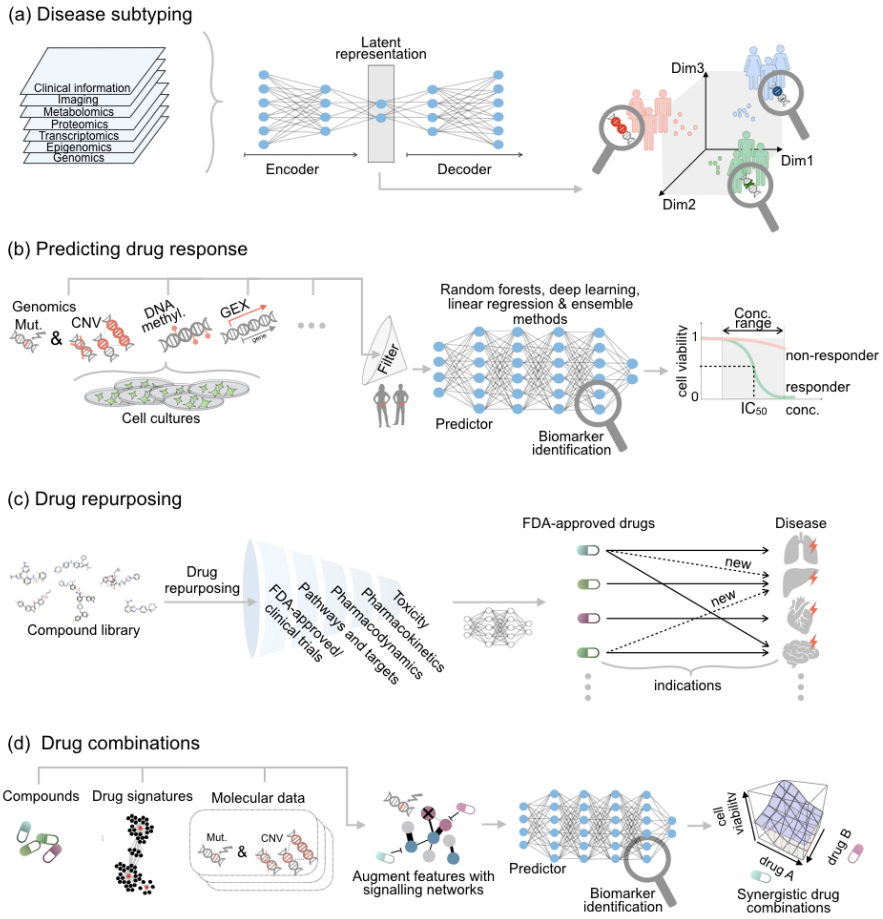

Figure 3.1: Computational models described in this chapter are here represented as neural networks for sake of simplicity. a) ML approaches can exploit different molecular layers to identify clinically relevant tumor subtypes. Typical subtyping approaches try to define low-dimensional representations (extracted with, e.g., an autoencoder) of the phenomenon of interest. b) ML algorithms can be used to predict drug response from multiple -omics layers and analyzed to identify biomarkers of drug response. c) AI techniques can be used to identify new targets for drugs compounds previously approved for different diseases in the framework of drug repurposing. d) ML methods can identify new effective drug combinations by combining different layers of information on the compounds of interest and available molecular layers.

3.2 Disease subtyping

As discussed in the introduction, a cancer type is not a unique disease but rather a heterogeneous class of sub-malignancies that, despite having their origin in the same tissue type, may differ in cell of origin, etiology, micro-, and macroenvironment [223]. Such differences

make these groups, typically referred to as subtypes, unique in their molecular characteristics, prognostic outcomes, and sensitivity to targeted therapies [224]. Various subtyping efforts led to the identification of clinically relevant tumor subtypes, as in colorectal cancer [225], bladder cancer [226], breast cancer [227], and pancreatic cancer [228] (see Chapter 1). A plethora of methods, both computational and non-, has been developed to identify clinically relevant subtypes, e..g., histopathological analyses [229, 230]. I here give an overview of current approaches for tumor subtyping, focusing on the most frequently implemented techniques and on the datasets that allowed the discovery or validation of tumor subtypes. The reader can find a more comprehensive overview of tumor subtyping applications at [231, 232].

Standard tumor subtyping methods are based on unsupervised ML techniques such as non-negative matrix factorization (NMF) and independent component analysis (ICA), particularly useful since they result in subtype-specific signatures that are biologically interpretable. The growing availability of large-scale omics datasets, such as The Cancer Genome Atlas (TCGA) [233], the International Cancer Genome Consortium (ICGC) [234], and the Pan-Cancer Analysis of Whole Genomes (PCAWG) [235], led to the application of deep learning strategies to tumor subtyping tasks. For instance, variational autoencoders have been designed to stratify non-small cell lung cancer patients based on methylation patterns [236]. Similar approaches aim at exploiting multiple molecular layers, or even data modalities, to further boost precision oncology [237] and uncover potentially interesting patterns in neuroblastoma [238], lung adenocarcinoma [239] and breast cancer [240].

3.3 High-throughput drug screens

in vitro and *in vivo* models have always had an important role in oncology and cancer research, with the first *in vitro* cultures being used over a century ago [241]. Later studies used these models to investigate sensitivity to chemotherapeutics using a broad range of readouts such as proliferation rates or viability, with results that drove, for example, the definition of many treatment regimens today [242]. Despite their known limitations and questionable role in the identification of clinically relevant biomarkers [243, 244], the growing amount of information on the heterogeneity of cancer genomes, the increasing availability of active compounds, and

the introduction of new models such as 3D patient-derived organoids, have renewed the push for preclinical models. In particular, *in vitro* models offer the chance to systematically compare the effect of large libraries of therapeutic drugs on large cohorts in a high throughput fashion. When complemented with deep molecular characterization of the samples of interest, such efforts offer the chance to identify clinically relevant drug response biomarkers [245]. 2D tumor-derived cell cultures have always been the workhorse of these projects, with the first high-throughput screening efforts exploiting rather limited cohorts of 60 cell lines [246] and the more recent ones expanding the screened samples up to 1000 cell lines representing more than 30 cancer types screened with hundreds of compounds (e.g., the Genomics of Drugs Sensitivity in Cancer (GDSC) project [247], and the Cancer Target Discovery and Development (CTD2) initiative [248]). 3D self-organizing tumor cell cultures, i.e., organoids, have been identified as the up-and-coming model for HTSs [249], as they are better at capturing the heterogeneity observed in tumors and at conserving architecture and cell-type composition of the tissue of origin [250].

Traditionally in HTS projects, drug efficacy has been quantified via "static" approaches that measured proliferation, survival, or viability, with metrics such as the half-maximal inhibitory concentration (IC50) or by the area under the dose-response curve (AUC). The advent of new technologies paved the way to more "functional" approaches, able to measure perturbations of living cells and to incorporate analysis at single-cell resolution to evaluate, for example, sub-clonal phenotypic effects [251, 252]. These technologies could overcome the intrinsic limitations of traditional tumor models by simulating the presence of the tumor microenvironment or the pharmacokinetics and mode of drug delivery with innovative organ-on-a-chip technologies [253, 254].

Pharmacogenomic analyses try to find patterns and links between drug response data and existing -omics datasets. Typical approaches rely on the approximation of functions mapping drug response read-outs, e.g., AUC or IC50, onto gene expression, gene methylation, mutations, etc. [255]. More advanced approaches, such as transfer learning, have been used to leverage information from large HTS datasets to predict drug response in smaller, proprietary datasets collected from a single cancer type [256]. Moreover, representation

learning techniques have been implemented to define low-dimensional embeddings associated with drug-relevant modules, e.g., constrained matrix factorization [102] or manifold learning [257]. Deep learning methods have proven to be successful in the prediction of HTS results, using traditional architectures such as convolutional neural networks [258] or autoencoders [259].

While HTS analyses typically use baseline molecular profiles to predict drug response, drug signatures can be derived by measuring changes in omics profiles before and after treatment, as in the case of the connectivity map (C-Map) [260, 261]. Similar strategies can enable drug repurposing, an approach based on identifying and prioritizing drugs that have already successfully undergone clinical and safety trials as new treatment strategies for diseases different from the ones they were originally designed for, thus accelerating drug and clinical development pipelines [262]. Such approaches find their strength in the availability of extensive chemical datasets and databases such as ChEMBL [263] and PubChem [264], collecting information and biological and chemical properties (e.g., toxicity, pharmacokinetics, and pharmacodynamic profiles) of thousands of compounds. Moreover, the availability of experimentally validate interaction networks, such as protein-protein interaction (PPI) networks [265], has created a fertile ground for the integration of AI for drug repurposing applications [266].

3.4 Drug combinations

One of the main complications in cancer care is the surfacing of resistance to treatment, a phenomenon that leads to cancer cells becoming less tolerant to the administered cure. Drug resistance in cancer has historically been categorized either as intrinsic, or primary, and acquired, or secondary, drug resistance [267]. While the former implies that some or all tumor cells are observed not to respond to the chosen treatment, in the latter one tumor cells initially responding to treatment show a decrease in treatment efficacy in later stages [268]. Multiple biological mechanisms have been associated with drug resistance, such as tumor intrinsic factors, e.g., tumor burden tumor [269], tumor heterogeneity [270], or rewiring of

intercellular pathways (as described in chapter 1), or tumor extrinsic ones, e.g., the influence of the tumor microenvironment [271].

Drug combinations, or multi-drug therapies, offer an alternative to address both, primary resistance as well as the appearance of secondary drug resistance by exploiting drug synergies [272]. For example, colorectal cancers harboring *BRAF* mutations are known to activate negative feedback loops to EGFR when *BRAF* is inhibited [273]. Simultaneous activation of *BRAF* and *EGFR* has shown to be highly synergistic [274]. In typical approaches, optimal drug combinations are identified via two different approaches: i) "double-hit" strategies where both compounds target the same signaling pathway, or ii) targeting of two independent mechanisms or pathways [272, 275].

High-throughput strategies coupled with computational analysis approaches have proven to be the optimal method to systematically identify and prioritize effective drug combinations in different cancer types [276, 272]. Notably, algorithms can reduce the complexity of the searchable combination space (growing with a complexity of $(n^2\text{-}n)/2$, where n is the number of considered monotherapies) while taking into account key aspects such as the toxicity of the identified cocktail [277].

Various techniques have been implemented in this setting, focusing on the similarity of drug signatures to identify optimal combinations [278] or exploiting various computational approaches, as in the case of the AstraZeneca-DREAM crowd-sourcing challenge [272], where the best-performing methods utilized a combination of prior-knowledge and vanilla machine learning (random forest algorithm in this case) or various deep learning applications [279, 280]. Recently, single-cell sequencing technologies have given further impulse to the field, by offering the possibility to identify potential drug combinations based on the expression of specific receptors on the cell surface via algorithmic approaches [281].

3.5 Summary

Precision medicine aims at overcoming the limited effects of the traditional "one-drug-fits-all" medical paradigm by tailoring treatment choices to patients' clinical and molecular profiles.

In particular, stratified medicine tries to exploit the wealth of clinical and molecular data being generated to identify groups of observations that share defined disease characteristics to design mechanism-based diagnostic, prognostic, and therapeutic strategies. In oncology, multiple treatment strategies have been collected under the umbrella of precision medicine, exploiting different characteristics of cancer biology. For example, targeted therapies and different forms of immunotherapy have proven their efficacy against multiple cancer types.

Artificial intelligence is believed to have the potential to give further impulse to the advancement of precision medicine, by offering a way to analyze and mine large datasets to identify biologically meaningful patterns. For example, fields like medical imaging and biomedical signal processing have already benefited from the power of supervised and unsupervised computational strategies. These technologies have also created fertile ground for new scientific discoveries at the basic and translational level, as in the case of the analysis of multidimensional and complex omics data that can be exploited for patient stratification and biomarker identification purposes. They have had an influence on the investigation of tumor subtypes and in applications related to the prediction of drug response and optimal drug combinations and catalyzed further scientific and clinically relevant discoveries. In the next chapters, I will show two examples of such applications and describe how computational techniques can be pivotal to uncover and prioritize new potential biomarkers for patient stratification.

4 Patient-specific ceRNA modules can elucidate the cancer miRNA regulatory landscape

4.1 Declaration of contributions

This chapter is the result of a project started in the Big Data in Biomedicine group at the Technical University of Munich (Freising, Germany) under the supervision of Dr. Markus List, and in collaboration with the Universidade Federal do Paraná (Curitiba, Brazil) and the BC Cancer Genome Sciences Centre (Vancouver, Canada). The work described here has been driven by me and Markus Hoffmann, doctoral student in the Big Data in Biomedicine Group, who has equally contributed to it. The related manuscript has been submitted to the proceedings of the European Conference in Computational Biology (ECCB) 2022 on April 15th 2022 [282].

4.2 Introduction

The growing availability of large sequencing datasets, together with advancements in computational techniques and an increased understanding of the mechanisms driving gene expression regulation shed new light on the importance of non-coding RNAs as potential biomarkers and added a new information-rich layer for precision oncology approaches, moving past the traditional analysis of genome aberrations and gene expression alterations.

In particular, microRNAs (miRNAs) have been identified as important players in gene regulation, both in healthy and cancerous tissues [283, 284] and as important mediators in competing endogenous RNA (ceRNA) networks (see Chapter 1). lncRNAs have recently

received particular attention in the context of ceRNA networks, with recent works suggesting that one of the roles of lncRNAs is to indirectly regulate the expression of mRNAs via competition for the same miRNAs (see Chapter 1) [285].

Experimental identification and validation of miRNA-target interactions have proven to be extremely costly and laborious. Computational approaches have shown the potential to be a valid substitute, with many different approaches being implemented to identify important miRNAs and targeted genes. While the ultimate goal of these methods is the generation of a handful of hypotheses suitable for experimental validation, they are typically designed to infer complex regulatory networks comprising thousands of interactions. Such complexity hinders the discovery of portions of these networks that might assume an important role in the disease under analysis. The identification of network functional units, or modules, is a key aspect of biological network analysis [286] and assumes an even bigger role in the framework of ceRNA networks, where the identification of modules could point out discrepancies in regulation between healthy and disease statuses and eventually lead to the definition of novel diagnostic or prognostic biomarkers or new potential therapeutic targets.

Recent works have tackled the ceRNA network module identification problem, exploiting a broad range of computational approaches such as community detection algorithms, network-based clustering, and matrix factorization techniques [287]. Despite showing interesting results, these methods often result in a small number of modules containing a large number of edges [288] related to very broad pathways typically associated with cancer, making it difficult to identify robust hypotheses for further experimental validation. Moreover, while recent techniques have focused on the inference of patient specific-networks and on the identification of "aberrant" edges that deviate from the norm [285], none of the ceRNA module identification methods are, to the extent of my knowledge, able to compute sample-specific or patient-specific scores summarising the information contained in the identified modules. Such a summary can be extremely valuable, in that it might offer a straightforward way to link computationally identified modules with their biological functions while offering a starting point for further downstream modeling steps.

In this chapter, I describe spongEffects, a tool able to infer ceRNA modules from pre-

computed ceRNA networks, like the ones inferred by SPONGE [110]. In addition, spongEffects offers the chance to quantitatively estimate the regulatory activity of the inferred modules using single sample enrichment score-inspired frameworks and thus building a platform for the comparison of ceRNA modules across different groups. The general pipeline is presented in Figure 4.1. Using gene expression data and pre-computed ceRNA networks, spongEffects can i) find important nodes in the network via the calculation of different node centrality metrics, ii) define modules centered around high degree nodes, iii) perform gene set enrichment to calculate module- and patient-specific scores, and eventually iv) use the calculated scores for downstream machine learning tasks. I here show an example of how to use spongEffects to retrieve insights into the biology of breast cancer subtypes.

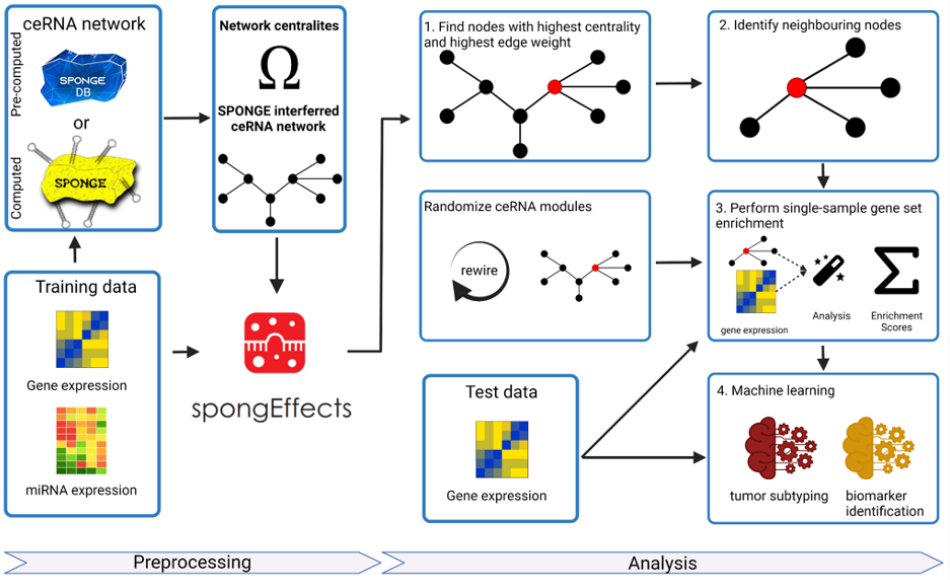

Figure 4.1: SpongEffects requires a gene expression matrix and a ceRNA network as input. Once these are provided, it 1) preprocesses the network and computes weighted centrality scores, 2) defines modules 3) calculates modules' enrichment scores (spongEffects scores), and 4) formats the data for further downstream tasks. Figure from [282].

4.3 Material and methods

4.3.1 Data sources and preprocessing

SpongEffects relies on previous work on ceRNA networks from the Big Data in Biomedicine Group. In particular, we envision it as an add-on to "Sparse Partial correlation on Gene Expression" (SPONGE), a data-driven approach able to infer ceRNA networks from gene- and miRNA- level expression data [110]. SPONGE ceRNA networks for 22 cancer types have been computed and made freely available, together with accompanying information and analyses via SPONGEdb, an online resource for the investigation of ceRNA networks [289]. Log2-transformed tpm-normalized RNA-Seq data for the TCGA breast cancer dataset (TCGA-BRCA) were downloaded together with associated miRNA expression levels and clinical metadata from the XENA Browser [290]. Furthermore, we downloaded log-transformed Illumina microarray data for the 1st and 2nd Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohorts [291] and used them as an independent dataset to validate our findings.

We selected in both cohorts all patients with tumor gradings annotated as stage I, II, III, and IV, and removed all the samples not associated with any of the five subtypes investigated here, namely LumA, LumB, Her2, Basal, Normal-like. After the preprocessing step, we obtained a training cohort of 944 patients (TCGA-BRCA) and a validation cohort of 1699 patients (METABRIC). The TCGA-BRCA specific ceRNA network was downloaded via SPONGEdb (http://sponge.biomedical-big-data.de/) and filtered down (mscor > 0.1, adjusted p-value < 0.01) from 3x107 to 702.026 edges to preserve all connections with significant effect size (see [110] for more information). SPONGE networks and RNA-seq data were used as input data for spongEffects to calculate module-level enrichment scores, i.e., spongEffects scores.

4.3.2 Identification of ceRNA modules

Centrality measures are a pivotal step in the analysis of complex networks, given their potential in providing essential clues about the organization of biological graphs [292, 293]. Degree, closeness, and betweenness, standard centrality measures introduced for the first time by

Freeman et al. [294], have been extensively found to be able to capture and identify important nodes in biological networks [295, 296, 297]. While originally designed for applications in unweighted networks, they have been generalized to be applied in weighted network frameworks [298, 299, 300]. In this section, I focus on degree centrality, with closeness and betweenness considered as equally important and powerful but outside the scope of this work. In particular, given an unweighted network comprising N nodes, the adjacency matrix X_{ij} is a binary matrix containing the description of the connection between node i and node j, with $x_{ij} = 1$ if node i and node j are connected, and $x_{ij} = 0$ otherwise. In the case of a weighted network, the associated weight matrix W_{ij} is a matrix with elements $w_{ij} > 0$ if node i and node j are connected and values corresponding to the weight of the edge between them.

Degree centrality can be defined as the number of edges connecting to node i. More formally, degree centrality can be calculated as:

$$Degree_i = \sum_j^N x_{i,j} \tag{1}$$

The weighted counterpart of degree centrality, here called node strength, can be formalized as:

$$Strength_i = \sum_j^N w_{i,j} \tag{2}$$

Further solutions have been proposed to combine the two measures and strike a trade-off between the influence of the number of edges and the scale of the weights on the definition of important nodes. In particular, Opsahl et al. introduced the following [301]:

$$Centrality_i^{\alpha} = Degree_i \times (\frac{Strength_i}{Degree_i})^{\alpha} = (Degree_i)^{(1-\alpha)} \times Strength_i^{\alpha} \tag{3}$$

Where α is described as "a positive tuning parameter that can be set according to the research setting and data. If this parameter is between 0 and 1, then having a high degree is taken as favorable, whereas if it is set above 1, a low degree is favorable" [301]. In this chapter, I implement the definition of weighted centrality as described in Opsahl et al. and as implemented in the R package *tnet* (version 3.0.16) [302], with the $\alpha = 1$ to

prioritize the identification of ceRNAs with high involvement (i.e. high node strength) in the ceRNA network, where multiple sensitivity correlation values are considered as the edge weights/effect sizes. Identified ceRNAs with high weighted centrality scores are considered to be the central nodes of the sponge modules, defined as all the first-degree neighbors of the central ceRNA nodes.

4.3.3 spongEffects scores

Various gene enrichment methods have been introduced to combine the information from multiple genes, belonging e.g. to a pathway, gene set, or, as in this case, to sponge modules, in a unique score. They are typically grouped under the umbrella "unsupervised single sample enrichment tools" [303], given the fact that they do not rely on *a priori* knowledge or the existence of specific phenotype groups and result in sample-specific aggregated scores. I implemented three of these methods in spongEffects: i) single sample Gene Set Enrichment Analysis (ssGSEA), ii) Gene Set Variation Analysis (GSVA) algorithms (both added as implemented in the *GSVA* package (version 1.34.0) [303]), and iii) Overall Expression (OE) [304]. While the choice of the optimal approach is closely related to the task and available datasets and can hardly be defined *a priori*, as observed in the original GSVA publication [303], I highlight here a shared benefit derived from the implementation of these methods. Namely, they all allow the calculation of spongEffects scores independently of the fact that all the genes in the modules are also present in the gene expression matrix used as input. This is particularly important in validation scenarios, where the validation matrix is likely to contain expression of a set of genes only partially overlapping with the ones part of the original training matrix and part of the modules.

4.3.4 Random Forest for subtype classification

SpongeEffects scores hold the potential to be used in a wide range of downstream tasks. In this chapter, I showcase their use in a classification setting, where spongEffects scores are used as inputs to classify tumor samples in their respective annotated subtypes. To do so, I exploit Random Forest for classification, an ensemble tree-based algorithm that classifies

samples via majority voting [305]. In particular, I used Random Forest as implemented in the *caret* R package (version 6.0.90) [306]. Hyperparameter optimization is achieved via repeated (3x) 5-fold cross-validation, as implemented in the same R package. *Ex-post* identification of sponge modules driving subtype prediction is achieved via calculation of the Gini index, as implemented in the *randomForest* package (version 4.6.14) [307].

4.3.5 Quality control of the classification model via module randomization

As typical in any computational analysis, we evaluated the quality of the classification model by comparing its accuracy to the performance of a model built on randomly defined modules. This step is important to understand if the ceRNA modules capture random noise or covariance structures that are not biologically meaningful or directly related to the differences in subtypes. To define the random modules, we randomly sampled the ceRNA network. More specifically, we defined for each ceRNA module (see above) a random module containing the same number of genes. These were randomly selected from the ceRNA network. Finally, we calculated the spongEffects scores and calibrated a classification model as previously described.

4.3.6 Implementation and data availability

We implemented spongEffects in R (version 3.6.2), and we made it publicly available as a new function in the SPONGE package in Bioconductor at:
https://www.bioconductor.org/packages/release/bioc/html/SPONGE.html

spongEffects source code is available at:
https://www.bioconductor.org/packages/release/bioc/html/SPONGE.html

SPONGE is available at:
https://github.com/biomedbigdata/SPONGE

SPONGEdb is available at:
http://sponge.biomedical-big-data.de/.

The spongEffects scores for each TCGA datasets are available at:

https://doi.org/10.6084/m9.figshare.19328885.v1

4.4 Results and discussion

4.4.1 SPONGE modules are predictive of breast cancer subtypes

I here present an example of the potential use of the spongEffects methods for cancer subtyping and biomarker identification. In particular, I focus on breast ductal carcinoma (see chapter 1). This is just a specific example, as we envision spongEffects being utilized in different scenarios and for various cancer types in which miRNA-mediated post-transcriptional regulation might have an effect on the observed phenotype.

Alterations of miRNA regulation are a known factor in breast cancer [308] and have been proposed as potential disease biomarkers [309]. The newly developed method spongEffects introduced in this chapter can be used to analyze such alterations and how they characterize different breast cancer subtypes. To do so, we used two large publicly available breast cancer datasets, TCGA-BRCA and METABRIC, respectively, containing 944 and 1699 samples after preprocessing (see Materials and methods paragraph above). We used the TCGA-BRCA dataset as training set and the METABRIC one as external validation set, as standard in any ML pipeline.

I calculated SPONGE modules using the TCGA-BRCA ceRNA network available at http://sponge.biomedical-big-data.de/ [289], preprocessed as described in the Materials and methods section. Weighted centrality values were calculated for all the ceRNAs in the network that could be classified as lncRNA after annotation with the R package *biomaRt* (version 2.42.1) [310]. This choice derived from the potential of this RNA family to be used as biomarkers [285], and from their validated importance in breast cancer subtypes [311]. The top 750 lncRNAs (ranked on by their weighted centrality scores) were further selected as central nodes to define the sponge modules, using the first-neighbor approach described above. Only modules containing between 10 and 200 genes were considered, in a filtering step recommended in similar endeavors [303].

We calculated the SpongEffects scores for the two datasets independently, using the three different single sample enrichment methods described above. Given the differences in the three methods, we were interested in comparing how different enrichment choices impacted the performances of models calibrated on spongEffects scores calculated via GSVA, ssGSEA, and OE. Interestingly, the three approaches showed very comparable performances (Figure 4.2), hinting at the robustness of our approach for the definition of sponge modules.

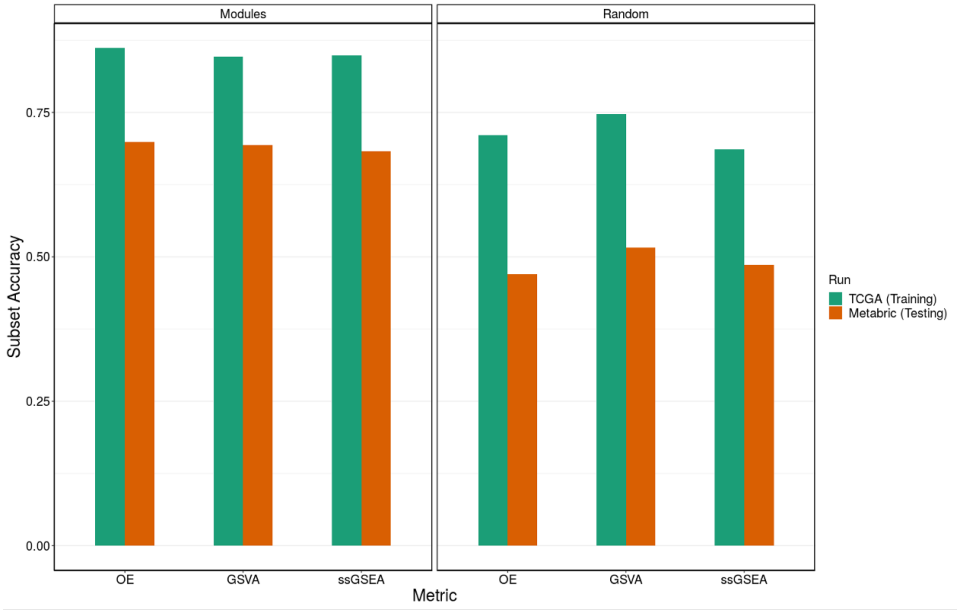

Figure 4.2: Comparison of model performances based on spongEffects scores calculated on the TCGA (training, in green) and METABRIC (validation, orange) datasets using the three different single-sample enrichment tools, Overall Expression (OE), Gene Set Variation Analysis (GSVA), and Single-Sample Gene Set Enrichment Analysis (ssGSEA), implemented in the package. Subset accuracies were evaluated on ceRNA modules (left) and randomly defined gene sets (right). Figure from [282].

I here focus on the results of the spongEffects calculated via OE, given the way this method was used in the original publication on similar bulk transcriptional data [304]. OE-based spongEffects scores are designed to be normally distributed [304]. Discrepancies from such distribution can point at the presence of subgroups of patients/observations potentially different from the rest of the class representatives. This holds true for the spongEffects scores calculated for the samples belonging to the different breast cancer subtypes (Figure 4.3).

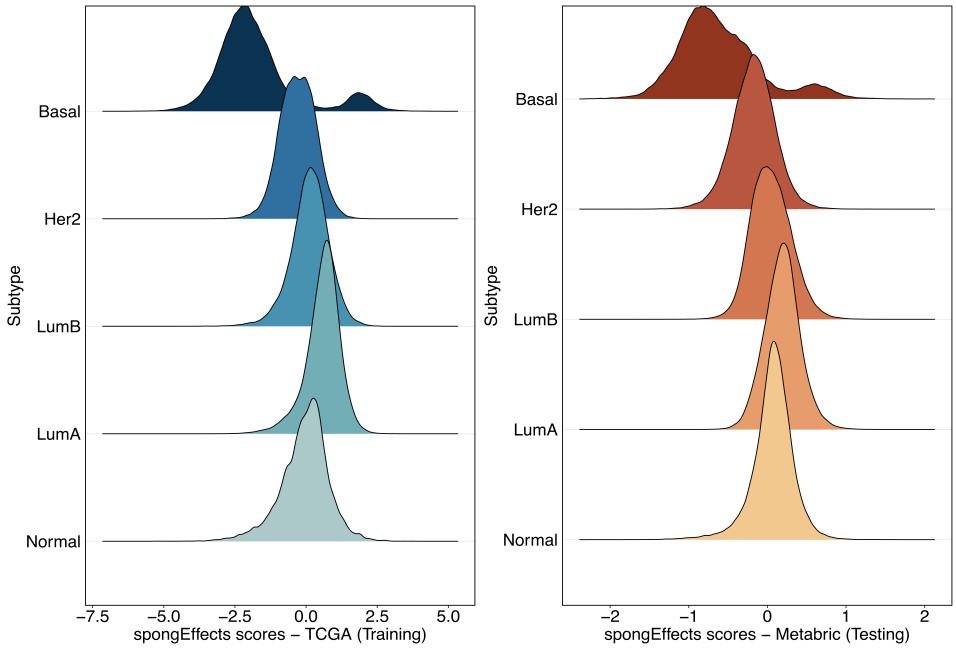

Figure 4.3: Distribution of the spongEffects scores in training (TCGA, left) and testing (METABRIC, right) datasets divided by tumor subtypes. All classes show normal-like distributions apart from the Basal subtype. Figure from [282].

Indeed, while the majority of the subtypes seem to follow a normal-like distribution in both cohorts, the basal samples show a bimodal distribution. The samples belonging to this class can be further modeled via model-based clustering, as implemented in the R package *mixtools* (version 1.2.0) [312], in two subpopulations that show differences in purity and stroma content as calculated via ESTIMATE [313] (Figure 4.4.a) and an over-enrichment of extracellular matrix-associated genes that have been observed to be involved in typical basal invasion programs [314] (Figure 4.4.b). The differences could potentially hint at the role of miRNA regulation in the crosstalk between tumors and their microenvironment. Similar studies have hypothesized the existence of multiple subgroups in Basal cancers based on different data types [315], laying the ground for the investigation of new biomarkers that could help to better stratify breast cancer patients. This emphasizes the potential of spongEffects for the identification of subgroups of patients with potential prognostic value.

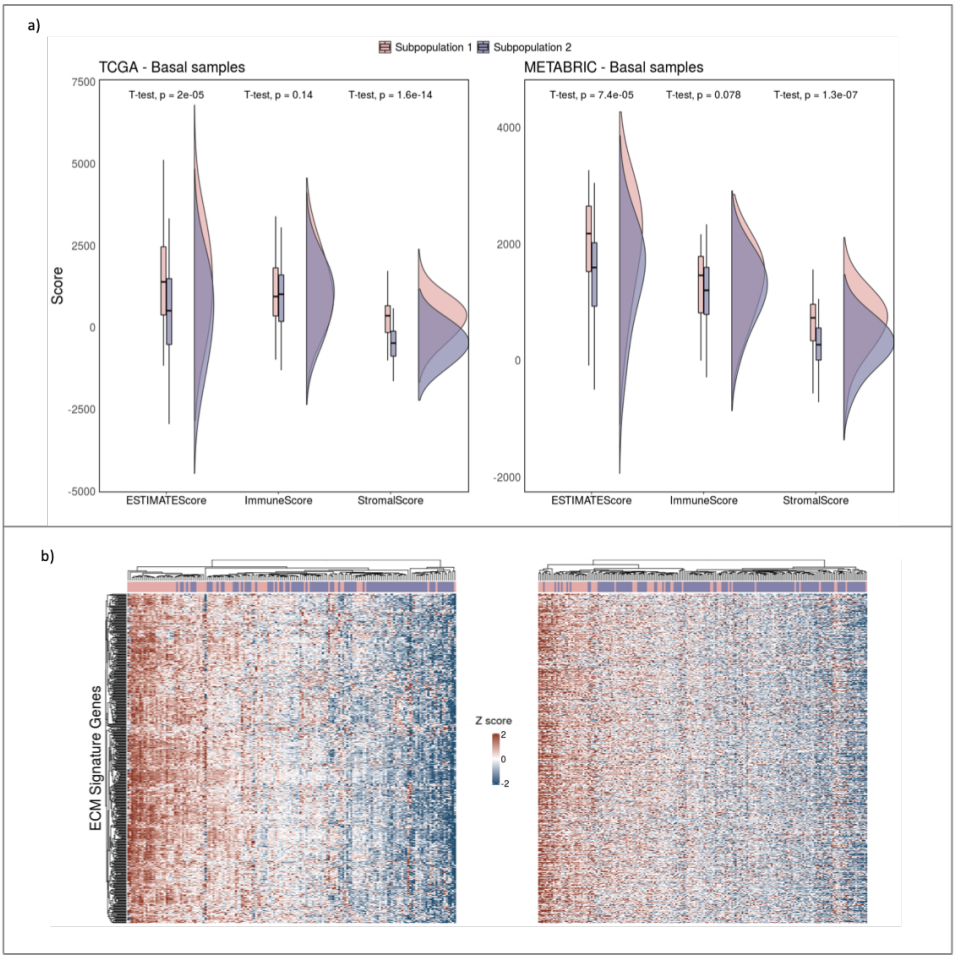

Figure 4.4: Gaussian mixture modeling applied to the spongEffects scores for the Basal samples pinpoints the existence of two subpopulations of patients with different characteristics and points at the role of miRNA regulation in the tumor microenvironment. a) ESTIMATE scores related to purity and stromal content are significantly different between the identified subpopulations. b) Heatmaps showing genes belonging to a validated ECM signature in TCGA (left) and METABRIC (right). Figure from [282].

We calibrated a Random Forest classifier algorithm on the TCGA-BRCA dataset, using sponge modules as input features and annotated subtypes classes as labels. After training, the model was evaluated with multiple metrics. Overall accuracy was evaluated using the exact match ratio, also known as subset accuracy and often used in multiclass classification tasks, while standard singe-glass measures such as sensitivity and specificity were calculated to check the behavior of the model for subtypes traditionally tricky to distinguish and separate from the others, such as Luminal B. Furthermore, we compared the spongEffects-based model to one calibrated on randomly defined modules and on a baseline model trained on the

expression of the central ceRNAs alone. spongEffects scores outperformed the performance of the other approaches in both training and testing and preserved good performance across all subtypes (Figure 4.5).

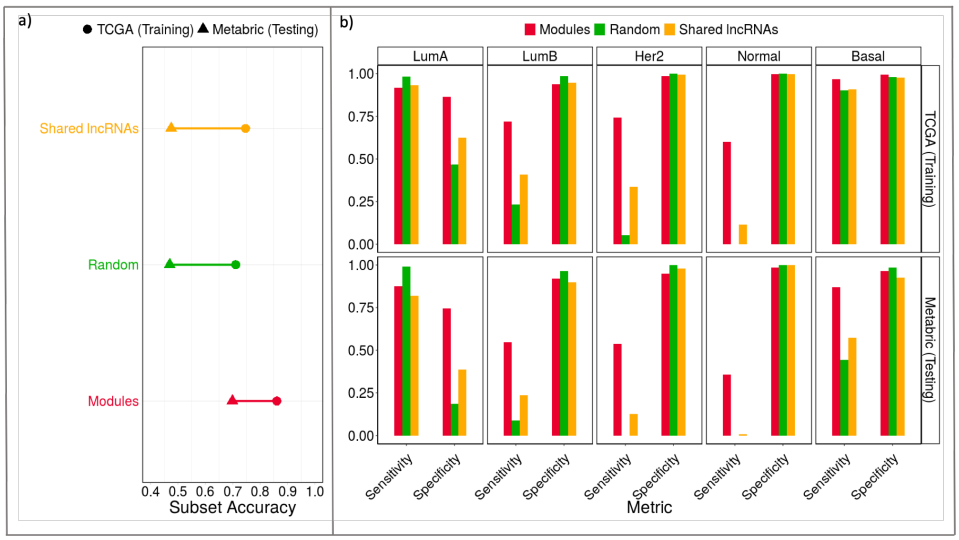

Figure 4.5: Visualization of the performances of the Random Forest models trained on SPONGE modules (red), random modules (green), and central genes only (yellow) on the training and test datasets. a) Subset accuracy values for the three types of models in training and testing. b) Sensitivity and specificity for the three types of models across the 5 breast cancer subtypes taken into consideration. Sponge modules preserve good performance for all the subtypes. Figure from [282].

4.4.2 Interpretation of spongEffect scores

spongEffects scores are designed to summarise the contribution of two different post-translational regulatory mechanisms, namely regulation at the ceRNA network level and miRNA regulation. The method explained in this chapter can summarise the effect of the two different layers on the expression measurements of genes that have the potential to be involved in key mechanisms in the biology of cancer. Purely computational approaches have limited capabilities to disentangle the two effects unless expression data are paired with miRNA one, as is the case for TCGA data shown in the next paragraphs.

SpongEffects scores are hypothesized to be the result of two possible scenarios, shown in Figure 4.6. The first scenario describes the above-mentioned contributions in the case

of increased spongEffects scores. Specifically, these can result from i) upregulation of the central ceRNA, which in turn drives upregulation of the target ceRNAs part of the module independently of the miRNA expression levels, or ii) downregulation of miRNAs and subsequent decreased post-translational regulation. In the opposite situation, i.e., decreased spongEffects scores, downregulation of central ceRNAs may lead to similar effects for the target genes in the modules, or upregulation of the miRNAs could lead to higher regulation.

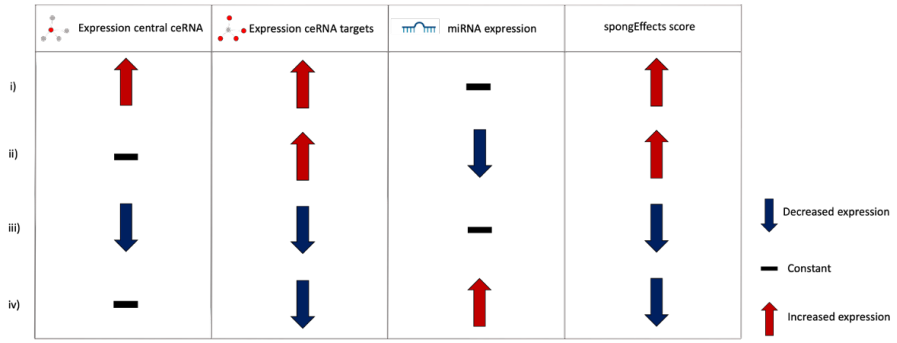

Figure 4.6: Interpretation of increases and reductions of spongEffects scores. i) Increased expression levels of the central ceRNA or ii) lower the expression of miRNAs may result in increased expression of the ceRNAs in a module and thus to higher enrichment scores. On the contrary, iii) lower expression of the central ceRNA and iii) higher expression of targeting miRNAs might lead to lower expression levels of the genes in the modules and overall decreased spongEffects scores. Figure from [282].

4.4.3 ceRNA modules identify fundamental biological mechanisms

In machine learning, the term "feature importance" relates to a group of techniques able to score the variables used to train models to quantify their impact on the final prediction. In biology, features identified with these methods can offer a glimpse into the biology of the system of interest. Here, we used the Gini Index, one of the main feature importance methods applied in Random Forest tasks, to analyze the top 25 ceRNA modules driving subtype prediction in breast cancer (Figure 4.7).

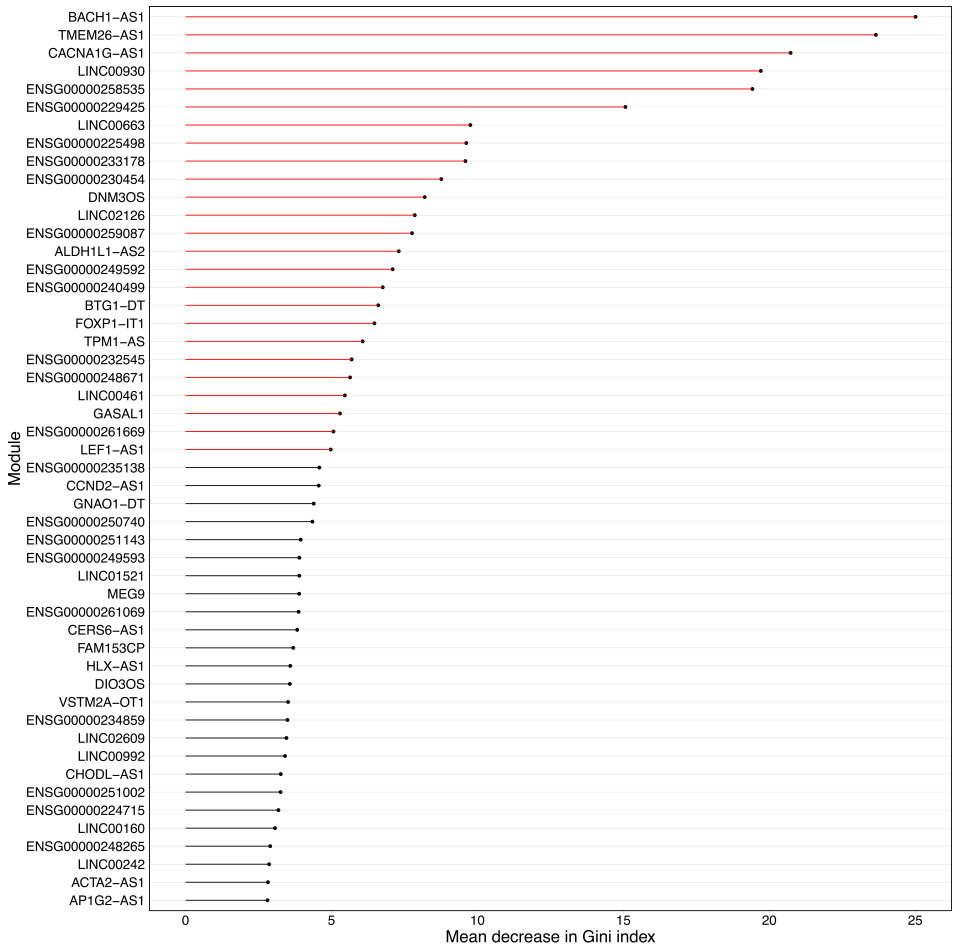

Figure 4.7: Visualization of the Gini indexes for the top 50 most predictive sponge Modules for breast cancer subtype classification tasks. The top25 modules further analyzed in this work are highlighted in red. Figure from [282].

SpongeEffects scores of these modules show clear differences between basal samples and the remaining breast cancer subtypes (Figure 4.8).

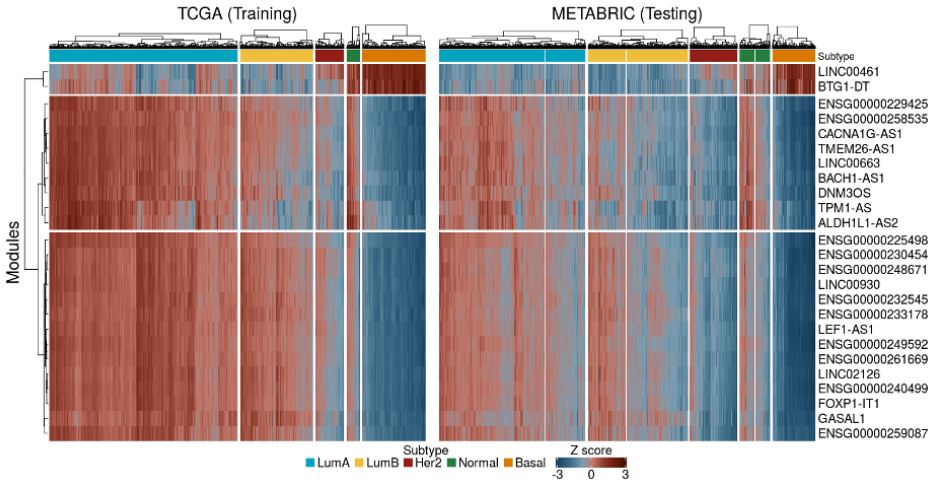

Figure 4.8: Visualization of the spongeEffects scores of the 25 most predictive modules for the training (TCGA, left) and testing (METABRIC, right) cohorts. Basal samples' scores are clearly different from the remaining subtypes, hinting at the potential role of miRNA-based post-transcriptional regulation in the etiology of this aggressive disease. Figure from [282].

Particularly interesting is the case of modules centered around lncRNAs that have been experimentally shown to act as miRNA sponges, such as CACNA1G-AS1, DNM3OS, TPM1-AS, whose modules seem to be downregulated in basal samples, or LINC00461, enriched in the basal subtype. All of these modules have been validated as markers of aggressiveness, proliferation, and migration in multiple cancer types (including breast cancer) [316, 317, 318, 319], thus assuming relevance in this framework.

As described earlier, spongEffects scores are designed to summarise the independent or combined effect of miRNA regulation and ceRNA-target regulation. While the two layers are generally difficult to disentangle, it is possible to gain a qualitative understanding of the relative contributions if matched gene-miRNA expression data are available for the cohort of interest, as is the case for the TCGA-BRCA dataset. In order to do so, we analyzed how many times different miRNAs were predicted by SPONGE to target the genes in the most predictive modules mentioned above. The results for the 51 most representative modules are represented in Figure 4.9.a as the number of genes in a module targeted by the same miRNA divide by the size of the modules and hint at certain over-represented miRNAs and miRNA families that might have an important role in breast cancer biology (more on this in

the original publication [282]). Interestingly, 13 of these are also driving prediction in baseline classification models calibrated on miRNA expression alone and show important differences in expression between the different subtypes (Figure 4.9.b).

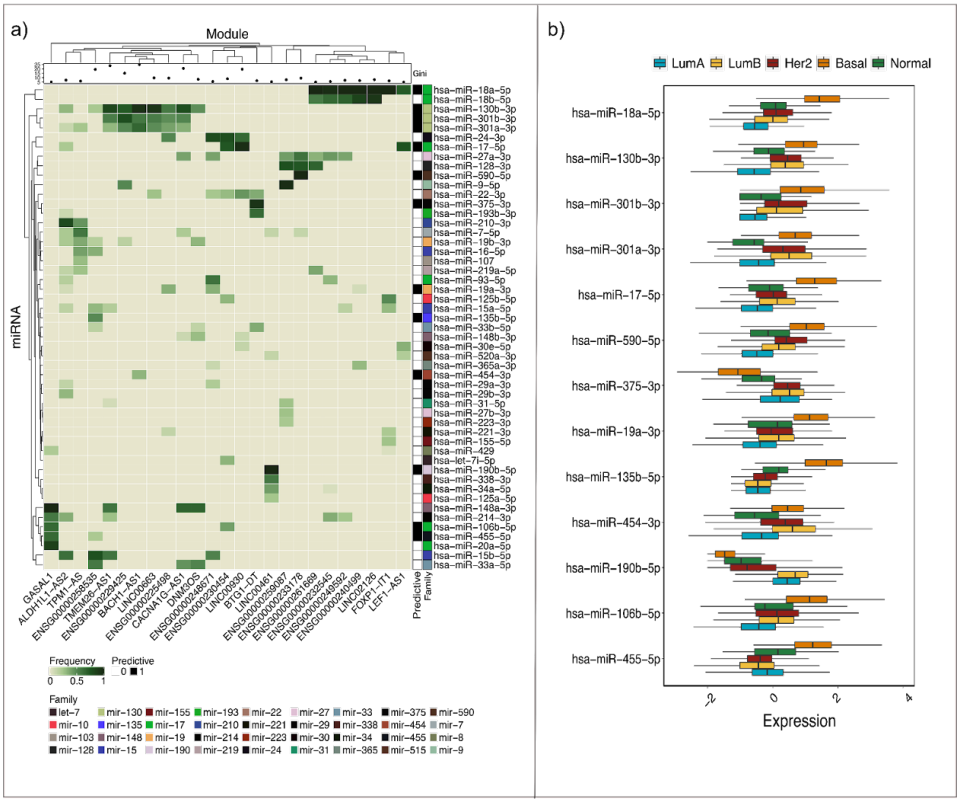

Figure 4.9: The availability of miRNA expression data can help in interpreting the spongEffects scores. a) miRNAs that are predicted to target the most predictive ceRNA modules.Colour coding refers to the number of times a miRNA was predicted to target the genes in the module. Additionally,miRNA families and subtype-predictive miRNAs are highlighted b) Normalized expression of the subtype-predictive miRNAs. Figure from [282].

In order to showcase the interpretation of spongEffects explained above, I here focus on two specific modules, CACNA1G-AS1 and LINC00461, and on the target ceRNAs part of the modules that have been experimentally tested for their role in basal breast cancer. The first module, showing lower spongEffects scores in basal samples in comparison to other subtypes, is composed of genes that are known to have lower expression in basal cancers (Figure 4.10.a and b), such as *TBC1D9* [320], *MYB*, or *ZBTB16* [321, 322]. miRNAs predicted to target the majority of the genes in this module (Figure 4.10.c), such as hsa-miR-301b-3p

and hsa-miR-130b-3p have higher expression in the Basal subtypes, offering a potential way to interpret the resulting scores. Module LINC00461 contains genes that have been found to be highly expressed in basal cancer and to be instrumental to its observed phenotypes (Figure 4.10.d and e), such as *CRYAB* [323, 324, 325], *RARRES1* [326], *BCL11A* [327], *IGF2BP2* [328], and *CDK6* [329] and are regulated by miRNA miR-190b-5p (1q21.3), showing lower expression in Basal samples (Figure 4.10.f).

Figure 4.10: Contribution of miRNA regulation on breast cancer subtypes for the CACNA1G-AS1 and LINC00461 modules. a) 3 experimentally validated genes in the CACNA1G-AS1 module and their shared targeting miRNAs. b) The three genes under analysis part of the CACNA1G-AS1 module (TBC1D9, ZBTB16, and MYB) show different expression levels in the different subtypes. c) Expression of miRNAs targeting the genes in panel b, divided by subtypes. d) 4 experimentally validated genes in the LINC00461 and their shared targeting miRNAs. e) The four genes under analysis part of the LINC00461 module (IGF2BP2, CKD6, RARRES1, and BCL11A) show different expression levels in the different subtypes. f) Expression of miRNAs targeting the genes in panel e. Figure from [282].

4.5 Conclusion and outlook

In this chapter, I introduced spongEffects, a newly developed method able to infer ceRNA modules from available ceRNA networks and calculate sample-specific scores that recapitulate the regulatory activity of ceRNAs and associated miRNAs. By applying it to two large breast cancer transcriptional datasets, I showcase how spognEffects can elucidate regulatory mechanisms in breast cancer subtypes. Importantly, I show how learned modules and sample-specific scores generalize well to new datasets, even if based on different sequencing platforms (e.g., RNA-seq or microarrays). Moreover, I show how ceRNA modules inferred from existing ceRNA networks can be validated on datasets that are missing miRNA expression data. I hypothesize that spongEffects scores recapitulate two different regulation mechanisms, i.e., ceRNA regulation and miRNA regulation, and explain how disentangling the two is possible only in the presence of miRNA data.

I focus on lncRNAs and their role in breast cancer subtypes to elucidate their regulatory mechanisms in combinations with miRNAs. SpongEffects is able to identify important lncRNAs that are known to have an impact on the biology of different cancer types, thus offering the chance to prioritize them in future validation experiments. Significant for future endeavors will be the investigation of lncRNAs' mode of action. For example, it is currently unclear whether lncRNAs are carried outside of the nucleus, transport that would be required for them to take part in the Argonaute-dependent mechanisms of miRNA regulation [70]. Such advancements would lead to the validation of the ceRNA hypothesis and the role of lncRNAs as potential biomarkers or therapeutic targets.

Notably, while this chapter was about post-transcriptional regulation and ceRNA networks inferred via SPONGE, the same framework can in principle be applied to different ceRNA networks and, more generally, gene regulatory networks where similar approaches have been implemented [44, 330].

Finally, I foresee two potential new research avenues. First, spongEffects could be integrated with current methods able to infer transcription factor activity [44, 45], thus combining two different regulatory levels. Second, the increased availability of single-cell datasets, now able

to capture miRNAs' and lncRNAs' expression levels [331, 332], is opening new directions of research for the study of regulatory mechanisms at a higher resolution [333, 334, 335] and has the potential to drive the development of new tools able to disentangle the complexity of regulation in cancer biology.

5 A pharmacogenomics analysis for the identification of biomarkers of drug response in pancreatic cancer

5.1 Declaration of contributions

This project is the result of a collaborative effort as part of the Pancreatic Cancer Collaborative Research Center (SFB 1321) and has been mainly led by Prof. Dr. Dieter Saur and Prof. Dr. Günter Schneider. Hannah Jakubowsky has performed the screening experiments and contributed to the validation of the results of the computational analysis together with Christian Schneeweis. Chiara Falcomatà drove the experimental validation and biological interpretation and performed the functional genomics screens. I designed and implemented the pharmacogenomics analysis, ran the analysis pipeline, and performed the data analysis. In this chapter, I illustrate the technical details related to the implementation of the pipeline. Validation and further experiments related to this project are going to be available in the related publication.

5.2 Introduction

Pancreatic ductal adenocarcinoma (PDAC) is an aggressive and deadly disease, projected to become the 2nd leading cause of cancer-related deaths by 2030 in the US [336]. Unlike other solid tumors, whose prognosis has significantly improved in the past few decades [9], patients diagnosed with PDAC still suffer from very poor outcomes, with 1% of PDAC patients surviving 10 years [8]. The development of targeted therapies offered new hope

for PDAC patients, with retrospective studies on a small cohort of patients (n = 46), which received matched therapies to actionable molecular alteration, showing improved median survival [337]. Such efforts highlight the potential for molecularly-driven therapies in PDAC relying on the mechanistic understanding of drug action, biomarkers of drug sensitivity, and pathways driving resistance, as discussed in this chapter.

Standard PDAC treatment strategies involve surgery as first-line treatment, suitable for only 15% of patients and often combined with adjuvant regimens [338]. Chemotherapy-based therapies typically involve cycles of 5-fluorouracil, leucovorin (folinic acid), irinotecan, and oxaliplatin (FOLFIRINOX), or gemcitabine with or without nab-paclitaxel. Approved targeted therapies for PDAC currently include gemcitabine/erlotinib, inhibitors of poly-ADP-ribose polymerase (PARP) for patients with germline *BRCA1/2* mutations, and immune-checkpoint blockade (ICB) for microsatellite unstable or mismatch repair-deficient tumors [167, 339], with new potential strategies aiming at the targeting of specific PDAC subtypes [168]. The causes for the limited success of targeted therapies in PDAC are multifaceted and have been linked to the high heterogeneity of this disease (see Chapter 1), whose analysis is limited by the low number of tissue and culture resources publicly available and often confounded by the high stroma content, a hallmark of PDAC, hampering molecular profiling and playing a role in the immunosuppressive phenotypes often observed for this disease [340].

Genetically engineered mouse models (GEMMs) have been shown to be a suitable option to overcome the limited availability of PDAC tissues and to offer a route to discover actionable biomarkers for PDAC treatment. GEMMs can be bred in large numbers to fully represent the genetic and molecular heterogeneity observed in PDAC patients, including examples of advanced, highly aggressive, and metastatic tumors that are often not surgically resectable. Furthermore, GEMMs have been shown to recapitulate the main feature of human PDAC, such as the complexity of its microenvironment, while allowing the flexibility of controlling and manipulating fundamental genes involved in PDAC [341, 342]. In this work, I exploit access to the world's largest cohort of 2D cell cultures derived from PDAC GEMMs. The cell cultures were isolated from GEMMs harboring mutations in commonly observed oncogenes such as *KRAS*, *BRAF*, *MEK*, and *PIK3CA*, often combined with loss-of-function alleles from known

tumor suppressors such as *TP53*, *CDKN2A*, *ARF*, *CDKN1B*, or *SMAD4*. Such alterations induce tumors that recapitulate the mutational landscape found in human PDAC tumors, together with salient histopathological and evolutionary features [163, 342].

High-throughput screens offer a chance to systematically investigate drug response in PDAC cell cultures and identify biomarkers of drug resistance or sensitivity (see chapter 2). Existing large scale screening efforts screened and characterized only a small number of human pancreatic cancer cell lines, 48 and 40 by the Cancer Cell Line Encyclopedia [343] and The Genomics of Drug Sensitivity in Cancer (GDSC) [247] respectively, limiting the potential for the identification of robust biomarkers given the limited sample size and statistical power. Moreover, the high passage number of these cultures, often lacking matched normal samples, led to the accumulation of mutations that further confound biomarker-identification approaches, despite the large molecular characterization undergone by these consortia.

Typical pharmacogenomic settings involve the prediction of drug response, quantified e.g. via IC50 or AUC, from basal molecular features such as genomics (e.g. copy number variation or point mutations) or transcriptomics. While different methods have been implemented to tackle this prediction problem (see chapter 2 for an overview of available methods), it has been observed that prediction performances tend to perform similarly independently of the complexity of the ML techniques used [344, 345]. Recent efforts focused on the integration of multiple omics layers for drug response prediction [346, 347] or the addition of *a priori* knowledge [348] to improve predictions.

The latter has shown particularly meaningful results, not only in terms of increased predictive performances but also in terms of enhanced interpretability of the trained models. For instance, combinations of genomic data and chemical structures have shown promising results in the identification of pathways involved in response to mTOR and CDK4/6 inhibitors in breast cancer [349]. Manually annotated and validated gene sets, as the ones collected in the Molecular Signature Database [350], can be a powerful source of *a priori* knowledge, given the advantage they offer in supplying an intuitive and interpretable way to evaluate biological activity and in shifting the focus from the role of single genes to the coordination of multiple gene groups and, potentially, disease mechanisms [303]. Furthermore, the use of

gene sets rather than single genes as input features tackles one of the main problems often encountered in pharmacogenomic projects, namely the differences in complexity between HTS results, often available in limited sample sizes, and heterogeneous and information-rich sequencing data [351, 352]. Finally, gene sets can be easily integrated with RNA-seq data to obtain sample-specific summaries of gene set activity in the cohort by using single-sample gene set enrichment methods such as single-sample Gene Set Enrichment Analysis (ssGSEA) [303, 353].

Network-based approaches have given further impulse to the use of gene sets and provided a solid ground for the characterization of the relationship between drugs and diseases. For instance, different network-based approaches have been implemented in the framework of drug repurposing (see chapter 2), where supervised and unsupervised methods have used networks to investigate target similarity between drugs initially developed for different diseases [354, 355]. In addition, it has been demonstrated that network analyses can result in the identification of clusters of genes associated with treatment outcomes, in particular when information about the proximity of drug targets and disease-relevant subnetworks are taken into account [356, 357]. Similar approaches can be implemented to select clusters of genes of relevance for the drug of interest, thus performing a network-based feature selection step that reduces the complexity of the system under analysis.

In parallel to these methodological advancements, in-depth analysis of existing drug screening efforts, e.g., CCLE, GDSC mentioned above and the Cancer Therapeutics Response Portal (CTRP) [358], and smaller-scale ex vivo functional drug testing efforts showed that cell lines undergoing high-throughput drug screens tend to display comparable resistant behaviors across different drugs [359, 360]. This phenomenon, which I will refer to as "General Response across Drugs" (GRD, as in [359]), has been hypothesized to be related to multi-drug resistance, typically observed in the clinical setting [360]. Multi-drug resistance occurs when tumors display mechanisms of resistance that confer protection against compounds that are structurally and functionally different. Multi-drug resistance has been traced back to different causes, such as pathways rewiring and over-activation or inhibition of mechanisms inducing apoptosis [361, 362]. GRD may play an important role in high-throughput drug screens and

has been shown to confound the identification of biomarkers of drug response [359, 360].

In this chapter, I introduce an innovative pharmacogenomic pipeline for drug response prediction and biomarker discovery in murine PDAC cell lines. We performed high-throughput drug screens on 251 murine PDAC cell lines using an extensive drug library comprising 416 compounds ranging from chemotherapeutics to targeted therapies. Baseline transcriptional profiles of the cell lines were measured via RNA-sequencing and used to find associations between drug response and gene expression in order to uncover biomarkers of drug sensitivity or resistance. In order to do so, the pipeline builds on in-house generated data, namely RNA-seq and drug response data, and publicly available information such as protein-protein interaction (PPI) networks and manually annotated gene sets related to cellular processes, signaling pathways, and regulatory mechanisms (Figure 5.1.a). I exploit the gene sets to drive an *a priori* feature selection step implemented via a network-based approach, thus overcoming the limitations imposed by typical techniques used in this framework, e.g., the instability of feature selection in elastic-net and lasso regression (Figure 5.1.b). In parallel to the feature selection step, I investigate the HTS results data to calculate estimates of general mechanisms of resistance (i.e., GRD) (Figure 5.1.c). Finally, I calculate the single sample enrichment scores for the selected gene sets and use them, together with the GRD estimates, as covariates in the pharmacogenomic model to predict drug response values (Figure 5.1.d). To my knowledge, this is the first example where information from the drug space, i.e., estimation of GRD, and expression space, i.e., a combination of RNA-seq data and gene sets, are unified in a unique pipeline. Combination of the *a priori* feature selection step and GRD estimate results in models that help identify biomarkers of drug sensitivity and resistance that can be experimentally validated (Figure 5.1.e).

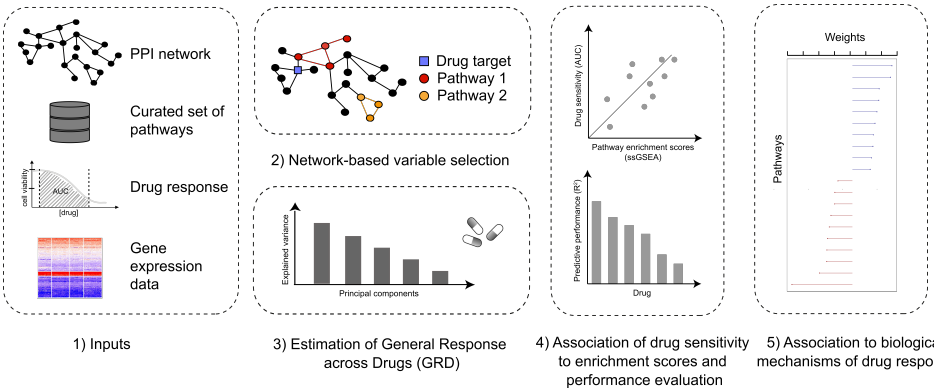

Figure 5.1: Schematic overview of the pharmacogenomics analysis designed in this chapter. a) The pipeline uses both in-house generated data, i.e., gene expression data and drug response values, and publicly available information, i.e., PPI networks and curated gene sets, as inputs. b) A network-based feature selection step is introduced to reduce the dimensionality and complexity of the analysis. c) General Response across Drugs (GRD) is taken into account as a potential confounding source to control for hypersensitive cell lines responding to all drugs. d) Training of penalized linear regression models to associate selected features with drug response values. e) Interpretation and validation of predictive features lead to the identification of potential mechanistic biomarkers or pathways linked to drug response or resistance.

5.3 Material and methods

5.3.1 Primary PDAC cell cultures

Primary low-passaged 2D mPDAC cell cultures were isolated from a large cohort of genetically engineered PDAC mouse models, with various different genotypes [363, 364, 365, 366, 342]. Endogenous tumors are initiated by various oncogenic drivers, such as $KRAS^{G12D}$, $BRAF^{V600E}$, $MEK^{1S218D/S222D}$, and $PIK3CA^{H1047R}$ combined with loss and gain of function alleles for >30 genes and gene combinations, which recapitulate the spectrum of genetic alterations in human PDAC, such as *TP53*, *CDKN2A*, *INK4a*, *ARF*, *SMAD4*, *TGFBR2*. All samples were in culture for <30 passages, genotyped and quality tested for Mycoplasma contaminations, as described in [168].

5.3.2 Automated high-throughput drug screening

Automated drug screening of the 2D cell cultures was performed as described in [367]. In particular, tumor cells were seeded in 96-well plates (750/2000 cells/well) using a Multidrop™

Combi Reagent Dispenser (Thermo Fisher Scientific). After overnight incubation, the drugs were added to the cells using a CyBio® FeliX pipetting platform (Analytik Jena, Jena, Germany). The drug library consisted of 416 drugs, all obtained from SelleckChem targeting a variety of cancer-relevant pathways in clinical and preclinical development. Cells were treated at 7 different concentrations defined via serial dilutions (3x), with minimum and maximum concentration values set at 10 nm and 10 µM respectively. Cell viability was measured with CellTiter-Glo® Luminescent Cell Viability Assay after 72 hours of treatment. Dose-response curves and traditional measures of drug sensitivity, i.e., half-maximal effective concentration (EC50), efficacy (Emax), area under the curve (AUC), and half-maximal inhibitory concentration (IC50), were generated with the *GRmetrics* R package (version 1.12.2) [368, 369]. Every cell line was treated in 2 replicates to obtain reliable metrics of drug response.

5.3.3 Gene expression profiling and pathway data

RNA isolation was performed as described in [168]. RNA-seq library preparation and sequencing were done as described in [163]. RNA sequencing data were normalized and log-stabilized with the *DESeq2* R package (version 1.26.0). Single-sample gene set enrichment analysis (ssGSEA) [370] was performed with the *GSVA* R package (version 1.34.0) [303] using standard parameters on the normalized gene expression data to obtain sample-specific scores of the pathways of interest. The PID pathways [371] and 50 cancer hallmark gene sets with mouse genome annotation were downloaded via the *msigdbr* R package (version 7.4.1)[372].

5.3.4 Quantification of drug target-pathway proximity

A network-based feature selection approach was implemented prior to the model calibration step to identify pathways potentially related to the targets of the monotherapy, following the procedure presented in [348]. Drug target-pathway associations were assessed on the shortest distance between genes in the gene sets/pathways and drug targets within a protein-protein interaction network. More specifically, distances were defined as the average of the shortest paths *d(g,t)* between genes t annotated as drug targets G_t and genes g in the gene set G_s, as described in [348]:

$$d_c = \frac{1}{|G_T|} \sum_{t \epsilon G_t} min_{g \epsilon G_s} d(g,t) \quad (1)$$

Significance of the gene-target distance d(s,t) was assessed via 10,000 bootstrapping iterations of random genes, selected by maintaining the degree of the original drug target and gene-set genes. Such a procedure resulted in a control distribution that was used to calculate z-scores of the resulting distances calculated as in equation 4. Gene sets resulting in z-scores lower than -1.286 (i.e. alpha = 0.9) were considered as proximal to the drug targets. Such procedure was based on the implementation at: https://github.com/emreg00/toolbox. The protein-protein interaction network used for the proximity search was downloaded from STRING (version 11) [373], while the drug targets were downloaded from DrugBank or Proteome DB (downloaded on 17.01.2022) [374, 375].

5.3.5 General Response across Drugs (GRD)

Patterns of resistance across multiple drugs have been previously identified in high-throughput drug screening efforts (see Introduction). It has been shown that it is possible to estimate them via the analysis of the drug screening space, to obtain sample specific estimates that can be included as covariates in pharmacogenomic models. In this work, I referred to them as General Response across Drugs (GRD, as done in [359]) and estimated them similarly to what was done in [360]. For each drug, I selected a set of unrelated drugs by applying the following two selection steps. First, I selected all the drugs not sharing the same targets as the drug of interest. Targets annotated by the drug producing company were used in this step. Second, I calculated Pearson correlation coefficients between the Area Under the Curve (AUC) values of the drugs selected in the first step and the ones of the drug of interest. I ranked the correlation coefficients and removed the 10 drugs with the highest correlation. The resulting drugs are seen as "negative-control" and used to calculate GRD without the risk of taking into account any signal specific to the drug of interest [360]. I estimated the GRD via principal component analysis as implemented in the *prcomp* built-in R function and selected the first 5 principal components as similarly done in the original publication. These were then included as covariates in the penalized linear regression model.

5.3.6 Penalized linear regression

Penalized linear models were calibrated on PID pathway enrichment scores to predict drug response values, here represented by the Area Under the Curve (AUC). Each drug was predicted separately, including the first 5 components from the GRD estimation as covariates, as done in [359, 360]. Thus, the resulting models had the following form for each compound c:

$$AUC_c = \sum_{i=1}^{5} GRD_i + \sum_{g \epsilon G_s} \beta_g x_g \quad (2)$$

Model coefficients were penalized with ridge regularization, to constrain model weights and avoid overfitting [376] while minimizing the following penalized sum of squares:

$$\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda(\sum_{j=1}^{p} \beta_j^2) \quad (3)$$

All the models were implemented with the *glmnet* R package (version 4.1-2) with mixing parameter alpha = 0 to force l2 regularization [377]. For each drug, the full cohort was split before the training step, with 90% of the samples being allocated for model calibration and 10% as an external validation set. Calibration and optimization of the lambda parameter, defining the constraints on the model weights, were achieved through 5-fold cross-validation on the training samples. The whole process was repeated 1000 times to assess model transferability and to obtain robust estimations and resulted in 1000 different models for each drug in the library. For each iteration, I calculated the Pearson correlation between predicted and observed AUC values in the external validation set and defined the model performance as the median coefficient across the 100 iterations. I considered drug-specific models to be predictive if i) the resulting Pearson correlation coefficient was above 0.3 and ii) their performance was consistently better than models built after randomization of the response variable.

Finally, I compared the performance of these models to baseline ones trained on the gene expression data alone and on the pathway enrichment scores without the addition of the GRD covariates.

5.3.7 Whole-genome CRISPR–Cas9 screens

The whole-genome CRISPR–Cas9 screen was performed as described in [168]. In short, the screen was performed in the clonal 9091 Cas9-expressing cell line, using the genome-wide Brie library (pLenti-guide puro). The cells were infected with pLenti Cas9-2A-BSD (Addgene) and selected with BlasticidinS (Invivogen; 10 µg ml1). After dilution and testing for Cas9 expression, cells were treated with different doses of trametinib (from 1.25to 20 nM, 2x dilutions) the cell lines were assessed for cell proliferation and ERK1/2 phosphorylation at the indicated doses of trametinib. Thereby, we identified a concentration of 5 nM trametinib as the optimal concentration to perform the CRISPR/Cas9 negative selection screen. Cas9-expressing cells were transduced with the Brie whole-genome library and selected in puromycin-containing media (by Sigma-Aldrich). After puromycin withdrawal, the cells were left to recover and subsequently treated with either DMSO (control arm of the screen) or 5 nM trametinib (experimental arm). The cells were treated for two weeks and passaged every 4 days. On the final day, genomic DNA was extracted after harvesting of the cells using the DNeasy Blood Tissue kit or the Blood Cell Culture DNA Maxi Kit (according to the manufacturer's instructions).

Downstream analysis was performed with *MAGeCK* (version 0.5.9.4) [378]. Specifically, reads were aligned using sgRNA sequences as references and counted. β-scores were estimated for each gene via maximum likelihood estimation. β-scores represent enrichment (β-score > 0) or depletion (β-score <0) of the sgRNAs with respect to their initial abundance. Scores falling 2 standard deviations away from the mean of the overall distribution were considered to be related to genes conferring resistance to Trametinib.

5.4 Results and discussion

5.4.1 Gene expression reveals significant heterogeneity in transcriptional states of mPDAC 2D cell cultures

Pancreatic cancer is an extremely heterogeneous disease, presenting a variety of phenotypes that may impact clinical decisions. Transcriptional profiling has proven to be an important tool

to study tumor heterogeneity and has been often implemented as the main molecular layer for tumor subtyping approaches (see Chapter 2). Here, I exploited baseline RNA-sequencing data derived from 251 primary low-passaged murine PDAC 2D cell cultures harboring activating mutations in multiple oncogenic drivers (i.e. *Kras*, *Braf*, or *Pi3k*) in combination with different tumor suppressors (Figure 5.2.a) to identify signatures and signaling pathways enriched in PDAC subtypes.

Dimensionality reduction techniques such as Principal Component Analysis (Figure 5.2.b) show the presence of a gradient along with the first principal component, explaining more than 20% of the total variance, which can be linked to the differences between epithelial and mesenchymal subtypes. This confirms the existence of a continuum of transcriptional states that cover the different PDAC subtypes while highlighting the limitations in defining discrete tumor groups [379].

I further characterized the cohort by performing single-sample Gene Set Enrichment Analysis (ssGSEA) [370] using the 50 cancer hallmark gene sets from the MsigDB [380] (Figure 5.2.c). Differences in epithelial and mesenchymal phenotypes appear to be associated with clear discrepancies in enriched pathways. In particular, while mesenchymal cells show higher enrichment scores for pathways related to inflammation and epithelial to mesenchymal transition, epithelial cell lines are more metabolically active. Interestingly, a small cluster of epithelial cells shows high enrichment of pathways related to enhanced transcription, such as the MYC-related ones. Finally, it is possible to observe that the main oncogene drivers dictate the transcriptional states of the cell lines, with the clearest distinction being between the PI3K- and the KRAS-driven tumors thus emphasizing the role of these two interconnected signaling cascades.

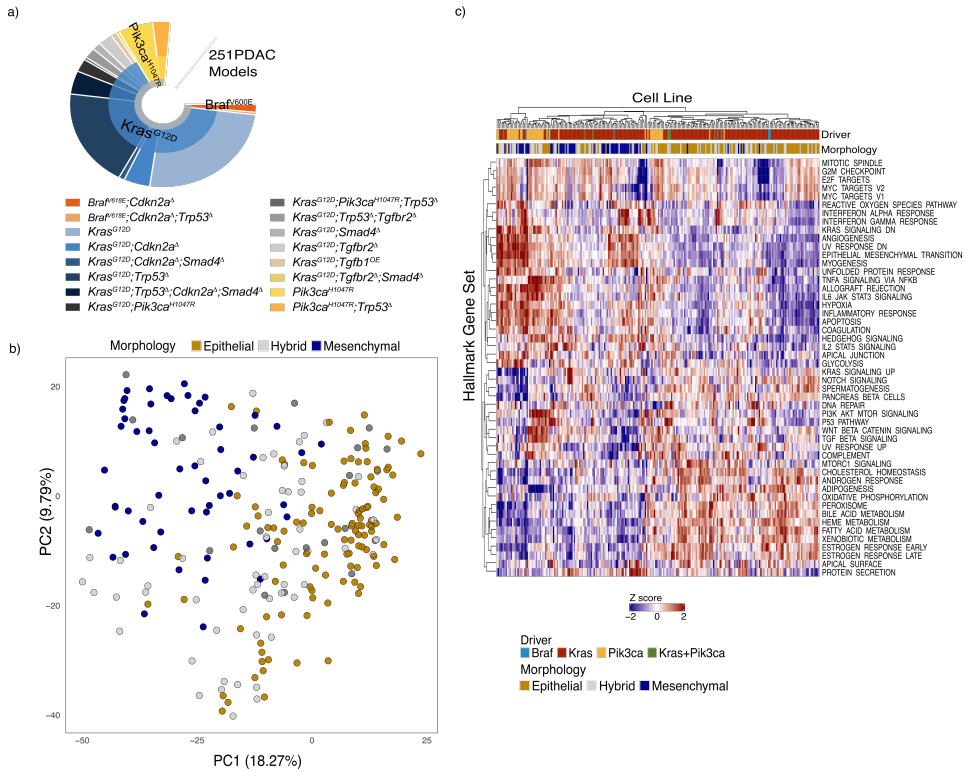

Figure 5.2: Overview of the screened murine cell cultures and exploratory analysis of the related transcriptional profiles. a) Circular plot showing the overall distribution of genetic background of the genetically engineered mouse models that originated the screened PDAC cell lines. b) Principal Component Analysis (PCA) of the 251 baseline RNA-seq profiles of the cohort in analysis. c) Heatmap of the enrichment scores from the 50 Hallmarks of cancer gene sets, annotated with main oncogenic and morphology of the cell lines.

5.4.2 HTSs highlight variability in levels of drug sensitivity

The genetic and phenotypic heterogeneity of PDAC tumors results in a high degree of variability in response to therapy, with patients presenting highly aggressive tumors often not responding to therapeutic interventions. We performed high throughput drug screening on murine PDAC cell cultures to assess drug sensitivity across the full spectrum of PDAC heterogeneity. We used an extensive library of 416 compounds, consisting of drugs approved for clinical use (31%), in clinical trials (29%), or in pre-clinical development (40%), and targeting key pathways and molecular mechanisms altered in cancer (Figure 5.3.a). We observed high variability of drug sensitivity values across the cohort, confirming the high heterogeneity in drug response that characterizes pancreatic cancer (Figure 5.3.b). In order

to carry out the analysis and filter out not informative results (e.g., general cytotoxic or not-effective drugs), I selected the top 102 drugs that presented a median AUC between 0.2 and 0.8 (Figure 5.3.c bottom) and a median absolute deviation higher 0.05 (Figure 5.3.c top). For each of the selected drugs, I calculated their mean response across cell lines (MRC) and analyzed whether drugs targeting similar pathways had similar AUC values (Figure 5.3.d). Similarly, I calculated for each cell line its median response across drugs (a proxy for the GRD), to quantify whether I could observe a similar response across all drugs for different groups of samples. A small group tended to respond relatively poorly to the filtered drugs (red, top part of Figure 5.3.d). Similar observations have already been made in previous works [359, 360], and motivate the need to take into account this phenomenon in any downstream modeling step.

AUC values were, in general, positively correlated across the selected drugs (Figure 5.3.e), with 84% of Spearman's rank correlation values being positive, showing that the screened cell lines had comparable responses to treatment, independently of the different modes of actions of the drugs in the library and backing the notion that GRD plays a role in high-throughput screens. Valuable insights from further analysis of drug-drug correlations can be extracted by focusing on, e.g., negative correlation values. For example, a group of drugs targeting epigenetic mechanisms or kinases shows the highest anticorrelation with agents targeting metabolic or ubiquitin-related pathways (Figure 5.3.e, bottom right in the red square). Interestingly, these are the same drugs whose AUCs show statistically significant differences when compared between epithelial and mesenchymal cell lines (adjusted P-values calculated via ANOVA testing, effect scores are differences in mean AUCs between subtypes). This comparison highlights already known associations, such as the high effectiveness of HDAC inhibitors in mesenchymal cell lines (Figure 5.3.f, left) [381] or the relevance of MEK inhibitors in the epithelial subtypes (Figure 5.3.f, right). These discoveries offer a possible way to stratify patients better towards different treatment strategies based on the relevant PDAC subtype.

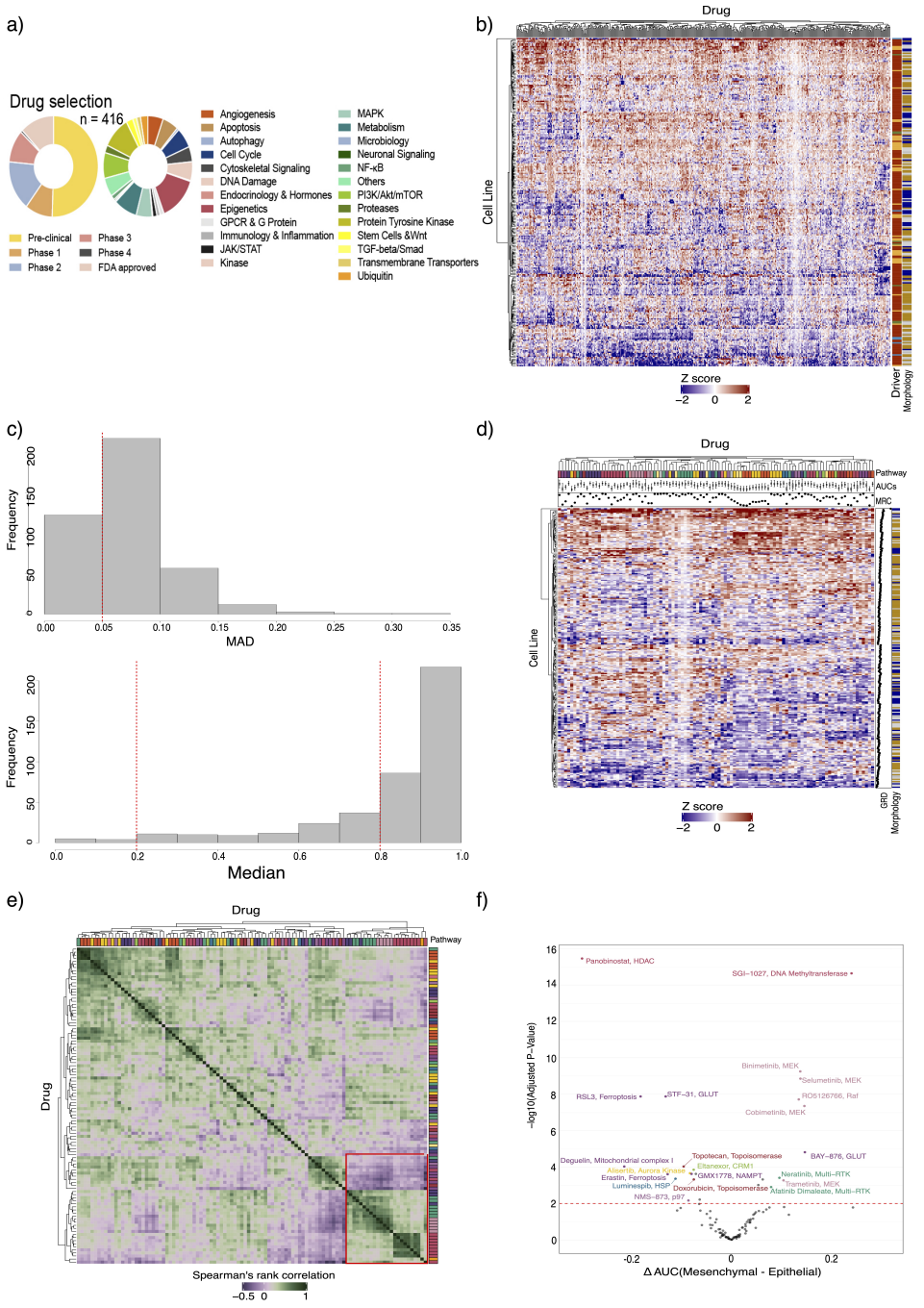

Figure 5.3: Figure caption in the following page

Figure 5.3: Exploratory analysis of the drug space. a) Overview of the drug compound library used for the screens. b) Transformed AUC values for all the drugs in the library across the full cell line cohort. c) Overview of two descriptive statistics, median and median absolute deviation (mad), of the drug response distribution across the whole cohort. The red lines define the cut-offs used to select the 102 drugs of interest for further downstream processing. Drugs showing mad values lower than 0.5 (top) and median outside of the 0.2-0.8 range (bottom) were excluded. d) Transformed AUC values for the selected drugs across the full cohort. Pathways, AUC distributions, and median response across cell lines (MRC) are used to annotate the different drugs (columns). Morphology and median response across drugs (taken as a proxy for GRD) are used to annotate the cell lines (rows). e) Correlation matrix of the 102 selected drugs. Colors represent the Spearman's rank correlation coefficient, ranging from 0.5 (purple) to 1 (green). Rows and columns are annotated with the pathways targeted by the drugs taken into account. A specific group of agents showing interesting negative correlation values is highlighted with a red square. f) Volcano plot on the results of two-way ANOVA test comparing AUC values between epithelial and mesenchymal cell lines, with the x-axis representing the differences in AUCs and the y-axis representing the adjusted p-value. Each dot represents one of the drugs showing negative correlation values and highlighted in panel e. The epithelial morphology was chosen as reference. The color legend is at the bottom.

5.4.3 The addition of *a priori* knowledge and GRD improves predictive performances and interpretability of pharmacogenomic models

High-throughput drug screens allow for the systematic analysis of the therapeutic effects of a vast number of compounds across many samples, in particular when drug response data are integrated with extensive molecular characterization of the screened cohorts. Computational algorithms have the potential to disentangle the complexity of pharmacogenomics interactions to predict drug response and identify potential biomarkers of sensitivity or resistance [196].

The limited success of computational approaches to identify robust biomarkers of drug response can be linked to the heterogeneity and high collinearity of the molecular features, e.g. RNA-seq data, often used as inputs of these pharmacogenomic models, and to the complex and non-linear relationship between omics layers and drug response [348]. Moreover, advanced computational models often lack interpretability, i.e., it is not always possible to understand how and why they reached a solution [382], limiting the possibility of investigating the molecular mechanisms driving drug sensitivity or resistance. Here, I introduce a two-step pipeline designed to overcome these limitations.

First, I implemented a network-based feature selection approach. This class of methods has been shown to significantly improve drug response prediction [356, 357] while offering the chance to use *a priori* knowledge in the shape of validated gene sets and known protein-

protein interactions to overcome the instability of feature selection approaches such as elastic net or LASSO, typically used in pharmacogenomics pipelines. This feature selection method is based on the identification of potential biological pathways that can be associated with drug response and is subsequently used as input features for the prediction of drug response values.

In the second step, variability in general drug response (GRD) is taken into account. GRD has been shown to confound the identification of robust biomarkers and to be an important covariate in the modeling of drug response values [359, 360]. I calculated GRD levels for each drug as described in the Materials and methods section and added them as covariates in the prediction model.

I used ridge regression to associate drug response values to the enrichment scores of the selected pathways, in order to regularize the model coefficients and decrease the chances of overfitting. The use of the selected pathways as inputs for the model drastically reduces the size of the input space, while increasing the possibility of obtaining more stable and reproducible models than those typically built using LASSO or elastic net regression using the expression of single genes as features [383]. RNA-seq-based expression profiles of the 251 murine PDAC 2D cell cultures were transformed into pathway enrichment scores via single-sample enrichment analysis. I used the Pathway Interaction Database (PID) to extract gene sets of interest [371]. It contains 196 manually curated gene sets collecting genes part of key cellular processes, molecular signaling pathways, and regulatory structures. For each drug, I calculated the pathways considered as proximal to its target(s) and used them as inputs to train the ML model together with the GRD estimation via 5-fold cross-validation to predict AUC values. I compared the results of these models to three baseline ones, respectively obtained by calibrating elastic net linear models on gene expression data, alone and with the addition of the estimated GRD, and on ridge regression models calibrated on proximal pathways without the addition of GRD. The addition of GRD has a clear effect on model performances, both in models using gene expression data and pathway enrichment scores, reinforcing the hypothesis that GRD is a useful confounder to take into account in this framework (Figure 5.4.a). A comparison of the performances of models built on single gene

expression and on pathway enrichment scores shows that the performances of models built on different input features may vary based on the drug of interest (Figure 5.4.b). While I am focusing on the pathway-based models for the remainder of this chapter, further work is needed to understand which factors drive better predictive performance based on different inputs and whether there is a biological reason, e.g., the mode of action of the compound, behind it. Moreover, similar works comparing performances of gene-based and pathway-based models showed that the first ones tend to perform better than the others on the training dataset but lose predictive performance on independent datasets [384].

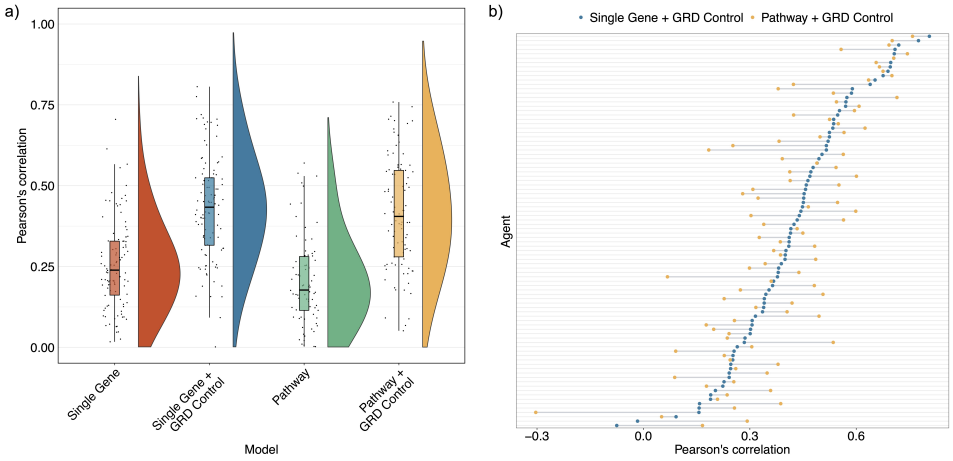

Figure 5.4: Comparison of the performances of the different classes of calibrated models. a) Distribution of the Pearson's correlation values for the different models calibrated in this step of the analysis, i.e., elastic net models based on gene expression alone, with (blue) and without (orange) the addition of GRD control, and ridge regression models calibrated on pathway enrichment scores, with (yellow) and without (green) the addition of GRD control. Each dot represents the performance of a model built on one of the 102 drugs of interest. Higher correlation values correspond to better predictive performance. b) Comparison of the performances of two classes of models resulting from the addition of GRD estimates, i.e., gene expression-based models (blue) and pathway-based ones (yellow). Pearson's correlation values, used as estimates of model performance, are on the x-axis while the 102 drugs under analysis are represented on the y-axis.

5.4.4 Computational models identify important biomarkers of drug response

The approach described earlier resulted in 102 models that can be further investigated to analyze whether they are able to capture mechanisms of drug response and how to use these to derive biomarkers of sensitivity or resistance. I showcase how to do so by illustrating an

example based on the pathway-based model trained to predict response to Trametinib, a highly selective MEK inhibitor found to be a valid anchor for combination therapies in one of our previous works [168]. In principle, the same analysis can be applied to the remaining models and will be part of a follow-up publication.

Robust and hence informative models can be identified by comparing their predictive performance with models calibrated on the same set of input features but randomized drug response values. This approach is designed to identify models that capture meaningful relationships between inputs and outputs, as opposed to models that capture random noise or non-informative signals. To do so, I compare the true distribution of Pearson correlation values to a background distribution of Pearson correlation values from 1,000 random models, see Materials and methods section. For Trametinib, this procedure results in the two distributions depicted in Figure 5.5.a. The difference between the means of the distributions, equal to 0.554, and the p-values resulting from a t-test, equal to 3.2e-145, hint that the model is able to capture informative associations between pathway enrichment scores and response to Trametinib.

Furthermore, investigation of the most predictive features, identified by their coefficients in the model (Figure 5.5.b), can be useful to pinpoint pathways that were found to be positively or negatively associated with drug response. In this framework, positive associations point to the fact that increasing pathway enrichment scores correspond to an increment in AUC values, i.e., to higher resistance. For example, the response of the screened cell cultures to Trametinib showed a strong positive association with KIT-, ERBB1-, ERBB3-, and MYC-related pathways, suggesting that an increase in the activity of these pathways might confer MEKi resistance to tumor cells (Figure 5.5.c), and targeting these pathways might sensitize them towards Trametinib. On the other hand, negative association values correspond to inverse relationships between pathway enrichment scores and drug response values, as in the case of the RAS signaling, for which decreased enrichment scores, i.e., lower evidence of pathway activity, is associated with higher resistance to Trametinib, a compound specifically targeting Ras downstream signaling and MEK-EKR signaling (see Chapter 1).

These dependencies were validated via a pooled genome-wide CRISPR/Cas9-based nega-

tive selection (viability) screen, where PDAC cell lines were screened upon or in absence of Trametinib treatment (Figure 5.5.d), similarly to what was done in [168]. Inferred β-scores, see Materials and methods section for more details, were used to investigate which genes influenced the response to the administered treatment. We focused on genes presenting higher β-scores in the control arm when compared to treatment one, to identify enhanced depletion upon treatment (Figure 5.5.e). The screen allowed us to functionally validate the role of pathways such as ERBB and KIT in driving response to MEK inhibitors (Figure 5.5.f).

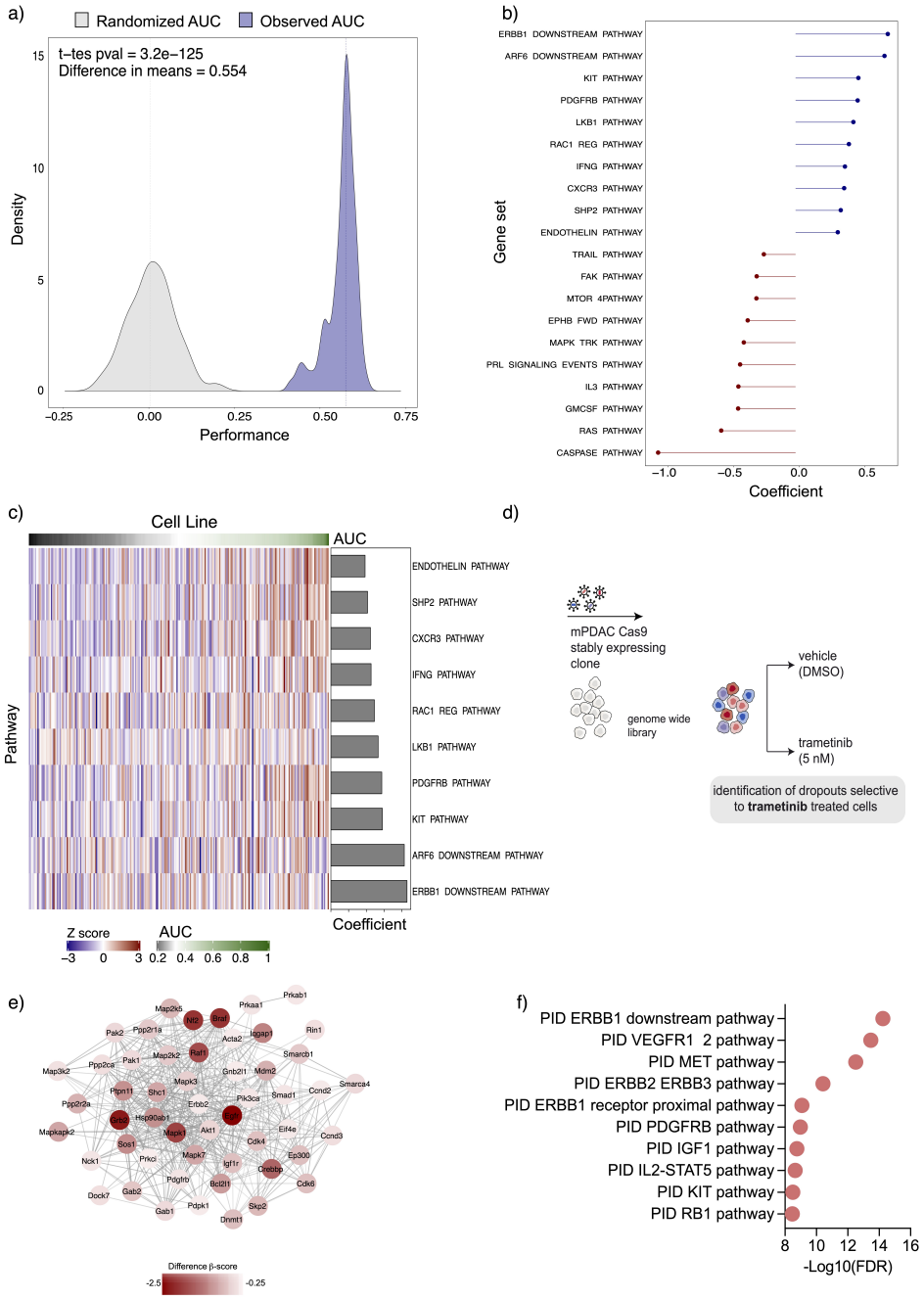

Figure 5.5: Figure caption in the following page

Figure 5.5: Analysis and validation of the model calibrated to predict response to Trametinib treatment. a) Comparison of the distributions of the Pearson correlation values, used as model performance metric, resulting from training a ridge regression model 1000 times to predict observed (violet, right) and randomized (grey, left) AUC values calculated upon Trametinib treatment. P-value, as calculated via a t-test, and difference in means of the distributions hint that the trained model was able to capture potentially informative relationships between inputs and outputs and not pure noise. b) Visualization of the model coefficients for the pathways associated with drug response. The top 10 positively (blue) and top 10 negatively (red) associated pathways are shown. c) Visualization of the single-sample enrichment (ssGSEA) scores for the pathways found to be positively correlated with Trametinib response, stratified by AUC values. d) Design of the whole-genome CRISPR-Cas9 experiment, as presented in [168]. PDAC cell lines were transfected with Cas9 expressing lentivirus and treated in two treatment arms to identify gene dropouts selective to Trametinib treatment. e) Network-based visualization of the genes associated with Trametinib response. Nodes are colored by differences in beta scores between treatment and control arm. Negative differences point to increased depletion upon treatment. f) Ranking of the pathways involved with the development of resistance to Trametinib, resulting from the enrichment of the genes shown in panel e.

5.5 Conclusion

In this chapter, I presented the results of a pharmacogenomics analysis performed on a large yet unpublished murine PDAC 2D cell culture cohort. 251 2D murine PDAC cell cultures have been screened with 416 different compounds in a high-throughput fashion and sequenced to collect their baseline expression profiles. I highlight how the analysis of drug response alone can help the stratification of tumor cells towards specific classes of inhibitors. Moreover, I show how the implementation of a pharmacogenomic pipeline associating drug response to RNA-seq data can uncover meaningful mechanisms of drug response and resistance and point at potential biomarkers. I do so by integrating *a priori* knowledge in the form of gene sets and by applying a network-based feature selection method. Finally, I demonstrate that the inclusion of covariates related to the general response of a drug across the screened cohort drastically improves the predictive performance of the pharmacogenomic model and allows the identification of pathways associated with drug response and resistance that have been successfully validated in independent *in vitro* functional screens.

While this chapter focused on one drug only, the MEKi Trametinib, this analysis lays the ground for the systematic investigation of multiple pathways of drug resistance and can be used as starting point for the design of new, effective, and personalized combination therapies [272, 385]. To achieve this, a few more steps are required in the future. First, the addition of an independent dataset is going to be important to test the results shown in this work

and to check whether pathway-based models are indeed able to capture the biology of drug response better than gene-based ones, as stated in [384, 383]. Moreover, the results must be validated in existing human pharmacogenomic datasets in order to identify associations with the potential of being translated. We are currently in the process of generating such a resource, which will appear in the related publication. Second, *in vitro* models are not able to recapitulate *in vivo* drug action and efficacy. For example, the use of 2D cell cultures does not take into account the effect of the tumor-microenvironment on drug response. More accurate estimations of drug response could be achieved by perturbing 3D organoid or organ-on-a-chip cultures [253, 254], or by using system modeling approaches to model *in silico* tumor cells and their response to drug perturbations [386, 387]. Third, the presented pipeline suffers from limitations given by the chosen feature selection method, which biases the analysis towards known drug targets and ignores potential unknown off-target effects that could explain drug response. In addition, the hard thresholds identified here (i.e., the removal of the 10 drugs with the highest correlation and the selection of the first 5 principal components to include in the model, see Materials and methods) are not optimal and will need to be defined appropriately and separately for each compound in future iterations of this work.

Finally, while the use of transcriptional data has been shown to be beneficial in these types of applications, the addition of multiple molecular layers of characterization may offer the chance to identify biological mechanisms driving resistance or sensitivity not necessarily captured in transcriptional changes across the cohort. Furthermore, the availability of multiple omics layers would pave the way for the use of advanced modeling techniques that would offer the chance to move past the simple associations built in this chapter and better approximate the real relationship between molecular processes and drug response. For example, the integration of proteomic and genomic data with transcriptional profiles, as done in [388], has the potential to offer new insights into the mechanisms of drug response and mode of action of different drugs.

6 General discussion and outlook

6.0.1 Declaration of contributions

This chapter is a personal take on the main topics discussed in this thesis and possible future developments in computational biology and medicine.

In the past decade, advancements in computational biology, machine learning, and artificial intelligence have been the catalysts for the beginning of a new age of discoveries in medicine and biology. This has been possible thanks to the development of new technologies that allowed the generation of large collections of data with different modalities, e.g. imaging or sequencing technologies, at different resolution scales, e.g., at the bulk or single-cell level, often in a high-throughput fashion and at a constantly decreasing price [389].

Precision medicine is one of the fields that are expected to benefit the most from this transformation, given the possibility of using large datasets to find patterns and similarities across different molecular layers to better drive diagnosis, prognosis, and clinical interventions. The benefits of computational techniques have already been established at the level of basic and translational research, where they have become an integral part of the scientific endeavor and contributed to important discoveries.

The aim of this thesis was to investigate applications of machine learning and computational biology to analyze large datasets, with the goal of elucidating mechanisms that play a role in cancer biology. In Chapter 3, I showed how to investigate post-transcriptional regulation at the patient-specific level, with a particular focus on small RNAs such as miRNAs and lncRNAs. I used publicly available human datasets and created a framework for the application of the newly introduced method, spongEffects, to better stratify incoming patients and

identify meaningful prognostic biomarkers or therapeutic targets. In Chapter 4, I presented the implementation of a pharmacogenomics pipeline, designed to associate drug response values generated via high-throughput drug screens to the transcriptional profiles of 251 murine pancreatic cancer cell lines. This is the biggest available pancreatic cancer cohort and lays the ground for the characterization of this extremely aggressive and heterogeneous disease. I showed how the integration of *a priori* knowledge combined with the estimation of confounding factors related to the general effects of the drugs on the screened samples results in highly predictive models and the identification of meaningful biomarkers of drug sensitivity. While both projects are not strictly related to the clinical setting and need further experimental validation, they resulted in the generation of a significant amount of results that can be exploited for patient stratification and the prioritization of biomarkers, thus paving the way towards new precision medicine approaches.

These projects are part of a more general process that is becoming more and more integrated into the way research is performed. While the possibility of measuring thousands of variables, e.g., genes, loci, or genomic regions, with new sequencing technologies has been the main driver of the first large consortia and sequencing efforts such as the ones mentioned in this thesis (e.g., TCGA, CCLE or GDSC), the appearance of single-cell technologies offered a way to increase the resolution by collecting thousands of measurement for hundreds of thousands of observations, i.e. cells, and created an ideal ground for the application of ML and AI technologies in biomedicine [390]. The appearance of the first organ-wide atlases (see for example, [391]) collecting hundreds of thousands of cells from different organs in health and disease are going to be the pivot point for these applications and hold the promise to lead to important discoveries with high translational potential.

The use of AI-based techniques as tools in biomedical research presents different problems than the ones typically associated with the use of these technologies, e.g., fairness or transparency (see Chapter 2 for an overview). Indeed, the applications require predictive performance to be associated with the ability to capture biological phenomena that must be testable and falsifiable. Biological systems are extremely complex and heterogeneous, where observable outputs, i.e. measurable and observable phenotypes, are the results of different

internal processes happening a different scales (e.g., nucleotides form DNA sequences, linear sequences encode 3D proteins, proteins create signaling pathways, etc.). Models, by definition and independently of their nature (i.e., mechanistic or data-driven [382]), can only represent part of this complexity. Recent works suggested the importance of adding prior knowledge and leveraging understanding of the biological processes under analysis when designing and developing new tools [392, 393, 394, 395, 396]. While this idea is mainly applied to advanced machine learning techniques such as deep neural networks, it can be exploited for more basic approaches to gain useful insights, as shown in this work. While not trivial, successfully embedding prior knowledge in computational models will have two advantages, far more important and impactful, from my point of view, than a mere increase in predictive performances: i) increase generalization capabilities of the trained models by introducing inductive biases [392], ii) increased model interpretability. I see these developments as pivotal steps toward the possibility for computational methods to reach their full potential and drive a new era of scientific discoveries.

References

[1] Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed: 2022-2-25.

[2] Freddie Bray, Mathieu Laversanne, Elisabete Weiderpass, and Isabelle Soerjomataram. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, 127(16):3029–3030, August 2021.

[3] Pina Brianti, Eduardo De Flammineis, and Santo Raffaele Mercuri. Review of HPV-related diseases and cancers. *New Microbiol.*, 40(2):80–85, April 2017.

[4] Claudio Pelucchi, Silvano Gallus, Werner Garavello, Cristina Bosetti, and Carlo La Vecchia. Cancer risk associated with alcohol and tobacco use: focus on upper aero-digestive tract and liver. *Alcohol Res. Health*, 29(3):193–198, 2006.

[5] Michelle C Turner, Zorana J Andersen, Andrea Baccarelli, W Ryan Diver, Susan M Gapstur, C Arden Pope, 3rd, Diddier Prada, Jonathan Samet, George Thurston, and Aaron Cohen. Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *CA Cancer J. Clin.*, August 2020.

[6] Reginald D Tucker-Seeley. Social determinants of health and disparities in cancer care for black people in the united states. *JCO Oncol Pract*, 17(5):261–263, May 2021.

[7] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, 2021.

[8] Manuela Quaresma, Michel P Coleman, and Bernard Rachet. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each

cancer in england and wales, 1971-2011: a population-based study. *Lancet*, 385(9974): 1206–1218, March 2015.

[9] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA Cancer J. Clin.*, 70(1):7–30, January 2020.

[10] Annual report to the nation: Cancer deaths continue to drop. `https://www.cancer.gov/news-events/press-releases/2021/annual-report-nation-2021`, July 2021. Accessed: 2022-2-26.

[11] Olivier Elemento, Christina Leslie, Johan Lundin, and Georgia Tourassi. Artificial intelligence in cancer research, diagnosis and therapy. *Nat. Rev. Cancer*, 21(12):747–752, December 2021.

[12] Norman E Sharpless and Anthony R Kerlavage. The potential of AI in cancer care and research. *Biochim. Biophys. Acta Rev. Cancer*, 1876(1):188573, August 2021.

[13] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017.

[14] Adam Yala, Peter G Mikhael, Fredrik Strand, Gigin Lin, Siddharth Satuluru, Thomas Kim, Imon Banerjee, Judy Gichoya, Hari Trivedi, Constance D Lehman, Kevin Hughes, David J Sheedy, Lisa M Matthis, Bipin Karunakaran, Karen E Hegarty, Silvia Sabino, Thiago B Silva, Maria C Evangelista, Renato F Caron, Bruno Souza, Edmundo C Mauad, Tal Patalon, Sharon Handelman-Gotlib, Michal Guindy, and Regina Barzilay. Multi-Institutional validation of a Mammography-Based breast cancer risk model. *J. Clin. Oncol.*, page JCO2101337, November 2021.

[15] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.*, 53(3):354–366, March 2021.

[16] Geoff Fudenberg, David R Kelley, and Katherine S Pollard. Predicting 3D genome folding from DNA sequence with akita. *Nat. Methods*, 17(11):1111–1117, November 2020.

[17] Alireza Karbalayghareh, Merve Sahin, and Christina S Leslie. Chromatin interaction aware gene regulatory modeling with graph attention networks. December 2021.

[18] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, Soledad Perez-Amodio, Pierluigi Strippoli, and Silvia Canaider. An estimation of the number of cells in the human body. *Ann. Hum. Biol.*, 40(6):463–471, November 2013.

[19] George E Billman. Homeostasis: The underappreciated and far too often ignored central organizing principle of physiology. *Front. Physiol.*, 11:200, March 2020.

[20] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R Glassman, Andy DeGiovanni, Jose H Pereira, Andria V Rodrigues, Alberdina A van Dijk, Ana C Ebrecht, Diederik J Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K Rathinaswamy, Udit Dalwadi, Calvin K Yip, John E Burke, K Christopher Garcia, Nick V Grishin, Paul D Adams, Randy J Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021.

[21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis.

Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[22] Yoo-Ah Kim, Dong-Yeon Cho, and Teresa M Przytycka. Understanding Genotype-Phenotype effects in cancer via network approaches. *PLoS Comput. Biol.*, 12(3):e1004747, March 2016.

[23] Chee Seng Ku, En Yun Loy, Agus Salim, Yudi Pawitan, and Kee Seng Chia. The discovery of human genetic variations and their use as disease markers: past, present and future. *J. Hum. Genet.*, 55(7):403–415, July 2010.

[24] Meng Ma, Ying Ru, Ling-Shiang Chuang, Nai-Yun Hsu, Li-Song Shi, Jörg Hakenberg, Wei-Yi Cheng, Andrew Uzilov, Wei Ding, Benjamin S Glicksberg, and Rong Chen. Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics*, 16 Suppl 8:S3, June 2015.

[25] Feng Zhang and James R Lupski. Non-coding genetic variants in human disease. *Hum. Mol. Genet.*, 24(R1):R102–10, October 2015.

[26] R A Weinberg and R A Weinberg. The biology of cancer. 2006.

[27] Douglas Hanahan. Hallmarks of cancer: New dimensions. *Cancer Discov.*, 12(1):31–46, January 2022.

[28] D Hanahan and R A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, January 2000.

[29] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

[30] Eva Y H P Lee and William J Muller. Oncogenes and tumor suppressor genes. *Cold Spring Harb. Perspect. Biol.*, 2(10):a003236, October 2010.

[31] C Giacinti and A Giordano. RB and cell cycle progression. *Oncogene*, 25(38):5220–5227, August 2006.

[32] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, 11(5):345–355, May 2010.

[33] Jia Qian Wu, Lukas Habegger, Parinya Noisa, Anna Szekely, Caihong Qiu, Stephen Hutchison, Debasish Raha, Michael Egholm, Haifan Lin, Sherman Weissman, Wei Cui, Mark Gerstein, and Michael Snyder. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 107(11):5254–5259, March 2010.

[34] Auinash Kalsotra and Thomas A Cooper. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.*, 12(10):715–729, September 2011.

[35] Jonathan D Ellis, Miriam Barrios-Rodiles, Recep Colak, Manuel Irimia, Taehyung Kim, John A Calarco, Xinchen Wang, Qun Pan, Dave O'Hanlon, Philip M Kim, Jeffrey L Wrana, and Benjamin J Blencowe. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol. Cell*, 46(6):884–892, June 2012.

[36] Yuanjiao Zhang, Jinjun Qian, Chunyan Gu, and Ye Yang. Alternative splicing and cancer: a systematic review. *Signal Transduct Target Ther*, 6(1):78, February 2021.

[37] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. *Cell*, 172(4):650–665, February 2018.

[38] Kanchan Vishnoi, Navin Viswakarma, Ajay Rana, and Basabi Rana. Transcription factors in cancer development and therapy. *Cancers*, 12(8), August 2020.

[39] Toshinori Ozaki and Akira Nakagawara. Role of p53 in cell death and human s. *Cancers*, 3(1):994–1013, March 2011.

[40] Yen-Nien Liu, Wassim Abou-Kheir, Juan Juan Yin, Lei Fang, Paul Hynes, Orla Casey, Dong Hu, Yong Wan, Victoria Seng, Heather Sheppard-Tillman, Philip Martin, and Kathleen Kelly. Critical and reciprocal regulation of KLF4 and SLUG in transforming growth factor β-initiated prostate cancer epithelial-mesenchymal transition. *Mol. Cell. Biol.*, 32(5):941–953, March 2012.

[41] Jennifer L Yori, Emhonta Johnson, Guangjin Zhou, Mukesh K Jain, and Ruth A Keri. Kruppel-like factor 4 inhibits epithelial-to-mesenchymal transition through regulation of e-cadherin gene expression. *J. Biol. Chem.*, 285(22):16854–16863, May 2010.

[42] Sovana Adhikary and Martin Eilers. Transcriptional regulation and transformation by myc proteins. *Nat. Rev. Mol. Cell Biol.*, 6(8):635–645, August 2005.

[43] Sheyla Trefflich, Rodrigo J S Dalmolin, José Miguel Ortega, and Mauro A A Castro. Which came first, the transcriptional regulator or its target genes? an evolutionary perspective into the construction of eukaryotic regulons. *Biochim. Biophys. Acta Gene Regul. Mech.*, 1863(6):194472, June 2020.

[44] Mauro A A Castro, Ines de Santiago, Thomas M Campbell, Courtney Vaughn, Theresa E Hickey, Edith Ross, Wayne D Tilley, Florian Markowetz, Bruce A J Ponder, and Kerstin B Meyer. Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.*, 48(1):12–21, January 2016.

[45] Michael N C Fletcher, Mauro A A Castro, Xin Wang, Ines de Santiago, Martin O'Reilly, Suet-Feung Chin, Oscar M Rueda, Carlos Caldas, Bruce A J Ponder, Florian Markowetz, and Kerstin B Meyer. Master regulators of FGFR2 signalling and breast cancer risk. *Nat. Commun.*, 4:2464, 2013.

[46] R D Kornberg. Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(4139):868–871, May 1974.

[47] K Luger, A W Mäder, R K Richmond, D F Sargent, and T J Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–260, September 1997.

[48] Sean D Taverna, Haitao Li, Alexander J Ruthenburg, C David Allis, and Dinshaw J Patel. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.*, 14(11):1025–1040, November 2007.

[49] Marc A Morgan and Ali Shilatifard. Chromatin signatures of cancer. *Genes Dev.*, 29(3): 238–249, February 2015.

[50] Yuan Cheng, Cai He, Manni Wang, Xuelei Ma, Fei Mo, Shengyong Yang, Junhong Han, and Xiawei Wei. Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduct Target Ther*, 4:62, December 2019.

[51] Angelo Ferraro. Altered primary chromatin structures and their implications in cancer development. *Cell. Oncol.*, 39(3):195–210, June 2016.

[52] Boris G Wilson, Xi Wang, Xiaohua Shen, Elizabeth S McKenna, Madeleine E Lemieux, Yoon-Jae Cho, Edward C Koellhoffer, Scott L Pomeroy, Stuart H Orkin, and Charles W M Roberts. Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer Cell*, 18(4):316–328, October 2010.

[53] Jeremy Schwartzentruber, Andrey Korshunov, Xiao-Yang Liu, David T W Jones, Elke Pfaff, Karine Jacob, Dominik Sturm, Adam M Fontebasso, Dong-Anh Khuong Quang, Martje Tönjes, Volker Hovestadt, Steffen Albrecht, Marcel Kool, Andre Nantel, Carolin Konermann, Anders Lindroth, Natalie Jäger, Tobias Rausch, Marina Ryzhova, Jan O Korbel, Thomas Hielscher, Peter Hauser, Miklos Garami, Almos Klekner, Laszlo Bognar, Martin Ebinger, Martin U Schuhmann, Wolfram Scheurlen, Arnulf Pekrun, Michael C Frühwald, Wolfgang Roggendorf, Christoph Kramm, Matthias Dürken, Jeffrey Atkinson, Pierre Lepage, Alexandre Montpetit, Magdalena Zakrzewska, Krzystof Zakrzewski, Pawel P Liberski, Zhifeng Dong, Peter Siegel, Andreas E Kulozik, Marc Zapatka, Abhijit Guha, David Malkin, Jörg Felsberg, Guido Reifenberger, Andreas von Deimling, Koichi Ichimura, V Peter Collins, Hendrik Witt, Till Milde, Olaf Witt, Cindy Zhang, Pedro Castelo-Branco, Peter Lichter, Damien Faury, Uri Tabori, Christoph Plass, Jacek Majewski, Stefan M Pfister, and Nada Jabado. Driver mutations in histone h3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature*, 482(7384):226–231, January 2012.

[54] Adrian Bird. DNA methylation patterns and epigenetic memory. *Genes Dev.*, 16(1):6–21, January 2002.

[55] Malcolm V Brock, James G Herman, and Stephen B Baylin. Cancer as a manifestation of aberrant chromatin structure. *Cancer J.*, 13(1):3–8, January 2007.

[56] Andrew P Feinberg, Rolf Ohlsson, and Steven Henikoff. The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.*, 7(1):21–33, January 2006.

[57] S A Belinsky, K J Nikula, W A Palmisano, R Michels, G Saccomanno, E Gabrielson, S B Baylin, and J G Herman. Aberrant methylation of p16(INK4a) is an early event in lung cancer and a potential biomarker for early diagnosis. *Proc. Natl. Acad. Sci. U. S. A.*, 95 (20):11891–11896, September 1998.

[58] J G Herman, A Merlo, L Mao, R G Lapidus, J P Issa, N E Davidson, D Sidransky, and S B Baylin. Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. *Cancer Res.*, 55(20):4525–4530, October 1995.

[59] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian A Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Nadav S Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J Luo, Eddie Park, Kimberly Persaud, Jonathan B Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E Antonarakis, Gregory Hannon, Morgan C Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, September 2012.

[60] ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, October 2004.

[61] Alexander F Palazzo and Eliza S Lee. Non-coding RNA: what is functional and what is junk? *Front. Genet.*, 6:2, January 2015.

[62] H Ling, K Vincent, M Pichler, R Fodde, I Berindan-Neagoe, F J Slack, and G A Calin. Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene*, 34(39): 5003–5011, September 2015.

[63] Derek de Rie, Imad Abugessaisa, Tanvir Alam, Erik Arner, Peter Arner, Haitham Ashoor, Gaby Åström, Magda Babina, Nicolas Bertin, A Maxwell Burroughs, Ailsa J Carlisle, Carsten O Daub, Michael Detmar, Ruslan Deviatiiarov, Alexandre Fort, Claudia Gebhard, Daniel Goldowitz, Sven Guhl, Thomas J Ha, Jayson Harshbarger, Akira Hasegawa, Kosuke Hashimoto, Meenhard Herlyn, Peter Heutink, Kelly J Hitchens, Chung Chau Hon, Edward Huang, Yuri Ishizu, Chieko Kai, Takeya Kasukawa, Peter Klinken, Timo Lassmann, Charles-Henri Lecellier, Weonju Lee, Marina Lizio, Vsevolod Makeev, Anthony Mathelier, Yulia A Medvedeva, Niklas Mejhert, Christopher J Mungall, Shohei Noma, Mitsuhiro Ohshima, Mariko Okada-Hatakeyama, Helena Persson, Patrizia Rizzu, Filip Roudnicky, Pål Sætrom, Hiroki Sato, Jessica Severin, Jay W Shin, Rolf K Swoboda, Hiroshi Tarui, Hiroo Toyoda, Kristoffer Vitting-Seerup, Louise Winteringham, Yoko Yamaguchi, Kayoko Yasuzawa, Misako Yoneda, Noriko Yumoto, Susan Zabierowski, Peter G Zhang, Christine A Wells, Kim M Summers, Hideya Kawaji, Albin Sandelin, Michael Rehli, FANTOM Consortium, Yoshihide Hayashizaki, Piero Carninci, Alistair R R Forrest, and Michiel J L de Hoon. An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.*, 35(9):872–878, September 2017.

[64] Jinju Han, Yoontae Lee, Kyu-Hyun Yeom, Young-Kook Kim, Hua Jin, and V Narry Kim. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.*, 18 (24):3016–3027, December 2004.

[65] Yoontae Lee, Kipyoung Jeon, Jun-Tae Lee, Sunyoung Kim, and V Narry Kim. MicroRNA

maturation: stepwise processing and subcellular localization. *EMBO J.*, 21(17):4663–4670, September 2002.

[66] Minju Ha and V Narry Kim. Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, 15(8):509–524, August 2014.

[67] Luca F R Gebert and Ian J MacRae. Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.*, 20(1):21–37, January 2019.

[68] Kioomars Saliminejad, Hamid Reza Khorram Khorshid, Shahrzad Soleymani Fard, and Seyed Hamidollah Ghaffari. An overview of microRNAs: Biology, functions, therapeutics, and analysis methods. *J. Cell. Physiol.*, 234(5):5451–5465, May 2019.

[69] Laura B Chipman and Amy E Pasquinelli. miRNA targeting: Growing beyond the seed. *Trends Genet.*, 35(3):215–222, March 2019.

[70] Sean E McGeary, Kathy S Lin, Charlie Y Shi, Thy M Pham, Namita Bisaria, Gina M Kelley, and David P Bartel. The biochemical basis of microRNA targeting efficacy. *Science*, 366(6472), December 2019.

[71] Timothy K K Kamanu, Aleksandar Radovanovic, John A C Archer, and Vladimir B Bajic. Exploration of miRNA families for hypotheses generation. *Sci. Rep.*, 3:2940, October 2013.

[72] Anthony Mathelier and Alessandra Carbone. Large scale chromosomal mapping of human microRNA structural clusters. *Nucleic Acids Res.*, 41(8):4392–4408, April 2013.

[73] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19(1):92–105, January 2009.

[74] Lu Zhang, Yi Liao, and Liling Tang. MicroRNA-34 family: a potential tumor suppressor and therapeutic candidate in cancer. *J. Exp. Clin. Cancer Res.*, 38(1):53, February 2019.

[75] Joseph Eniafe and Shuai Jiang. MicroRNA-99 family in cancer and immunity. *Wiley Interdiscip. Rev. RNA*, 12(3):e1635, May 2021.

[76] Luisa Statello, Chun-Jie Guo, Ling-Ling Chen, and Maite Huarte. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.*, 22(2): 96–118, February 2021.

[77] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, Julien Lagarde, Lavanya Veeravalli, Xiaoan Ruan, Yijun Ruan, Timo Lassmann, Piero Carninci, James B Brown, Leonard Lipovich, Jose M Gonzalez, Mark Thomas, Carrie A Davis, Ramin Shiekhattar, Thomas R Gingeras, Tim J Hubbard, Cedric Notredame, Jennifer Harrow, and Roderic Guigó. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.*, 22(9): 1775–1789, September 2012.

[78] Chun-Jie Guo, Xu-Kai Ma, Yu-Hang Xing, Chuan-Chuan Zheng, Yi-Feng Xu, Lin Shan, Jun Zhang, Shaohua Wang, Yangming Wang, Gordon G Carmichael, Li Yang, and Ling-Ling Chen. Distinct processing of lncRNAs contributes to non-conserved functions in stem cells. *Cell*, 181(3):621–636.e22, April 2020.

[79] Marta Melé, Kaia Mattioli, William Mallard, David M Shechner, Chiara Gerhardinger, and John L Rinn. Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.*, 27(1):27–37, January 2017.

[80] Lesca M Holdt, Steve Hoffmann, Kristina Sass, David Langenberger, Markus Scholz, Knut Krohn, Knut Finstermeier, Anika Stahringer, Wolfgang Wilfert, Frank Beutner, Stephan Gielen, Gerhard Schuler, Gabor Gäbel, Hendrik Bergert, Ingo Bechmann, Peter F Stadler, Joachim Thiery, and Daniel Teupser. Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS Genet.*, 9(7):e1003588, July 2013.

[81] Stefanie Rosa, Susan Duncan, and Caroline Dean. Mutually exclusive sense-antisense transcription at FLC facilitates environmentally induced gene repression. *Nat. Commun.*, 7:13031, October 2016.

[82] Oskar Marín-Béjar, Aina M Mas, Jovanna González, Dannys Martinez, Alejandro Athie, Xabier Morales, Mikel Galduroz, Ivan Raimondi, Elena Grossi, Shuling Guo, Ana Rouzaut, Igor Ulitsky, and Maite Huarte. The human lncRNA LINC-PINT inhibits tumor cell invasion through a highly conserved sequence element. *Genome Biol.*, 18(1): 202, October 2017.

[83] Andrea Piunti and Ali Shilatifard. The roles of polycomb repressive complexes in mammalian development and cancer. *Nat. Rev. Mol. Cell Biol.*, 22(5):326–345, May 2021.

[84] Kyoko L Yap, Side Li, Ana M Muñoz-Cabello, Selina Raguz, Lei Zeng, Shiraz Mujtaba, Jesús Gil, Martin J Walsh, and Ming-Ming Zhou. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell*, 38(5):662–674, June 2010.

[85] Manuela Portoso, Roberta Ragazzini, Živa Brenčič, Arianna Moiani, Audrey Michaud, Ivaylo Vassilev, Michel Wassef, Nicolas Servant, Bruno Sargueil, and Raphaël Margueron. PRC2 is dispensable for HOTAIR-mediated transcriptional repression. *EMBO J.*, 36(8):981–994, April 2017.

[86] Abhinav K Jain, Yuanxin Xi, Ryan McCarthy, Kendra Allton, Kadir C Akdemir, Lalit R Patel, Bruce Aronow, Chunru Lin, Wei Li, Liuqing Yang, and Michelle C Barton. LncPRESS1 is a p53-regulated LncRNA that safeguards pluripotency by disrupting SIRT6-Mediated de-acetylation of histone H3K56. *Mol. Cell*, 64(5):967–981, December 2016.

[87] Paulina A Latos, Florian M Pauler, Martha V Koerner, H Başak Şenergin, Quanah J Hudson, Roman R Stocsits, Wolfgang Allhoff, Stefan H Stricker, Ruth M Klement, Katarzyna E Warczok, Karin Aumayr, Pawel Pasierbek, and Denise P Barlow. Airn transcriptional overlap, but not its lncRNA products, induces imprinted igf2r silencing. *Science*, 338(6113):1469–1472, December 2012.

[88] Lovorka Stojic, Malwina Niemczyk, Arturo Orjalo, Yoko Ito, Anna Elisabeth Maria Ruijter, Santiago Uribe-Lewis, Nimesh Joseph, Stephen Weston, Suraj Menon, Duncan T Odom, John Rinn, Fanni Gergely, and Adele Murrell. Transcriptional silencing of

long noncoding RNA GNG12-AS1 uncouples its transcriptional and product-related functions. *Nat. Commun.*, 7:10406, February 2016.

[89] Philippe Thebault, Geneviève Boutin, Wajid Bhat, Anne Rufiange, Joseph Martens, and Amine Nourani. Transcription regulation by the noncoding RNA SRG1 requires spt2-dependent chromatin deposition in the wake of RNA polymerase II. *Mol. Cell. Biol.*, 31(6):1288–1300, March 2011.

[90] Karen Yap, Svetlana Mukhina, Gen Zhang, Jason S C Tan, Hong Sheng Ong, and Eugene V Makeyev. A short tandem Repeat-Enriched RNA assembles a nuclear compartment to control alternative splicing and promote cell survival. *Mol. Cell*, 72(3): 525–540.e13, November 2018.

[91] Ailone Tichon, Rotem Ben-Tov Perry, Lovorka Stojic, and Igor Ulitsky. SAM68 is required for regulation of pumilio by the NORAD long noncoding RNA. *Genes Dev.*, 32 (1):70–78, January 2018.

[92] Aaron Arvey, Erik Larsson, Chris Sander, Christina S Leslie, and Debora S Marks. Target mRNA abundance dilutes microRNA and siRNA activity, 2010.

[93] Iain M Dykes and Costanza Emanueli. Transcriptional and post-transcriptional gene regulation by long non-coding RNA. *Genomics Proteomics Bioinformatics*, 15(3):177–186, June 2017.

[94] Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A ceRNA hypothesis: the rosetta stone of a hidden RNA language? *Cell*, 146(3):353–358, August 2011.

[95] Yvonne Tay, Lev Kats, Leonardo Salmena, Dror Weiss, Shen Mynn Tan, Ugo Ala, Florian Karreth, Laura Poliseno, Paolo Provero, Ferdinando Di Cunto, Judy Lieberman, Isidore Rigoutsos, and Pier Paolo Pandolfi. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell*, 147(2):344–357, October 2011.

[96] Ugo Ala. Competing endogenous RNAs, Non-Coding RNAs and diseases: An intertwined story. *Cells*, 9(7), June 2020.

[97] Olga Plotnikova, Ancha Baranova, and Mikhail Skoblov. Comprehensive analysis of human microRNA-mRNA interactome. *Front. Genet.*, 10:933, October 2019.

[98] Sonia C DaSilva-Arnold, Che-Ying Kuo, Viralkumar Davra, Yvonne Remache, Peter C W Kim, John P Fisher, Stacy Zamudio, Abdulla Al-Khan, Raymond B Birge, and Nicholas P Illsley. ZEB2, a master regulator of the epithelial-mesenchymal transition, mediates trophoblast differentiation. *Mol. Hum. Reprod.*, 25(2):61–75, February 2019.

[99] Yu-Ru Lee, Ming Chen, and Pier Paolo Pandolfi. The functions and regulation of the PTEN tumour suppressor: new modes and prospects. *Nat. Rev. Mol. Cell Biol.*, 19(9): 547–562, September 2018.

[100] Florian A Karreth, Yvonne Tay, Daniele Perna, Ugo Ala, Shen Mynn Tan, Alistair G Rust, Gina DeNicola, Kaitlyn A Webster, Dror Weiss, Pedro A Perez-Mancera, Michael Krauthammer, Ruth Halaban, Paolo Provero, David J Adams, David A Tuveson, and Pier Paolo Pandolfi. In vivo identification of tumor- suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma. *Cell*, 147(2):382–395, October 2011.

[101] Yifei Cai, Ziling Sun, Huizhen Jia, Hongxue Luo, Xiaoyang Ye, Qi Wu, Yi Xiong, Wei Zhang, and Jun Wan. Rpph1 upregulates CDC42 expression and promotes hippocampal neuron dendritic spine formation by competing with mir-330-5p. *Front. Mol. Neurosci.*, 10:27, February 2017.

[102] Lin Wang, Xiaozhong Li, Louxin Zhang, and Qiang Gao. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer*, 17(1):513, August 2017.

[103] Xiaolei Zhang, Jiaming Zhang, Kailun Zheng, Heng Zhang, Xixiang Pei, Zhi Yin, Duancheng Wen, and Qingran Kong. Long noncoding RNAs sustain high expression

levels of exogenous octamer-binding protein 4 by sponging regulatory microRNAs during cellular reprogramming. *J. Biol. Chem.*, 294(47):17863–17874, November 2019.

[104] Yang Yu, Ying Chen, Xiaolei Zhang, Xiaolang Lu, Jianjun Hong, Xiaoshan Guo, and Dongsheng Zhou. Knockdown of lncRNA KCNQ1OT1 suppresses the adipogenic and osteogenic differentiation of tendon stem cell via downregulating mir-138 target genes PPARγ and RUNX2. *Cell Cycle*, 17(19-20):2374–2385, October 2018.

[105] Yan Nie, Xiang Liu, Shaohua Qu, Erwei Song, Hua Zou, and Chang Gong. Long non-coding RNA HOTAIR is an independent prognostic marker for nasopharyngeal carcinoma progression and survival. *Cancer Sci.*, 104(4):458–464, April 2013.

[106] Ming-Zhe Ma, Chun-Xiao Li, Yan Zhang, Ming-Zhe Weng, Ming-Di Zhang, Yi-Yu Qin, Wei Gong, and Zhi-Wei Quan. Long non-coding RNA HOTAIR, a c-myc activated driver of malignancy, negatively regulates miRNA-130a in gallbladder cancer. *Mol. Cancer*, 13:156, June 2014.

[107] Xiang-Hua Liu, Ming Sun, Feng-Qi Nie, Ying-Bin Ge, Er-Bao Zhang, Dan-Dan Yin, Rong Kong, Rui Xia, Kai-Hua Lu, Jin-Hai Li, Wei De, Ke-Ming Wang, and Zhao-Xia Wang. Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging mir-331-3p in gastric cancer. *Mol. Cancer*, 13:92, April 2014.

[108] Hua-Sheng Chiu, David Llobet-Navas, Xuerui Yang, Wei-Jen Chung, Alberto Ambesi-Impiombato, Archana Iyer, Hyunjae Ryan Kim, Elena G Seviour, Zijun Luo, Vasudha Sehgal, Tyler Moss, Yiling Lu, Prahlad Ram, José Silva, Gordon B Mills, Andrea Califano, and Pavel Sumazin. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. *Genome Res.*, 25(2):257–267, February 2015.

[109] Pavel Sumazin, Xuerui Yang, Hua-Sheng Chiu, Wei-Jen Chung, Archana Iyer, David Llobet-Navas, Presha Rajbhandari, Mukesh Bansal, Paolo Guarnieri, Jose Silva, and Andrea Califano. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell*, 147(2):370–381, October 2011.

[110] Markus List, Azim Dehghani Amirabad, Dennis Kostka, and Marcel H Schulz. Large-scale inference of competing endogenous RNA networks with sparse partial correlation. *Bioinformatics*, 35(14):i596–i604, July 2019.

[111] Reece, Jane B., Campbell, and Neil A. *Campbell biology*. Benjamin Cummings / Pearson, Boston, 2011.

[112] Nicola Normanno, Antonella De Luca, Caterina Bianco, Luigi Strizzi, Mario Mancino, Monica R Maiello, Adele Carotenuto, Gianfranco De Feo, Francesco Caponigro, and David S Salomon. Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*, 366(1):2–16, January 2006.

[113] Jonas Cicenas, Egle Zalyte, Amos Bairoch, and Pascale Gaudet. Kinases and cancer. *Cancers*, 10(3), March 2018.

[114] Fatima Ardito, Michele Giuliani, Donatella Perrone, Giuseppe Troiano, and Lorenzo Lo Muzio. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (review). *Int. J. Mol. Med.*, 40(2):271–280, August 2017.

[115] G Manning, D B Whyte, R Martinez, T Hunter, and S Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, December 2002.

[116] P C John, M Mews, and R Moore. Cyclin/Cdk complexes: their involvement in cell cycle progression and mitotic division. *Protoplasma*, 216(3-4):119–142, 2001.

[117] Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K Chatila, Augustin Luna, Konnor C La, Sofia Dimitriadoy, David L Liu, Havish S Kantheti, Sadegh Saghafinia, Debyani Chakravarty, Foysal Daian, Qingsong Gao, Matthew H Bailey, Wen-Wei Liang, Steven M Foltz, Ilya Shmulevich, Li Ding, Zachary Heins, Angelica Ochoa, Benjamin Gross, Jianjiong Gao, Hongxin Zhang, Ritika Kundra, Cyriac Kandoth, Istemi Bahceci, Leonard Dervishi, Ugur Dogrusoz, Wanding Zhou, Hui Shen, Peter W Laird, Gregory P Way, Casey S Greene, Han Liang, Yonghong Xiao, Chen Wang, Antonio Iavarone, Alice H Berger, Trever G Bivona, Alexander J Lazar, Gary D Hammer, Thomas Giordano, Lawrence N Kwong, Grant McArthur, Chenfei Huang, Aaron D Tward,

Mitchell J Frederick, Frank McCormick, Matthew Meyerson, Cancer Genome Atlas Research Network, Eliezer M Van Allen, Andrew D Cherniack, Giovanni Ciriello, Chris Sander, and Nikolaus Schultz. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337.e10, April 2018.

[118] Joan Massagué. TGFbeta in cancer. *Cell*, 134(2):215–230, July 2008.

[119] Pixu Liu, Hailing Cheng, Thomas M Roberts, and Jean J Zhao. Targeting the phosphoinositide 3-kinase pathway in cancer. *Nat. Rev. Drug Discov.*, 8(8):627–644, August 2009.

[120] Alfonso Quintás-Cardama and Srdan Verstovsek. Molecular pathways: Jak/STAT pathway: mutations, inhibitors, and resistance. *Clin. Cancer Res.*, 19(8):1933–1940, April 2013.

[121] Raphaela Fritsche-Guenther, Franziska Witzel, Anja Sieber, Ricarda Herr, Nadine Schmidt, Sandra Braun, Tilman Brummer, Christine Sers, and Nils Blüthgen. Strong negative feedback from erk to raf confers robustness to MAPK signalling. *Mol. Syst. Biol.*, 7:489, May 2011.

[122] Sung-Young Shin, Oliver Rath, Sang-Mok Choo, Frances Fee, Brian McFerran, Walter Kolch, and Kwang-Hyun Cho. Positive- and negative-feedback regulations coordinate the dynamic behavior of the Ras-Raf-MEK-ERK signal transduction pathway. *J. Cell Sci.*, 122(Pt 3):425–435, February 2009.

[123] A S Dhillon, S Hagan, O Rath, and W Kolch. MAP kinase signalling pathways in cancer. *Oncogene*, 26(22):3279–3290, May 2007.

[124] Naoya Fujita, Saori Sato, and Takashi Tsuruo. Phosphorylation of p27kip1 at threonine 198 by p90 ribosomal protein S6 kinases promotes its binding to 14-3-3 and cytoplasmic localization. *J. Biol. Chem.*, 278(49):49254–49260, December 2003.

[125] Rana Anjum and John Blenis. The RSK family of kinases: emerging roles in cellular signalling. *Nat. Rev. Mol. Cell Biol.*, 9(10):747–758, October 2008.

[126] Charles Y Lin, Jakob Lovén, Peter B Rahl, Ronald M Paranal, Christopher B Burge, James E Bradner, Tong Ihn Lee, and Richard A Young. Transcriptional amplification in tumor cells with elevated c-myc. *Cell*, 151(1):56–67, September 2012.

[127] Zuqin Nie, Gangqing Hu, Gang Wei, Kairong Cui, Arito Yamane, Wolfgang Resch, Ruoning Wang, Douglas R Green, Lino Tessarollo, Rafael Casellas, Keji Zhao, and David Levens. c-myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, 151(1):68–79, September 2012.

[128] Renumathy Dhanasekaran, Anja Deutzmann, Wadie D Mahauad-Fernandez, Aida S Hansen, Arvin M Gouw, and Dean W Felsher. The MYC oncogene - the grand orchestrator of cancer growth and immune evasion. *Nat. Rev. Clin. Oncol.*, 19(1):23–36, January 2022.

[129] Reiko Sugiura, Ryosuke Satoh, and Teruaki Takasaki. ERK: A Double-Edged sword in cancer. ERK-Dependent apoptosis as a potential therapeutic strategy for cancer. *Cells*, 10(10), September 2021.

[130] Carlos L Arteaga and Jeffrey A Engelman. ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. *Cancer Cell*, 25(3):282–303, March 2014.

[131] Annette S Little, Kathryn Balmanno, Matthew J Sale, Scott Newman, Jonathan R Dry, Mark Hampson, Paul A W Edwards, Paul D Smith, and Simon J Cook. Amplification of the driving oncogene, KRAS or BRAF, underpins acquired resistance to MEK1/2 inhibitors in colorectal cancer cells. *Sci. Signal.*, 4(166):ra17, March 2011.

[132] N Nakayama, K Nakayama, S Yeasmin, M Ishibashi, A Katagiri, K Iida, M Fukumoto, and K Miyazaki. KRAS or BRAF mutation status is a useful predictor of sensitivity to MEK inhibition in ovarian cancer. *Br. J. Cancer*, 99(12):2020–2028, December 2008.

[133] Georgina V Long, Daniil Stroyakovskiy, Helen Gogas, Evgeny Levchenko, Filippo de Braud, James Larkin, Claus Garbe, Thomas Jouary, Axel Hauschild, Jean Jacques Grob, Vanna Chiarion Sileni, Celeste Lebbe, Mario Mandalà, Michael Millward, Ana

Arance, Igor Bondarenko, John B A G Haanen, Johan Hansson, Jochen Utikal, Virginia Ferraresi, Nadezhda Kovalenko, Peter Mohr, Volodymyr Probachai, Dirk Schadendorf, Paul Nathan, Caroline Robert, Antoni Ribas, Douglas J DeMarini, Jhangir G Irani, Michelle Casey, Daniele Ouellet, Anne-Marie Martin, Ngocdiep Le, Kiran Patel, and Keith Flaherty. Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma. *N. Engl. J. Med.*, 371(20):1877–1888, November 2014.

[134] Jeffrey A Engelman, Liang Chen, Xiaohong Tan, Katherine Crosby, Alexander R Guimaraes, Rabi Upadhyay, Michel Maira, Kate McNamara, Samanthi A Perera, Youngchul Song, Lucian R Chirieac, Ramneet Kaur, Angela Lightbown, Jessica Simendinger, Timothy Li, Robert F Padera, Carlos García-Echeverría, Ralph Weissleder, Umar Mahmood, Lewis C Cantley, and Kwok-Kin Wong. Effective use of PI3K and MEK inhibitors to treat mutant kras G12D and PIK3CA H1047R murine lung cancers. *Nat. Med.*, 14(12):1351–1356, December 2008.

[135] X Yang and M E Lippman. BRCA1 and BRCA2 in breast cancer. *Breast Cancer Res. Treat.*, 54(1):1–10, March 1999.

[136] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, October 2012.

[137] Adrienne G Waks and Eric P Winer. Breast cancer treatment: A review. *JAMA*, 321(3): 288–300, January 2019.

[138] Ramona G Dumitrescu. Interplay between genetic and epigenetic changes in breast cancer subtypes. In Ramona G Dumitrescu and Mukesh Verma, editors, *Cancer Epigenetics for Precision Medicine : Methods and Protocols*, pages 19–34. Springer New York, New York, NY, 2018.

[139] C M Perou, T Sørlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, A L Børresen-Dale, P O Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, August 2000.

[140] A Goldhirsch, E P Winer, A S Coates, R D Gelber, M Piccart-Gebhart, B Thürlimann, H-J Senn, and Panel members. Personalizing the treatment of women with early breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2013. *Ann. Oncol.*, 24(9):2206–2223, September 2013.

[141] Hee Kyung Kim, Kyung Hee Park, Youjin Kim, Song Ee Park, Han Sang Lee, Sung Won Lim, Jang Ho Cho, Ji-Yeon Kim, Jeong Eon Lee, Jin Seok Ahn, Young-Hyuck Im, Jong Han Yu, and Yeon Hee Park. Discordance of the PAM50 intrinsic subtypes compared with Immunohistochemistry-Based surrogate in breast cancer patients: Potential implication of genomic alterations of discordance. *Cancer Res. Treat.*, 51(2):737–747, April 2019.

[142] Joel S Parker, Michael Mullins, Maggie C U Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F Quackenbush, Inge J Stijleman, Juan Palazzo, J S Marron, Andrew B Nobel, Elaine Mardis, Torsten O Nielsen, Matthew J Ellis, Charles M Perou, and Philip S Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, 27 (8):1160–1167, March 2009.

[143] Praveen-Kumar Raj-Kumar, Jianfang Liu, Jeffrey A Hooke, Albert J Kovatich, Leonid Kvecher, Craig D Shriver, and Hai Hu. PCA-PAM50 improves consistency between breast cancer intrinsic and clinical subtyping reclassifying a subset of luminal a tumors as luminal B. *Sci. Rep.*, 9(1):7956, May 2019.

[144] Seokhyun Yoon, Hye Sung Won, Keunsoo Kang, Kexin Qiu, Woong June Park, and Yoon Ho Ko. Hormone Receptor-Status prediction in breast cancer using gene expression profiles and their macroscopic landscape. *Cancers*, 12(5), May 2020.

[145] S Cascinu, M Falconi, V Valentini, S Jelic, and ESMO Guidelines Working Group. Pancreatic cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.*, 21 Suppl 5:v55–8, May 2010.

[146] D J Brat, K D Lillemoe, C J Yeo, P B Warfield, and R H Hruban. Progression of

pancreatic intraductal neoplasias to infiltrating adenocarcinoma of the pancreas. *Am. J. Surg. Pathol.*, 22(2):163–169, February 1998.

[147] Ilse Rooman and Francisco X Real. Pancreatic ductal adenocarcinoma and acinar cells: a matter of differentiation and development? *Gut*, 61(3):449–458, March 2012.

[148] Peter Storz. Acinar cell plasticity and development of pancreatic ductal adenocarcinoma. *Nat. Rev. Gastroenterol. Hepatol.*, 14(5):296–304, May 2017.

[149] Andrea Costamagna, Dora Natalini, Maria Del Pilar Camacho Leal, Matilde Simoni, Luca Gozzelino, Paola Cappello, Francesco Novelli, Chiara Ambrogio, Paola Defilippi, Emilia Turco, Elisa Giovannetti, Emilio Hirsch, Sara Cabodi, and Miriam Martini. Docking protein p130cas regulates acinar to ductal metaplasia during pancreatic adenocarcinoma development and pancreatitis. *Gastroenterology*, December 2021.

[150] John P Morris Iv, David A Cano, Shigeki Sekine, Sam C Wang, and Matthias Hebrok. β-catenin blocks kras-dependent reprogramming of acini into pancreatic cancer precursor lesions in mice. *J. Clin. Invest.*, 120(2):508–520, 2010.

[151] Kirsten L Bryant, Joseph D Mancias, Alec C Kimmelman, and Channing J Der. KRAS: feeding pancreatic cancer proliferation. *Trends Biochem. Sci.*, 39(2):91–100, February 2014.

[152] Jorg Kleeff, Murray Korc, Minoti Apte, Carlo La Vecchia, Colin D Johnson, Andrew V Biankin, Rachel E Neale, Margaret Tempero, David A Tuveson, Ralph H Hruban, and John P Neoptolemos. Pancreatic cancer. *Nat Rev Dis Primers*, 2:16022, April 2016.

[153] M Schutte, R H Hruban, J Geradts, R Maynard, W Hilgers, S K Rabindran, C A Moskaluk, S A Hahn, I Schwarte-Waldhoff, W Schmiegel, S B Baylin, S E Kern, and J G Herman. Abrogation of the rb/p16 tumor-suppressive pathway in virtually all pancreatic carcinomas. *Cancer Res.*, 57(15):3126–3130, August 1997.

[154] Kyle M LaPak and Christin E Burd. The molecular balancing act of p16(INK4a) in cancer and aging. *Mol. Cancer Res.*, 12(2):167–183, February 2014.

[155] M S Redston, C Caldas, A B Seymour, R H Hruban, L da Costa, C J Yeo, and S E Kern. p53 mutations in pancreatic carcinoma and evidence of common involvement of homocopolymer tracts in DNA microdeletions. *Cancer Res.*, 54(11):3025–3033, June 1994.

[156] Sunil R Hingorani, Lifu Wang, Asha S Multani, Chelsea Combs, Therese B Deramaudt, Ralph H Hruban, Anil K Rustgi, Sandy Chang, and David A Tuveson. Trp53R172H and KrasG12D cooperate to promote chromosomal instability and widely metastatic pancreatic ductal adenocarcinoma in mice. *Cancer Cell*, 7(5):469–483, May 2005.

[157] Milind Javle, Yanan Li, Dongfeng Tan, Xiaoqun Dong, Ping Chang, Siddhartha Kar, and Donghui Li. Biomarkers of TGF-β signaling pathway and prognosis of pancreatic cancer. *PLoS One*, 9(1):e85942, January 2014.

[158] Agnieszka K Witkiewicz, Elizabeth A McMillan, Uthra Balaji, Guemhee Baek, Wan-Chi Lin, John Mansour, Mehri Mollaee, Kay-Uwe Wagner, Prasad Koduru, Adam Yopp, Michael A Choti, Charles J Yeo, Peter McCue, Michael A White, and Erik S Knudsen. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.*, 6:6744, April 2015.

[159] T Golan, Z S Kanji, R Epelbaum, N Devaud, E Dagan, S Holter, D Aderka, S Paluch-Shimon, B Kaufman, R Gershoni-Baruch, D Hedley, M J Moore, E Friedman, and S Gallinger. Overall survival and clinical characteristics of pancreatic cancer in BRCA mutation carriers. *Br. J. Cancer*, 111(6):1132–1138, September 2014.

[160] Nicholas J Roberts, Alexis L Norris, Gloria M Petersen, Melissa L Bondy, Randall Brand, Steven Gallinger, Robert C Kurtz, Sara H Olson, Anil K Rustgi, Ann G Schwartz, Elena Stoffel, Sapna Syngal, George Zogopoulos, Syed Z Ali, Jennifer Axilbund, Kari G Chaffee, Yun-Ching Chen, Michele L Cote, Erica J Childs, Christopher Douville, Fernando S Goes, Joseph M Herman, Christine Iacobuzio-Donahue, Melissa Kramer, Alvin Makohon-Moore, Richard W McCombie, K Wyatt McMahon, Noushin Niknafs, Jennifer Parla, Mehdi Pirooznia, James B Potash, Andrew D Rhim, Alyssa L Smith, Yuxuan Wang, Christopher L Wolfgang, Laura D Wood, Peter P Zandi, Michael Goggins, Rachel Karchin, James R Eshleman, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vo-

gelstein, Ralph H Hruban, and Alison P Klein. Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. *Cancer Discov.*, 6(2):166–175, February 2016.

[161] Ibrahim H Sahin, Christine A Iacobuzio-Donahue, and Eileen M O'Reilly. Molecular signature of pancreatic adenocarcinoma: an insight from genotype to phenotype and challenges for targeted therapy. *Expert Opin. Ther. Targets*, 20(3):341–359, 2016.

[162] Michelle Chan-Seng-Yue, Jaeseung C Kim, Gavin W Wilson, Karen Ng, Eugenia Flores Figueroa, Grainne M O'Kane, Ashton A Connor, Robert E Denroche, Robert C Grant, Jessica McLeod, Julie M Wilson, Gun Ho Jang, Amy Zhang, Anna Dodd, Sheng-Ben Liang, Ayelet Borgida, Dianne Chadwick, Sangeetha Kalimuthu, Ilinca Lungu, John M S Bartlett, Paul M Krzyzanowski, Vandana Sandhu, Hervé Tiriac, Fieke E M Froeling, Joanna M Karasinska, James T Topham, Daniel J Renouf, David F Schaeffer, Steven J M Jones, Marco A Marra, Janessa Laskin, Runjan Chetty, Lincoln D Stein, George Zogopoulos, Benjamin Haibe-Kains, Peter J Campbell, David A Tuveson, Jennifer J Knox, Sandra E Fischer, Steven Gallinger, and Faiyaz Notta. Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. *Nat. Genet.*, 52(2):231–240, February 2020.

[163] Sebastian Mueller, Thomas Engleitner, Roman Maresch, Magdalena Zukowska, Sebastian Lange, Thorsten Kaltenbacher, Björn Konukiewitz, Rupert Öllinger, Maximilian Zwiebel, Alex Strong, Hsi-Yu Yen, Ruby Banerjee, Sandra Louzada, Beiyuan Fu, Barbara Seidler, Juliana Götzfried, Kathleen Schuck, Zonera Hassan, Andreas Arbeiter, Nina Schönhuber, Sabine Klein, Christian Veltkamp, Mathias Friedrich, Lena Rad, Maxim Barenboim, Christoph Ziegenhain, Julia Hess, Oliver M Dovey, Stefan Eser, Swati Parekh, Fernando Constantino-Casas, Jorge de la Rosa, Marta I Sierra, Mario Fraga, Julia Mayerle, Günter Klöppel, Juan Cadiñanos, Pentao Liu, George Vassiliou, Wilko Weichert, Katja Steiger, Wolfgang Enard, Roland M Schmid, Fengtang Yang, Kristian Unger, Günter Schneider, Ignacio Varela, Allan Bradley, Dieter Saur, and Roland Rad. Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. *Nature*, 554(7690):62–68, February 2018.

[164] Peter Bailey, David K Chang, Katia Nones, Amber L Johns, Ann-Marie Patch, Marie-Claude Gingras, David K Miller, Angelika N Christ, Tim J C Bruxner, Michael C Quinn, Craig Nourse, L Charles Murtaugh, Ivon Harliwong, Senel Idrisoglu, Suzanne Manning, Ehsan Nourbakhsh, Shivangi Wani, Lynn Fink, Oliver Holmes, Venessa Chin, Matthew J Anderson, Stephen Kazakoff, Conrad Leonard, Felicity Newell, Nick Waddell, Scott Wood, Qinying Xu, Peter J Wilson, Nicole Cloonan, Karin S Kassahn, Darrin Taylor, Kelly Quek, Alan Robertson, Lorena Pantano, Laura Mincarelli, Luis N Sanchez, Lisa Evers, Jianmin Wu, Mark Pinese, Mark J Cowley, Marc D Jones, Emily K Colvin, Adnan M Nagrial, Emily S Humphrey, Lorraine A Chantrill, Amanda Mawson, Jeremy Humphris, Angela Chou, Marina Pajic, Christopher J Scarlett, Andreia V Pinho, Marc Giry-Laterriere, Ilse Rooman, Jaswinder S Samra, James G Kench, Jessica A Lovell, Neil D Merrett, Christopher W Toon, Krishna Epari, Nam Q Nguyen, Andrew Barbour, Nikolajs Zeps, Kim Moran-Jones, Nigel B Jamieson, Janet S Graham, Fraser Duthie, Karin Oien, Jane Hair, Robert Grützmann, Anirban Maitra, Christine A Iacobuzio-Donahue, Christopher L Wolfgang, Richard A Morgan, Rita T Lawlor, Vincenzo Corbo, Claudio Bassi, Borislav Rusev, Paola Capelli, Roberto Salvia, Giampaolo Tortora, Debabrata Mukhopadhyay, Gloria M Petersen, Australian Pancreatic Cancer Genome Initiative, Donna M Munzy, William E Fisher, Saadia A Karim, James R Eshleman, Ralph H Hruban, Christian Pilarsky, Jennifer P Morton, Owen J Sansom, Aldo Scarpa, Elizabeth A Musgrove, Ulla-Maja Hagbo Bailey, Oliver Hofmann, Robert L Sutherland, David A Wheeler, Anthony J Gill, Richard A Gibbs, John V Pearson, Nicola Waddell, Andrew V Biankin, and Sean M Grimmond. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52, March 2016.

[165] Noboru Ideno, Hiroshi Yamaguchi, Bidyut Ghosh, Sonal Gupta, Takashi Okumura, Dana J Steffen, Catherine G Fisher, Laura D Wood, Aatur D Singhi, Masafumi Nakamura, J Silvio Gutkind, and Anirban Maitra. $GNAS^{R201C}$ induces pancreatic cystic neoplasms in mice that express activated KRAS by inhibiting YAP1 signaling. *Gastroenterology*, 155(5):1593–1607.e12, November 2018.

[166] Nicola Waddell, Marina Pajic, Ann-Marie Patch, David K Chang, Karin S Kassahn,

Peter Bailey, Amber L Johns, David Miller, Katia Nones, Kelly Quek, Michael C J Quinn, Alan J Robertson, Muhammad Z H Fadlullah, Tim J C Bruxner, Angelika N Christ, Ivon Harliwong, Senel Idrisoglu, Suzanne Manning, Craig Nourse, Ehsan Nourbakhsh, Shivangi Wani, Peter J Wilson, Emma Markham, Nicole Cloonan, Matthew J Anderson, J Lynn Fink, Oliver Holmes, Stephen H Kazakoff, Conrad Leonard, Felicity Newell, Barsha Poudel, Sarah Song, Darrin Taylor, Nick Waddell, Scott Wood, Qinying Xu, Jianmin Wu, Mark Pinese, Mark J Cowley, Hong C Lee, Marc D Jones, Adnan M Nagrial, Jeremy Humphris, Lorraine A Chantrill, Venessa Chin, Angela M Steinmann, Amanda Mawson, Emily S Humphrey, Emily K Colvin, Angela Chou, Christopher J Scarlett, Andreia V Pinho, Marc Giry-Laterriere, Ilse Rooman, Jaswinder S Samra, James G Kench, Jessica A Pettitt, Neil D Merrett, Christopher Toon, Krishna Epari, Nam Q Nguyen, Andrew Barbour, Nikolajs Zeps, Nigel B Jamieson, Janet S Graham, Simone P Niclou, Rolf Bjerkvig, Robert Grützmann, Daniela Aust, Ralph H Hruban, Anirban Maitra, Christine A Iacobuzio-Donahue, Christopher L Wolfgang, Richard A Morgan, Rita T Lawlor, Vincenzo Corbo, Claudio Bassi, Massimo Falconi, Giuseppe Zamboni, Giampaolo Tortora, Margaret A Tempero, Australian Pancreatic Cancer Genome Initiative, Anthony J Gill, James R Eshleman, Christian Pilarsky, Aldo Scarpa, Elizabeth A Musgrove, John V Pearson, Andrew V Biankin, and Sean M Grimmond. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518 (7540):495–501, February 2015.

[167] Eric A Collisson, Peter Bailey, David K Chang, and Andrew V Biankin. Molecular subtypes of pancreatic cancer. *Nat. Rev. Gastroenterol. Hepatol.*, 16(4):207–220, April 2019.

[168] Chiara Falcomatà, Stefanie Bärthel, Sebastian A Widholz, Christian Schneeweis, Juan José Montero, Albulena Toska, Jonas Mir, Thorsten Kaltenbacher, Jeannine Heetmeyer, Jonathan J Swietlik, Jing-Yuan Cheng, Bianca Teodorescu, Oliver Reichert, Constantin Schmitt, Kathrin Grabichler, Andrea Coluccio, Fabio Boniolo, Christian Veltkamp, Magdalena Zukowska, Angelica Arenas Vargas, Woo Hyun Paik, Moritz Jesinghaus, Katja Steiger, Roman Maresch, Rupert Öllinger, Tim Ammon, Olga Baranov, Maria S Robles, Julia Rechenberger, Bernhard Kuster, Felix Meissner, Maximilian

Reichert, Michael Flossdorf, Roland Rad, Marc Schmidt-Supprian, Günter Schneider, and Dieter Saur. Selective multi-kinase inhibition sensitizes mesenchymal pancreatic cancer to immune checkpoint blockade by remodeling the tumor microenvironment. *Nat Cancer*, January 2022.

[169] Cristian Nogales, Zeinab M Mamdouh, Markus List, Christina Kiel, Ana I Casas, and Harald H H W Schmidt. Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends Pharmacol. Sci.*, 43(2):136–150, February 2022.

[170] Nicholas J Schork. Personalized medicine: Time for one-person trials. *Nature*, 520(7549): 609–611, April 2015.

[171] Graeme P Currie, Daniel K C Lee, and Brian J Lipworth. Long-acting beta2-agonists in asthma: not so SMART? *Drug Saf.*, 29(8):647–656, 2006.

[172] Jack Wilkinson, Kellyn F Arnold, Eleanor J Murray, Maarten van Smeden, Kareem Carr, Rachel Sippy, Marc de Kamps, Andrew Beam, Stefan Konigorski, Christoph Lippert, Mark S Gilthorpe, and Peter W G Tennant. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health*, 2(12):e677–e680, December 2020.

[173] Lori M Millner and Lindsay N Strotman. The future of precision medicine in oncology. *Clin. Lab. Med.*, 36(3):557–573, September 2016.

[174] Eoghan R Malone, Marc Oliva, Peter J B Sabatini, Tracy L Stockley, and Lillian L Siu. Molecular profiling for precision cancer therapies. *Genome Med.*, 12(1):8, January 2020.

[175] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, 47(D1): D941–D947, January 2019.

[176] AACR Project GENIE Consortium. AACR project GENIE: Powering precision medicine through an international consortium. *Cancer Discov.*, 7(8):818–831, August 2017.

[177] Brendan Reardon, Nathanael D Moore, Nicholas S Moore, Eric Kofman, Saud H Al-Dubayan, Alexander T M Cheung, Jake Conway, Haitham Elmarakeby, Alma Imamovic, Sophia C Kamran, Tanya Keenan, Daniel Keliher, David J Konieczkowski, David Liu, Kent W Mouw, Jihye Park, Natalie I Vokes, Felix Dietlein, and Eliezer M Van Allen. Integrating molecular profiles into clinical frameworks through the molecular oncology almanac to prospectively guide precision oncology. *Nat Cancer*, 2(10):1102–1112, October 2021.

[178] Davide Torti and Livio Trusolino. Oncogene addiction as a foundational rationale for targeted anti-cancer therapy: promises and perils. *EMBO Mol. Med.*, 3(11):623–636, November 2011.

[179] Sai-Hong Ignatius Ou, Cynthia Huang Bartlett, Mari Mino-Kenudson, Jean Cui, and A John Iafrate. Crizotinib for the treatment of ALK-rearranged non-small cell lung cancer: a success story to usher in the second decade of molecular targeted therapy in oncology. *Oncologist*, 17(11):1351–1375, September 2012.

[180] David L Deremer, Celalettin Ustun, and Kavita Natarajan. Nilotinib: a second-generation tyrosine kinase inhibitor for the treatment of chronic myelogenous leukemia. *Clin. Ther.*, 30(11):1956–1975, November 2008.

[181] Alexander M Menzies, Georgina V Long, and Rajmohan Murali. Dabrafenib and its potential for the treatment of metastatic melanoma. *Drug Des. Devel. Ther.*, 6:391–405, December 2012.

[182] Nancy Martinez-Montiel, Nora Hilda Rosas-Murrieta, Maricruz Anaya Ruiz, Eduardo Monjaraz-Guzman, and Rebeca Martinez-Contreras. Alternative splicing as a target for cancer treatment. *Int. J. Mol. Sci.*, 19(2), February 2018.

[183] Judith A Seidel, Atsushi Otsuka, and Kenji Kabashima. Anti-PD-1 and Anti-CTLA-4

therapies in cancer: Mechanisms of action, efficacy, and limitations. *Front. Oncol.*, 8:86, March 2018.

[184] Terry J Fry, Nirali N Shah, Rimas J Orentas, Maryalice Stetler-Stevenson, Constance M Yuan, Sneha Ramakrishna, Pamela Wolters, Staci Martin, Cindy Delbrook, Bonnie Yates, Haneen Shalabi, Thomas J Fountaine, Jack F Shern, Robbie G Majzner, David F Stroncek, Marianna Sabatino, Yang Feng, Dimiter S Dimitrov, Ling Zhang, Sang Nguyen, Haiying Qin, Boro Dropulic, Daniel W Lee, and Crystal L Mackall. CD22-targeted CAR T cells induce remission in B-ALL that is naive or resistant to CD19-targeted CAR immunotherapy. *Nat. Med.*, 24(1):20–28, January 2018.

[185] Alex D Waldman, Jill M Fritz, and Michael J Lenardo. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat. Rev. Immunol.*, 20(11):651–668, November 2020.

[186] Norbert Hilf, Sabrina Kuttruff-Coqui, Katrin Frenzel, Valesca Bukur, Stefan Stevanović, Cécile Gouttefangeas, Michael Platten, Ghazaleh Tabatabai, Valerie Dutoit, Sjoerd H van der Burg, Per Thor Straten, Francisco Martínez-Ricarte, Berta Ponsati, Hideho Okada, Ulrik Lassen, Arie Admon, Christian H Ottensmeier, Alexander Ulges, Sebastian Kreiter, Andreas von Deimling, Marco Skardelly, Denis Migliorini, Judith R Kroep, Manja Idorn, Jordi Rodon, Jordi Piró, Hans S Poulsen, Bracha Shraibman, Katy McCann, Regina Mendrzyk, Martin Löwer, Monika Stieglbauer, Cedrik M Britten, David Capper, Marij J P Welters, Juan Sahuquillo, Katharina Kiesel, Evelyna Derhovanessian, Elisa Rusch, Lukas Bunse, Colette Song, Sandra Heesch, Claudia Wagner, Alexandra Kemmer-Brück, Jörg Ludwig, John C Castle, Oliver Schoor, Arbel D Tadmor, Edward Green, Jens Fritsche, Miriam Meyer, Nina Pawlowski, Sonja Dorner, Franziska Hoffgaard, Bernhard Rössler, Dominik Maurer, Toni Weinschenk, Carsten Reinhardt, Christoph Huber, Hans-Georg Rammensee, Harpreet Singh-Jasuja, Ugur Sahin, Pierre-Yves Dietrich, and Wolfgang Wick. Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature*, 565(7738):240–245, January 2019.

[187] Derin B Keskin, Annabelle J Anandappa, Jing Sun, Itay Tirosh, Nathan D Mathew-

son, Shuqiang Li, Giacomo Oliveira, Anita Giobbie-Hurder, Kristen Felt, Evisa Gjini, Sachet A Shukla, Zhuting Hu, Letitia Li, Phuong M Le, Rosa L Allesøe, Alyssa R Richman, Monika S Kowalczyk, Sara Abdelrahman, Jack E Geduldig, Sarah Charbonneau, Kristine Pelton, J Bryan Iorgulescu, Liudmila Elagina, Wandi Zhang, Oriol Olive, Christine McCluskey, Lars R Olsen, Jonathan Stevens, William J Lane, Andres M Salazar, Heather Daley, Patrick Y Wen, E Antonio Chiocca, Maegan Harden, Niall J Lennon, Stacey Gabriel, Gad Getz, Eric S Lander, Aviv Regev, Jerome Ritz, Donna Neuberg, Scott J Rodig, Keith L Ligon, Mario L Suvà, Kai W Wucherpfennig, Nir Hacohen, Edward F Fritsch, Kenneth J Livak, Patrick A Ott, Catherine J Wu, and David A Reardon. Neoantigen vaccine generates intratumoral T cell responses in phase ib glioblastoma trial. *Nature*, 565(7738):234–239, January 2019.

[188] Patrick A Ott, Zhuting Hu, Derin B Keskin, Sachet A Shukla, Jing Sun, David J Bozym, Wandi Zhang, Adrienne Luoma, Anita Giobbie-Hurder, Lauren Peter, Christina Chen, Oriol Olive, Todd A Carter, Shuqiang Li, David J Lieb, Thomas Eisenhaure, Evisa Gjini, Jonathan Stevens, William J Lane, Indu Javeri, Kaliappanadar Nellaiappan, Andres M Salazar, Heather Daley, Michael Seaman, Elizabeth I Buchbinder, Charles H Yoon, Maegan Harden, Niall Lennon, Stacey Gabriel, Scott J Rodig, Dan H Barouch, Jon C Aster, Gad Getz, Kai Wucherpfennig, Donna Neuberg, Jerome Ritz, Eric S Lander, Edward F Fritsch, Nir Hacohen, and Catherine J Wu. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature*, 547(7662):217–221, July 2017.

[189] Patrick A Ott, Siwen Hu-Lieskovan, Bartosz Chmielowski, Ramaswamy Govindan, Aung Naing, Nina Bhardwaj, Kim Margolin, Mark M Awad, Matthew D Hellmann, Jessica J Lin, Terence Friedlander, Meghan E Bushway, Kristen N Balogh, Tracey E Sciuto, Victoria Kohler, Samantha J Turnbull, Rana Besada, Riley R Curran, Benjamin Trapp, Julian Scherer, Asaf Poran, Dewi Harjanto, Dominik Barthelme, Ying Sonia Ting, Jesse Z Dong, Yvonne Ware, Yuting Huang, Zhengping Huang, Amy Wanamaker, Lisa D Cleary, Melissa A Moles, Kelledy Manson, Joel Greshock, Zakaria S Khondker, Ed Fritsch, Michael S Rooney, Mark DeMario, Richard B Gaynor, and Lakshmi Srinivasan. A phase ib trial of personalized neoantigen therapy plus Anti-PD-1 in patients with advanced

melanoma, non-small cell lung cancer, or bladder cancer. *Cell*, 183(2):347–362.e24, October 2020.

[190] Ugur Sahin, Evelyna Derhovanessian, Matthias Miller, Björn-Philipp Kloke, Petra Simon, Martin Löwer, Valesca Bukur, Arbel D Tadmor, Ulrich Luxemburger, Barbara Schrörs, Tana Omokoko, Mathias Vormehr, Christian Albrecht, Anna Paruzynski, Andreas N Kuhn, Janina Buck, Sandra Heesch, Katharina H Schreeb, Felicitas Müller, Inga Ortseifer, Isabel Vogler, Eva Godehardt, Sebastian Attig, Richard Rae, Andrea Breitkreuz, Claudia Tolliver, Martin Suchan, Goran Martic, Alexander Hohberger, Patrick Sorn, Jan Diekmann, Janko Ciesla, Olga Waksmann, Alexandra-Kemmer Brück, Meike Witt, Martina Zillgen, Andree Rothermel, Barbara Kasemann, David Langer, Stefanie Bolte, Mustafa Diken, Sebastian Kreiter, Romina Nemecek, Christoffer Gebhardt, Stephan Grabbe, Christoph Höller, Jochen Utikal, Christoph Huber, Carmen Loquai, and Özlem Türeci. Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature*, 547(7662):222–226, July 2017.

[191] Elaine R Mardis. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.*, 6: 287–303, 2013.

[192] Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter, and Frank Stahl. Transcriptome analysis using next-generation sequencing. *Curr. Opin. Biotechnol.*, 24(1):22–30, February 2013.

[193] Klaas Mensaert, Simon Denil, Geert Trooskens, Wim Van Criekinge, Olivier Thas, and Tim De Meyer. Next-generation technologies and data analytical approaches for epigenomics. *Environ. Mol. Mutagen.*, 55(3):155–170, April 2014.

[194] Mia Yang Ang, Teck Yew Low, Pey Yee Lee, Wan Fahmi Wan Mohamad Nazarie, Victor Guryev, and Rahman Jamal. Proteogenomics: From next-generation sequencing (NGS) and mass spectrometry-based proteomics to precision medicine. *Clin. Chim. Acta*, 498: 38–46, November 2019.

[195] Valeria D'Argenio. The High-Throughput analyses era: Are we ready for the data struggle? *High Throughput*, 7(1), March 2018.

[196] Fabio Boniolo, Emilio Dorigatti, Alexander J Ohnmacht, Dieter Saur, Benjamin Schubert, and Michael P Menden. Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opin. Drug Discov.*, 16(9):991–1007, September 2021.

[197] Ali Torkamani, Kristian G Andersen, Steven R Steinhubl, and Eric J Topol. High-Definition medicine. *Cell*, 170(5):828–843, August 2017.

[198] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, February 2022.

[199] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878): 672–677, September 2021.

[200] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P Agapiou, Max Jaderberg, Alexander S Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level

in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, November 2019.

[201] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. AI in health and medicine. *Nat. Med.*, 28(1):31–38, January 2022.

[202] Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology - new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.*, 16(11):703–715, November 2019.

[203] Dejun Zhou, Fei Tian, Xiangdong Tian, Lin Sun, Xianghui Huang, Feng Zhao, Nan Zhou, Zuoyu Chen, Qiang Zhang, Meng Yang, Yichen Yang, Xuexi Guo, Zhibin Li, Jia Liu, Jiefu Wang, Junfeng Wang, Bangmao Wang, Guoliang Zhang, Baocun Sun, Wei Zhang, Dalu Kong, Kexin Chen, and Xiangchun Li. Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer. *Nat. Commun.*, 11(1):2961, June 2020.

[204] Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S Bitterman, Steven F Petit, Daphne A Haas-Kogan, Benjamin Kann, Hugo J W L Aerts, and Raymond H Mak. Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.*, 17(12):771–781, December 2020.

[205] Mihaela Porumb, Saverio Stranges, Antonio Pescapè, and Leandro Pecchia. Precision medicine and artificial intelligence: A pilot study on deep learning for hypoglycemic events detection based on ECG. *Sci. Rep.*, 10(1):170, January 2020.

[206] Jan Claassen, Kevin Doyle, Adu Matory, Caroline Couch, Kelly M Burger, Angela Velazquez, Joshua U Okonkwo, Jean-Rémi King, Soojin Park, Sachin Agarwal, David Roh, Murad Megjhani, Andrey Eliseyev, E Sander Connolly, and Benjamin Rohaut. Detection of brain activation in unresponsive patients with acute brain injury. *N. Engl. J. Med.*, 380(26):2497–2505, June 2019.

[207] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W R Nelson, Alex Bridgland, Hugo

Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, January 2020.

[208] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach, Global Edition*. Pearson Education Limited, July 2016.

[209] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. A brief survey of deep reinforcement learning. August 2017.

[210] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*, 3(11):e745–e750, November 2021.

[211] Jürgen Schmidhuber. Deep learning in neural networks: an overview. *Neural Netw.*, 61: 85–117, January 2015.

[212] Christine M Cutillo, Karlie R Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, Kenneth D Mandl, and MI in Healthcare Workshop Working Group. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med*, 3:47, March 2020.

[213] Sara Gerke, Boris Babic, Theodoros Evgeniou, and I Glenn Cohen. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit Med*, 3:53, April 2020.

[214] Jessica Morley, Caio Machado, Christopher Burr, Josh Cowls, Mariarosaria Taddeo, and Luciano Floridi. The debate on the ethics of AI in health care: A reconstruction and critical review. November 2019.

[215] W Nicholson Price, 2nd, Sara Gerke, and I Glenn Cohen. Potential liability for physicians using artificial intelligence. *JAMA*, 322(18):1765–1766, November 2019.

[216] Gillian K Hadfield. Explanation and justification: AI decision-making, law, and the rights of citizens —. `https://srinstitute.utoronto.ca/news/hadfield-justifiable-ai`, May 2021. Accessed: 2022-3-21.

[217] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, June 2020.

[218] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci. U. S. A.*, 117(23):12592–12594, June 2020.

[219] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, October 2019.

[220] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. *N. Engl. J. Med.*, 383(9):874–882, August 2020.

[221] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in health care. September 2020.

[222] Francisco Azuaje. Artificial intelligence for precision oncology: beyond patient stratification. *NPJ Precis Oncol*, 3:6, February 2019.

[223] Felipe De Sousa E Melo, Louis Vermeulen, Evelyn Fessler, and Jan Paul Medema. Cancer heterogeneity–a multifaceted view. *EMBO Rep.*, 14(8):686–695, August 2013.

[224] R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer*, 108(3):479–485, February 2013.

[225] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien de Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, Brian M Bot, Jeffrey S Morris, Iris M Simon, Sarah Gerster, Evelyn Fessler, Felipe

De Sousa E Melo, Edoardo Missiaglia, Hena Ramay, David Barras, Krisztian Homicsko, Dipen Maru, Ganiraju C Manyam, Bradley Broom, Valerie Boige, Beatriz Perez-Villamil, Ted Laderas, Ramon Salazar, Joe W Gray, Douglas Hanahan, Josep Tabernero, Rene Bernards, Stephen H Friend, Pierre Laurent-Puig, Jan Paul Medema, Anguraj Sadanandam, Lodewyk Wessels, Mauro Delorenzi, Scott Kopetz, Louis Vermeulen, and Sabine Tejpar. The consensus molecular subtypes of colorectal cancer. *Nat. Med.*, 21(11): 1350–1356, November 2015.

[226] Anne Biton, Isabelle Bernard-Pierrot, Yinjun Lou, Clémentine Krucker, Elodie Chapeaublanc, Carlota Rubio-Pérez, Nuria López-Bigas, Aurélie Kamoun, Yann Neuzillet, Pierre Gestraud, Luca Grieco, Sandra Rebouissou, Aurélien de Reyniès, Simone Benhamou, Thierry Lebret, Jennifer Southgate, Emmanuel Barillot, Yves Allory, Andrei Zinovyev, and François Radvanyi. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.*, 9(4):1235–1245, November 2014.

[227] Agata Szymiczek, Amna Lone, and Mohammad R Akbari. Molecular intrinsic versus clinical subtyping in breast cancer: A comprehensive review. *Clin. Genet.*, 99(5):613–637, May 2021.

[228] Richard A Moffitt, Raoud Marayati, Elizabeth L Flate, Keith E Volmar, S Gabriela Herrera Loeza, Katherine A Hoadley, Naim U Rashid, Lindsay A Williams, Samuel C Eaton, Alexander H Chung, Jadwiga K Smyla, Judy M Anderson, Hong Jin Kim, David J Bentrem, Mark S Talamonti, Christine A Iacobuzio-Donahue, Michael A Hollingsworth, and Jen Jen Yeh. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.*, 47(10):1168–1178, October 2015.

[229] Kezia Gaitskell, Jane Green, Kirstin Pirie, Isobel Barnes, Carol Hermon, Gillian K Reeves, Valerie Beral, and Million Women Study Collaborators. Histological subtypes of ovarian cancer associated with parity and breastfeeding in the prospective million women study. *Int. J. Cancer*, 142(2):281–289, January 2018.

[230] Jaafar Makki. Diversity of breast carcinoma: Histological subtypes and clinical relevance. *Clin. Med. Insights Pathol.*, 8:23–31, December 2015.

[231] Shihua Zhang. Computational methods for subtyping of tumors and their applications for deciphering tumor heterogeneity. *Methods Mol. Biol.*, 1878:193–207, 2019.

[232] Lan Zhao, Victor H F Lee, Michael K Ng, Hong Yan, and Maarten F Bijlsma. Molecular subtyping of cancer: current status and moving toward clinical applications. *Brief. Bioinform.*, 20(2):572–584, March 2019.

[233] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, 19(1A):A68–77, 2015.

[234] International Cancer Genome Consortium, Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, Alan Guttmacher, Mark Guyer, Fiona M Hemsley, Jennifer L Jennings, David Kerr, Peter Klatt, Patrik Kolar, Jun Kusada, David P Lane, Frank Laplace, Lu Youyong, Gerd Nettekoven, Brad Ozenberger, Jane Peterson, T S Rao, Jacques Remacle, Alan J Schafer, Tatsuhiro Shibata, Michael R Stratton, Joseph G Vockley, Koichi Watanabe, Huanming Yang, Matthew M F Yuen, Bartha M Knoppers, Martin Bobrow, Anne Cambon-Thomsen, Lynn G Dressler, Stephanie O M Dyke, Yann Joly, Kazuto Kato, Karen L Kennedy, Pilar Nicolás, Michael J Parker, Emmanuelle Rial-Sebbag, Carlos M Romeo-Casabona, Kenna M Shaw, Susan Wallace, Georgia L Wiesner, Nikolajs Zeps, Peter Lichter, Andrew V Biankin, Christian Chabannon, Lynda Chin, Bruno Clément, Enrique de Alava, Françoise Degos, Martin L Ferguson, Peter Geary, D Neil Hayes, Thomas J Hudson, Amber L Johns, Arek Kasprzyk, Hidewaki Nakagawa, Robert Penny, Miguel A Piris, Rajiv Sarin, Aldo Scarpa, Tatsuhiro Shibata, Marc van de Vijver, P Andrew Futreal, Hiroyuki Aburatani, Mónica Bayés, David D L Botwell, Peter J Campbell, Xavier Estivill, Daniela S Gerhard, Sean M Grimmond, Ivo Gut, Martin Hirst, Carlos López-Otín, Partha Majumder, Marco Marra, John D McPherson, Hidewaki Nakagawa, Zemin Ning, Xose S Puente, Yijun Ruan, Tatsuhiro

Shibata, Michael R Stratton, Hendrik G Stunnenberg, Harold Swerdlow, Victor E Velculescu, Richard K Wilson, Hong H Xue, Liu Yang, Paul T Spellman, Gary D Bader, Paul C Boutros, Peter J Campbell, Paul Flicek, Gad Getz, Roderic Guigó, Guangwu Guo, David Haussler, Simon Heath, Tim J Hubbard, Tao Jiang, Steven M Jones, Qibin Li, Nuria López-Bigas, Ruibang Luo, Lakshmi Muthuswamy, B F Francis Ouellette, John V Pearson, Xose S Puente, Victor Quesada, Benjamin J Raphael, Chris Sander, Tatsuhiro Shibata, Terence P Speed, Lincoln D Stein, Joshua M Stuart, Jon W Teague, Yasushi Totoki, Tatsuhiko Tsunoda, Alfonso Valencia, David A Wheeler, Honglong Wu, Shancen Zhao, Guangyu Zhou, Lincoln D Stein, Roderic Guigó, Tim J Hubbard, Yann Joly, Steven M Jones, Arek Kasprzyk, Mark Lathrop, Nuria López-Bigas, B F Francis Ouellette, Paul T Spellman, Jon W Teague, Gilles Thomas, Alfonso Valencia, Teruhiko Yoshida, Karen L Kennedy, Myles Axton, Stephanie O M Dyke, P Andrew Futreal, Daniela S Gerhard, Chris Gunter, Mark Guyer, Thomas J Hudson, John D McPherson, Linda J Miller, Brad Ozenberger, Kenna M Shaw, Arek Kasprzyk, Lincoln D Stein, Junjun Zhang, Syed A Haider, Jianxin Wang, Christina K Yung, Anthony Cros, Yong Liang, Saravanamuttu Gnaneshan, Jonathan Guberman, Jack Hsu, Martin Bobrow, Don R C Chalmers, Karl W Hasel, Yann Joly, Terry S H Kaan, Karen L Kennedy, Bartha M Knoppers, William W Lowrance, Tohru Masui, Pilar Nicolás, Emmanuelle Rial-Sebbag, Laura Lyman Rodriguez, Catherine Vergely, Teruhiko Yoshida, Sean M Grimmond, Andrew V Biankin, David D L Bowtell, Nicole Cloonan, Anna deFazio, James R Eshleman, Dariush Etemadmoghadam, Brooke B Gardiner, James G Kench, Aldo Scarpa, Robert L Sutherland, Margaret A Tempero, Nicola J Waddell, Peter J Wilson, John D McPherson, Steve Gallinger, Ming-Sound Tsao, Patricia A Shaw, Gloria M Petersen, Debabrata Mukhopadhyay, Lynda Chin, Ronald A DePinho, Sarah Thayer, Lakshmi Muthuswamy, Kamran Shazand, Timothy Beck, Michelle Sam, Lee Timms, Vanessa Ballin, Youyong Lu, Jiafu Ji, Xiuqing Zhang, Feng Chen, Xueda Hu, Guangyu Zhou, Qi Yang, Geng Tian, Lianhai Zhang, Xiaofang Xing, Xianghong Li, Zhenggang Zhu, Yingyan Yu, Jun Yu, Huanming Yang, Mark Lathrop, Jörg Tost, Paul Brennan, Ivana Holcatova, David Zaridze, Alvis Brazma, Lars Egevard, Egor Prokhortchouk, Rosamonde Elizabeth Banks, Mathias Uhlén, Anne Cambon-Thomsen, Juris Viksna, Fredrik

Ponten, Konstantin Skryabin, Michael R Stratton, P Andrew Futreal, Ewan Birney, Ake Borg, Anne-Lise Børresen-Dale, Carlos Caldas, John A Foekens, Sancha Martin, Jorge S Reis-Filho, Andrea L Richardson, Christos Sotiriou, Hendrik G Stunnenberg, Giles Thoms, Marc van de Vijver, Laura van't Veer, Fabien Calvo, Daniel Birnbaum, Hélène Blanche, Pascal Boucher, Sandrine Boyault, Christian Chabannon, Ivo Gut, Jocelyne D Masson-Jacquemier, Mark Lathrop, Iris Pauporté, Xavier Pivot, Anne Vincent-Salomon, Eric Tabone, Charles Theillet, Gilles Thomas, Jörg Tost, Isabelle Treilleux, Fabien Calvo, Paulette Bioulac-Sage, Bruno Clément, Thomas Decaens, Françoise Degos, Dominique Franco, Ivo Gut, Marta Gut, Simon Heath, Mark Lathrop, Didier Samuel, Gilles Thomas, Jessica Zucman-Rossi, Peter Lichter, Roland Eils, Benedikt Brors, Jan O Korbel, Andrey Korshunov, Pablo Landgraf, Hans Lehrach, Stefan Pfister, Bernhard Radlwimmer, Guido Reifenberger, Michael D Taylor, Christof von Kalle, Partha P Majumder, Rajiv Sarin, T S Rao, M K Bhan, Aldo Scarpa, Paolo Pederzoli, Rita A Lawlor, Massimo Delledonne, Alberto Bardelli, Andrew V Biankin, Sean M Grimmond, Thomas Gress, David Klimstra, Giuseppe Zamboni, Tatsuhiro Shibata, Yusuke Nakamura, Hidewaki Nakagawa, Jun Kusada, Tatsuhiko Tsunoda, Satoru Miyano, Hiroyuki Aburatani, Kazuto Kato, Akihiro Fujimoto, Teruhiko Yoshida, Elias Campo, Carlos López-Otín, Xavier Estivill, Roderic Guigó, Silvia de Sanjosé, Miguel A Piris, Emili Montserrat, Marcos González-Díaz, Xose S Puente, Pedro Jares, Alfonso Valencia, Heinz Himmelbauer, Victor Quesada, Silvia Bea, Michael R Stratton, P Andrew Futreal, Peter J Campbell, Anne Vincent-Salomon, Andrea L Richardson, Jorge S Reis-Filho, Marc van de Vijver, Gilles Thomas, Jocelyne D Masson-Jacquemier, Samuel Aparicio, Ake Borg, Anne-Lise Børresen-Dale, Carlos Caldas, John A Foekens, Hendrik G Stunnenberg, Laura van't Veer, Douglas F Easton, Paul T Spellman, Sancha Martin, Anna D Barker, Lynda Chin, Francis S Collins, Carolyn C Compton, Martin L Ferguson, Daniela S Gerhard, Gad Getz, Chris Gunter, Alan Guttmacher, Mark Guyer, D Neil Hayes, Eric S Lander, Brad Ozenberger, Robert Penny, Jane Peterson, Chris Sander, Kenna M Shaw, Terence P Speed, Paul T Spellman, Joseph G Vockley, David A Wheeler, Richard K Wilson, Thomas J Hudson, Lynda Chin, Bartha M Knoppers, Eric S Lander, Peter Lichter, Lincoln D Stein, Michael R Stratton, Warwick Anderson, Anna D Barker, Cindy Bell, Martin Bobrow, Wylie Burke,

Francis S Collins, Carolyn C Compton, Ronald A DePinho, Douglas F Easton, P Andrew Futreal, Daniela S Gerhard, Anthony R Green, Mark Guyer, Stanley R Hamilton, Tim J Hubbard, Olli P Kallioniemi, Karen L Kennedy, Timothy J Ley, Edison T Liu, Youyong Lu, Partha Majumder, Marco Marra, Brad Ozenberger, Jane Peterson, Alan J Schafer, Paul T Spellman, Hendrik G Stunnenberg, Brandon J Wainwright, Richard K Wilson, and Huanming Yang. International network of cancer genome projects. *Nature*, 464 (7291):993–998, April 2010.

[235] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C Boutros, David D Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D Morris, PCAWG Evolution & Heterogeneity Working Group, Paul T Spellman, David C Wedge, Peter Van Loo, and PCAWG Consortium. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, February 2020.

[236] Zhenxing Wang and Yadong Wang. Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders. *BMC Bioinformatics*, 20(Suppl 18):568, November 2019.

[237] Kevin M Boehm, Pegah Khosravi, Rami Vanguri, Jianjiong Gao, and Sohrab P Shah. Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer*, 22(2):114–126, February 2022.

[238] Li Zhang, Chenkai Lv, Yaqiong Jin, Ganqi Cheng, Yibao Fu, Dongsheng Yuan, Yiran Tao, Yongli Guo, Xin Ni, and Tieliu Shi. Deep Learning-Based Multi-Omics data integration reveals two prognostic subtypes in High-Risk neuroblastoma. *Front. Genet.*, 9:477, October 2018.

[239] Tzong-Yi Lee, Kai-Yao Huang, Cheng-Hsiang Chuang, Cheng-Yang Lee, and Tzu-Hao Chang. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. *Comput. Biol. Chem.*, 87:107277, May 2020.

[240] Runpu Chen, Le Yang, Steve Goodison, and Yijun Sun. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics*, 36(5): 1476–1483, March 2020.

[241] P Rous. A SARCOMA OF THE FOWL TRANSMISSIBLE BY AN AGENT SEPARABLE FROM THE TUMOR CELLS. *J. Exp. Med.*, 13(4):397–411, April 1911.

[242] Vincent T DeVita, Jr and Edward Chu. A history of cancer chemotherapy. *Cancer Res.*, 68(21):8643–8653, November 2008.

[243] Harold J Burstein, Pamela B Mangu, Mark R Somerfield, Deborah Schrag, David Samson, Lawrence Holt, Debra Zelman, Jaffer A Ajani, and American Society of Clinical Oncology. American society of clinical oncology clinical practice guideline update on the use of chemotherapy sensitivity and resistance assays. *J. Clin. Oncol.*, 29(24): 3328–3330, August 2011.

[244] Deborah Schrag, Harinder S Garewal, Harold J Burstein, David J Samson, Daniel D Von Hoff, Mark R Somerfield, and ASCO Working Group on Chemotherapy Sensitivity and Resistance Assays. American society of clinical oncology technology assessment: chemotherapy sensitivity and resistance assays. *J. Clin. Oncol.*, 22(17):3631–3638, September 2004.

[245] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and Shanrong Zhao. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.*, 18(6):463–477, June 2019.

[246] Robert H Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, 6(10):813–823, October 2006.

[247] Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S Gray, Daniel A Haber, Michael R Stratton, Cyril H Benes, Lodewyk F A Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J Garnett. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3): 740–754, July 2016.

[248] Mahmoud Ghandi, Franklin W Huang, Judit Jané-Valbuena, Gregory V Kryukov, Christopher C Lo, E Robert McDonald, 3rd, Jordi Barretina, Ellen T Gelfand, Craig M Bielski, Haoxin Li, Kevin Hu, Alexander Y Andreev-Drakhlin, Jaegil Kim, Julian M Hess, Brian J Haas, François Aguet, Barbara A Weir, Michael V Rothberg, Brenton R Paolella, Michael S Lawrence, Rehan Akbani, Yiling Lu, Hong L Tiv, Prafulla C Gokhale, Antoine de Weck, Ali Amin Mansour, Coyin Oh, Juliann Shih, Kevin Hadi, Yanay Rosen, Jonathan Bistline, Kavitha Venkatesan, Anupama Reddy, Dmitriy Sonkin, Manway Liu, Joseph Lehar, Joshua M Korn, Dale A Porter, Michael D Jones, Javad Golji, Giordano Caponigro, Jordan E Taylor, Caitlin M Dunning, Amanda L Creech, Allison C Warren, James M McFarland, Mahdi Zamanighomi, Audrey Kauffmann, Nicolas Stransky, Marcin Imielinski, Yosef E Maruvka, Andrew D Cherniack, Aviad Tsherniak, Francisca Vazquez, Jacob D Jaffe, Andrew A Lane, David M Weinstock, Cory M Johannessen, Michael P Morrissey, Frank Stegmeier, Robert Schlegel, William C Hahn, Gad Getz, Gordon B Mills, Jesse S Boehm, Todd R Golub, Levi A Garraway, and William R Sellers. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757): 503–508, May 2019.

[249] Lisa Liu, Lei Yu, Zhichao Li, Wujiao Li, and Weiren Huang. Patient-derived organoid (PDO) platforms to facilitate clinical decision making. *J. Transl. Med.*, 19(1):40, January 2021.

[250] Jarno Drost and Hans Clevers. Organoids in cancer research. *Nat. Rev. Cancer*, 18(7): 407–418, July 2018.

[251] Adam A Friedman, Anthony Letai, David E Fisher, and Keith T Flaherty. Precision medicine for cancer with next-generation functional diagnostics. *Nat. Rev. Cancer*, 15 (12):747–756, December 2015.

[252] Anthony Letai, Patrick Bhola, and Alana L Welm. Functional precision oncology: Testing tumors with drugs to identify vulnerabilities and novel combinations. *Cancer Cell*, 40(1):26–35, January 2022.

[253] Anna Herland, Ben M Maoz, Debarun Das, Mahadevabharath R Somayaji, Rachelle Prantil-Baun, Richard Novak, Michael Cronce, Tessa Huffstater, Sauveur S F Jeanty, Miles Ingram, Angeliki Chalkiadaki, David Benson Chou, Susan Marquez, Aaron Delahanty, Sasan Jalili-Firoozinezhad, Yuka Milton, Alexandra Sontheimer-Phelps, Ben Swenor, Oren Levy, Kevin K Parker, Andrzej Przekwas, and Donald E Ingber. Quantitative prediction of human pharmacokinetic responses to drugs via fluidically coupled vascularized organ chips. *Nat Biomed Eng*, 4(4):421–436, April 2020.

[254] Alexandra Sontheimer-Phelps, Bryan A Hassell, and Donald E Ingber. Modelling cancer in microfluidic human organs-on-chips. *Nat. Rev. Cancer*, 19(2):65–81, February 2019.

[255] Francisco Azuaje. Computational models for predicting drug responses in cancer research. *Brief. Bioinform.*, 18(5):820–829, September 2017.

[256] T Turki, Z Wei, and J T L Wang. Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, 5:7381–7393, 2017.

[257] Fatemeh Ahmadi Moughari and Changiz Eslahchi. ADRML: anticancer drug response prediction using manifold learning. *Sci. Rep.*, 10(1):14245, August 2020.

[258] Yoosup Chang, Hyejin Park, Hyun-Jin Yang, Seungju Lee, Kwee-Yum Lee, Tae Soon Kim, Jongsun Jung, and Jae-Min Shin. Cancer drug response profile scan (CDRscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.*, 8(1):8857, June 2018.

[259] Ladislav Rampášek, Daniel Hidru, Petr Smirnov, Benjamin Haibe-Kains, and Anna Goldenberg. Dr.VAE: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*, 35(19):3743–3751, October 2019.

[260] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, David L Lahr, Jodi E Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M Enache, Federica Piccioni, Sarah A Johnson, Nicholas J Lyons, Alice H Berger, Alykhan F Shamji, Angela N Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S Gray, Paul A Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J Haggarty, Lucienne V Ronco, Jesse S Boehm, Stuart L Schreiber, John G Doench, Joshua A Bittker, David E Root, Bang Wong, and Todd R Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171 (6):1437–1452.e17, November 2017.

[261] Wei Zhao, Jun Li, Mei-Ju M Chen, Yikai Luo, Zhenlin Ju, Nicole K Nesser, Katie Johnson-Camacho, Christopher T Boniface, Yancey Lawrence, Nupur T Pande, Michael A Davies, Meenhard Herlyn, Taru Muranen, Ioannis K Zervantonakis, Erika von Euw, Andre Schultz, Shwetha V Kumar, Anil Korkut, Paul T Spellman, Rehan Akbani, Dennis J Slamon, Joe W Gray, Joan S Brugge, Yiling Lu, Gordon B Mills, and Han Liang. Large-Scale characterization of drug responses of clinically relevant proteins in cancer cell lines. *Cancer Cell*, October 2020.

[262] Sepideh Sadegh, Julian Matschinske, David B Blumenthal, Gihanna Galindez, Tim Kacprowski, Markus List, Reza Nasirigerdeh, Mhaned Oubounyt, Andreas Pichlmair, Tim Daniel Rose, Marisol Salgado-Albarrán, Julian Späth, Alexey Stukalov, Nina K Wenke, Kevin Yuan, Josch K Pauling, and Jan Baumbach. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat. Commun.*, 11(1):3518, July 2020.

[263] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(Database issue):D1100–7, January 2012.

[264] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H Bryant. PubChem substance and compound databases. *Nucleic Acids Res.*, 44(D1):D1202–13, January 2016.

[265] Azam Peyvandipour, Nafiseh Saberian, Adib Shafi, Michele Donato, and Sorin Draghici. A novel computational approach for drug repurposing using systems biology. *Bioinformatics*, 34(16):2817–2825, August 2018.

[266] Bin Chen, Lana Garmire, Diego F Calvisi, Mei-Sze Chua, Robin K Kelley, and Xin Chen. Harnessing big 'omics' data and AI for drug discovery in hepatocellular carcinoma. *Nat. Rev. Gastroenterol. Hepatol.*, 17(4):238–251, April 2020.

[267] Neil Vasan, José Baselga, and David M Hyman. A view on drug resistance in cancer. *Nature*, 575(7782):299–309, November 2019.

[268] Jacinta Simasi, Andreas Schubert, Christopher Oelkrug, Adrian Gillissen, and Karen Nieber. Primary and secondary resistance to tyrosine kinase inhibitors in lung cancer. *Anticancer Res.*, 34(6):2841–2850, June 2014.

[269] J H Goldie and A J Coldman. The genetic origin of drug resistance in neoplasms: implications for systemic therapy. *Cancer Res.*, 44(9):3643–3653, September 1984.

[270] Behzad Mansoori, Ali Mohammadi, Sadaf Davudian, Solmaz Shirjang, and Behzad Baradaran. The different mechanisms of cancer drug resistance: A brief review. *Adv Pharm Bull*, 7(3):339–348, September 2017.

[271] Padmanee Sharma, Siwen Hu-Lieskovan, Jennifer A Wargo, and Antoni Ribas. Primary, adaptive, and acquired resistance to cancer immunotherapy. *Cell*, 168(4):707–723, February 2017.

[272] Michael P Menden, Dennis Wang, Mike J Mason, Bence Szalai, Krishna C Bulusu, Yuanfang Guan, Thomas Yu, Jaewoo Kang, Minji Jeon, Russ Wolfinger, Tin Nguyen, Mikhail Zaslavskiy, AstraZeneca-Sanger Drug Combination DREAM Consortium, In Sock Jang, Zara Ghazoui, Mehmet Eren Ahsen, Robert Vogel, Elias Chaibub Neto, Thea Norman, Eric K Y Tang, Mathew J Garnett, Giovanni Y Di Veroli, Stephen Fawell, Gustavo Stolovitzky, Justin Guinney, Jonathan R Dry, and Julio Saez-Rodriguez. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat. Commun.*, 10(1):2674, June 2019.

[273] Anirudh Prahallad, Chong Sun, Sidong Huang, Federica Di Nicolantonio, Ramon Salazar, Davide Zecchin, Roderick L Beijersbergen, Alberto Bardelli, and René Bernards. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. *Nature*, 483(7387):100–103, January 2012.

[274] Ryan B Corcoran, Hiromichi Ebi, Alexa B Turke, Erin M Coffee, Michiya Nishino, Alexandria P Cogdill, Ronald D Brown, Patricia Della Pelle, Dora Dias-Santagata, Kenneth E Hung, Keith T Flaherty, Adriano Piris, Jennifer A Wargo, Jeffrey Settleman, Mari Mino-Kenudson, and Jeffrey A Engelman. EGFR-mediated re-activation of MAPK signaling contributes to insensitivity of BRAF mutant colorectal cancers to RAF inhibition with vemurafenib. *Cancer Discov.*, 2(3):227–235, March 2012.

[275] Xuan Wang, Haiyun Zhang, and Xiaozhuo Chen. Drug resistance and combating drug resistance in cancer. *Cancer Drug Resist*, 2:141–160, 2019.

[276] Hui Liu, Wenhao Zhang, Bo Zou, Jinxian Wang, Yuanyuan Deng, and Lei Deng. DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res.*, 48(D1):D871–D881, January 2020.

[277] Lianlian Wu, Yuqi Wen, Dongjin Leng, Qinglong Zhang, Chong Dai, Zhongming Wang, Ziqi Liu, Bowei Yan, Yixin Zhang, Jing Wang, Song He, and Xiaochen Bo. Machine learning methods, databases and tools for drug combination prediction. *Brief. Bioinform.*, 23(1), January 2022.

[278] Lei Huang, Fuhai Li, Jianting Sheng, Xiaofeng Xia, Jinwen Ma, Ming Zhan, and Stephen T C Wong. DrugComboRanker: drug combination discovery based on target network analysis. *Bioinformatics*, 30(12):i228–36, June 2014.

[279] Kristina Preuer, Richard P I Lewis, Sepp Hochreiter, Andreas Bender, Krishna C Bulusu, and Günter Klambauer. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 34(9):1538–1546, May 2018.

[280] Fangfang Xia, Maulik Shukla, Thomas Brettin, Cristina Garcia-Cardona, Judith Cohn, Jonathan E Allen, Sergei Maslov, Susan L Holbeck, James H Doroshow, Yvonne A Evrard, Eric A Stahlberg, and Rick L Stevens. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics*, 19(Suppl 18):486, December 2018.

[281] Saba Ahmadi, Pattara Sukprasert, Rahulsimham Vegesna, Sanju Sinha, Fiorella Schischlik, Natalie Artzi, Samir Khuller, Alejandro A Schäffer, and Eytan Ruppin. The landscape of receptor-mediated precision cancer combination therapy via a single-cell perspective. *Nat. Commun.*, 13(1):1613, March 2022.

[282] Fabio Boniolo, Markus Hoffmann, Norman Roggendorf, Bahar Tercan, Jan Baumbach, Mauro Castro, A Gordon Robertson, Dieter Saur, and Markus List. spongeffects: ceRNA modules offer patient-specific insights into the miRNA regulatory landscape. March 2022.

[283] Hongyu Liu, Cheng Lei, Qin He, Zou Pan, Desheng Xiao, and Yongguang Tao. Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. *Mol. Cancer*, 17(1):64, February 2018.

[284] Yong Peng and Carlo M Croce. The role of MicroRNAs in human cancer. *Signal Transduct Target Ther*, 1:15004, January 2016.

[285] Peng Wang, Qiuyan Guo, Yue Qi, Yangyang Hao, Yue Gao, Hui Zhi, Yuanfu Zhang, Yue Sun, Yakun Zhang, Mengyu Xin, Yunpeng Zhang, Shangwei Ning, and Xia Li. LncACTdb 3.0: an updated database of experimentally supported ceRNA interactions

and personalized networks contributing to precision medicine. *Nucleic Acids Res.*, 50 (D1):D183–D189, January 2022.

[286] Sarvenaz Choobdar, Mehmet E Ahsen, Jake Crawford, Mattia Tomasoni, Tao Fang, David Lamparter, Junyuan Lin, Benjamin Hescott, Xiaozhe Hu, Johnathan Mercer, Ted Natoli, Rajiv Narayan, DREAM Module Identification Challenge Consortium, Aravind Subramanian, Jitao D Zhang, Gustavo Stolovitzky, Zoltán Kutalik, Kasper Lage, Donna K Slonim, Julio Saez-Rodriguez, Lenore J Cowen, Sven Bergmann, and Daniel Marbach. Assessment of network module identification across complex diseases. *Nat. Methods*, 16(9):843–852, September 2019.

[287] Junpeng Zhang, Lin Liu, Taosheng Xu, Wu Zhang, Jiuyong Li, Nini Rao, and Thuc Duy Le. Time to infer miRNA sponge modules. *Wiley Interdiscip. Rev. RNA*, page e1686, August 2021.

[288] Olga Lazareva, Jan Baumbach, Markus List, and David B Blumenthal. On the limits of active module identification. *Brief. Bioinform.*, 22(5), September 2021.

[289] M Hoffmann, E Pachl, M Hartung, V Stiegler, and others. SPONGEdb: a pan-cancer resource for competing endogenous RNA interactions. *Narodonaselenie*, 2021.

[290] Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, Jingchun Zhu, and David Haussler. Visualizing and interpreting cancer genomics data via the xena platform. *Nat. Biotechnol.*, 38(6):675–678, June 2020.

[291] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, April 2012.

[292] Francisco Aparecido Rodrigues. Network centrality: An introduction, 2019.

[293] Fang Zheng, Le Wei, Liang Zhao, and Fuchuan Ni. Pathway network analysis of complex diseases based on multiple biological networks. *Biomed Res. Int.*, 2018:5670210, July 2018.

[294] Linton C Freeman. Centrality in social networks conceptual clarification. *Soc. Networks*, 1(3):215–239, January 1978.

[295] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genet.*, 2(6):e88, June 2006.

[296] Xin Lu, Vipul V Jain, Patricia W Finn, and David L Perkins. Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol. Syst. Biol.*, 3:98, April 2007.

[297] Xiaowei Zhu, Mark Gerstein, and Michael Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev.*, 21(9):1010–1024, May 2007.

[298] A Barrat, M Barthélemy, R Pastor-Satorras, and A Vespignani. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.*, 101(11):3747–3752, March 2004.

[299] Ulrik Brandes. A faster algorithm for betweenness centrality. *J. Math. Sociol.*, 25(2): 163–177, June 2001.

[300] M E Newman. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 64(1 Pt 2):016132, July 2001.

[301] Tore Opsahl, Filip Agneessens, and John Skvoretz. Node centrality in weighted networks: Generalizing degree and shortest paths. *Soc. Networks*, 32(3):245–251, July 2010.

[302] Tore Opsahl. *Structure and evolution of weighted networks*. PhD thesis, Queen Mary, University of London, 2009.

[303] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14:7, January 2013.

[304] Livnat Jerby-Arnon, Parin Shah, Michael S Cuoco, Christopher Rodman, Mei-Ju Su, Johannes C Melms, Rachel Leeson, Abhay Kanodia, Shaolin Mei, Jia-Ren Lin, Shu Wang, Bokang Rabasha, David Liu, Gao Zhang, Claire Margolais, Orr Ashenberg, Patrick A Ott, Elizabeth I Buchbinder, Rizwan Haq, F Stephen Hodi, Genevieve M Boland, Ryan J Sullivan, Dennie T Frederick, Benchun Miao, Tabea Moll, Keith T Flaherty, Meenhard Herlyn, Russell W Jenkins, Rohit Thummalapalli, Monika S Kowalczyk, Israel Cañadas, Bastian Schilling, Adam N R Cartwright, Adrienne M Luoma, Shruti Malu, Patrick Hwu, Chantale Bernatchez, Marie-Andrée Forget, David A Barbie, Alex K Shalek, Itay Tirosh, Peter K Sorger, Kai Wucherpfennig, Eliezer M Van Allen, Dirk Schadendorf, Bruce E Johnson, Asaf Rotem, Orit Rozenblatt-Rosen, Levi A Garraway, Charles H Yoon, Benjamin Izar, and Aviv Regev. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*, 175(4):984–997.e24, November 2018.

[305] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[306] Max Kuhn. Building predictive models inRUsing thecaretpackage. *J. Stat. Softw.*, 28(5), 2008.

[307] Andy Liaw, Matthew Wiener, and Others. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[308] Laoighse Mulrane, Sharon F McGee, William M Gallagher, and Darran P O'Connor. miRNA dysregulation in breast cancer. *Cancer Res.*, 73(22):6554–6562, November 2013.

[309] Rimi Hamam, Dana Hamam, Khalid A Alsaleh, Moustapha Kassem, Waleed Zaher, Musaad Alfayez, Abdullah Aldahmash, and Nehad M Alajez. Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death Dis.*, 8(9):e3045, September 2017.

[310] Steffen Durinck, Paul T Spellman, Ewan Birney, and Wolfgang Huber. Mapping

identifiers for the integration of genomic datasets with the R/Bioconductor package biomart. *Nat. Protoc.*, 4(8):1184–1191, July 2009.

[311] Xiaoping Su, Gabriel G Malouf, Yunxin Chen, Jianping Zhang, Hui Yao, Vicente Valero, John N Weinstein, Jean-Philippe Spano, Funda Meric-Bernstam, David Khayat, and Francisco J Esteva. Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. *Oncotarget*, 5(20):9864–9876, October 2014.

[312] Tatiana Benaglia, Didier Chauveau, David R Hunter, and Derek S Young. mixtools: An R package for analyzing mixture models. *J. Stat. Softw.*, 32:1–29, 2010.

[313] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, Scott L Carter, Gad Getz, Katherine Stemke-Hale, Gordon B Mills, and Roel G W Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, 4:2612, 2013.

[314] Christopher J Hanley, Elodie Henriet, Orit Katarina Sirka, Gareth J Thomas, and Andrew J Ewald. Tumor-Resident stromal cells promote breast cancer invasion through regulation of the basal phenotype. *Mol. Cancer Res.*, 18(11):1615–1622, November 2020.

[315] Karama Asleh, Gian Luca Negri, Sandra E Spencer Miko, Shane Colborne, Christopher S Hughes, Xiu Q Wang, Dongxia Gao, C Blake Gilks, Stephen K L Chia, Torsten O Nielsen, and Gregg B Morin. Proteomic analysis of archival breast cancer clinical specimens identifies biological subtypes with distinct survival outcomes. *Nat. Commun.*, 13(1):896, February 2022.

[316] Jinghui Yang, Chunsheng Li, Hang Li, and Changyong E. LncRNA CACNA1G-AS1 facilitates hepatocellular carcinoma progression through the miR-2392/C1orf61 pathway. *J. Cell. Physiol.*, 234(10):18415–18422, August 2019.

[317] P-F Yu, A-R Kang, L-J Jing, and Y-M Wang. Long non-coding RNA CACNA1G-AS1 promotes cell migration, invasion and epithelial-mesenchymal transition by HNRNPA2B1

in non-small cell lung cancer. *Eur. Rev. Med. Pharmacol. Sci.*, 22(4):993–1002, February 2018.

[318] Guo-Wei Huang, Ying-Li Zhang, Lian-Di Liao, En-Min Li, and Li-Yan Xu. Natural antisense transcript TPM1-AS regulates the alternative splicing of tropomyosin I through an interaction with RNA-binding motif protein 4. *Int. J. Biochem. Cell Biol.*, 90:59–67, September 2017.

[319] Lifeng Dong, Junbin Qian, Fangfang Chen, Yangfan Fan, and Jingpei Long. LINC00461 promotes cell migration and invasion in breast cancer through mir-30a-5p/integrin β3 axis. *J. Cell. Biochem.*, 120(4):4851–4862, April 2019.

[320] Charu Kothari, Alisson Clemenceau, Geneviève Ouellette, Kaoutar Ennour-Idrissi, Annick Michaud, René C-Gaudreault, Caroline Diorio, and Francine Durocher. TBC1D9: An important modulator of tumorigenesis in breast cancer. *Cancers*, 13(14), July 2021.

[321] Jin He, Mingjun Wu, Lei Xiong, Yijia Gong, Renjie Yu, Weiyan Peng, Lili Li, Li Li, Shaorong Tian, Yan Wang, Qian Tao, and Tingxiu Xiang. BTB/POZ zinc finger protein ZBTB16 inhibits breast cancer proliferation and metastasis through upregulating ZBTB28 and antagonizing BCL6/ZBTB27. *Clin. Epigenetics*, 12(1):82, June 2020.

[322] J Devon Roll, Ashley G Rivenbark, Rupninder Sandhu, Joel S Parker, Wendell D Jones, Lisa A Carey, Chad A Livasy, and William B Coleman. Dysregulation of the epigenome in triple-negative breast cancers: basal-like and claudin-low breast cancers express aberrant DNA hypermethylation. *Exp. Mol. Pathol.*, 95(3):276–287, December 2013.

[323] Jose V Moyano, Joseph R Evans, Feng Chen, Meiling Lu, Michael E Werner, Fruma Yehiely, Leslie K Diaz, Dmitry Turbin, Gamze Karaca, Elizabeth Wiley, Torsten O Nielsen, Charles M Perou, and Vincent L Cryns. AlphaB-crystallin is a novel oncoprotein that predicts poor clinical outcome in breast cancer. *J. Clin. Invest.*, 116(1):261–270, January 2006.

[324] Stephanie M Sitterding, William R Wiseman, Carol L Schiller, Chunyan Luan, Feng Chen, Jose V Moyano, William G Watkin, Elizabeth L Wiley, Vincent L Cryns, and

Leslie K Diaz. AlphaB-crystallin: a novel marker of invasive basal-like and metaplastic breast carcinomas. *Ann. Diagn. Pathol.*, 12(1):33–40, February 2008.

[325] Zhijuan Chen, Qing Ruan, Song Han, Lei Xi, Wenguo Jiang, Huabei Jiang, David A Ostrov, and Jun Cai. Discovery of structure-based small molecular inhibitor of αB-crystallin against basal-like/triple-negative breast cancer development in vitro and in vivo. *Breast Cancer Res. Treat.*, 145(1):45–59, May 2014.

[326] Krysta M Coyle, J Patrick Murphy, Dejan Vidovic, Ahmad Vaghar-Kashani, Cheryl A Dean, Mohammad Sultan, Derek Clements, Melissa Wallace, Margaret L Thomas, Amos Hundert, Carman A Giacomantonio, Lucy Helyer, Shashi A Gujar, Patrick W K Lee, Ian C G Weaver, and Paola Marcato. Breast cancer subtype dictates DNA methylation and ALDH1A3-mediated expression of tumor suppressor RARRES1. *Oncotarget*, 7(28): 44096–44112, July 2016.

[327] Walid T Khaled, Song Choon Lee, John Stingl, Xiongfeng Chen, H Raza Ali, Oscar M Rueda, Fazal Hadi, Juexuan Wang, Yong Yu, Suet-Feung Chin, Mike Stratton, Andy Futreal, Nancy A Jenkins, Sam Aparicio, Neal G Copeland, Christine J Watson, Carlos Caldas, and Pentao Liu. BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat. Commun.*, 6:5987, January 2015.

[328] A Barghash, V Helms, and S M Kessler. Overexpression of IGF2 mRNA-Binding protein 2 (IMP2/p62) as a feature of basal-like breast cancer correlates with short survival. *Scand. J. Immunol.*, 82(2):142–143, August 2015.

[329] Yi-Hsin Hsu, Jun Yao, Li-Chuan Chan, Ting-Jung Wu, Jennifer L Hsu, Yueh-Fu Fang, Yongkun Wei, Yun Wu, Wen-Chien Huang, Chien-Liang Liu, Yuan-Ching Chang, Ming-Yang Wang, Chia-Wei Li, Jia Shen, Mei-Kuang Chen, Aysegul A Sahin, Anil Sood, Gordon B Mills, Dihua Yu, Gabriel N Hortobagyi, and Mien-Chie Hung. Definition of PKC-α, CDK6, and MET as therapeutic targets in triple-negative breast cancer. *Cancer Res.*, 74(17):4822–4835, September 2014.

[330] Vinicius S Chagas, Clarice S Groeneveld, Kelin G Oliveira, Sheyla Trefflich, Rodrigo C de Almeida, Bruce A J Ponder, Kerstin B Meyer, Steven J M Jones, A Gordon Robertson,

and Mauro A A Castro. RTNduals: an R/Bioconductor package for analysis of co-regulation and inference of dual regulons. *Bioinformatics*, 35(24):5357–5358, December 2019.

[331] Sarah M Hücker, Tobias Fehlmann, Christian Werno, Kathrin Weidele, Florian Lüke, Anke Schlenska-Lange, Christoph A Klein, Andreas Keller, and Stefan Kirsch. Single-cell microRNA sequencing method comparison and application to cell lines and circulating lung tumor cells. *Nat. Commun.*, 12(1):4316, July 2021.

[332] Siyuan John Liu, Tomasz J Nowakowski, Alex A Pollen, Jan H Lui, Max A Horlbeck, Frank J Attenello, Daniel He, Jonathan S Weissman, Arnold R Kriegstein, Aaron A Diaz, and Daniel A Lim. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol.*, 17:67, April 2016.

[333] Claudia Skok Gibbs, Christopher A Jackson, Giuseppe-Antonio Saldi, Andreas Tjärnberg, Aashna Shah, Aaron Watters, Nicholas De Veaux, Konstantine Tchourine, Ren Yi, Tymor Hamamsy, Dayanne M Castro, Nicholas Carriero, Bram L Gorissen, David Gresham, Emily R Miraldi, and Richard Bonneau. High performance single-cell gene regulatory network inference at scale: The inferelator 3.0. February 2022.

[334] Anjun Ma, Cankun Wang, Yuzhou Chang, Faith H Brennan, Adam McDermaid, Bingqiang Liu, Chi Zhang, Phillip G Popovich, and Qin Ma. IRIS3: integrated cell-type-specific regulon inference server from single-cell RNA-Seq. *Nucleic Acids Res.*, 48(W1): W275–W286, July 2020.

[335] Bram Van de Sande, Christopher Flerin, Kristofer Davie, Maxime De Waegeneer, Gert Hulselmans, Sara Aibar, Ruth Seurinck, Wouter Saelens, Robrecht Cannoodt, Quentin Rouchon, Toni Verbeiren, Dries De Maeyer, Joke Reumers, Yvan Saeys, and Stein Aerts. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.*, 15(7):2247–2276, July 2020.

[336] Lola Rahib, Mackenzie R Wehner, Lynn M Matrisian, and Kevin T Nead. Estimated projection of US cancer incidence and death to 2040. *JAMA Netw Open*, 4(4):e214708, April 2021.

[337] Michael J Pishvaian, Edik M Blais, Jonathan R Brody, Emily Lyons, Patricia DeArbeloa, Andrew Hendifar, Sam Mikhail, Vincent Chung, Vaibhav Sahai, Davendra P S Sohal, Sara Bellakbira, Dzung Thach, Lola Rahib, Subha Madhavan, Lynn M Matrisian, and Emanuel F Petricoin, 3rd. Overall survival in patients with pancreatic cancer receiving matched therapies following molecular profiling: a retrospective analysis of the know your tumor registry trial. *Lancet Oncol.*, 21(4):508–518, April 2020.

[338] Daniel Ansari, Adam Gustafsson, and Roland Andersson. Update on the management of pancreatic cancer: surgery is not enough. *World J. Gastroenterol.*, 21(11):3157–3165, March 2015.

[339] Michael P Menden, Francesco Paolo Casale, Johannes Stephan, Graham R Bignell, Francesco Iorio, Ultan McDermott, Mathew J Garnett, Julio Saez-Rodriguez, and Oliver Stegle. The germline genetic component of drug sensitivity in cancer cell lines. *Nat. Commun.*, 9(1):3385, August 2018.

[340] Günter Schneider, Marc Schmidt-Supprian, Roland Rad, and Dieter Saur. Tissue-specific tumorigenesis: context matters. *Nat. Rev. Cancer*, 17(4):239–253, April 2017.

[341] Chiara Falcomatà, Stefanie Bärthel, Günter Schneider, Dieter Saur, and Christian Veltkamp. Deciphering the universe of genetic context-dependencies using mouse models of cancer. *Curr. Opin. Genet. Dev.*, 54:97–104, February 2019.

[342] Nina Schönhuber, Barbara Seidler, Kathleen Schuck, Christian Veltkamp, Christina Schachtler, Magdalena Zukowska, Stefan Eser, Thorsten B Feyerabend, Mariel C Paul, Philipp Eser, Sabine Klein, Andrew M Lowy, Ruby Banerjee, Fangtang Yang, Chang-Lung Lee, Everett J Moding, David G Kirsch, Angelika Scheideler, Dario R Alessi, Ignacio Varela, Allan Bradley, Alexander Kind, Angelika E Schnieke, Hans-Reimer Rodewald, Roland Rad, Roland M Schmid, Günter Schneider, and Dieter Saur. A next-generation dual-recombinase system for time- and host-specific targeting of pancreatic cancer. *Nat. Med.*, 20(11):1340–1347, November 2014.

[343] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov,

Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F Berger, John E Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H Engels, Jill Cheng, Guoying K Yu, Jianjun Yu, Peter Aspesi, Jr, Melanie de Silva, Kalpana Jagtap, Michael D Jones, Li Wang, Charles Hatton, Emanuele Palescandolo, Supriya Gupta, Scott Mahan, Carrie Sougnez, Robert C Onofrio, Ted Liefeld, Laura MacConaill, Wendy Winckler, Michael Reich, Nanxin Li, Jill P Mesirov, Stacey B Gabriel, Gad Getz, Kristin Ardlie, Vivien Chan, Vic E Myer, Barbara L Weber, Jeff Porter, Markus Warmuth, Peter Finan, Jennifer L Harris, Matthew Meyerson, Todd R Golub, Michael P Morrissey, William R Sellers, Robert Schlegel, and Levi A Garraway. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, March 2012.

[344] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Muhammad Ammad-ud din, Petteri Hintsanen, Suleiman A Khan, John-Patrick Mpindi, Olli Kallioniemi, Antti Honkela, Tero Aittokallio, Krister Wennerberg, NCI DREAM Community, James J Collins, Dan Gallahan, Dinah Singer, Julio Saez-Rodriguez, Samuel Kaski, Joe W Gray, and Gustavo Stolovitzky. A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.*, 32(12):1202–1212, December 2014.

[345] In Sock Jang, Elias Chaibub Neto, Juistin Guinney, Stephen H Friend, and Adam A Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac. Symp. Biocomput.*, pages 63–74, 2014.

[346] Vidhi Malik, Yogesh Kalakoti, and Durai Sundar. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics*, 22(1):214, March 2021.

[347] Hossein Sharifi-Noghabi, Olga Zolotareva, Colin C Collins, and Martin Ester. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*, 35(14):i501–i509, July 2019.

[348] Jungho Kong, Heetak Lee, Donghyo Kim, Seong Kyu Han, Doyeon Ha, Kunyoo Shin, and Sanguk Kim. Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. *Nat. Commun.*, 11(1):5485, October 2020.

[349] Brent M Kuenzi, Jisoo Park, Samson H Fong, Kyle S Sanchez, John Lee, Jason F Kreisberg, Jianzhu Ma, and Trey Ideker. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 38(5):672–684.e6, November 2020.

[350] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, June 2011.

[351] Jake Crawford and Casey S Greene. Incorporating biological structure into machine learning models in biomedicine. *Curr. Opin. Biotechnol.*, 63:126–134, June 2020.

[352] Michael K Yu, Jianzhu Ma, Jasmin Fisher, Jason F Kreisberg, Benjamin J Raphael, and Trey Ideker. Visible machine learning for biomedicine. *Cell*, 173(7):1562–1565, June 2018.

[353] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43): 15545–15550, October 2005.

[354] Seth I Berger and Ravi Iyengar. Network analyses in systems pharmacology. *Bioinformatics*, 25(19):2466–2472, October 2009.

[355] Zikai Wu, Yong Wang, and Luonan Chen. Network-based drug repositioning. *Mol. Biosyst.*, 9(6):1268–1281, June 2013.

[356] Adrià Fernández-Torras, Miquel Duran-Frigola, and Patrick Aloy. Encircling the regions of the pharmacogenomic landscape that determine drug response. *Genome Med.*, 11(1): 17, March 2019.

[357] Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barábasi. Network-based in silico drug efficacy screening. *Nat. Commun.*, 7:10331, February 2016.

[358] Matthew G Rees, Brinton Seashore-Ludlow, Jaime H Cheah, Drew J Adams, Edmund V Price, Shubhroz Gill, Sarah Javaid, Matthew E Coletti, Victor L Jones, Nicole E Bodycombe, Christian K Soule, Benjamin Alexander, Ava Li, Philip Montgomery, Joanne D Kotz, C Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Daniel A Haber, Clary B Clish, Joshua A Bittker, Michelle Palmer, Bridget K Wagner, Paul A Clemons, Alykhan F Shamji, and Stuart L Schreiber. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.*, 12(2):109–116, February 2016.

[359] Paul Geeleher, Nancy J Cox, and R Stephanie Huang. Cancer biomarker discovery is improved by accounting for variability in general levels of drug sensitivity in pre-clinical models. *Genome Biol.*, 17(1):190, September 2016.

[360] Brian S White, Suleiman A Khan, Mike J Mason, Muhammad Ammad-Ud-Din, Swapnil Potdar, Disha Malani, Heikki Kuusanmäki, Brian J Druker, Caroline Heckman, Olli Kallioniemi, Stephen E Kurtz, Kimmo Porkka, Cristina E Tognon, Jeffrey W Tyner, Tero Aittokallio, Krister Wennerberg, and Justin Guinney. Bayesian multi-source regression and monocyte-associated gene expression predict BCL-2 inhibitor resistance in acute myeloid leukemia. *NPJ Precis Oncol*, 5(1):71, July 2021.

[361] Michael M Gottesman. Mechanisms of cancer drug resistance. *Annu. Rev. Med.*, 53: 615–627, 2002.

[362] Olivier Trédan, Carlos M Galmarini, Krupa Patel, and Ian F Tannock. Drug resistance and the solid tumor microenvironment. *J. Natl. Cancer Inst.*, 99(19):1441–1454, October 2007.

[363] Federica Catalanotti, Gloria Reyes, Veronika Jesenberger, Gergana Galabova-Kovacs, Ricardo de Matos Simoes, Oliviero Carugo, and Manuela Baccarini. A Mek1-Mek2 heterodimer determines the strength and duration of the erk signal. *Nat. Struct. Mol. Biol.*, 16(3):294–303, March 2009.

[364] Sunil R Hingorani, Emanuel F Petricoin, Anirban Maitra, Vinodh Rajapakse, Catrina King, Michael A Jacobetz, Sally Ross, Thomas P Conrads, Timothy D Veenstra, Ben A Hitt, Yoshiya Kawaguchi, Don Johann, Lance A Liotta, Howard C Crawford, Mary E Putt, Tyler Jacks, Christopher V E Wright, Ralph H Hruban, Andrew M Lowy, and David A Tuveson. Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell*, 4(6):437–450, December 2003.

[365] E L Jackson, N Willis, K Mercer, R T Bronson, D Crowley, R Montoya, T Jacks, and D A Tuveson. Analysis of lung tumor initiation and progression using conditional expression of oncogenic k-ras. *Genes Dev.*, 15(24):3243–3248, December 2001.

[366] Hassan Nakhai, Saadettin Sel, Jack Favor, Lidia Mendoza-Torres, Friedrich Paulsen, Gernot I W Duncker, and Roland M Schmid. Ptf1a is essential for the differentiation of GABAergic and glycinergic amacrine cells and horizontal cells in the mouse retina. *Development*, 134(6):1151–1160, March 2007.

[367] Katja Peschke, Hannah Jakubowsky, Arlett Schäfer, Carlo Maurer, Sebastian Lange, Felix Orben, Raquel Bernad, Felix N Harder, Matthias Eiber, Rupert Öllinger, Katja Steiger, Melissa Schlitter, Wilko Weichert, Ulrich Mayr, Veit Phillip, Christoph Schlag, Roland M Schmid, Rickmer F Braren, Bo Kong, Ihsan Ekin Demir, Helmut Friess, Roland Rad, Dieter Saur, Günter Schneider, and Maximilian Reichert. Identification of treatment-induced vulnerabilities in pancreatic cancer patients using functional model systems. *EMBO Mol. Med.*, page e14876, February 2022.

[368] Nicholas A Clark, Marc Hafner, Michal Kouril, Elizabeth H Williams, Jeremy L Muhlich, Marcin Pilarczyk, Mario Niepel, Peter K Sorger, and Mario Medvedovic. GRcalculator: an online tool for calculating and mining dose-response data. *BMC Cancer*, 17(1):698, October 2017.

[369] Marc Hafner, Mario Niepel, Mirra Chung, and Peter K Sorger. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat. Methods*, 13(6):521–527, June 2016.

[370] David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Fröhling, Edmond M Chan, Martin L Sos, Kathrin Michel, Craig Mermel, Serena J Silver, Barbara A Weir, Jan H Reiling, Qing Sheng, Piyush B Gupta, Raymond C Wadlow, Hanh Le, Sebastian Hoersch, Ben S Wittner, Sridhar Ramaswamy, David M Livingston, David M Sabatini, Matthew Meyerson, Roman K Thomas, Eric S Lander, Jill P Mesirov, David E Root, D Gary Gilliland, Tyler Jacks, and William C Hahn. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–112, November 2009.

[371] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. PID: the pathway interaction database. *Nucleic Acids Res.*, 37(Database issue):D674–9, January 2009.

[372] Igor Dolgalev. MSigDB gene sets for multiple organisms in a tidy data format [r package msigdbr version 7.4.1]. May 2021.

[373] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, 47(D1):D607–D613, January 2019.

[374] Patroklos Samaras, Tobias Schmidt, Martin Frejno, Siegfried Gessulat, Maria Reinecke, Anna Jarzab, Jana Zecha, Julia Mergner, Piero Giansanti, Hans-Christian Ehrlich, Stephan Aiche, Johannes Rank, Harald Kienegger, Helmut Krcmar, Bernhard Kuster, and Mathias Wilhelm. ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic Acids Res.*, 48(D1):D1153–D1163, January 2020.

[375] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank

5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46(D1): D1074–D1082, January 2018.

[376] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, November 2016.

[377] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

[378] Wei Li, Han Xu, Tengfei Xiao, Le Cong, Michael I Love, Feng Zhang, Rafael A Irizarry, Jun S Liu, Myles Brown, and X Shirley Liu. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.*, 15 (12):554, 2014.

[379] Rémy Nicolle, Yuna Blum, Pauline Duconseil, Charles Vanbrugghe, Nicolas Brandone, Flora Poizat, Julie Roques, Martin Bigonnet, Odile Gayet, Marion Rubis, Nabila Elarouci, Lucile Armenoult, Mira Ayadi, Aurélien de Reyniès, Marc Giovannini, Philippe Grandval, Stephane Garcia, Cindy Canivet, Jérôme Cros, Barbara Bournet, Louis Buscail, BACAP Consortium, Vincent Moutardier, Marine Gilabert, Juan Iovanna, and Nelson Dusetti. Establishment of a pancreatic adenocarcinoma molecular gradient (PAMG) that predicts the clinical outcome of pancreatic cancer. *EBioMedicine*, 57:102858, July 2020.

[380] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*, 1(6):417–425, December 2015.

[381] Lukas Krauß, Bettina C Urban, Sieglinde Hastreiter, Carolin Schneider, Patrick Wenzel, Zonera Hassan, Matthias Wirth, Katharina Lankes, Andrea Terrasi, Christine Klement, Filippo M Cernilogar, Rupert Öllinger, Niklas de Andrade Krätzig, Thomas Engleitner, Roland M Schmid, Katja Steiger, Roland Rad, Oliver H Krämer, Maximilian Reichert, Gunnar Schotta, Dieter Saur, and Günter Schneider. HDAC2 facilitates pancreatic cancer metastasis. *Cancer Res.*, 82(4):695–707, February 2022.

[382] Sophie Bekisz and Liesbet Geris. Cancer modeling: From mechanistic to data-driven approaches, and from fundamental insights to clinical applications. *J. Comput. Sci.*, 46: 101198, October 2020.

[383] Xuewei Wang, Zhifu Sun, Michael T Zimmermann, Andrej Bugrim, and Jean-Pierre Kocher. Predict drug sensitivity of cancer cells with pathway activity inference. *BMC Med. Genomics*, 12(Suppl 1):15, January 2019.

[384] Nicolas Alcaraz, Markus List, Richa Batra, Fabio Vandin, Henrik J Ditzel, and Jan Baumbach. De novo pathway-based biomarker identification. *Nucleic Acids Res.*, 45(16): e151, September 2017.

[385] Deborah Plana, Adam C Palmer, and Peter K Sorger. Independent drug action in combination therapy: Implications for precision oncology. *Cancer Discov.*, 12(3):606–624, March 2022.

[386] Sandeep C Pingle, Zeba Sultana, Sandra Pastorino, Pengfei Jiang, Rajesh Mukthavaram, Ying Chao, Ila Sri Bharati, Natsuko Nomura, Milan Makale, Taher Abbasi, et al. In silico modeling predicts drug sensitivity of patient-derived cancer cells. *Journal of translational medicine*, 12(1):1–13, 2014.

[387] Marco Berghoff, Jakob Rosenbauer, Felix Hoffmann, and Alexander Schug. Cells in silico–introducing a high-performance framework for large-scale tissue modeling. *BMC bioinformatics*, 21(1):1–21, 2020.

[388] Martin Frejno, Chen Meng, Benjamin Ruprecht, Thomas Oellerich, Sebastian Scheich, Karin Kleigrewe, Enken Drecoll, Patroklos Samaras, Alexander Hogrebe, Dominic Helm, Julia Mergner, Jana Zecha, Stephanie Heinzlmeir, Mathias Wilhelm, Julia Dorn, Hans-Michael Kvasnicka, Hubert Serve, Wilko Weichert, and Bernhard Kuster. Proteome activity landscapes of tumor cell lines determine drug responses. *Nat. Commun.*, 11(1): 3639, July 2020.

[389] Gaye Lightbody, Valeriia Haberland, Fiona Browne, Laura Taggart, Huiru Zheng, Eileen Parkes, and Jaine K Blayney. Review of applications of high-throughput sequencing

in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Brief. Bioinform.*, 20(5):1795–1811, September 2019.

[390] Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.*, 20(7):389–403, July 2019.

[391] L Sikkema, D Strobl, L Zappia, E Madissoon, N S Markov, L Zaragosi, M Ansari, M Arguel, L Apperloo, C Bécavin, M Berg, E Chichelnitskiy, M Chung, A Collin, A C A Gay, B Hooshiar Kashani, M Jain, T Kapellos, T M Kole, C Mayr, M von Papen, L Peter, C Ramírez-Suástegui, J Schniering, C Taylor, T Walzthoeni, C Xu, L T Bui, C de Donno, L Dony, M Guo, A J Gutierrez, L Heumos, N Huang, I Ibarra, N Jackson, P Kadur Lakshminarasimha Murthy, M Lotfollahi, T Tabib, C Talavera-Lopez, K Travaglini, A Wilbrey-Clark, K B Worlock, M Yoshida, Lung Biological Network Consortium, T Desai, O Eickelberg, C Falk, N Kaminski, M Krasnow, R Lafyatis, M Nikolíc, J Powell, J Rajagopal, O Rozenblatt-Rosen, M A Seibold, D Sheppard, D Shepherd, S A Teichmann, A Tsankov, J Whitsett, Y Xu, N E Banovich, P Barbry, T E Duong, K B Meyer, J A Kropski, D Pe'er, H B Schiller, P R Tata, J L Schultze, A V Misharin, M C Nawijn, M D Luecken, and F Theis. An integrated cell atlas of the human lung in health and disease. March 2022.

[392] Mohammed AlQuraishi and Peter K Sorger. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nat. Methods*, 18(10):1169–1180, October 2021.

[393] Haitham A Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, Taylor E Arnoff, Jihye Park, William C Hahn, and Eliezer M Van Allen. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, October 2021.

[394] Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K Min, Kelly Brock, Yarin Gal, and Debora S Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, November 2021.

[395] Mohammad Lotfollahi, Sergei Rybakov, Karin Hrovatin, Soroor Hediyeh-zadeh, Carlos Talavera-López, Alexander V Misharin, and Fabian J Theis. Biologically informed deep learning to infer gene program activity in single cells. February 2022.

[396] Zifeng Wang, Aria Masoomi, Zhonghui Xu, Adel Boueiz, Sool Lee, Tingting Zhao, Russell Bowler, Michael Cho, Edwin K Silverman, Craig Hersh, Jennifer Dy, and Peter J Castaldi. Improved prediction of smoking status via isoform-aware RNA-seq deep learning models. *PLoS Comput. Biol.*, 17(10):e1009433, October 2021.

List of Publications

Boniolo, Fabio*, Markus Hoffmann*, Norman Roggendorf, Bahar Tercan, Jan Baumbach, Mauro Castro, A. Gordon Robertson, Dieter Saur, and Markus List. "spongEffects: ceRNA modules offer patient-specific insights into the miRNA regulatory landscape." bioRxiv (2022). * equal contribution

Boniolo, Fabio*, Emilio Dorigatti*, Alexander J. Ohnmacht*, Dieter Saur, Benjamin Schubert, and Michael P. Menden. "Artificial intelligence in early drug discovery enabling precision medicine." Expert opinion on drug discovery 16, no. 9 (2021): 991-1007.* equal contribution

In addition to the first author publications, I also published as a contributing author in peer-reviewed journals (not part of this dissertation):

Orben, Felix, Katharina Lankes, Christian Schneeweis, Zonera Hassan, Hannah Jakubowsky, Lukas Krauß, **Fabio Boniolo** et al. "Epigenetic drug screening defines a PRMT5 inhibitor sensitive pancreatic cancer subtype." JCI insight (2022).

Falcomatà, Chiara, Stefanie Bärthel, Sebastian A. Widholz, Christian Schneeweis, Juan José Montero, Albulena Toska, Jonas Mir et al. "Selective multi-kinase inhibition sensitizes mesenchymal pancreatic cancer to immune checkpoint blockade by remodeling the tumor microenvironment." Nature Cancer 3, no. 3 (2022): 318-336.

Falcomatà, Chiara, Stefanie Bärthel, Angelika Ulrich, Sandra Diersch, Christian Veltkamp, Lena Rad, **Fabio Boniolo** et al. "Genetic screens identify a context-specific PI3K/p27Kip1

node driving extrahepatic biliary cancer." Cancer discovery 11, no. 12 (2021): 3158-3177.

List of Figures