Exploring the Intersection of Multi-Omics and Machine Learning in Cancer Research

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy
of Freie Universität Berlin

by

Artem Baranovskii

Berlin 2023

The following work was performed from June 2019 until March 2023 under the supervision of Dr. Altuna Akalin at the Berlin Institute for Medical Systems Biology (BIMSB), Max Delbrück Center for Molecular Medicine, Hannoversche Str. 28, 10115 Berlin-Mitte.

1st reviewer: Dr. Altuna Akalin
Berlin Institute for Medical Systems Biology (BIMSB)
Max-Delbrück-Centrum für Molekulare Medizin
Hannoversche Str. 28, 10115 Berlin

2nd reviewer: Prof. Dr. Irmtraud Meyer
Freie Universität Berlin, Special Professor
Berlin Institute for Medical Systems Biology (BIMSB)
Max-Delbrück-Centrum für Molekulare Medizin
Hannoversche Str. 28, 10115 Berlin

Date of thesis defence: 18.01.2024

Acknowledgements

Although personal experiences may differ, the road to a doctorate is rarely ordinary. Wading, a student often finds oneself lost and searches waymarks and those one finds, one needs to know to trust. For this, I want to leave in ink a heartwarming gratitude to the people who served me in this role: Altuna Akalin, Irmtraud Meyer, Dmitrii Pervoushine, and Michaela Herzig. Alongside them, I want to thank Vedran Franke, Bora Uyar, Nicolai von Kügelgen and Inga Lödige for all the amusing talk and advice that helped me learn better. Finally, there are people who were around to lean on, not necessarily in the domain of science. Those are Samantha Mendonsa and Erik Becher; thank you for this.

Declaration of Independence

Herewith I certify that I have prepared and written my thesis independently and that I have not used any sources and aids other than those indicated by me. I also declare that I have not submitted the dissertation in this or any other form to any other institution as a dissertation. Artem Baranovskii 24.07.2023

Foreword

This thesis is cumulative. It includes two works that have been published in peer-reviewed journals. These publications are reproduced in Chapters 3 and 4 of this thesis.

Publication I (Chapter 4.1):

Jan Dohmen*, *Artem Baranovskii**, Jonathan Ronen, Bora Uyar, Vedran Franke, and Altuna Akalin, Identifying tumor cells at the single-cell level using machine learning, Genome Biology 2022, 23, 123, https://doi.org/10.1186/s13059-022-02683-1

* These authors contributed equally to the work

Publication II (Chapter 4.2):

*Artem Baranovskii**, Irem Gündüz*, Vedran Franke, Bora Uyar & Altuna Akalin, Multi-Omics Alleviates the Limitations of Panel Sequencing for Cancer Drug Response Prediction, Cancers 2022, 14(22), 5604,
https://doi.org/10.3390/cancers14225604

* These authors contributed equally to the work

Contents

1 Summary

Cancer biology and machine learning represent two seemingly disparate yet intrinsically linked fields of study. Cancer biology, with its complexities at the cellular and molecular levels, brings up a myriad of challenges. Of particular concern are the deviations in cell behaviour and rearrangements of genetic material that fuel transformation, growth, and spread of cancerous cells. Contemporary studies of cancer biology often utilise wide arrays of genomic data to pinpoint and exploit these abnormalities with an end-goal of translating them into functional therapies.

Machine learning allows machines to make predictions based on the learnt data without explicit programming. It leverages patterns and inferences from large datasets, making it an invaluable tool in the modern era of large scale genomics. To this end, this doctoral thesis is underpinned by three themes: the application of machine learning, multi-omics, and cancer biology. It focuses on employment of machine learning algorithms to the tasks of cell annotation in single-cell RNA-seq datasets and drug response prediction in pre-clinical cancer models.

In the first study, the author and colleagues developed a pipeline named Ikarus to differentiate between neoplastic and healthy cells within single-cell datasets, a task crucial for understanding the cellular landscape of tumours. Ikarus is designed to construct cancer cell-specific gene signatures from expert-annotated scRNA-seq datasets, score these genes, and distribute the scores to neighbouring cells via network propagation. This method successfully circumvents two common challenges in single-cell annotation: batch effects and unstable clustering. Furthermore, Ikarus utilises a multi-omic approach by incorporating CNVs inferred from scRNA-seq to enhance classification accuracy.

The second study investigated how multi-omic analysis could enhance drug response prediction in pre-clinical cancer models. The research suggests that the typical practice of panel sequencing — a deep profiling of select, validated genomic features — is limited in its predictive power. However, incorporating transcriptomic features into the model significantly improves predictive ability across a variety of cancer models and is especially effective for drugs with collateral effects. This implies that the combined use of genomic and transcriptomic data has potential advantages in the pharmacogenomic arena.

This dissertation recapitulates the findings of two aforementioned studies, which were published in Genome Biology and Cancers journals respectively. The two studies illustrate the application of machine learning techniques and multi-omic approaches to address conceptually distinct

problems within the realm of cancer biology.

1 Zusammenfassung

Die Krebsbiologie und das maschinelle Lernen sind zwei scheinbar konträre, aber intrinsisch verbundene Forschungsbereiche. Insbesondere die Krebsbiologie ist auf zellulärer und molekularer Ebene hoch komplex und stellt den Forschenden vor eine Vielzahl von Herausforderungen. Zu verstehen wie abweichendes Zellverhalten und die Umstrukturierung genetischer Komponente die Transformation, das Wachstum und die Ausbreitung von Krebszellen antreiben, ist hierbei eine besondere Herausforderung. Gleichzeitig bestrebt die Krebsbiologie diese Abnormalitäten zu nutzen zu machen, Wissen aus ihnen zu gewinnen und sie so in funktionale Therapien umzusetzen.

Maschinelles Lernen ermöglicht es Vorhersagen auf der Grundlage von gelernten Daten ohne explizite Programmierung zu treffen. Es erkennt Muster in großen Datensätzen, erschließt sich so Erkenntnisse und ist deswegen ein unschätzbar wertvolles Werkzeug im modernen Zeitalter der Hochdurchsatz Genomforschung. Aus diesem Grund ist maschinelles Lernen eines der drei Haupthemen dieser Doktorarbeit, neben Multi-Omics und Krebsbiologie. Der Fokus liegt hierbei insbesondere auf dem Einsatz von maschinellen Lernalgorithmen zum Zweck der Zellannotation in Einzelzell-RNA-Sequenzdatensätzen und der Vorhersage der Arzneimittelwirkung in präklinischen Krebsmodellen.

In der ersten, hier präsentierten Studie, entwickelten der Autor und seine Kollegen eine Pipeline namens Ikarus. Diese kann zwischen neoplastischen und gesunden Zellen in Einzelzell-Datensätzen unterscheiden. Eine Aufgabe, die für das Verständnis der zellulären Landschaft von Tumoren entscheidend ist. Ikarus ist darauf ausgelegt, krebszellenspezifische Gensignaturen aus expertenanotierten scRNA-seq-Datensätzen zu konstruieren, diese Gene zu bewerten und die Bewertungen über Netzwerkverbreitung auf benachbarte Zellen zu verteilen. Diese Methode umgeht erfolgreich zwei häufige Herausforderungen bei der Einzelzellannotation: den Chargeneffekt und die instabile Clusterbildung. Darüber hinaus verwendet Ikarus, durch das Einbeziehen von scRNA-seq abgeleiteten CNVs, einen Multi-Omic-Ansatz der die Klassifikationsgenauigkeit verbessert.

Die zweite Studie untersuchte, wie Multi-Omic-Analysen die Vorhersage der Arzneimittelwirkung in präklinischen Krebsmodellen optimieren können. Die Forschung legt nahe, dass die übliche Praxis des Panel-Sequenzierens - die umfassende Profilierung ausgewählter, validierter genomischer Merkmale - in ihrer Vorhersagekraft begrenzt ist. Durch das Einbeziehen transkriptomischer Merkmale in das Modell konnte jedoch die Vorhersagefähigkeit bei verschiedenen Krebsmodellen signifikant verbessert werden, ins besondere für Arzneimittel mit Nebenwirkungen.

Diese Dissertation fasst die Ergebnisse der beiden oben genannten Studien zusammen, die jeweils in Genome Biology und Cancers Journalen veröffentlicht wurden. Die beiden Studien veranschaulichen die Anwendung von maschinellem Lernen und Multi-Omic-Ansätzen zur Lösung konzeptionell unterschiedlicher Probleme im Bereich der Krebsbiologie.

2 Introduction

2.1 Early Genomics

Back in 1986, in Bethesda, on the eve of a well-sized international meeting, a group of geneticists, including Dr Thomas H. Roderick, attended an unassuming downtown bar to discuss the name of a new genome-oriented journal to be started. Naturally, Genome was already in use, announced as a new name for the Canadian Journal of Genetics and Cytology in its next issue. Compound names with Genomes were not appealing either. Somewhen in the pitcher, he came up with a historical idea — Genomics. Such wise, the first -omics discipline, Genomics, took its name from a witty proposal in a pitcher to name a new journal [1]. Later in 1988, when asked about his definition of Genomics, Dr Roderick gave an excellent commentary: "We thought of the genome as a functioning whole beyond just single genes or sequences spread around a chromosome" [1]. The author believes it underlines the unifying idea upon which all later -omics disciplines are built — the shift of focus from an individual element, in this case, a gene, to a system or an ensemble, in this case, a genome.

Undoubtedly, the systematic approach postulated in the definition of Genomics requires whole genome sequences to leverage their full potential. But what is this potential? To correctly answer this question, the author recommends to further back into the pre-genome era. The basic premise of medical genetics is that most, if not every, with the exclusion of trauma, the disease carries a genetic background. Which, of course, even back then was not new since those tend to cluster in families and related individuals. But how to identify the genetics behind the disease? Let's look at the first hereditary disease for which the causal gene was identified — Chronic Granulomatous Disease (CGD). Now we know that the condition stems from insufficient production of reactive oxygen species in specific immune cells that is rooted in a defective NAPDH-oxidase. CGD was first characterised in children suffering from recurrent infections; however, in 1954, the basis for children's susceptibility was not identified [2]. More than a decade after the first record of the disease, in 1966, researchers established the cellular nature of the affliction: the phagocytes were involved [3]. A year further, solid physiological and biochemical studies succeeded in mapping the molecular mechanism and linked it to a dysfunctional oxidase complex that failed to oxidise reporter molecules [4]. Altogether, by the end of the 70s, scientists figured out what was functionally wrong within the disease and ventured forth in search of the gene-disorder connection. It took a stunning 20 years to locate and clone the gene [5].

Another hereditary disease has a discovery story very much alike — Cystic Fibrosis (CF), a Mendelian, i.e. monogenic, a disease with variable

degrees of symptoms among those affected. First characterised in 1938, a lot was known about CF's physiology in 1970 [6].

Nonetheless, the effort to find the CF gene took more than 11 years, spanning most of the eighties and cost roughly 50 million dollars before the causal gene, appropriately named Cystic Fibrosis Transmembrane Regulator (CFTR), was identified in 1989 [7, 8]. In both cases, the technique known as positional cloning was used to locate and clone the causal gene [9]. In short, this method relies on linkage analysis, for which, at a time, a technique called Restriction Fragment Length Polymorphism (RFLP) was used. In the latter, DNA from the sample is digested, Southern blotted to estimate fragment length and then hybridised to random DNA probes. Because restriction enzymes favour specific restriction sites, i.e., DNA sequence motifs, DNA fragments will be cut at different positions if a polymorphism disrupts a restriction site, producing fragments of different lengths. If the size of the fragment differs between individuals, then the sequences are not identical. In the case of CF, the discovered polymorphism was a humble deletion of three nucleotides in the CF gene of affected individuals that were otherwise present in the healthy genotype. The deletion of these three nucleotides, CTT, is the most common cause of CF, a disease of enormous complexity and one that causes a great deal of human suffering. The timeframe mentioned above, and the costs well describe how tedious a procedure like this is when employed to locate a polymorphism in a single gene on the scale of the human genome, which is 3 billion letters long. A task that fits the analogy made by David Botstein, the method's inventor: "coming in from outer space towards Chicago". However, instead of searching blindly, one could use a map.

Herein lies the pivotal importance of the Human Genome Project (HGP), which was finished in 2001 simultaneously by public and private ventures [10, 11]. The initial sequencing of the human genome chartered the genomic landscape and highlighted the marked variation in the distribution of genomic features, such as transposable elements, CpG islands, and, most importantly, genes. Of the latter, 30,000 to 40,000 protein-coding genes were presumed to exist at the time. Since our genomes are 99% the same, the initial sequence of the genome provided an invaluable reference for genetic research and sped up biological research to an unimaginable degree. Alongside, nearly one and a half million single nucleotide polymorphisms (SNPs) in the human genome were characterised and assembled in a database called dbSNP by The SNP Consortium (TSC) [12]. In 2003, only two years after the human genome was published, the HapMap project was initiated to discover genetic factors contributing to common diseases. As ambitious as HGP covered the entirety of the human genome, "including 99.9% where we are all the same", HapMap set out to describe "the common patterns within that 0.1% where we differ from each other" [13]. By then, the number

of SNPs in dbSNP had grown nearly ten times to a staggering 9 million. Association studies, i.e. studies that assess the correlation between a genetic variant in a population and a particular phenotype of interest, that were common in the early days of genomics underwent a qualitative elaboration following an inflow of unprecedented amount of information, something that was considered futuristic only a few years ago. SNPs, copy number variants (CNVs), and new genomic variants were deposited to databases in millions, and millions more were to be recorded.

2.2 Early Transcriptomics

Alongside Genomics, other -omics developed. Establishing SAGE (Serial Analysis of Gene Expression) heralded the beginning of early transcriptomics [14]. SAGE allowed the simultaneous capture of 60 thousand sequences transcribed from 4655 genes in S. cerevisiae, a breakthrough from existing techniques [15]. Expression Sequence Tag (EST) -based methods, a workhorse at the time, were instrumental for gene identification, but like Northern blotting, RT-qPCR, and RNase protection assays were tedious in implementation and lacked in yield to cover more than a handful of genes in one run.

Alongside technological advancement, the publication of this research saw the definition of "transcriptome" first worded — "identity of each expressed gene and its level of expression for a defined population of cells" [15], therefore, making S. cerevisiae the first living eukaryotic organism in history with a sequenced transcriptome and breaking the ground for another -omics discipline, Transcriptomics (worth to mention, that a year before the transcriptome was finished, in 1996, yeast genome was completely sequenced [16], thus making S. cerevisiae a forerunner for both omics among the eukaryotic species).

Soon, another method became available — Massive Parallel Signature Sequencing (MPSS) [17]. It was developed by the team of Sydney Brenner, the famous populariser of nematode Caenorhabditis elegans, in Lynx therapeutics, one of the forerunners of sequencing technologies. Their approach relied on microscopic beads to which the DNA template, or "signature sequence", was hybridised. These beads are assembled on a planar array, and the templates' sequences are read through ligation cycles and cleavage by labelled adapters. After each cycle, the image is taken, which records a tagged adapter and stored for analysis. Gene expression, i.e., fractional abundance, could be estimated from the recorded sequence data. While capacity-wise similar to the early SAGE, MPSS followed a conceptually different approach, somewhat reminiscent of the microarray technologies that will be discussed later. Cap Analysis of Gene Expression (CAGE) [18] is a notable modification of the SAGE technique that added another modality to the gene expression data.

In contrast to SAGE, which identified short tags from the 3´-ends of transcripts, CAGE sequenced tags from the 5´-end. It, therefore, allowed us to gather information on promoters and identify transcription start sites (TSSs) of mRNA transcripts. Although CAGE never outcompeted its progenitor technology in popularity, its importance could hardly be overlooked. This technique was the workhorse of the FANTOM Consortium that published a gene expression atlas covering 975 human and 379 mouse samples, including

tissues, primary cells and cell lines [19, 20]. Due to its technological similarities with SAGE, this technique never really became widespread, remaining a niche technology to study the regulation of transcription. Due to this, CAGE remained in use much longer than its progenitor technology until it was overshadowed by more potent Next Generation Sequencing (NGS) methods.

In the time of early transcriptomics, high-density oligonucleotide arrays were the central pillar of transcriptomic research [21]. Array-based methods to quantify mRNA expression existed before, namely spotted arrays. However, they carried a disadvantage that limited their throughput: to prepare the array, many cDNAs must be amplified, purified, catalogued and spotted on a specialised surface, either modified microscope slides [22] or a nylon membrane [23]. Therein lay the main disadvantage of spotted arrays before oligonucleotide arrays. For the latter method, the preparation and handling of cDNA and PCR products were no longer necessary. Instead, 20-mer oligonucleotides (probes) based on sequences of interest are designed in silico and synthesised directly on an array in known locations. A standard 1.28 x 1.28 cm array contains 300,000 probing oligonucleotides and can record the expression of roughly 40,000 human genes and expressed sequence tags (ESTs) [24, 25]. The versatility of custom probe synthesis and array design also allowed for quantifying specific exons and unique splice isoforms [26]. Higher yield is generally a positive quality, yet it makes the data analysis more complex and requires an accompanying computational framework. The popularity of the method guaranteed that it would shortly follow. In 2003, a seminal paper by Rafael Irizarry standardised the approach to data analysis of Affymetrix arrays, cementing its status as a "go-to" method for gene expression analysis [27]. First introduced in 1996, oligonucleotide arrays overtook other competing transcriptomics techniques and remained an unparalleled leader in the field until the advent of high-throughput RNA sequencing.

2.3 The need for annotation

Naturally, a more readily available platform for transcriptome characterisation led to the accumulation of transcriptomic data. Most of the data were generated on human and mouse chips, giving extensive characterisations of transcriptomes for different tissues and cell types of these organisms [28]. The majority of these datasets focused on particular biological processes and perturbations. However, the sheer amount of information proved hard to characterise; with human and mouse genomes sequenced, tens of thousands of genomic features were now recorded in a transcriptome snapshot for researchers to quantify and analyse. Technological advancements were outpacing biological conceptualisation and staggered the interpretations of sequencing results. The situation called for a unified and descriptive gene annotation system to bridge the data and knowledge gap. An Ontology was needed, and one was already in development since 2000 — Gene Ontology [29].

Gene Ontology Consortium (GOC) aimed to produce a common vocabulary for gene and protein roles in cells for all eukaryotes. An ambitious goal, but not anymore unrealistic. As was mentioned earlier, the yeast genome was sequenced in 1996, the genome of the nematode Caenorhabditis elegans was finished in 1998 [30], and the fruit fly Drosophila melanogaster done a year after [31], both famous model organisms that deserve their chapters outside of this thesis. The initial sequence of the human genome was done in 2001 and was shortly followed by a draft of a mouse genome [32]. An early comparison study between yeast and worm genomes should have an unexpected level of genetic conservation. About a tenth of a worm's protein-coding genes can be functionally annotated by their putative orthologues in the yeast. A three-way comparison study of human, worm, and fruit fly genomes similarly showed substantial conservation [33]. The same year, before sequences of human and mouse genomes were published, a group identified 1,185 orthologous gene pairs between humans and mice that shared approximately 85% of the coding sequence [34]. For the sake of brevity, the author will not iterate over all metazoan genomes sequenced or close to completion at the time but only mention that by 2003, 25 were already made public [35]. Such an arsenal at the hands of comparative genomics guaranteed the success of GOC that aimed at integrating many gene- and protein-keyword databases, including SwissPROT, EMBL, Pfam, and Genbank [36, 37, 38, 39].

Ultimately, GOC assembled three main categories to be used, among others, in the annotation of gene expression data that we all now know by heart: Cellular compartment, Biological process and Molecular function. Many ontologies and semantic entities annotate groups of gene-protein pairs related to a specific cellular compartment, biological process, or molecular

function within these three categories. Ontologies assembled had a graph-based structure, meaning that entities within were interconnected, reflecting a reality where one protein can fulfil a wide array of functions.

Another notable database, the Kyoto Encyclopedia of Genes and Genomes (KEGG), was built upon the pre-existing idea of molecular pathways [40]. On top of looking for conservation in coding sequences, KEGG founders grouped genes into sets based on the involvement of their protein products in a specific cell function, i.e. pathway. Functional transcriptomic studies of yeast metabolism well showcase this idea. A recurrent and well-known cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism that occurs when fermentable glucose in the media is depleted, and yeast starts to metabolise ethanol, a by-product of the previous step, as a carbon source. Well-executed time-series analysis with transcriptome capture by microarrays allowed us to identify genes active in both processes, i.e. pathways [41]. Because the same gene-protein pairs can be involved in multiple pathways, KEGG compiled genes and protein products into a connected molecular pathway-oriented graph, building upon orthology and data from functional studies.

Others, like GenMAPP, provided the community with the tools to visualise gene expression data in a pathway context and assemble their own gene sets to test [42]. The importance of these ontological constructs is well exemplified by the fact that now, 20 years after, these are still in active use. Fundamental ideas behind these databases underlined the growing idea in transcriptomics — to study functionally related gene sets instead of individual genes or whole transcriptomes. There was, however, another issue to solve.

Microarrays have been very successful in studying diseases like cancer, where genomic aberrations often lead to large changes in the expression of individual genes [43]. However, unlike modern RNA-sequencing methods sensitive to tiny changes in the transcriptome when sequenced deep enough, arrays struggled to detect modest perturbations in gene expression. In some cases, many genes show a unidirectional change in expression between conditions. Although common sense suggests the biological meaning behind this shift, these results will most likely be discarded as non-significant after the correction for multiple hypotheses testing [44]. It was clear that perturbations occur in individual genes and groups of functionally related genes and gene sets. However, a statistical framework was lacking to measure and test for differences between the expression of entire gene sets.

It didn't take long for this major bottleneck to be resolved. The basic idea was that by combining measurements across multiple genes, one could more reliably detect subtle but coordinated changes in gene expression.

One of the first attempts devised an approach that examined the enrichment of gene set members among the top-ranking genes selected by a user-defined cut-off. Those were later tested against a null distribution where the genes were picked randomly [45]. Looking at this method, one can see that the definition of top-ranking genes restricts it, and thus the method can vary in its results, potentially leading to false positives. However, a major drawback was that this tool was embedded into a large and unwieldy microarray analysis software suite, which the author believes prevented its popularity.

In contrast, an alternative method called Gene Set Enrichment Analysis (GSEA) suggested a lightweight solution [46, 47]. The idea behind GSEA is elegant in its simplicity. For pairwise comparison, imagine a list L containing genes ranked and ordered according to an appropriate metric of difference, such as a logged ratio between intensities or a correlation between gene expression and class labels when many samples are available. The Null hypothesis is that genes in the list L are randomly distributed and not associated with the compared classes' diagnostic categories. In the case of the original publication, individuals with normal glucose tolerance were compared to those with diabetes. An alternative hypothesis is that the order of genes in the list L is not random and associated with the categorisation of the samples. Then, for each a priori defined gene set S, a running-sum statistic called Enrichment Score (ES) is calculated by walking down the list L and increasing the ES for every encounter of a gene from S and decreasing for every gene that is not in S. The magnitude of the increment depends on the rank of a gene in L. The final score of the gene set S is assigned to the Maximum ES recorded during the random walk. The significance of the ES is estimated from the empirical permutation test. Namely, a null distribution of ESs is constructed by re-calculating the ES for the gene set S in randomly permuted data, against which the original ES is later tested. GSEA readily proved to be a great complement to single-gene studies: reanalysis of published studies from the new perspective generated more biological insight. While the original method was shipped together with 1325 pre-defined gene sets, nothing prevented it from operating with other gene sets assembled by mentioned GO and KEGG databases. To summarise, GSEA provided a user-friendly tool that allowed gene expression analysis on a higher level of biological organisation, finally bridging the gap between single-gene and gene set studies. Generally speaking, tens of thousands of dimensions in the gene expression matrix could be collapsed into several hundred (dozen or thousand, depending on the question of the study) biologically meaningful pathway scores. Abstraction of this sort enabled analysis of microarray transcriptomic data from a genome-wide perspective.

2.4 Cancer transcriptomics

It would be only natural to think that diseases of high complexity would be the most welcome target for an approach that works in abstractions of higher order. What can be more complex than cancer? A tumour-bearing disease that arises from different tissues, different organs, and different cells of an organism. Inherently variable by the mutational nature of the tumour's forebear cells and the microenvironment where they reside, cancer cells and the tumours they compose should still bear a resemblance. After all, cancer can be viewed as a consequence of the dysregulation of finely tuned cellular pathways that typically coerce cells to co-exist as a multicellular organism and, when highjacked, improve cancer cells' relative fitness. Following this perspective, cancer cells should exhibit specific traits, derivative of dysregulated pathways, distinguishing them from normal cells of the organism while remaining common between different tumours. After a quarter century of cancer research, these traits have been concisely dried into six hallmarks of cancer: avoidance of apoptosis, sustained proliferative signalling, evasion of growth suppressors, induction of angiogenesis, replicative immortality, and ability to metastasise [48]. Because cancer is a disease of the genome, the number of genetic lesions causal to either of these processes can be uncountable. Worth to mention, that the further expansion of cancer biology in the 21st century reflected in the widening of cancer hallmarks; the latest instalment of hallmarks of cancer included six new characteristics and totalled twelve [49].

Nevertheless, similarly to how cancer traits can be condensed into six hallmarks, these genetic lesions, although numerous and heterogeneous, should lead to similar functional outcomes or, more precisely, shapes of the transcriptome, i.e. patterns of coordinated gene expression, something very fitting to gene set analysis. Built on this idea, many laboratories worldwide rigorously interrogated the accumulated corpus of microarray gene expression data [50, 51, 52, 53]. In one study, Authors pulled together 1975 gene expression datasets covering primary tumours of 22 distinct types from 26 published studies [54]. Employing curated gene sets from GO, KEGG and GeneMAPP databases mentioned beforehand, they further narrowed input to "activated" and "repressed" gene modules by extracting the core expression cluster of the gene set. These modules corresponded to specific tumour types, subtypes, and malignant processes. For example, an osteoblastic module was significantly responsive in some cases of hepatocellular carcinoma (HCC), lung, and breast cancers, all known to metastasise in bones, suggesting that this gene expression pattern is associated with metastatic processes. Although this finding might sound trivial today, in the early 2000s, the metastatic potential of primary tumours was still debatable [55, 56]. Overall, this period in the history of cancer transcriptomics can be described as a "gene set rush",

assumingly alluding to the gold rushes of the 19th century. Worth mentioning an allusion that was first made towards the rush to clone human genes in the 1980s [57]. To not downplay the importance of that period, it is worth saying that an incredible amount of knowledge was generated from these studies. Hundreds of gene sets describing cellular processes contributed to the public databases ensuring their sustained growth and future application.

2.5 Early Genome-wide Association Studies (GWASs)

Transcriptomics was not the only -omic discipline reshaped by array technology; successes of the array designs readily permeated into Genomics as well leading to a revolution in genetic screening. However, everything in order. First of all, why the genetic screening is important? For the knowledge of identified genetic risk variants to benefit an individual, these variants must be detected, i.e. screened for, in patients. The screening allows to gain insight into genetic predisposition to disease and helps to select treatments.

In already discussed examples of CF and CGD, relatively uncommon Mendelian disorders, a fistful of causal mutations produce the disease phenotype. And through tremendous effort, those mutations have been mapped. The outcome of these mutations is categorical; whether one has a disease or not, early genotyping methods could still suffice to test those at risk. Let's, however, look at the example of breast cancer. Unlike CF, where the 3-base deletion accounts for most cases, the majority of breast cancer cases stem from an ensemble of mutations in two genes — BRCA1 and BRCA2 [58, 59]. There are unique groups, like the Ashkenazi Jewish population, where there are only a couple of common mutations in BRCA1 and BRCA2, which makes it easier to test. However, in most of the population representing more outbred groups, one must scan the whole gene to gather clinically helpful information. Already in the 1990s, we knew 200 risk-inducing mutations scattered all over these two very long genes (110Kb and 80Kb, respectively). And those do not convey categorical information on the occurrence of breast cancer, but rather a probabilistic estimate like 45% or 75% higher risk than the baseline. To be clinically valuable, one needs a test that covers the entirety of the genes.

In the 1990s, the mentioned RFLP method was based on early genetic fingerprinting. Naturally, a common theme among different genome fingerprinting approaches ran along enzymatic digestion of the DNA sample, followed by an adaptor ligation at the restriction sites and amplification with a common primer. One method called Amplified Fragment Length Polymorphism (AFLP) uses a specific set of primers to amplify genomic loci [60]. This approach was successfully employed by TSC, albeit its methodological limitations required a lot of automation and laborious sample preparation, thereby preventing genotyping on a large scale. In 2003, Kennedy et al. described a technology behind constructing large-scale SNP arrays, or chips — Whole Genome Sampling Analysis (WGSA) [61]. Building on the millions of identified SNPs, this array allows rapid interrogation of both alleles for over 10 thousand SNPs.

Further enlargement of databases and improvement of this technology increased the number of SNPs on an array, from 100,000 by 2004 and later to

500,000 SNPs [62]. Most importantly, when designed, these chips can be mass-produced and, as a natural consequence, decrease the price of genotyping. First, this technology allowed affordable genotyping (an estimate from the 1990s was around 10 to 20 dollars per test for breast cancer) for diseases of higher genetic complexity in the clinic. Second, it was now feasible to truly genotype large cohorts for association studies of qualitatively new kinds. The stage was set for the appearance of GWASs, association studies that survey most of the genome for correlated genetic variants [63].

Unlike approaches focused on identifying rare mutations responsible for Mendelian diseases, GWAS allowed researchers to investigate common genetic variations responsible for complex traits and diseases along the whole genome in large groups of individuals. By comparing individuals with and without a particular trait, GWASs identified thousands of genetic variants associated with everything from height and weight to cardiovascular diseases and cancer [64, 65]. Early GWASs led to significant advances in the latter by identifying genetic variants associated with increased risk for various types of cancer. For example, a landmark GWAS study published in 2007 identified common genetic variants associated with an increased risk for breast cancer [66]. An international group stretching across continents, including several research institutions in US and UK, conducted a multi-stage GWAS: first, 4398 breast cancer cases and 4316 controls were genotyped to scan for SNP candidates, and then 30 selected SNPs were validated on a cohort of 21,860 patients and 22,587 controls. This colossal scale was allowed mainly by the advances and expanse of microarray technology (in this study, a 550 thousand human SNP array by Illumina was used) and impressive data-sharing initiatives of the Breast Cancer Association Consortium (BCAC). The genotyping data for the SNP confirmation stage was pulled together from 22 studies. Before this research was published, the known susceptibility genes — such as BRCA1 and BRCA2 — accounted only for a quarter of the hereditary risk of breast cancer [67]. While previously studied genes corresponded to DNA repair, the new candidate genes from the study - FGFR2, TNRC9, MAP3K1 and LSP1 — were more related to cell growth and signalling. Breast cancer as a complex disease proved to be more complex than anticipated. Nevertheless, the study demonstrated the power of GWAS to identify genetic variants associated with complex diseases like cancer. It identified novel associated genes that provided new avenues for researching the causes behind breast cancer.

This success spurred a great deal of attention, and more studies followed. Genomics of cancer was still mostly an uncharted territory and benefitted greatly from GWASs. In the two years from 2007 to 2009, other studies have identified novel genetic variants associated with an increased risk for colorectal, prostate, and lung cancers, among others [68, 69, 70]. In 2009,

the group behind the first breast cancer GWAS refined the tentative genetic association identified in the original GWAS through a large replication study. Combining data from BCAC and Cancer Genetic Markers of Susceptibility (CGEMS) collaborations, researchers pulled together 37,012 cases and 40,069 controls, surpassing the initial research's power. While demonstrating the promise behind large-scale association studies, the authors hinted at the limitation that plagues association studies: "The power to have detected these associations with this strategy was still limited, suggesting that other breast cancer loci should be detectable by further large GWAS". The expectation has been that the enlargement of the sample size will detect more true cancer driver genes, i.e. increased sensitivity, while better separating the true signal from the background noise, i.e. increased specificity. We will see later that, particularly in the case of cancer, this line of thought led to results rather opposite to expectations [71].

Nevertheless, cumulative findings of early GWASs challenged the prevailing view that cancer susceptibility was driven primarily by rare, highly penetrant mutations. They highlighted the importance of common genetic variation in cancer risk. Even more valuable were the millions of cancer-associated genomic aberrations, SNPs, and CNVs that were identified and deposited into public databases to be readily used in future studies.

2.6 Next Generation Sequencing (NGS)

Sometime in 1970, in a debate over creationism, an evolutionary biologist Theodosius Dobzhansky memorably said: “Nothing in biology makes sense except in the light of evolution” [72]. And what is a genome if not the record of evolution? Creation of the methodology to read the genome or, more commonly, to sequence the DNA brought unseen prospects to biology. The history behind the development of DNA sequencing was rich and full of wonders, and it had its trailblazer [73]. Since its inception in 1977, Sanger sequencing has long been the dominant approach in DNA sequencing [74]. Its position was further solidified with the development of the first automated Sanger sequencer by Applied Biosystems in 1986 [75]. Decades of gradual improvement guaranteed exceptional accuracy over read lengths up to 1000 base pairs (bps), making it a sure shot for the tasks that necessitate fidelity.

These qualities keep Sanger sequencing in use to this day despite its modest throughput — an automated Sanger sequencing machine can generate around 100 Kilobases (Kb) of sequence per run, largely making it tedious and expensive for massive sequencing [76]. While both Sanger and NGS sequencing methodologies rely on the elongation of a single-strand DNA template, they are principally different in how the sequence is recorded. In that regard, Sanger’s approach is sequencing by termination. Elongation proceeded with modified di-deoxynucleotides that lack hydroxy-moiety at the 3’ position. The latter prevented the formation of the bond between nucleotides terminating the elongation and creating a truncated fragment with a labelled base in the last position, which is to be recorded.

Next-generation sequencing, on the other hand, mostly follows sequencing by synthesis (SBS). In that approach, every nucleotide added by a DNA polymerase initiates a fluorescent process that is recorded. The first attempts to pursue this principle were made in the early 1990s by researchers at the Royal Technical University, Sweden [77]. They were working on a technology that would later be developed into ”pyrosequencing”. The initial approach employed luciferase that emitted light upon nucleotide addition. However, it lacked throughput because, after each addition of the nucleotide mixture, the solution must be washed away before the next nucleotide mixture is added. This bottleneck was resolved some five years after by the addition of an enzyme apyrase, hence the name “pyrosequencing”, that degraded the remaining nucleotides in the mixture, allowing to keep the enzymes from the initial mixture [78].

Finally, by 2005, a microfluidic-based, fully automated pyrosequencing apparatus was showcased by sequencing and de novo assembly of a full genome of bacteria Mycoplasma genitalium (550kb) in one run [79]. When commercialised, a standard pyrosequencing machine was capable of

sequencing up to 120 Megabases (Mb) per run [76]. More importantly, it demonstrated reliable sequencing accuracy with short reads; for pyrosequencing, they are within 100 to 300bps, something to be shared among all second-generation sequencing methods.

Years following the publication of this technology saw an explosion as companies continued to innovate and improve their platforms. In 2005, another method was published called polony sequencing, a portmanteau to polymerase colony [80]. This technique implemented sequencing by ligation principle that proceeds in sequential rounds of hybridisation and ligation of different nonameric oligonucleotides. In detail, each nonamer is tagged by one of the four colour-coding fluorescent dyes that distinguish a central base of the nonamer (A, C, T, or G). A tagged nonamer mixture is added to the ligation reaction, where a successfully ligated nonamer - exactly matching the DNA template — records the central base of the sequence. The ligation cycles are then completed until the template is decoded. Polony sequencing was commercialised in the same year by Applied Biosystems as a supported oligonucleotide ligation and detection (SOLiD) system. It improved throughput, capable of sequencing up to 1.2Mb per run but suffered from very short reads (35bp) [76].

There was another SBS technology, which built upon the success of microarrays, developed by a company called Solexa. Their method followed a slightly different approach and aimed at sequencing of individual DNA molecules that were covalently bound in clusters on a planar surface of a microarray. The clustering ensued an in situ amplification of the DNA templates via a pair of primers bound to an array [81]. Sequencing was then carried out by adding tagged reversible terminators and DNA polymerase, which resulted in adding one nucleotide to the template and the emission of the fluorescent signal depending on the added base. The signal is recorded, the block is removed, and the cycle repeats until all the templates are sequenced. In 2006, Solexa launched their first sequencer apparatus — Genome Analyser (GA) - capable of generating approximately 1.3Mb of sequence with 40bp reads in a single run [82].

Shortly after, in 2007, Solexa was bought by Illumina. The GA line was expanded with the release of GA II and GA IIx, further improving on the capabilities of the original, now allowing sequencing of the paired-end reads with up to 75bp length. The explosion of DNA sequencing technologies was an enormous leap forward for biological research. In a matter of days, whole genomes could now be sequenced on the base of one facility. A task that a decade ago required international collaboration. Gradually, the understanding of this capacity changed the scope in which future biological projects could be conceptualised. Far from the position where the data was

scarce, researchers now held a panoply of datasets at hand, the potential of which only seemed to be restricted by the analytical capacity of that day's computers [83].

2.7 Ribonucleic Acid Sequencing (RNA-seq)

As much as RNA is inseparably tethered to DNA through transcription, DNA is likewise connected to RNA through a reverse process unambiguously called reverse transcription. A mechanism that founds the lifecycle of reverse transcribing viruses that carry an RNA genome while relying on DNA intermediate reverse transcribed from their RNA genome in order to replicate [84]. Alike to early sequence-based methods SAGE, CAGE, and MPSS, the application of NGS DNA sequencing to RNA — RNA sequencing — is based on the same process of reverse transcription [85].

As previously defined, the transcriptome is the "identity of each expressed gene and its expression level for a defined population of cells". Transcriptome serves as a functional link between the genome and the current phenotype of the cell or a group of cells, which was touched upon in the discussion on early cancer transcriptomics in section 2.4. Therefore, it is essential to record the transcriptome to characterise a cell's or tissue's current functional status and to understand disease or development in the same regard. The most popular methods for early transcriptomics have already been described: SAGE, MPSS, CAGE, and oligonucleotide arrays. The latter drove an early expansion of transcriptomics but also suffered certain drawbacks. First, arrays relied on a set of hybridised sequences predefined by already assembled genomes, thus restricting the de novo assembly of a transcriptome and making the search for new isoforms tedious. Additionally, arrays suffered from background signals coming from cross-hybridisation [86]. RNA-seq methods showed improvement over microarrays in the majority of the drawbacks of the latter. To better illustrate, the author will summarise a standard RNA-seq library preparation protocol for coding poly-A selected RNAs as performed in 2008 [87]. After the poly-A selection, RNA molecules are digested by hydrolysis to the average size selected by the researcher. Digesting RNAs into smaller fragments breaks down RNA secondary structures and reduces the cDNA amplification bias in the next step. After RNAs are sheared, they are reverse transcribed and amplified by random priming. The generated library is then sequenced on a DNA sequencer. The first thing to catch the eye is that in the case of poly-A RNA, RNA-seq captures the snapshot of the transcriptome together with the sequences of the coding regions and UTRs, allowing the discovery of new isoforms, i.e. splicing analysis [88]. Additionally, with relatively deep coverage, genetic variations like SNPs could be called directly from RNA-seq data [89]. Last, RNA-seq greatly improves the dynamic range of estimated gene expression compared to microarrays. In the former, dynamic range is limited to the number of probes dedicated to a specific feature, usually ranging between one to two hundred, while the depth of sequencing generally borders the dynamic range of RNA-seq; for example, in the study of yeast

transcriptome with RNA-seq that sequenced 16 million reads the dynamic range was 9000 [90].

Nevertheless, the biggest advantage of RNA-seq over microarrays is the interpretability and integrability of the results between different experiments run over different tissues, cells, and even model organisms without sophisticated normalisation methods [91, 92][90, 91]. Nevertheless, as with every new technology, it comes with its challenges. First, the effects of conditions of library preparation were not fully explored. Some earlier approaches, instead of digesting RNA, proceeded directly to reverse transcription and only then followed with DNA digestion. This tends to create an uneven coverage profile over a transcript with a heavy 5' bias [90]. Moreover, preparing stranded libraries that reliably detect transcripts from overlapping regions of opposite strands was tedious to produce [92].

On the other side, there are computational challenges. First of all, due to the immense amount of data generated, data formats in use were not feasible anymore. The development of the now universal SAM (Sequence Alignment/Map) format and its binary equivalent, BAM, largely solved this problem by 2009 [93]. Second, traditional alignment tools like Blast and Blat couldn't cope with the scale of generated data [94, 95]. Other more efficient methods struggled with aligning short reads that were signature of NGS RNA-seq [96]. Similar to the age of microarrays, the rapid development of computational software followed the explosion of generated data, but not many of those stood the test of time [97, 98, 99]. It is worth mentioning those still in use today, such as BowTie, TopHat and STAR, the latter of which arguably remains the most popular alignment tool for standard RNA-seq analysis [100, 88, 101].

Overall, RNA-seq proved to be superior to microarrays for transcriptome analysis. Initial afflictions that restricted access to RNA-seq were solved throughout the second decade of the 2000s allowing the new method to gain momentum as more and more laboratories gained access to this approach. By the end of 2015, RNA-seq surpassed microarrays in popularity and effectively started a new era in transcriptomics (Figure 1).

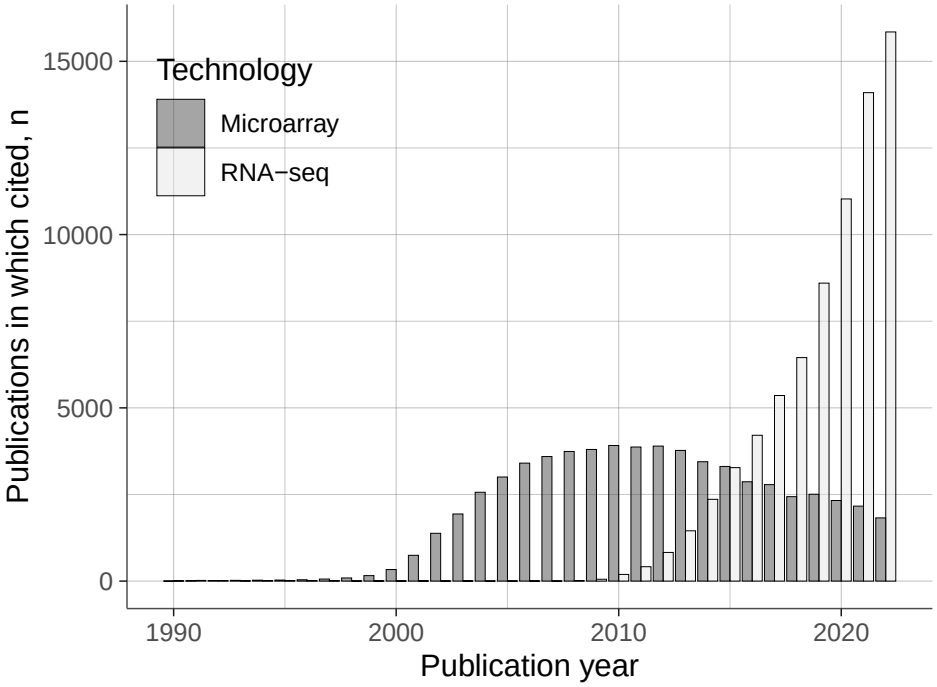


Figure 1: Expansion of RNA-seq technologies
Histograms showing the number of publications per year referencing either oligonucleotide microarrays or RNA-seq for gene expression assays from 1990 to 2022.

2.8 Human Genome Projects of Cancer

The success of the iterative and distributed approach of the Human Genome Project provided a vantage point to plan future large-scale projects in biology. When the human genome was finished, the interest in the cancer genome was second to none. As was quoted earlier, HGP was designed to make a reference genome that would be "where we are all the same". Now, the aim was to find genomic signatures that are alike within a cancer type. However, even within the same type of cancer, this is not an easy task to tackle due to the inherent mutational nature of cancer. Although we know that cancerogenic mutations, i.e. driver mutations, tend to accumulate in specific driver genes shared within the same cancer type, they only explain a fraction of genomic variability. They are accompanied by other mutations [102, 103]. Within each tumour are subpopulations of cancerous cells, each carrying in their genome a mutational record of clonal evolution to a malignant state [104, 105, 106]. Together with driver mutations, this assembly of genetic aberrations is called the mutational signature of cancer. It would be reasonable to assume that tissue microenvironments, where cancer originates, put specialised evolutionary constraints and favour specific mutational signatures. This would imply that many more than a single path to malignancy exists for a cell, i.e. a multitude of different mutational signatures can lead to the same malignant phenotype.

Consequently, it would take many cancer genomes to charter a comprehensive map of cancer signatures. With the distribution of high-throughput of sequencing, maiden steps were readily taken in this direction. In 2006, a consensus coding sequence derived from 11 breast and 11 colorectal cancer cases was published [107]. The majority of the reported mutations were single-base substitutions. 81% corresponded to missense, 7% to antisense and 4% to splice site alteration, while the remaining covered insertions, deletions and duplication events. Interestingly, although the fraction of single base substitutions was the same in both cancers, the nucleotide context was drastically different. More importantly, research manages to identify the genetic (mutational) basis for the genes previously suggested by earlier transcriptomic studies [108]. Further sequencing of different breast cancer genomes identified more and more novel somatic rearrangements [109, 110]. In 2007, a combined study of both genome and transcriptome of Acute Myeloid Leukaemia (AML) was published [111]. This study is of particular interest because of its overall controlled design.

First, to ease the identification of novel mutations, investigators selected a case that carried a remarkably normal cytogenotype, i.e. no inter- or intra- chromosomal translocation events were detected, limiting the background mutational burden. Second, investigators collected samples from a primary tumour, relapse tumour, and skin from the same individual,

thereby controlling for germline and background mutations, and sequenced all three in parallel. In the analysis, a total of 63 thousand tumour-specific SNVs in coding regions were recorded. And of those, ten non-synonymous somatic mutations are specific to the cancer genome. Two mutations were well known, and eight others were unknown, undetected by array-based methods. Half of the novel mutations in metabolic pathways were previously not associated with cancer by curated cancer genes. Another exciting study gathered patterns of genomic alterations signature to lung cancer with tobacco exposure [112].In contrast to previous studies, here, researchers expanded the scope of their genome screening outside coding regions and surveyed regulatory parts of the genome. To summarise, NGS allowed unbiased and genome-wide mutational screens to be carried out routinely, accumulating cancer genomic data on an unprecedented scale. An explosion was akin to the early transcriptomics boom of the microarray era.

At the turn of the millennium, the director of Wellcome Trust Sanger Institute, then called Sanger Institute, Michael Stratton, suggested a project under the same name (Cancer Genome Project or CGP) to map out unique mutations that distinguish the human genome from the cancer genome [113]. By 2002, the project saw its first success in identifying the BRAF gene, mutations of which are responsible for the lion's share of melanomas, followed by several impactful discoveries in cancer genomics [102, 114, 115, 116]. Besides their research, CGP curated and accumulated the scientific literature on known cancer genes and mutations. Work started shortly after the publication of the just-mentioned study on coding sequences from Breast and Colorectal cancer genomes [107]. In 2009, the CGP team released COSMiC, the Catalogue of Somatic Mutations in Cancer, which covered 7734 published articles, including all genome-wide screens [117]. With the ever-accelerating speed of DNA sequencing, this undertaking rivals Gene Ontology's value. It provided the centralised resource with curating schemas that accumulated and kept accumulating the great expanse of "Somatic Mutations in Cancer". The latter was not the last data-sharing initiative of CGP. One year later, an International Cancer Genome Consortium (ICGC) was announced [118]. This time, the goal was not to curate but to coordinate cumulative efforts of various global data centres towards characterisation and centralised deposition of genomic data on 50 types/subtypes of human tumours that bear societal weight. Looking backwards, an approach taken by ICGC proved to be the right one. Instead of heavy investments in the generation of datasets, ICGC built infrastructure for data to be easily accessed by the scientific community, furthering this by hosting a web platform for quick data access and analysis via cloud computing [119].

Alongside Cancer Genome Project and across the ocean, another initiative was conceptualised with the goal as ambitious. With more and

more affordable DNA sequencing looming ahead, an idea of a "Human Cancer Genome Project" was proposed, assumingly hinting at the scale and impact of HGP, with the goal of "obtaining a comprehensive understanding of the genomic alterations that underlie all major cancers" [120]. Following HGP's footsteps, the initiative put the ideas of public access and data-sharing at the core of the inceptive project. On the eve of 2005, a pilot project was initiated under The Cancer Genome Atlas (TCGA); the goals are set to map lung, brain, and ovarian cancers. Conceived as a large-scale project at its core, TCGA funded many genome centres to generate and process data in a unified format [121]. For the pilot, samples from selected tumours were to be screened in parallel for mutations, copy number variants and gene expression profiles. Herein the particular value of this undertaking — a multi-omic approach to cancer studies, in this case, profiling genomic and transcriptomic data from the same sample. In 2008, the first fruits were harvested, and the consortium published Glioblastoma characterisation [122]. Although the initial glioblastoma dataset was generated using pre-NGS technologies, the largely successful publication allowed the project to attract additional funding and guarantee its operations for years. More importantly, seven more new genomic sequencing centres were established, outfitted with NGS sequencers, allowing the application of the multi-omic approach on a truly grand scale. A plan was charted to cover 20 more cancer types and reprocess already collected samples, including whole genome sequencing (WGS), whole exome sequencing (WES), CNV and SNP profiling, DNA methylation, micro-RNA sequencing, RNA-sequencing, and proteomic data from RPPA arrays for a selected set of biomarkers. Three years after the initial publication, TCGA gathered 5000 cancer samples in their repository and followed the charted strategy of publishing a comprehensive study on ovarian cancer, already using NGS technologies [123].

The coming years saw a stream of publications covering six more cancer types [124, 125, 126, 127, 128, 129]. Alongside the publications, new data was readily deposited to the publicly accessible data portal in operation since 2012. The most intriguing, of course, was the now real possibility of performing a pan-cancer analysis to search for shared genomic and transcriptomics features across different cancers. It was already hypothesised that tissue lineage and surrounding microenvironment are likely to favour similar patterns of clonal evolution in different cancer types [130]. Next year, a pan-cancer project was drafted to "gain analytical breadth—defining commonalities, differences and emergent themes across cancer types and organs of origin" [131]. Although studies on individual cancers showed that most genomic aberrations are unique for a cancer type, they tend to be somehow related to similar pathways, i.e. hallmarks of cancer. However, mutations in the same pathway sometimes result in opposite effects in different tissues, and the prime example is the NOTCH family of genes, which are

inactivated in some squamous cell carcinomas but activated in leukaemia [131]. The idea behind integrating multimodal datasets was twofold: first, it would identify some convergence point to which different cancers gravitate; second, the bird-eye view at the micro cancer verse would better highlight the uniqueness of each tumour type. The overarching aim of uniting these two somewhat opposite tasks is to locate markers that would allow us to move further from histological classification. Strictly speaking, pan-cancer analysis is a clustering problem. By 2014 it was done and published, largely confirming many of the hypotheses but failing to identify the convergence point [132]. Methodologically, the analysis followed a rigid design: for each data modality, samples from all 12 tumours were pulled together and stratified by hierarchical clustering, using appropriate metrics for each data type. Then, samples were super clustered using the Cluster of Clusters Assignment (COCA) algorithm, which is a weighted hierarchical clustering based on previously assigned clusters with weights assigned depending on the total number of clusters in each variable (data modality) [133, 134].

Ultimately, this approach recreated the tissue architecture of the original histological annotation while successfully grouping squamous-like tumour types from the lung, bladder, head and neck [131]. The parallel analysis allowed better transcriptomic and genomic characterisation of squamous-like cancer phenotype. Looking backwards, the study carried certain limitations. First, the pan-cancer clustering did not utilise multi-omic modalities altogether, rather clustering them individually and using the obtained cluster vectors onwards. Given the nature of the data, it is reasonable to assume that some features in each data modality correlate, representing a flow of information from the genome to transcriptome to proteome. Therefore, independent clustering within each data modality effectively cuts this link. While maximising the between-cluster variance within each data modality, the links between data types are not considered resulting in suboptimal multi-omic clustering. Second, this study was likely limited by the linear nature of the clustering metric used. Re-clustering with a non-linear metric, as an example of support vector clustering with a non-linear kernel, could potentially recover a finer structure within the data. The other limitation comes from the innate heterogeneity of tumour samples, otherwise known as cellularity. In the case of even the most careful resection, the cell composition of the sample is not uniform and transcriptomic analysis was likely to be biased by the RNA from normal/healthy organismal cells residing in a tumour microenvironment [135]. Most of these shortcomings were addressed in the next iteration of pan-cancer analysis published four years later that further surpassed the scale of the initial incorporating a total of ten thousand tumour samples across 33 cancers [136]. In this instance, the authors used an integrative clustering approach iCluster that simultaneously optimises cluster structure across data modalities using Expectation Maximisation (EM) algorithm [137].

Cellularity was also considered; expectedly, a cellularity-driven structure underlying clusters suggested a heterogeneous cellular composition existed. Finally, the integrative clustering converged to a 28-cluster solution; among those, only a third were mono-tissue, and the other two-thirds corresponded to mixed-tissue subpopulations, greatly adding to the results of the previous analyses. Particularly peculiar are the dissimilar immune signatures expressed in some mixed clusters hinting at potential targets for immune modulation.

Another instance of particular interest is the second flagship paper, where the consortium assayed the alterations of ten oncogenic pathways across the same dataset of 33 cancer types [138]. There, great work was done to characterise the signatures of pathway alteration across this massive data compendium as comprehensively as possible. Pathway analyses from reports on individual cancer types were curated alongside public databases to refine the gene sets used to score the pathway in each tumour sample. Summed together, how a resource like this can be of use? This carefully curated data can be drawn into a map of actionable drug targets, often covering co-mutated oncogenic pathways. This brings to mind two possible ways to utilise this resource: first, it can be used as a base for the design of a test platform for tumour profiling; second, it can be of great help to customise a multi-drug cocktail specific to a tumour to maximise the efficiency of therapy. A project as massive as TCGA requires a foreword. Yet, skimming the publications accompanying this compendium over its ten years of history, the author struggles to narrow it down to a uniform conclusion. Undoubtedly, the contribution of publicly available uniformly processed multi-omic datasets to the scientific community is paramount. The TCGA consortium publications have been cumulatively cited roughly fifty thousand times [139]. Yet, cancer proved to be more complex than we expected. Even when integrated by a large consortium, an immense corpus of data seems to produce more questions than answers. As Robert Weinberg wrote some four years before the conclusion of TCGA: “The data that we now generate overwhelm our abilities of interpretation, and the attempts of the new discipline of “systems biology” to address this shortfall have to date produced few insights into cancer biology beyond those revealed by simple, home-grown intuition” [140].

2.9 Capturing the complexity with cancer models

The paragraphs above focused on the analysis of patient-derived tumour samples and the translation of tumour genomic and transcriptomic data into potential knowledge that can be utilised to combat cancer. There are, however, specific restrictions on the extraction of therapeutic insight from genomic data imposed by the nature of tumour biospecimens. Tumour samples must be resected and then stabilised before being stored in a biobank for profiling. Stabilisation often involves freezing and fixation in formalin, which carries on specific issues [141]. First, even in the most standardised conditions, tissues experience some degree of degradation [142]. Second, the assays are done, in essence, on dead tissue, which is largely unimportant for genomic characterisation but disallows screenings for drug sensitivity. Therefore, testing for drug sensitivity requires a model system to function as a proxy for a tumour. To this end, immortalised human cancer cell lines represent a centrepiece of preclinical cancer research and have been widely used to model tumour response to anti-cancer therapies [143, 144, 145]. Notably, this practice was pioneered in the early nineties with a panel covering sixty cancer cell lines called NCI-60 (National Cancer Institute 60) [146]. In fact, at the turn of the millennium, nearly 60,000 anti-cancer compounds were already screened against the NCI-60 panel [147]. Despite the large-scale pharmacologic characterisation of the cancer cell lines (CCL), these data were rarely generated in synergy with genomic and transcriptomic profiling, mainly utilised in studies on individual cancer types. However, with the increasing accessibility to sequencing technologies, generating large-scale pharmacogenomic datasets became a feasible enterprise. In 2012, two large compendiums of CCL genomic data were released side by side. The first one, Cancer Cell Line Encyclopaedia (CCLE), in its initial release, recorded pharmacogenomic profiles over 947 cancer cell lines encompassing 34 cancer types. All cell lines were profiled for mutations, CNVs, and gene expression with microarrays. In parallel, 24 anti-cancer compounds were tested on the larger part of the dataset (500 cell lines) [148]. Following a similar microarray-based design, the second dataset Genomics of Drug Sensitivity in Cancer (GDSC), went public with pharmacogenomic profiles of 639 cancer cell lines over 130 drugs [149, 150]. The main advantage of these systematic studies is that interrogation of multi-omic data over multiple CCL lineages, i.e. cancer types, allows for pinpointing the dependencies that are shared in a specific CCL lineage or by CCLs that carry a specific mutational signature, all of which would be impossible to detect in the individual CCL datasets. The initial forays into large-scale pharmacogenomic studies readily bore results. In pharmacogenomic studies, genomic and transcriptomic features can be mapped to a potential drug-gene (drug target) pair, allowing prediction of drug response from this signature. Expectedly, in many cases, known genomic and transcriptomic features directly associated with a drug target

appear at the top of the predictor's list, like mutations and overexpression of EGFR that lead to a heightened response to Erlotinib, a selective EGFR inhibitor [148]. There are, however, an abundance of peculiarities. For example, overexpression of the SLFN11 gene leads to a stronger response to topoisomerase inhibitors, particularly in Ewing's sarcomas [148]. Similarly, the GDSC core group mapped an Ewing's sarcoma characteristic genomic EWS-FLI1 rearrangement to sensitivity to a PARP inhibitor Olaparib [149]. Based on the initial successes, it would be only reasonable to consider that a more thorough genomic characterisation would uncover more dependencies between drug response and genomic features.

Nevertheless, despite all their positive sides, CCL assays are still tedious to perform *en masse* because CCLs cannot be mixed and therefore need to be cultured individually. Since the publication of the initial CCLE in 2012, a pooled strategy approach to simultaneous cell line profiling was conceptualised. This methodology relied on introducing specific barcodes to the genome of individual CCLs that could later be cultured in a mixture. The presence of a barcode allows for a quantitative read-out of the cell number that is proportional to the barcode signal. By 2015, the concept was reshaped into a proper high-throughput platform Profiling Relative Inhibition Simultaneously in Mixtures (PRISM), and the proof of principle paper was published in 2016 with promising results [151]. Although there are concerns regarding the interactions of the heterogeneous CCLs via paracrine and juxtacrine signalling, these are largely outweighed by the high-throughput capacity of the platform. To compare, in their initial releases, CCLE and GDSC screened 24 and 130 compounds, respectively, over roughly five hundred CCLs. PRISM profiled the responses of 103 CCLs to a stunning 8,400 compounds in its initial release. This technology was further employed in large screenings of non-oncology-related drugs for potential repurposing into anti-cancer therapies [152].

Another improvement could be made to the methodologies of dependency screenings themselves. So far, we have discussed only the drug screens where the drug effects are usually reported in three different values: half maximal inhibitory concertation (IC50), half maximal effective concentration (EC50), and activity area. All these metrics are estimated from a drug response curve that reflects the dependency between drug-induced growth inhibition relative to negative control and the concentration of the administered drug. Therefore, IC50 is the drug concentration necessary for half of the possible maximal biological response, i.e. half to the negative control. EC50 records the drug concentration corresponding to half of the recorded maximal biological effect, i.e. half to the inhibition biological response of the largest administered dose. Last, the activity curve is the area above the response curve. In this approach, drug-gene dependencies are sought

from the perspective of the drug. The alternative approach would be to look from the perspective of the gene, i.e. to conduct functional studies. To this end, a project Achilles was established that leveraged genome-wide loss of function screens using CRISPR-Cas9 and RNAi screens [153]. The ultimate shared contribution of all these initiatives is building and maintaining the infrastructure for public access to the respective databases. Since 2014, they have operated through a unified infrastructure under the Cancer Dependency Map consortium (DepMap) [154].

Under the umbrella of the DepMap consortium, another project was not discussed — Cancer Cell Line Factory (CCLF). While all aforementioned initiatives research methodologies and characteristics of CCL models, CCLF is fully dedicated to developing and characterising new cancer models. Since the establishment of DepMap, in its joint operations, somewhat 1500 CCLs have been cumulatively developed, characterised, and made publicly available, while somewhat 1000 CCLs exist in private collections [155]. In the earlier stages of contemporary cancer research, this number seemed sufficient. Yet, the discoveries of TCGA and ICGC show that the expanse of genomic and transcriptomic complexity of cancer greatly surpasses the variants of those 1500 CLLs in use. The genetic aberrations in the existing models are insufficient to cover human cancer's heterogeneity comprehensively [155]. Therefore, CCLF took upon the task of developing cancer models from patient samples with selected backgrounds, including those with rare cancers that are underrepresented in the CCL collections [156].

Another part of this initiative was directed at developing qualitatively new cancer models. Normally, mammalian cells, even cancer-derived cell lines, are known for their limited proliferation capacity in culture and early onset of senescence [157, 158]. In older methodologies, the senescence block is bypassed by the immortalisation process, i.e. unlocking the cells' unlimited proliferative capacity. It is achieved by introducing viral oncogenes or exogenous telomerase, which alters the phenotype [159]. Advances in cell culture technologies allowed the rapid expansion of stable CCL cultures from as little as a needle biopsy retaining the tumorigenic characteristics of the original tumour [160]. The new approach bypasses the need for immortalisation. It relies on introducing a specific mixture of ligands and signalling molecules to the extracellular matrix (ECM) to reprogram the cells to maintain their proliferative capacity. Cells grown in this way retain the ability to differentiate and, under the right conditions, can form organotypic, three-dimensional structures commonly called organoids or "organs in a dish". Unlike two-dimensional in vitro cultures, organoids resemble the original tissue in its architecture and composition, harbouring subpopulations of differentiated cells in proportions somewhat similar to those of the living tissue [161]. Another beneficial quality of this model is logistical: organoids

can be expanded indefinitely and cryopreserved in biobanks. Therefore, organoids bridge the gap between cell lines and in vivo models striking a fine balance between the feasibility of operations and the biological proximity of the model to the real tissues, which makes them an attractive platform for cancer models [162]. To this end, CCLF collaborated with Human Cancer Models Initiative (HCMI) to create the first thousand next-generation cancer models (NGCMs), including patient-derived organoids and reprogrammed cells. So far, in the collection of CCLF are 648 newly generated cancer models derived from 501 patients. Of those, over a third are rare malignancies, and 278 are organoid models from a joint effort in HCMI.

Among NGCM, another up-and-coming model is a xenograft — a tumour growing in the body of immunodeficient mice, a xenopatient. Although facing competition from organoids, this in vivo model comes closest to replicating the biological complexity of an entire tumour [163]. Of particular interest is that xenografts could faithfully recapitulate the drug response in patients from whom the xenografts were derived [164]. Studies in patient-derived xenografts (PDX) of colorectal cancer successfully identified HER2 as a potential target in a subpopulation of resistant tumours [165]. Based on these initial studies, collaboration with industry was struck to generate an encyclopaedia of xenografts. The initial dataset covered 1,000 xenografts with full genomic and transcriptomic characterisation alongside pharmacological data for 62 compounds [166]. Despite the positive discoveries made with PDX, there are certain difficulties with analysing the PDX models. While histological parameters remain largely similar between the donor tumour and xenograft, one report observed drastic changes in the transcriptomes of xenografts, hinting at the potential post-engraftment clonal evolution of tumour cells within the mouse host [166]. In particular, expression of the genes annotated to immune- and ECM- related pathways was downregulated in xenografts compared to primary tumours, while cell division pathways were up-regulated. Given our accumulated knowledge of cancer, the adaptation of tumour cells to the organism of the "xenopatient" is not surprising. Yet, the degrees of the phenotypical changes are not properly characterised. One could hypothesise that the immunodeficiency of the xenopatient relieves the evolutionary constraints imposed on the tumour cells of the primary tumour by the immune system of the original host and therefore promotes a clonal evolution towards different phenotypes. This can lead to dissimilarities in drug responses between PDX and the primary tumours. A parallel genomic and transcriptomics profiling of donors and xenografts could potentially delineate common xenograft adaptation patterns to xenopatient to control for those in data analyses.

Alongside developing new models, efforts were put into deepening the multi-omic profiling of the existing CCL models. By 2019, CCLE amassed

more than a thousand cancer cell lines profiled in eight data modalities (WGS, WES, DNA methylation and chromatic profiling, mRNA expression and splicing, miRNA expression, and protein assays by RPPA) [167]. A year later, the CCLE dataset was complete with the addition of the last unexplored data modality - proteome [168]. All of the datasets discussed before are covered by the infrastructure of the Dependency Map Consortium that allows unified access. It would be worthy to say that at this point, the sheer size of the database turns the analysis task into a "needle in the haystack" like search. Nevertheless, in the publications supporting the releases of the updated datasets, the authors of CCLE well showcased how each added data modality allowed for uncovering a new, unexpected cancer dependency. Therefore, it would take years of computational analysis to interrogate such massive a collection of data exhaustively.

2.10 Reduced representation models

DepMap was not the only consortium that tried to connect genomic and transcriptomic features with the effects of pharmacological agents. Similar in goal but different in concept, another consortium was working to charter a Connectivity map (CMap) of connections "among small molecules sharing a mechanism of action (MOA), chemicals and physiological processes, and diseases" [169]. Subprojects of the DepMap consortium were unified in their approach to generating translational insight from cancer models, and this approach was twofold. First, to generate cancer models that recapitulate the original tumour as closely as possible regarding architecture, environment, and physiology. Second, to characterise cancer models as comprehensively as possible using different data modalities generated by multi-omic assays. This approach can be termed "maximalist" as it strives to achieve the most wholesome recapitulation of the disease in the model and the most comprehensive disease characterisation by biological assays, i.e. multi-omics [167, 168]. The benefits of this approach are apparent and many, while the limitations are ultimately tied to the cost of data generation and the computational limits of data integration and analysis. Worth to mention that the latter tends to improve swiftly, thereby ensuring the value of the DepMap datasets in future. CMap's approach, on the contrary, can be termed "reductionist" as it aims to achieve a minimalistic yet comprehensive biological representation of the disease [169]. But let the introduction be done with decorum and with order.

Conceptually, CMap draws inspiration from a seminal work of Hughs et al. from the era of early transcriptomics [170]. That work showed how a single assay, that being whole-genome gene expression profiling, over a large compendium (300 profiles) allows to connect perturbations, be it the administration of a drug or an unknown mutation, to a specific state of the transcriptome, thereby allowing the identification of the functional impact of a perturbation by similarity to transcriptional profiles in the compendium. To better illustrate, let's look at the example of drug target identification from the original publication. Among the profiles in the compendium was a yeast culture treated with a common topical anaesthetic, Dyclonine [171]. The expression profile of the dyclonine treated yeast culture was strikingly similar to the expression profile of erg2 mutant yeast, suggesting the inhibition of erg2 by dyclonine. The human protein with the highest sequence similarity to yeast Erg2 protein is a sigma receptor, a human neurosteroid interacting protein that regulates potassium conductivity and binds several neuroactive drugs, thus suggesting it to be a target of dyclonine [172, 173]. This concept could be similarly applied to human gene expression signatures corresponding to genetic and pharmacologic perturbations, as similar signatures can disclose previously unrecognised connections between MOAs of two small molecules,

a small molecule and mutation or two mutations that perturb the same pathway.

CMap leveraged the non-parametric approach conceptually similar to the GSEA to assess the connectivity between gene expression profiles unambiguously. Every perturbagen in cell culture has its gene expression profiled in parallel to negative vehicle control, generating two gene expression datasets: perturbed dataset A and control dataset B. Then, each gene in the expression profile is ranked according to a difference between A and B using an appropriate metric, creating a rank-based gene expression "ladder" specific to the perturbagen. The whole dataset is then amassed from the individual perturbagen-control "ladders". These ladders are used to test for a biological state of interest, i.e. a gene signature of interest, which is measured against the ranks of the genes in a ladder to generate a connectivity score derived in the same way as the enrichment score in GSEA; the derivation of ES was outlined in section 2.3. For example, in the original publication, the authors showcased a test of a 13-gene-long signature of histone deacetylase (HDAC) inhibitors derived from another study on CCLs [169, 174]. When tested against the ladders, the Connectivity map revealed strong transcriptomic connectivity of this signature to the two structurally distinct perturbations, HC toxin and valproic acid, known for their HDAC-inhibitory qualities [175, 176]. If the gene signature, however, was not pre-emptively characterised, CMap would have uncovered that connection. The initial CMap release covered 164 perturbagen in 3 CCLs [169]. Over the years, the platform assisted with drug repurposing and generating new therapeutic hypotheses. Notably, CMap was employed to identify the potential of an anthelmintic drug, parbendazole, as an inducer of osteoclast differentiation and Celastrol as a treatment for obesity [177, 178]. That being said, CMap also performed well in some cancer studies [179, 180]. While the initial CMap proved the concept's applicability, the small scale of the dataset restricted its utility: it lacked a variety of cell types as well as pharmacological and genetic perturbations.

The next edition of CMap was released more than ten years later, commonly known as CMap2 or CMapL1000 [181]. Although it aimed to expand on all aspects of the earlier iteration, this dataset's peculiarity lies in the way the data was generated: as much as the dataset was expanded, the transcriptional redout was "reduced". The authors brought up the "reduced" transcriptome concept to break off from the cost limits imposed by extensive transcriptomic profiling. At its core, the idea was that most of the information recorded by a snapshot of the transcriptome could be narrowed down to a handful of information-heavy transcripts, referred to by authors as "landmark" transcripts, whilst the rest of the transcripts are largely non-informative and could be discarded. The authors pulled together

twelve thousand gene expression microarray datasets to identify the landmark transcripts, all using the same Affymetrix HGU133A platform to minimise the platform-dependent variability. The transcript recovery then proceeded: a combined, centred and scaled gene expression dataset was reduced to 386 principal components, cumulatively accounting for 90% of the variance. The authors did not report the initial number of transcripts in the combined dataset, but we can hypothesise with a decision that it covered from ten to fifteen thousand transcripts. Then, the transcripts were clustered in the reduced subspace of 386 principal components using a variant of k-means clustering called "tight clustering" [182]. In detail, the dataset was randomly subsampled with resampling into 100 partitions, with each partition covering 75% of the original dataset. These partitions were then each clustered by k-means with 20 to 100 centroids. The final output of the procedure yielded a consensus matrix that recorded the proportion of trials in which each pair of genes were in the same cluster. Genes surpassing the threshold of 80% of trials were recorded into the "landmark" transcripts. Finally, this analysis showed that as few as one thousand landmark transcripts could capture 82% of the variance in the initial dataset. Recording the gene expression of one thousand landmark transcripts should technically provide a comprehensive image of the transcriptome. Building upon this idea, the authors designed a "minimal" oligonucleotide array system L1000 to profile the expression of 978 selected landmark transcripts. However, measurements of these thousand transcripts won't suffice for the end goal of the CMap2 simply because gene signature queries are very likely to be underrepresented among the landmark transcripts, which will obstruct the scoring of the queries. Therefore, the next step after L1000 profiling and normalisation is an inference of the transcriptome from the gene expression of landmark genes. The inference follows a rather straightforward approach: expression for the remaining part of the transcriptome is predicted using a collection of linear regression models with weights derived from training on the initial collection of microarray datasets. This approach performed surprisingly well and reliably recovered the expression of 9200 genes out of the 11350 genes inferred when tested on the GTEx dataset [183]. However, questions about its reliability remain and will be addressed later.

The validation analyses performed by the authors cogently stressed the reliability of Cmap2 connectivity scores. In one case, the authors queried signatures derived from tumour gene expression profiling in patients receiving MEK-pathway inhibitors. Consistent with the clinical data, Cmap2 assigned a high positive connectivity score to MEK-pathway inhibition signatures when drugs were administered and high negative connectivity in patients who experienced relapse after the treatment. According to the authors, this minimal approach allowed for the reduction of the cost of one gene expression profile to a mere two dollars (for reagents), which in turn allowed

the screening of a vastly larger number of perturbations, totalling more than twenty thousand small molecules and a little more than five thousand gene perturbations including both gain- and loss- of function, altogether summing up to more than 1.3 million L1000 profiles. Concordant with the concept of the authors of Cmap2, it found its application suggesting therapeutic hypotheses and support existing evidence in cancer studies [184, 185]. Notably, it was used in the pan-cancer analysis of stemness by the mentioned TCGA consortium (section 2.8) to look up compounds connected with the identified stemness signatures of cancer [186]. As of 2022, Cmap2 is integrated into the web-based resource iLINCS which aggregates multi-omic signatures derived from a vast collection of datasets [187].

Nevertheless, the author cannot ignore the potential shortcomings of the reduced approach to perturbation models. The data-driven process for selecting landmark genes tries to minimise the loss of information by reducing the dimensionality of the gene expression profile. Consider an ideal case where a size m x n matrix can be reduced to a lower-dimensional representation without losing information. The latter is only possible when some of the dimensions of the initial matrix are functions of other dimensions, which an appropriate dimensionality reduction technique can recover. In the case of Cmap2, approximately 9000 dimensions were recovered from a reduced representation of a thousand features fitted into an array of linear models. Considering transcriptional noise and the error of measurements, it is hardly arguable that this proxy of gene expression introduces a bias to inferred transcripts, which is likely to depend on the complexity of the inferred transcriptome. Although authors validated the reliability of L1000 to recover expression measurements similar to RNA-seq data, others reported limited reproducibility between the original Cmap and Cmap2 expression measurements and connectivity estimates [188]. Additionally, subsequent analyses highlighted that the reliability of the reference signatures depends on the strength of the differential expression induced by a perturbagen [188]. In short, Cmap2 tends to miss the small changes in transcriptional profiles while reliably recovering strong effects, an observation somewhat expected given the mental framework of the reductionist approach. The author hypothesises that more sensitive approaches can recover untapped information from the generated landmark signatures. To this end, there were attempts to improve the inference of unmeasured genes [189]. In that particular report, authors managed to infer the expression of approximately 22,000 transcripts using a fully connected neural network. Despite the dropping costs of RNA-sequencing, the production of millions of whole-genome RNA-seq profiles in a uniform setting would still be economically unfeasible, therefore more precise transformation of L1000 profiles into RNA-seq-like data would likely improve the usability of Cmap2 for future applications.

2.11 From Breadth to Depth — understanding cancer evolution

Decades of multi-omic assays of tumours brought many insights into cancer biology. Mutational signatures for diagnosis, thoroughly characterised molecular pathways, multi-omic defined subtypes accompanying histological classifications, prognosis predictors, and drug response markers were catalogued in databases of enormous scale. Building upon this accumulated knowledge, new non-invasive test systems were designed, which assess tumour status from mutational signatures detected in the cell-free DNA from blood samples [190]. Other collectives developed protocols to detect mutational signatures from tumour DNA recovered from saliva [191]. Of particular interest is the test panel cancerSEEK that utilises the combined assays of genetic material and protein biomarkers from the blood samples; initial results give high promises as the test performs reliably along eight different cancer types (successful median detection of 70% in all eight cancer types) [192]. Authors also remark on potential limitations of these approaches: cases of less advanced diseases might pass the test undetected, resulting in a higher rate of false negatives, while the unaccounted background processes like non-cancer driven inflammation could lead to a false positive result. The massive influx of genomic data expanded the clinical use of targeted therapies, exemplifying the use of EGFR inhibitors such as Erlotinib in non-small cell lung cancer (NSCLC) and KIT inhibitors in gastrointestinal tumours and myeloid leukaemia [193, 194, 195].

Similarly, widespread genomic characterisation of CCL screens promoted the repurposing of existing medications to the oncological field producing new and unexpected lines of treatments [196, 197]. A better understanding of molecular pathways hijacked by cancer cells allowed the development of combined therapies that simultaneously target multiple elements in a pathway [198, 199, 200]. In the United States, prognoses have improved yearly since 1991 across all cancer types, with overall 5-year relative survival reaching 70% as of 2015. Survival rates of thyroid, testis, prostate, breast, and skin (melanoma) cancers currently exceed 90% [201].

Nevertheless, lung, colon or pancreas cancer examples show a much grimmer picture. Lung cancer claims more lives than any other and still poorly responds to treatment, while metastatic tumours generally remain mostly incurable. A portion of deaths can be attributed to an untimely diagnosis, particularly for slow-growing tumours like pancreatic cancer that are usually detected already in an advanced stage where surgical resection is no longer possible. While some other tumours, even when diagnosed on time, possess near miraculous ability to adapt to any line of medication put against them. This ability stems from a continuous evolution process that

underlies cancer progression.

To illustrate, let us consider a case of multiple myeloma consisting of a clonal population with BRAF activating mutation and several subclones that acquired additional mutations in KRAS and NRAS. The administration of BRAF inhibitors would effectively kill a bulk of BRAF-carrying cells but, at the same time, would increase the relative fitness of KRAS and NRAS subclones by removing competition from BRAF mutant clones that harbour wild-type KRAS and NRAS [202]. In the end, the expansion of the new clonal population, potentially more aggressive, would nullify the clinical benefits of the therapy. Although the theory of clonal evolution was conceptualised in the seventies, in the past, only a few experimental surveys in this direction were attempted, primarily limited by the available methodology, by which the author means NGS (see section 2.6).

The current standard of whole genome sequencing used to assay tumour samples generates 100x coverage of the genome, orders of magnitude smaller than the number of cancer cells in the sample. Considering the admixture of normal cells from the tumour microenvironment (TME), under this sequencing depth, only the genomic profile of the most abundant clone would be reliably assayed. At the same time, the subpopulations harbouring additional mutations would largely remain undetected [203]. From the evolutionary perspective, the genomic snapshot of bulk sequencing informs us of the most recent common ancestor in the cancer cell population that is already extinct in the ever-growing malignancy. To better understand which methodological approaches are more suited to dissect the clonality of tumours, the author will outline what is currently known of the evolutionary processes that underlie the genetic heterogeneity of cancer.

Cancer cells generally adhere to the rules of evolution regarding how cells mutate, adapt, grow, and die off. The enormous population size ranging billions of cells and confined environment makes cancer evolution processes akin to those in bacterial populations. Fundamentally, they follow the evolutionary mechanics of asexually reproducing species: "replication, heritable variation, genetic drift, selection and environmental changes" [204]. Based on these mechanisms, different modes of cancer evolution are hypothesised: selection, neutral evolution, branching evolution, linear evolution and punctuated evolution. When, under specific environmental conditions, one lineage is favoured over another and produces larger and more viable progeny, it adapts this population. This process is the main driving force of evolution and is called selection. Generally, a positive selection that increases the frequency of the more adapted lineage is the process that fuels tumour progression [204]. Negative selection, on the other hand, is the evolutionary process by which phenotypically unfit lineages are removed from the popula-

tion. In case of cancer progression, it can, for example, cleanse the population from neo-antigen-carrying cells, as the attraction of immune response to the neo-antigens reduces their fitness [205]. As much as selection depends on an environmental context, it cannot always be operative. For example, if the environment does not favour differential survival within the population, the absence of positive selection would lead to neutral evolution, where only mutation and genetic drift are at play. Neutral evolution usually occurs in the intervals between selection events that either prune or expand the phylogenetic tree of the population [206]. Suppose a mutation arises during the stochastic mutational process that increases the fitness of a particular clone. In that case, it can initiate a clonal sweep, a process when a specific clone with fitness advantage is positively selected for and displaces other lineages leading to reduced population diversity. Therefore, a mutational rate itself could be subject to selection. As much as a higher mutational rate diversifies the population, it increases the chances of both positive and deleterious mutations appearing, whereby the tumour growth could be hampered. Therefore, a fine balance of chromosomal instability should be selected, as too much instability leads to autonomous cell mortality, while too little would restrict the generation of new and potentially advantageous mutations. Consequences of this selection could be observed even on a macro- level, when tumours with medium levels of chromosomal instability correlate with overall poorer survival prognosis, while extreme and low levels confer an improved prognosis [207]. Some mathematical models suggest that "mutator phenotypes" are selected in the population of cancer cells because the stochastic acquisition of the positive mutations leads to a two-edged benefit in fitness: from the advantages of the clone's positive mutation and the deleterious mutations appearing in the rest of the population [208, 209].

Although the evolutionary process is always branched, the presence of mutator phenotypes makes it particularly true for cancer cells. Constant cell divisions and mutations lead to constant diversification of the population, i.e. branching. Random fluctuations in some lineages' birth and death rates lead to the overrepresentation of specific genetic variants in the population, a process called genetic drift. Genetic drift is considered a form of neutral evolution, as all the lineages are neutral in their chances to produce a surviving progeny [210]. Interestingly, similar evolutionary patterns are observed in healthy tissue, suggesting that branching is a natural consequence of cellular proliferation [211]. In the case of cancer, however, due to the dysregulation of control circuits managing cell proliferation, when a subclone acquires a mutation in a driver gene, it can quickly expand over the population driven by positive selection.

Similarly, parallel evolution can occur when multiple subclones acquire mutations in the driver genes and expand simultaneously within the

same tumour [212]. In a mock contrast to this, linear evolution postulates that only one lineage survives over time. This, however, does not imply that there was ever only one single lineage evolving step by step. It is problematic to infer linear evolution from genomic data due to the limitations in the resolution of NGS technologies.

Finally, the last mode of evolution to be discussed focuses on the sudden changes in the genotype, called punctuated evolution. This mode posits rapid bursts of adaptation followed by long periods of relative stasis in contrast to stepwise evolution; theories that explain evolutionary dynamics through sudden and sizeable events are called saltationists, from the Latin *salto*, which means to jump [213]. In general, contemporary saltationist evolutionary theories, including punctuated evolution, share a conceptual reference point in a seminal paper by Gould et al., originally conceived to explain gaps in fossil records [214]. This theory, called punctuated equilibrium, suggests that adaptation occurs in small geographically segregated niches and largely remains undetected before the more fit lineage leaves the niche and disperses through the population. Henceforth, punctuated equilibrium posits long periods of stasis interrupted by short periods of change. This theory, however, was designed to explain population dynamics of sexually reproducing species and drew the idea of evolutionary puncta, i.e. sudden and short periods of adaptation, as a direct consequence of allopatric speciation. The fact that cancer cells are not a mating population and are not separated by geographical borders, at least in the primary tumour, makes the theory of Gould et al., to the largest extent, unfit to explain the evolutionary dynamics of cancer. Nevertheless, the idea of evolutionary puncta fits well with the mutational bursts observed in cancer. Such puncta can occur, for example, after extreme rearrangements of the genome, e.g. loss, gain, fusion, translocation of the chromosomes and ultimately chromothripsis or chromoplexy, that give rise to an extremely adaptive clone that can sweep through the population [215, 216]. Much more fitting to this non-Darwinian evolutionary process is a theory under the same name that terms such clone a "hopeful monster", where monster refers to the colossal genomic aberrations and hopeful refers to the meagre likelihood of the genomic alterations to be non-lethal [217]. Overall, the phenomena of punctuated evolution were observed in multiple cancers. However, it is difficult to say if this process is common in specific or widespread across all tumour types [218, 219].

Drawing from the mathematical abstractions of population genetics, many mathematical models were developed to replicate the evolution of cancer populations [220, 221]. Some specialised models could reliably predict the time of appearance of resistant clones after administering the anti-cancer drug, i.e. the selection and expanse of the adaptive lineage [222]. In that study, a group retrospectively analysed a temporal regiment of serum samples

from 28 patients with chemorefractory metastatic colorectal cancer. The samples were taken 7+/- 2 weeks and 25+/-10 weeks after administering EGFR inhibitor panitumumab and screened for KRAS mutations [223]. A branching process model built with data-derived parameters predicted an appearance of resistant clones on average after 22 weeks after the administration of EGFR inhibitors, with a 95% confidence interval from 18 to 25 weeks. Other studies introduced a spatial component into the modelling of tumour growth. The model's inspiration was built upon observing growth patterns of hepatocellular carcinoma, which is shaped into "balls" of cancer cells separated by normal tissue [224]. Simplification of the 3-dimensional space of tumour growth into a sphere allowed a feasible opportunity to model clonal dynamics, such as the resurgence of resistant subpopulations after treatment. As much as this model used parameters from the EGFR resistance modelling, it is unsurprising to see that the predicted time to a resurgence of resistant clones is per the previous model (one month and 22 weeks).

Peculiarly, the model also predicts higher speeds of tumoral growth and regrowth after treatment for cancer cells with the increased capacity to move and migrate, not necessarily to the distant parts of the host organism. Even slight cellular movement and dispersion of cancer cells lead to drastic changes in the predicted morphology and growth rate. To this end, there were reports that the loss of E-cadherin, the gene encoding a cell-adhesion protein that allows cells to adhere to each other, forming organised superstructures, is associated with the more lethal phenotype of pancreatic cancer [225]. In another line of research, building on observations in colon cancer, Sattariva et al. conceptualised a „Big Bang" model that posits a massive adaptive event followed by the expansion of intermixed clones happens early in the timeline of the tumour and is responsible for most of the detected mutational signatures, ultimately resulting in "star-shaped" phylogenies [226]. Although new mutations would be continuously generated during the expansion, their effect would not be pervasive, as they will be diluted in the growing subclonal populations. The methodology of this study adds gravity to the results: Sattiriya et al. profiled the genomes derived from the single tumour glands (¡10,000 tumour cells) that were sampled in different locations across 15 tumours, totalling 359 genomic profiles, together with bulk tumour profiles. Considering the spatial positions of individual glands, analysis of their genomic profiles against the profile of the bulk tumour allowed the reliable reconstruction of the evolutionary timeline. Concordantly with other previous reports, this analysis also showed an absence of clonal sweeps after the initial expansion, suggesting that the clonal expansion is a relatively infrequent event in the evolution of colorectal cancer [226, 227, 228].

Additionally, the model's framework explains why the potentially

more aggressive and less frequent clones remain undetected until the bulk of the population is wiped out by a selection event such as the administration of anti-cancer drugs. Nevertheless, the model has implications in the origins of metastasis; why do some tumours grow large but remain indolent while others actively metastasise in other regions? Interestingly, in the analysed panel of colorectal cancers, all invasive tumours had a variegated pattern of clonal expansion, i.e. clones of the same lineage are observed in distant locations within the tumour. Although this observation could be a result of an early scattering of the clones during the initial expansion, it might also be an early sign of an invasive phenotype, suggesting that some tumours can just be "born to be bad", as was hypothesised in the past [225, 229, 230].

This review should give a bird's eye to the current perception of tumour evolutionary processes. Evolutionary dynamics and mutational patterns underlying tumour growth, resistance, and metastasis are vital to developing a framework for combinatorial targeted therapies or, eventually, gene therapies. The utilisation of liquid biopsies is a prospective approach. Genomic profiling of circulating tumour cells and cell-free tumour DNA opens venues for designing personalised therapeutic regimens to tackle the unique mutational background of a tumour. Knowledge of clonal dynamics inferred from sequenced genomes could help create multi-layered therapies preventive to the expansion of mutated and resistant clones [231, 232].

However, understanding processes underlying tumour growth and formation is lacking, mainly because the abundance of generated genomic data rarely follows the same tumours temporally, instead recording genomic snapshots of the most recent common ancestors in different tumours [203]. For some time, the only approach to properly assay tumoral evolution processes was simultaneous genomic profiling of the same tumour from different locations, as implemented in Sattariva et al. [226]. The application of this approach culminated in the TracerX (TRAcking Cancer Evolution through therapy (Rx)) consortium that planned to massively widen the scale of multi-region genome profiling of tumours to track common and specific evolutionary patterns [233]. To this end, the TracerX consortium initiated four parallel research venues to assay clonal dynamics in different cancers: renal, lung (NSCLC), prostate, and melanoma [234, 235, 236, 237]. As of now, renal cancer received the most comprehensive characterisation; therefore, the author will accentuate that particular tumour to describe the results of this consortium.

As a foreword, let us briefly introduce renal cancer, clear cell renal cell carcinoma (ccRCC) in particular. This tumour is believed to arise from epithelial cells of the proximal convoluted tube of the nephron [238]. ccRCC contributes 5% of all new cancer cases in the United States, with around 81

thousand more estimated to be diagnosed in 2023, of which approximately a quarter would die from a disease [201]. Adding to the value of this case study, the genome of ccRCC has a distinctive feature — the loss of the short arm on chromosome 3 (3p) is detected in more than 90% of all cases [129]. This opens possibilities for reconstruction of the timeline of tumour evolution with significant precision by taking the 3p loss as the reference point.

Thereby, the first out of three ccRCC studies focused on reconstructing the evolutionary timeline [239]. To this end, whole-genomes from 99 multi-regional biopsies from 33 patients were sequenced and analysed. Concordantly with other public results, most of the genomes experienced the 3p loss. Closer examinations of genomic rearrangement surrounding the loss of the 3p, of which the most common pattern is the reshuffling of the 3p with the long arm of chromosome 5 (5q) that creates a hybrid chromosome t(3,5), implies their origin in chromothripsis. It is fascinating to see the evidence that the most ubiquitous cancer-initiating event is chromothripsis, in full accordance with the "hopeful monster" hypothesis [217]. Of particular interest is that reconstruction of the time of chromothripsis events from mutational signatures suggested them to occur early in adolescence, decades earlier than the emergence of the most recent common ancestor in the population. It indicates that the accumulation of driver mutations takes decades before the disease manifests itself.

Peculiarly, a Hippel-Landau disease caused by a germline mutation in the VHL gene is characterised by the early onset of ccRCCs. Genomic analysis of ccRCC cases with concurrent Hippel-Landau disease showed the same pattern of 3p loss and t(3,5) evidence of the same chromosome catastrophe [240]. This soundly explains the later onset of sporadic ccRCC, as the gap years before the disease onset are needed to acquire a somatic mutation that would inactivate VHL. The model built from these genomic data suggests that as little as one hundred cells persist before the inactivation of VHL to initiate the tumour expansion. Taken together, this study uncovered stable evolutionary paths taken by ccRCC in its progression. These factors could facilitate early diagnosis and target the nascent cancer cells before inactivating VHL and other carcinogenic mutations. Particularly, targeting the essential genes on 3p could be beneficial.

The second venue of research assessed in detail how specific evolutionary trajectories taken after the onset of malignancy focuses the tumour growth [234]. This study operated an expanded dataset of 101 ccRCC tumours that passed stringent quality controls, for example, eliminating those with germline VHL mutation and containing 1208 multi-region samples. All samples were screened for driver mutations using bespoke panels with a 612x median depth of sequencing (ranging from 105x to 1,520x). Detailed

analysis of the clonal trajectories established their connection to the common ccRCC prognostic variables such as the presence of necrosis, TNM tumour stage, Fuhrman grade, a pathological grading scheme for tumours of the kidney that is based on the shape of the nucleus and presence of nucleoli, and overall tumour size. The number of acquired driver mutations positively correlated with all these variables. The quantity of branching events, i.e. intra-tumour heterogeneity, behaved similarly. Interestingly, the number of clonal populations existing in parallel showed a non-linear association with time: averaging four clonal populations per tumour (median), their number reaches a plateau at a tumour size of 10cm and starts to decrease. This observation might be an example of earlier discussed linear evolution that suggests the ultimate convergence of the cancer population to one lineage. Although parallel evolution was apparent in different clonal populations within some tumours, some genes were never co-mutated in the same clone (BAP1 and SETD2) while often co-occur within the same tumour. Finally, an unsupervised clustering based on evolutionary features such as evolutionary timing, order, and co-occurrence, identified seven clusters of tumours with distinct physiological features and clinical behaviour. Cancer cell populations within the most aggressive tumours were dominated by a small number (5) of clones, each harbouring multiple driving mutations, evidence of a history where a dominant clone achieved a sufficient selective advantage that initiated a clonal sweep over the population. Interestingly, another cluster of high-stage tumours progressed through drastically different evolutionary trajectories: they showed very high intra-tumour heterogeneity. Usually, they averaged 12 clones simultaneously growing within a tissue. These tumours, however, showed twice as longer time to relapse compared to the tumours with "multiple driving mutations" clones (11.7 months to 4.7 months, although not statistically significant), suggesting that branching evolution and selective pressure on subclones limits the speed of tumour growth. Other clusters comprised early-stage tumours populated by BAP1-only mutants and clones that only bypassed VHL and have not yet accumulated additional mutational drivers. This suggests an early stage before progression to the defined evolutionary trajectories. This work serves mainly as the experimental documentation of the deterministic nature of cancer evolution, which could be utilised to develop an evolutionary grading system based on biopsies.

As much as the discussed studies concerned the evolution within primary tumours, the third study focused on tracking the evolutionary routes to metastasis [241]. ccRCC, with its well-established genetic landscape within the primary tumour, is a compelling model to study clonal evolution to metastasis. It readily spreads through both lymphatic and hematogenous routes. It can colonise various tissues, including the lung, bone, liver, brain, pancreas, and soft tissues, with the worst prognosis usually associated with the liver [242, 243]. The spatial distribution of metastases is

also variable. It ranges from solitary metastasis in a single location to oligo metastases, defined as up to three of five invasion sites in limited locations and to widespread metastases over multiple regions. Although patients with synchronous solitary or oligometastatic tumours can be managed with local strategies, i.e. surgical removal of primary tumour and metastases or ablative therapi1es, a fifth of these show signs of tumour progression as early as one-month post operation [244]. Building upon seven conserved evolutionary subtypes characterised before, the endpoint of TracerX was to distinguish metastasis-competent clones and explore the routes and timing of metastasis to different anatomical sites. For a better glance at the potential power of this study, a total of 575 primary and 335 metastatic multi-region biopsies taken from 100 patients were analysed, which greatly exceeds the average biopsy to tumour ratio of similar studies, e.g. 2 to 3 biopsies per tumour on average [245]. Provided that metastases emerge late in the tumour evolution and are thought to be seeded by small groups of cancer cells, the loss in the heterogeneity in metastases is to be expected. However, questions remain whether the metastases grow as a monoclonal population that evolves on the new site from a single clone or whether the metastases mirror the clonal composition of the primary tumour, and if yes, does it mean that the majority if not all subclones share the metastatic potential [246, 247]. From the analysis of the TracerX cohort, it is evident that the clonal population in metastases are much more homogenous, with the proportion of clonal variants reaching an average of 87%, contrasting 32% in primary tumours. Similarly, cancer cells in metastatic lesions carried fewer somatic driver mutations than those in their matched primary tumour, averaging 9 and 12, respectively. Peculiarly, only 5.4% of driver mutations observed in metastases appeared de novo, indicating that metastatic competence was selected within the primary tumour. Those "selected" clones that progressed to other sites showed higher frequencies of somatic copy-number alterations, heightened proliferation, and loss of heterozygosity in HLA, suggesting the ability to evade the immune response. Peculiarly, the loss of either the short arm of chromosome 9 (9p) or the long arm of chromosome 14 (14q) or both of them together was a hallmark of the majority of metastatic cases (71%). Oppositely, these genomic alterations were absent in tumours without metastatic disease at the time of analysis. Could another catastrophic chromosome event be needed to pass the evolutionary bottleneck? Considering the loss of 14q and 9p as potential biomarkers for the detection of early metastasis, 14q and 9p were predominantly subclonal and might go undetected by single biopsies [234]. In retrospect to the study on primary tumours, the identified evolutionary clusters exhibited distinct metastatic potentials. The "multiple clonal drivers" and BAP1-driven populations that exhibited attenuated chromosomal instability and low intratumor heterogeneity, i.e. populations whose clonality and genetic composition are evident of punctuated evolution, drive the more aggressive disease that culminates in rapid progression to

multi-region metastasis. In these tumours, the metastatic potential is acquired already within the most recent common ancestor, which drives swift metastatic spread. Similar evolutionary tendencies have been observed in other cancers of pancreatic, breast and uveal origins, consistent with the tendency of some tumours to metastasise rapidly [248, 213, 249]. In contrast, tumours that grow through the branched evolutionary process and consist of highly heterogenous populations with moderate levels of genomic instability progress slowly to solitary or oligo metastases. One could hypothesise that clonal competition between the abundance of clones that limits the speed of primary tumour growth should similarly confine the progress towards metastasis.

The preceding section supplemented the earlier discussed modes of cancer evolution with experimental evidence. Some evolutionary trajectories of tumour progression are rigid and reliably replicated in tumours from different patients: even across cancer types, some tumours evolve in a similar mode of punctuated evolution through early and rapid speciation events that are followed by an expansion of a small variety of adapted clones, while others evolve stepwise in a better image of Darwin's phyletic gradualism. These conserved trajectories could provide a basis for a new method of clinical classification or guide therapeutic decision-making, e.g., drafting drug regiments to blockade clonal resistance or deciding on surgical intervention. The development of new sequencing methodologies and theoretical frameworks facilitates the capture of tumour mutational signatures from the circulating tumour cells and cell-free DNA derived from blood and even salivary samples [250, 251, 252, 253]. Although the resolution of these approaches still would not detect rare subpopulations, a combination of the detected signatures with known evolutionary trajectories could help better describe the clonal composition of a tumour. Large-scale attempts were made to pan-cancer studies of cancer evolution based on the massive WGS data from TCGA and ICGC [254]. Although informative, due to the shallow sequencing (average coverage is 30x), these results cannot pierce deep into phylogenetic trees of cancers and miss infrequent but clinically relevant clones. Other studies focusing on individual cancer types and utilising the deepest available WGS data (coverage of 80x) add to the overall knowledge of cancer evolution [255]. More temporal studies like this and TracerX that utilise deep sequencing would benefit from creating a reliable catalogue of stable evolutionary trajectories. In this regard, another large-scale longitudinal study soon finishes recruiting patients, and it would be interesting to see more cancer types profiled this way [256]. Despite the comprehensive characterisation of tumours from genomic, transcriptomic, and temporal perspectives, the main limiting factor in cancer studies comes from the complexity of the bulk sequencing data, which, however, might soon no longer be an issue with the maturation of single-cell sequencing technologies [257, 258].

2.12 The promise of single-cell sequencing

Previous sections followed the development of omic technologies since the inception of the earliest methods like SAGE, through the explosion of microarray genotyping and transcriptomics, to the advent and establishment of NGS. The applications of multi-omics to cancer research were scrutinised alongside the limitations of the existing approaches. As becomes evident in the deliberation on the evolution of cancer, the heterogeneity of cancer cell populations, i.e. presence of clones with different adaptive capabilities in variable proportions, is the force that drives the resistance to therapy as well as the progression of the disease in general. Despite numerous informative studies and databases of cancer genomic data that utilise bulk sequencing, translating these data into knowledge seems to have reached its methodological limit. Cellular heterogeneity and the spatial organisation of clones within the tumour influence the disease's progression and, therefore, the potential venues for therapy. Thus, cancer heterogeneity necessitates the methodology that captures genomic and transcriptomic data on a cellular level for its mysteries to be recorded, analysed, and re-purposed.

RNA sequencing in single cells has a long history, with individual cell profiles recorded in the early nineties [259]. In that study, Eberwine et al. profiled gene expression of single neurons from rat's hippocampus using a sophisticated methodology: selected neurons were dissociated, after which primers, nucleotides, and enzyme mixtures were microinjected directly into the cell. The libraries from RNA were created by in vivo reverse transcription (RT) coupled with amplification by in vitro transcription (IVT). This approach measured the relative expression of four mRNAs in fifteen cells. Although humble, yet still a beginning. The incorporation of PCR amplification improved both and allowed simultaneous quantification of the absolute number of RNA molecules transcribed from two dozen genes, still far away from transcriptome-wide studies, which was first achieved by applying new amplification methodologies that allowed the preparation of libraries from single cells to be profiled by hybridization-based microarrays [260, 261]. In 2012, the first Smart-seq protocol was published, providing the experimental framework to achieve nearly bulk RNA-seq quality. However, this approach was still limited to the assay of only a handful of cells due to universal reliance of early methodologies on isolating single cells in the individual tubes before lysis and library preparation, which essentially was a bottleneck for up-scaling the number of cells that can be profiled per run [262]. Some peculiar designs utilised bespoke cell-pickers to improve the throughput (number of cells assayed) [263]. The alternative was to couple the method in use with FACS sorting to automate the loading of cells into microwells [264]. The addition of post-cell capture automation further improved the capacity of FACS-based methods [265]. Nevertheless, the actual expanse

of the single-cell sequencing capacity happened when new approaches for passive cell capture were developed. The C1 chip from Fluidigm was the first to be commercialised. This chip allowed the passive capture of up to 96 cells and their delivery in an exact volume of enzymatic mixture to reactors where the cDNA is prepared [266]. In parallel, another method was designed for a random capture of cells in nanolitre drops of emulsion: in a microfluidic circuit, two channels flow, one delivering an enzymatic mixture with beads covered by polyT primers and the other cells in a buffer [267]. Then, the flows of these two liquids are combined, resulting in a drop containing the cells and all the reagents. This process is imperfect; sometimes, no cells or multiple cells are captured, causing the creation of a "doublet", which are deconvoluted computationally during data analysis. Instead of using the flow of two channels, the parallel approach used gravity to mix cells in picolitre reactors, with the cell mixtures calculated according to Poisson statistic to maximise the chance for a well to receive only one cell [268]. Today's most popular platform for single-cell assays, Chromium from 10x, utilises a similar approach to drop-seq [269]. The captured cell drops are then collected in oil within a tube, where cDNA is created and barcoded. The barcoding process is vital, allowing us to identify RNAs from individual cells.

Additionally, due to the random nature of cell capture, a fraction of cells will always be missed; therefore, the pool of available barcodes should always be larger than the number of cells in the mixture. To this end, combinatorial barcodes allow multiplexing of tens of thousands of cells simultaneously [268]. The combination of these approaches pushed ahead the capacity of single-cell methods, with the number of cells assayed per study increasing exponentially up to 2017 [270]. In a similar trend, the overall number of published articles that utilise single-cell RNA-sequencing (scRNA-seq) in one way or another grows to the bulk RNA sequencing, while the percentage of single-cell RNA-seq among all RNA-seq studies increasing from 5% in 2016 to more than a quarter in 2022 (26%) (Figure 2).

Outside of RNA-sequencing, WGS and WES of single cells are also of interest concerning cancer heterogeneity. As was mentioned in the discussion on cancer clonality, sequencing individual cancer cells from different parts of the tumour would provide an unprecedented snapshot into the evolutionary trajectory of that particular tumour. However, the single-cell resolution comes at a price. Valid for all single-cell sequencing methods, the main trade-off is the tiny amount of RNA/DNA available in each cell. Therefore, all genomic single-cell approaches rely on whole genome amplification techniques (WGA). These methods allow DNA amplification from a meagre starting amount but inherently introduce various biases.

Currently, there exist three main WGA methods: multiple displace-

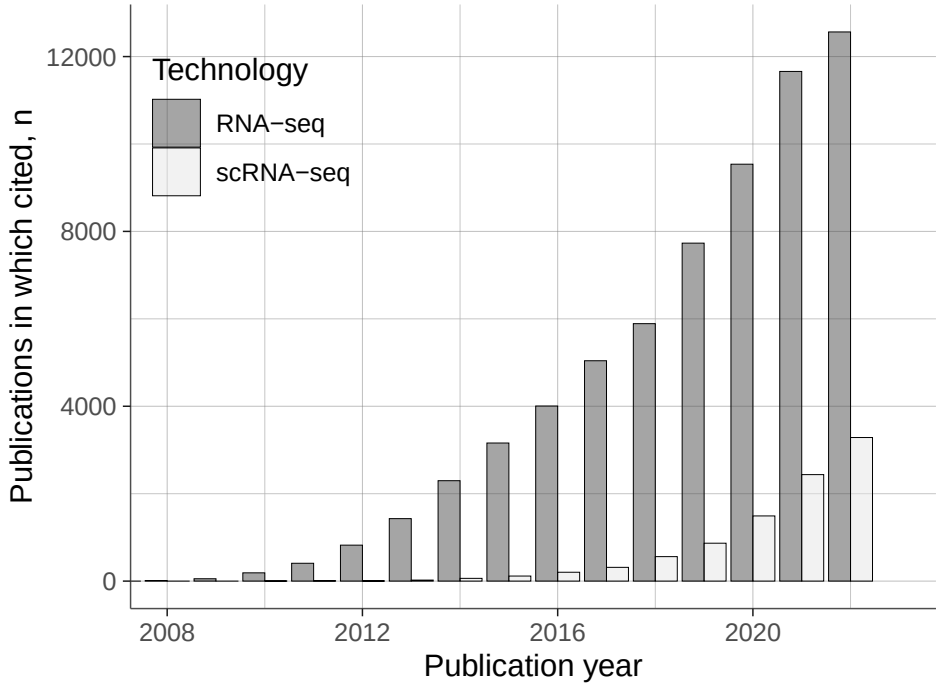


Figure 2: Shift from bulk to single-cell scope
Histograms showing the number of publications per year referencing RNA-seq or scRNA-seq.

ment alignment or isothermal amplification (MDA), degenerate nucleotide primed PCR (DOP-PCR), and hybrid methods that combine the usage of MDA with DOP-PCR [271].

MDA is based on annealing random hexamers with primers to the single-strand denatured DNA, followed by DNA fragment synthesis [272]. The synthesis proceeds through a DNA polymerase's sequential displacement of de novo synthesised DNA fragments with strand displacement activity. The most commonly used protocols utilise ϕ29 polymerase with high processivity and low error rates [273]. Displaced fragments are amplified by the newly annealed hexamers with primers, resulting in a network of branched DNA structures. The main shortcomings of MDA application in single cells are high rates of allelic dropout and preferential amplification. These two processes are connected and could exacerbate the adverse effects of each other. In that regard, allelic dropout corresponds to a random non-amplification of one of the alleles in a heterozygous sample, which can vary drastically between different studies: 25% and 60% rates have been reported [272, 274]. Preferential amplification, as evident from the name, refers to the relative overamplification of one of the alleles. It is still being determined if the process is random or systematically biased towards specific loci [271, 272]. Nevertheless, both processes may be a severe hindrance to quantitative genomics. Additionally, ϕ29 activity creates low levels of chimeric sequence side products [275]. The latter usually coincides with the branching process and can be mitigated by an endonuclease-driven debranching reaction [119].

As the name implies, the second group of methods uses PCR-based amplification with random priming [276]. In practice, PCR/based methods suffered a loss of signal from most of the genome during the amplification process due to differences in the density of common sequences and variability of PCR efficiency between different loci [271]. Further exacerbating these problems, these methods use thermostable DNA polymerases with higher error rates than thermolabile polymerases, resulting in a high rate of false negatives. To overcome the low coverage of PCR-based methods and an uneven signal from isothermal MDA methods, new hybrid approaches were developed that combined features of both. The first one, known as displacement DOP-PCR or PicoPLEX, achieves the initially limited amplification of the DNA by MDA, followed by PCR amplification of the fragments generated by the MDA. The PCR reaction is primed by the common sequences introduced into the amplicons by MDA [277]. This idea was further developed in a protocol named multiple annealing and looping-based amplification cycles, or MALBAC. MALBAC's principal difference from its predecessor is that it cycles the temperature regime during displacement amplification. This procedure promotes the looping of the isothermal displacement amplicons, inhibiting their further uneven amplification before the PCR step [278].

In practice, MDA and MALBAC are predominant. Although both were reported to perform similarly in human cells, some differences and trade-offs in the final amplification products are present. For example, in one report, MDA surpassed MALBAC in genome coverage (84% versus 52%) and, consequently, in the SNV detection rates to a similar degree. On the other hand, MALBAC produced more uniform coverage and fared better in detecting CNVs [279]. Other reported identical results: MDA shows a better range than MALBAC but suffers in uniformity of sequences covered compared to MALBAC [280].

Overall, an all-around-winning approach for preparing genome material from single cells has yet to be. Although whole genomes can be sequenced from single cells, it comes with the trade-off of increased rates of false-positive and false-negatives and the cost of the inability to cover the entirety of the genome reliably.

In the context of cancer heterogeneity, single-cell WGS is of value. Considering the drawbacks discussed, the phylogeny of cancer cells and clonal mutational profiles can still be recovered after imposing appropriate mutational rate cutoffs, similar to the standard procedure in bulk WGS [281]. In 2011, Navin et al., for the first time, utilised sequencing of whole genomes from single nuclei extracted from cancer cells of a breast tumour and related metastasis [282]. The group sequenced 100 cells from 5 to 12 different tumour regions and metastasis. It could detect patterns of punctuated clonal evolution in both cases, i.e. all detected clonal subpopulations were each distant from their root without observable branching and connecting subpopulations. A more recent analysis of single-cell genomes from metastatic colorectal cancer managed to reconstruct a phylogenetic tree and identify a clonal population that survived the treatment-induced evolutionary bottleneck and evolved for years under adjuvant therapy before expanding rapidly following the termination of treatment [283]. Similarly, circulating tumour cells can be isolated and interrogated using single-cell sequencing. Although there are questions about whether the cancer cells captured this way would be enough to characterise the diversity of the tumour, this possibility is very appealing for non-invasive diagnostics and disease monitoring [284, 285]. Alongside the experimental forays, new methods constantly develop to reliably infer evolutionary trajectories from single cells [286].

To conclude, single-cell genomics is positioned to be the technology that, with proper scaling, could allow chartering an exhaustive map of cancer evolution. Nevertheless, current methods suffer from methodological inefficiencies and low throughput. If hypothesised with enthusiasm, it may be that only a few years of honing the existing strategies divide us from breakthroughs.

2.13 Single-cell RNA-seq data normalisation

Together with incredible resolution, single-cell profiling brought along new computational issues. As of 2021, for every three studies involving single-cell RNA-seq profiling, two single-cell RNA-seq computational tools were published [269]. In this regard, single-cell RNA-seq is undergoing the same process as bulk RNA-seq and microarrays in their infancy. In other words, many different tools and approaches will be published before the field converges to a status quo. Currently, there are two venues to work with single-cell RNA-seq data: data transformation followed by general statistical methods and statistical modelling of the observed data.

Let us first consider the first approach. The common motivation for data transformation arises from the fact that distributions of mRNA counts are heteroscedastic, i.e. the variance increases along the mean of the observations. This is usually accounted for by so-called variance stabilising transformation or vst. Usually, mean-variance dependency is accounted for by assuming Negative Binomial (NB) distribution, the same as a Gamma-Poisson mixture. In this regard, scRNA-seq data is no different from bulk RNA-seq: almost all statistical tests devised to compare classes in categorical experiments utilise generalised linear models (GLM) that assume count data following either NB of similar distributions [287]. It is worth mentioning that differential expression between conditions is of only limited interest for single-cell studies that are more concerned about the identities of the groups of cells that might correspond to functionally different cell types (clustering analysis) and the inference of developmental trajectories [288].

Nevertheless, as long as most dimensionality reduction and clustering methods often use metrics based on Euclidian distance, which implies normality, the data transformation task stays on the agenda. Coming back to the data transformations, the mentioned NB distribution with mean μ and overdispersion parameter α implies a quadratic mean-variance relationship defined as follows:

$$\mathbb{V}\mathrm{ar}[\mathrm{Y}] = \nu(\mu) = \mu + \alpha\mu^2 \tag{1}$$

Here, α parameter defines the measure of overdispersion, whereas $\alpha = 0$ will result in a Poisson distribution. Assuming a functional form of a relationship between mean and variance as $Var(y) = g(\mu)$, the goal is to find a function g for which the standard deviation is constant:

$$\mathbb{S}\mathrm{d}[\mathrm{Y}] = \mathrm{const} \tag{2}$$

Delta method allows to approximate the standard deviation of a transformed random variable from the function of a random variable:

$$\mathbb{S}\mathrm{d}[\mathrm{g}(\mathrm{Y})] \approx |\mathrm{g}'(\mu)|\mathbb{S}\mathrm{d}[\mathrm{Y}] \tag{3}$$

Putting (2) and (3) together, one can set the requirement for (3) to be constant as in (2) and solve for $|g'(\mu)|$:

$$g'(\mu) = \frac{const}{\mathbb{S}\mathrm{d}[\mathrm{Y}]} = \frac{const}{\sqrt{\nu(\mu)}} \tag{4}$$

Since the constant does not affect variance stabilising properties, one can now derive a function form of this transformation by solving the integral [289]:

$$g(\mu) = \int \frac{1}{\sqrt{\nu(\mu)}} = \int \frac{1}{\sqrt{\mu + \alpha\mu^2}} = \frac{2}{\sqrt{\alpha}} asinh(\sqrt{\alpha\mu}) \tag{5}$$

Finally, (4) is our variance stabilising function for NB distribution. Peculiarly, these transformations of NB distribution were explored by Anscombe, who considered a well-working and familiar observation of the solution above [290]:

$$g(y) = log(y + \frac{1}{2\alpha}) \tag{6}$$

The latter approximation has the same form as the heuristic log transform, but instead of an arbitrarily picked pseudo count to mask zero values, a data-driven "pseudo count" is used. This transformation generally smoothens the variance for well (and strongly) expressed genes, while the small counts still exhibit a mild mean-variance dependency. Overall, no transformation works well on small counts, so the only option is to fare onward [291].

Another normalisation procedure involves scaling expression by size factors. Bulk-sequencing size factors are essential to account for heterogeneous sequencing depth across different samples. Although most scRNA-seq protocols nowadays use unique molecular identifiers (UMIs) that remove technical variation due to PCR amplification, technical variation due to stochastic molecular processes, such as the efficiency of reverse transcription or differences in cell lyses, remain unaccounted [292]. In this regard, one common approach is to use a normalisation factor, such as the total number of UMIs per cell multiplied by a scale factor. This approach, for example, is used in a standard workflow of Seurat [293]. Alternatively, the 10x genomics pipeline suggests normalising by scaling factor computed as a ratio between the median UMI count per cell and the total number of UMIs per cell [294].

The second approach utilises statistical modelling and can be subdivided into two categories based on the conceptual differences: observation and data-driven, or hypothesis-driven modelling [295].

Let us consider first the data-driven modelling. Models that follow this principle are hierarchical and try to simulate to the best extent two sources of variance: the true biological variability, i.e. true variance in expression of genes within/across cells, and the technical variability that

masks true expression levels from the ones observed in an experiment. Here, it is important to note one property of single-cell data: an abundance of zeros, or "dropouts", among the expression estimates. In early single-cell transcriptomics, an assumption prevailed that scRNA-seq data is zero-inflated, i.e., it contains more zero counts than expected by chance. This "dropout" phenomenon was first referenced in a study that analysed scRNA-seq data from low-throughput plate-based methods like Smart-seq [296]. It was noticed that differential expression between two cellular subpopulations assayed with scRNA-seq showed more zero values than was observed in comparison to typical bulk RNA-seq samples. How the idea of dropouts and zero inflation permeated the field of droplet-based methods so far remains unknown [297].

This notion led to the development of various methodologies to counter zero inflation. One concept suggested that an overabundance of zeros corresponds to missing data and that the latter can be predicted from the expression of other genes if there was no dropout [298, 299, 300]. The procedure to predict missing data based on existing observations is called imputation and is somewhat common in other machine-learning domains [301]. Over time, even imputation-based clustering methods were developed to work with single-cell data [302]. In an alternative approach, NB models were augmented with a probability to observe a zero in any given draw, i.e. "zero-inflated negative binomial" distribution-based modelling [303, 304]. However, closer investigations of droplet scRNA-seq counter the assumption of zero inflation as an inherent property of these data. Multiple studies found that NB can reliably model scRNA-seq gene expression values and didn't benefit from adding an inflation component [305, 306]. One peculiar paper suggested that a zero-inflation-like appearance could be simply a consequence of logarithmic data normalisation that introduces an artificial gap between zero and non-zero counts that can be alleviated by generalised PCA models [307]. When disregarded, the per-cell proportion of zeros can become the main source of variance in the dataset and bias the biological signal from gene expression. Overall, the compounding evidence from the multiple studies with scRNA-seq of negative controls with known mRNA contents suggests that the number of observed zeros is consistent with those expected from an NB distribution, and the additional inflated component is unnecessary. It is worth mentioning that the number of zeros per cell is also affected by the efficiency of RNA material capture after the lysis (10x Genomics V3 protocol estimates 30% capture efficiency); therefore, improvements in the sequencing depth per droplet would allow counting more molecules and sequentially reduce the number of zeros. On the other hand, Kim et al. noticed in the example of the PBMC dataset that genes annotated with immune-related terms are significantly zero-enriched compared to the rest of the population, attributing the biological variance to zero-fraction. They proposed to use a test for cellular heterogeneity using zero-fraction statistic [308].

Now that zero inflation is addressed, data-driven modelling can be further discussed. As was mentioned earlier, two models constitute the final model, one accounting for the expression variance and another for technical variance. As it should be evident by now, the most common choice is NB, a Gamma-Poisson mixture model, where expression is modelled by Gamma distribution and measurement error is modelled by Poisson distribution. Let us recite and consider the following terminology: the final model used is the Observational model, the model to account for true gene expression variance is the Expression model, and the model to account for technical variance is the Measurement model. This resource developed by Sarkar et al. holds a complete annotation of existing scRNA-seq methods with the associated models [309]. Similarly, the mentioned model that accounted for zero inflation in expression values used point-Gamma distribution as the expression model [304]. We have already discussed the connection between the NB model and the variance stabilising transformation to explain the feasibility of log transformation with a data-driven selection of a pseudo count. Data-driven modelling is sometimes also used for data transformation. Seurat's sctransform function exemplifies this approach, which uses Pearson residuals from the NB model [310]. This approach, however, was brought to question by Lause et al., who found evidence of over-specification in the model used by the authors. In detail, the authors fitted the NB expression model with three parameters, two of which were estimated gene-wise for each gene. Further examination showed that the two gene-wise estimated parameters show powerful correlations (Pearson's R = -0.91), especially for weakly expressed genes, altogether bringing up the evidence for overfitting [311]. The suggestions of Lause et al. were considered in the next release of the sctransform that now uses an offset model with only two parameters and a fixed slope. Benchmarking 59 real datasets, the authors conclude the overall applicability of the NB model for scRNA-seq data, inasmuch as all datasets exhibited overdispersion with the exclusion of the genes with minimal counts [312].

On the contrary, the hypothesis-driven approach tries to build the model describing observed gene expression around the mental concepts of actual biological processes. Therein originates the restriction that the estimated values of these models should represent meaningful physical quantities and be accompanied by estimate errors. One example of this approach is a Sanity framework, as abbreviated from "SAmpling-Noise-corrected Inference of Transcription activity" [313]. In this regard, Sanity follows the flow of information from the physical transcription process to the observed gene expression patterns. The model first considers the transcriptional activity of the cell as the sum of all gene-wise transcriptional activities, which individually represent a weighted average of recent transcription and mRNA decay rates. Each mRNA count is then defined as a Poisson sample with

the mean equal to the transcriptional activity of a related gene. Finally, the detected number of UMI is a Poisson sample with a mean of transcriptional activity multiplied by a capture probability. Although authors state that the probability of observing mRNA molecule given by Poisson distribution is independent of how the transcription and decay rates fluctuated over time, others report that non-independent, i.e. targeted, mRNA degradation and promoter fluctuations resulting in transcriptional bursts lead to deviation from Poisson statistics, whereby "Sanity" could show bias in its estimates [314]. On the contrary, other models are built specifically around the patterns of transcriptional bursting [315, 316, 317]. The model of Jiang et al., for example, utilises two hierarchical Beta-Poisson models to estimate true allele-specific expression and then models the technical noise from the hierarchical mixture of Poisson and Bernoulli random process [317]. To conclude, at this point, there exists no universal solution to model scRNA-seq ideally and for some tasks, the complications of elaborate models are time-ineffective, e.g. gene-wise fitting of overdispersion parameter in NB takes around 40 minutes for an average-sized scRNA-seq dataset, and not truly necessary, therein bringing us back full circle to the most widely used approach of log-transformation and depth-normalisation [287].

2.14 Computational suite for single-cell RNA-seq analysis

As mentioned in the previous paragraph, scRNA-seq data are primarily of interest for the search for new functional states of the cells and their developmental trajectories. Methods whereby the scRNA-seq data is analysed towards the mentioned goals, are not lacking in abundance. For clarity, the author will concentrate on the most popular techniques from the following computational tasks: feature selection, dimensionality reduction, clustering, and trajectory inference.

Commonly encountered scRNA-seq datasets assay the expression of roughly ten to twenty thousand genes, which, for the most part, do not carry valuable informational load and, if retained, may interfere with downstream analysis. Therefore, feature selection (FS), or, to be more precise, gene selection, occupies a key position in a scRNA-seq analysis pipeline insomuch that every analysis implemented downstream, one way or another, derives its results from the set of features selected afore.

The basic FS methods follow the idea that genes whose expression vary the strongest across the cells within the dataset are the most informationally-heavy and capture the largest fraction of biological variation within the dataset [318]. The most straightforward way to select features would be to assess mean-variance relation at the gene level and pick those genes that exceed a certain threshold. A vigilant reader would notice a fallacy in this approach, inasmuch as mean-variance relation in scRNA-seq datasets is heteroscedastic, and selection by a threshold would by all odds introduce a bias towards highly expressed genes (see section 2.13). Therefore, variance-based FS methodologies are designed to account for mean-variance correlation by modelling mean-variance dependency: only the genes whose variance across cells surpasses the null-model are selected; these informationally-heavy genes are colloquially known as highly variable genes (HVGs) [319].

Although variance-based FS arguably stands as the most common appraoch in scRNA-seq analysis, there other methods that follow conceptually different strategies. One earlier tool extracted features with highest loadings from PCA reduction of the normalised dataset and desribed the extracted genes in a familiar way: high loading genes (HLVs) [320]. More recent publications leveraged graphs and gene-gene correlations to extract informative features, yet the comprehensive benchmarking is yet to be performed to reliably assess the performance of these tools [321, 322].

Dimensionality reduction is a standard procedure in every single-cell analysis pipeline. The importance of dimensionality reduction (DR) was secondary in the era of bulk sequencing, where the number of samples in

the dataset, i.e. the number of dimensions, was small enough to be handled by methods like PCA or MDS. On the contrary, single-cell datasets are composed of tens of thousands of cells, i.e., dimensions, and the problem of DR is acute. First of all, why is DR important? In bulk RNA-seq, DR is used to assess the quality of the experiment visually, e.g. differences between conditions and similarities between replicates, before proceeding with the downstream analysis. In the case of a single-cell experiment, the DR is applied with a similar aim, i.e. to assess the cells in the experiment visually. Yet, the downstream consequences of this visualisation carry much more gravity. In other words, its penultimate task is to produce a visualisation, or rather a map, of the dataset in question that is tractable by the human eye and minimal in terms of informational loss. In the downstream analysis, this map is used to identify assemblies of cells, i.e., clusters, that are further interrogated for their functional qualities, expressed markers, developmental trajectories, etc. It makes it evident that the chosen DR method should be able to preserve hidden local and global structures of the data for downstream analysis to bear meaning [323].

In the last years, the trial and error approach of the community outlined a popular standard workflow that often co-utilises linear DR by PCA to a few dozen dimensions and non-linear DR of the PCA output. The non-linear part is done by either t-SNE or UMAP [324, 325]. Let us first have a cursory view of these methods starting with t-SNE. Developed to visualise high-dimensional data, t-SNE quickly found its application in single-cell data analysis. In their 2019 work, Kobak&Berens reviewed in detail how the t-SNE is used in single-cell transcriptomics and advised on good practices and how to avoid potential pitfalls in one's analysis [326]. To give the basic intuition about t-SNE, let us informally outline the founding concepts of this method. In general terms, t-SNE maps the points in a high-dimension to a low-dimensional space so that neighbouring points remain neighbours and distant points remain distant. To achieve this, the algorithm randomly distributes all points over the low-dimensional space, allowing them to interact as if they were physical entities like molecules. Analogous to the physical world, the positions of each point after the initialisation are governed by two powers: repulsion from each other and attraction to the nearest neighbours. Attraction is controlled by the perplexity parameter that effectively decides how many nearest neighbours the point is attracted to, whereas the power of repulsion is effective against all points in the dataset, not only nearest neighbours, and is controlled by the ρ parameter. This definition is enough to develop an intuition on the effect of the balance between these powers: favoured repulsion would lead to a smaller number of bigger and more stable clusters, whilst the favoured attraction would be more sensitive and capture finer data structures.

One thing to remember using any stochastic neighbour embedding (SNE) based methods, which include both UMAP and t-SNE, is that, although, neighbouring points, i.e. clusters, are preserved in low dimensional representation, the geometrical arrangement of these clusters is not reliably captured; spatial positioning of the clusters relative to each other could be misleading. This drawback's severity scales with the size of the analysed dataset. When considered, these drawbacks can be accounted for and ameliorated by the proper analysis procedure, a recipe of sorts. The recipe of Kobak&Berens advises on PCA initialisation, multi-scale similarities, and increased learning rate [326]. In detail, PCA can identify the global architecture of the data, and the initialisation of the t-SNE algorithm with PCA coordinates instead of random positioning helps t-SNE to identify finer substructures while preserving the geometrical structures provided by the PCA [327]. Multi-scale similarities approach suggests using multiple perplexity metrics simultaneously and averaging them for the final decision on cell attraction. This allows the method to simultaneously consider adjacent cells and more distant neighbours, thereby better capturing the data structures [328]. Multiple studies suggested that the default learning rate of 200 is insufficient for large datasets and leads to suboptimal convergence, with some adopting the learning rate constant of 1000 in their t-SNE implementations [329, 330]. Finally, the authors advise using the following empirically devised rule of thumb for the learning rate: $\eta = n/12$ [329].

The second method, UMAP, or Uniform manifold approximation and projections, although published ten years later than t-SNE, attracted many followers due to its seemingly faster performance than the original t-SNE [331]. The latter difference is likely not to stand anymore, as contemporary t-SNE algorithms are reported to perform as fast as UMAP on similar hardware [332]. Conceptually, UMAP and t-SNE are similar as both use the same forces of attraction and repulsion. However, the nature of repulsion forces and sampling-based approach to optimisation differ between UMAP and t-SNE. Both methods seem to produce comparable results regarding the preservation of geometric structures, although slightly different visually. In one report, a particular tuning of attraction-repulsion forces (stronger attraction) in t-SNE generated close-to-identical reproduction of UMAP visualisation [333]. In the same report, the t-SNE algorithm geared towards extreme attraction can recover continuous superstructures from simulated data and produce results similar to graph-based methods such as ForceAtlas2 [334]. This may be a glimpse of the hidden potential of t-SNE's DR in the tasks of trajectory inference. Before that, however, this approach needs to be formalised.

Trajectory inference (TI), whereby an investigator can conclude the developmental or any other temporal trajectories within a population of

cells, is a technique inseparable from the contemporary scRNA-seq analysis. Across the fields of cancer and developmental biology, the dynamic changes of cell states and lineages underlie the core questions of the respective disciplines. Insofar as temporal studies are experimentally tricky due to the cell-destructive nature of the sequencing process, TI methods are of interest for a panoply of studies. Therefore, many dedicated their work to developing computational methods to infer the temporal trajectories from a static snapshot of cell states, scRNA-seq [335, 336, 337]. Early methods were mainly focused on aligning the data to fixed topologies, such as linear or bifurcating trajectories [338, 339, 340]. More specialised techniques looked for circular topologies to model cell cycle [341, 342]. Approaches developed afterwards also infer a topology without a fixed reference and extend available trajectory topologies with graphs, cycles, and disjointed topologies, which makes the problem even more computationally complex [335, 343, 344].

In contrast to DR, where the state-of-the-art approach converged in UMAP and t-SNE, the TI methods at one's disposal are as diverse as many. Many undertakings tried to apply different TI methods alongside each other to the same dataset in attempts to make a comprehensive characterisation of their strengths and weaknesses [345, 346]. To the author's knowledge, the largest attempt in this direction was made in 2019. It culminated in a web database dedicated to cataloguing and integrating the whole universe of TI methods in one place [347, 348]. Despite the differences in the structures they infer, most of the methods start with DR as the first step; it is worth noting that t-SNE or UMAP are not used as the goal here is not visualisation but an attempt to repel the curse of dimensionality. PCA is often used, while the other common alternatives are diffusion maps and linear embeddings [349, 350, 351]. After this, the methodologies bifurcate into those that first identify clusters, i.e. cluster-based, and others that construct a similarity graph, i.e. a graph where data points are vertices that are connected by edges built from a selected distance metric and infer topologies from the graph via various probabilistic methods such as Hidden Markov Model (HMM).

Clustering-based methods, as the name implies, first locate stable cell states, i.e. clusters, in the dataset and then draw a trajectory through these clusters. Different clustering methods are used, including hierarchical clustering, k-Nearest Neighbours (kNN), non-negative matrix factorisation (NMF), and Louvain clustering. Some algorithms do not come with a specific clustering method but rather take cluster annotation as input, allowing the user to decide on the approach [352]. Clusters are commonly connected by the minimum spanning tree (MST), a graph that connects all vertices (clusters) with a minimum possible total edge weight. In some methods, an additional step to construct principal curves is added after clustering and building an MST [336].

Graph-based methods start with building a graph representation of the data and use diffusion or traversal methods to construct a trajectory through the graph. Most methods build kNN-graph with Euclidian distance metric, while the approaches to trajectory constructions vary. PAGA partitions the graph based on a cluster annotation provided by the user (the original method used Louvain clustering). Then for each partition, a "PAGA connectivity measure" is defined, which is the ratio of inter-edges between clusters to an expected number of inter-edges under random assignment, to reveal connected and disconnected regions [335]. Pseudotime is inferred using a modification of the diffusion pseudotime (DPT) algorithm that allows it to deal with disjointed partitions. One recent method was designed to expand the TI framework to include lineage tracing data [353]. Considering the contemporary trends of multi-modal single-cell data analysis, new universal methods can be anticipated to infer TI from different kinds of omic data [354].

As much as single-cell analysis is concerned with studies of cell communities, it depends on correctly identifying these communities, which is a clustering problem. Cluster analysis is omnipresent in all single-cell analysis pipelines, and selecting the correct method is paramount for proper biological interpretation. Although impressive progress has been achieved in the last years in the development of new clustering approaches, the latter has not yet converged to a consensus.

Before clustering, many tools start with variants of feature selection and DR. In practice, selecting the most variable features (HVGs) is sufficient in most cases [318]. For DR, PCA and its variations are ubiquitous [355]. Finally, the distances are calculated from the obtained low-dimensional space. Many metrics are available, including Pearson's and Spearman's correlations, Euclidian distance, Cosine similarity and Jensen-Shannon divergence. For the clustering itself, the most popular algorithm is k-means. However, it should be used wisely as k-means results depend a lot on the initial positions of centroids and can converge to a local minimum. It is advised to re-run the algorithm multiple times to achieve a stable solution, as is implemented in the SC3 scRNA-seq analysis suite [355]. Another issue is that k-means is biased towards clusters of similar sizes, which results in the admixture of smaller clusters towards bigger groups. One way to solve it is to add an outlier detection function to the procedure [356]. Hierarchical clustering (HC) is another universal method that fares well for single-cell analysis. However, the method is costly regarding computational time and memory requirements that scale quadratically with the number of cells in the dataset because, for every split, the algorithm considers distances between many, if not all, data points. One peculiar expansion of HC that improves its ability to identify small clusters is to carry out DR after every split and merge [357].

Due to the mentioned limitations of k-means and HC, especially for large datasets, graph-based community detection algorithms are rapidly gaining popularity. The main conceptual difference is that these algorithms identify densely connected communities, i.e. vertices sharing many nodes, instead of neighbours packed densely. Another advantage of graph-based approaches is that they do not require an input of the predefined number of clusters. Naturally, these methods imply the construction of a k-NN graph, whereas the graph structure, the number and size of final clusters depend on the selected k number of neighbours. This is commonly resolved by reweighting the graph based on the shared nearest neighbours of each pair of cells. Although many algorithms for community detection are developed to study larger communities like social networks and the world web, only the Louvain algorithm has been extensively used in single-cell analysis [358]. The combination of Louvain clustering with shared-nearest-neighbour graphs became a preferred choice for a large share of the single-cell community and is now incorporated in both Seurat and Scanpy [359, 330]. In contrast to k-means and HC, however, this approach performs poorly in small datasets [360]. Overall, many other methods are in one way or another based on the concepts discussed but slightly deviate in their implementation and, therefore, might be better suited for specialised tasks [354]. As a forewarning, it is wise to remember that no universal clustering method achieves sound performance in all situations. Therefore, a clustering approach should be considered based on the data at hand.

Clustering leaves a researcher with a catalogue of functionally distinct cell groups, and unless the input cells were gated by FACS-sorting, the identities of these groups are not known. Thus, bringing to the front another computational problem — cell annotation. The veracity of cell annotation underlies all downstream biological interpretations of the patterns uncovered in the data. The task is further aggravated by the sheer size of an average scRNA-seq dataset that makes manual annotation unfeasible. Therein is the basis for a demand for automated annotation tools [361].

Observing the problem, one can soundly consider annotating clusters by referencing cluster-specific DEGs against some trustworthy reference. This approach, marker-based annotation, is among the most commonly used and is backed by dozens of large-scale scRNA-seq studies that aimed to create a reliable reference, i.e., cell atlases [362, 363, 364, 365]. The latter are available in different resolutions for most mouse and human tissues. Peculiarly, cellular landscapes of more exotic models were also chartered in recent years [366]. These datasets provide a faithful foundation to catalogue cell-type specific reference markers. Many resources readily utilised this foundation to build exhaustive and rigid reference databases [367, 368].

Similarly, a resource for identifying cancer-specific cellular processes, i.e. hallmarks of cancer, was developed building upon a compendium of cancer single-cell and bulk datasets [369]. The annotation tools such as scCATCH and SCSA provide the interface to these databases, which can automatically query the marker genes and score clusters based on their affinity to the reference cell [370, 371]. Another approach to annotation eliminates the marker gene intermediates and infers cluster identities directly from the correlation of query gene expression with the reference. One way to do this is to compare averages of the query clusters against an average expression of the reference cluster, a pseudo-bulk-based procedure in its essence. However, this upfronts the identified clusters' veracity, whereas unstable clusterisation of the query may lead to biased annotation. Additionally, it disallows cell-wise quantitative annotation, although the gains in annotation speed from moving from cell-wise to cluster-wise correlation are indubitable [372, 373].

Contrasting this approach, more fine-grained methods correlate the expression of individual cells to the reference. However, how the correlation or similarity of a cell to the reference is computed varies. In some, bulk-cell references are used, while others proceed with single-cell and bulk references. In these methods, feature selection plays a pivotal role in cell scoring, and, as a result, multiple strategies were devised, including the use of randomly selected genes, highly-variable genes, DEGs, and iterative correlation of random gene subsets against the reference [374, 375, 376, 377]. Although more computationally demanding, the ability to provide individual cell scores is an undeniable advantage. The same ability allows better identification of the bordering cell types in-between clusters.

Finally, the last group of methods utilises various machine-learning techniques to annotate cells. In essence, this task of supervised classification can be taken on from many different angles, starting from the way reference is constructed and ending with the selected model itself. In this regard, classic models like Random Forest (RF), Support Vector Machines, and kNN classifiers are utilised [378, 379, 380]. More recently, neural-network-based methods started gaining momentum [381, 382]. The enhanced capability of these methods to deal with the batch effect in the reference dataset is often highlighted [361]. In summary, the current status quo of annotation tools mirrors the one within the TI field, as evident from the large-scale benchmarks of annotation tools that often conclude by remarking on the situational superiority of some methods in different conditions [383, 384].

Another problem that arises whilst working with large masses of scRNA-seq seq data is the correction for batch effects. Inasmuch as the fight against batch effects is a long standing one in the bioinformatics field, it begs a question: could the methods developed earlier for bulk RNA-seq data be

readily, and, more importantly, succesfully, applied to the scRNA-seq data? The short answer is yes, but there is a nuance; even so, let us proceed with order.

The honing of scRNA-seq methods gradually led to the alleviation of constaints to the number of cells in the studies and promoted an establishment of large projects such as Mouse and Human cell atlases along with organ-specific atlases [362, 365, 265]. Naturally, the studies of that volume bring along issues of logistical nature: data is often generated by multiple operators, in different genomic centres, and, ultimately, by different hands. All of this inevitably introduces batch effects that will mask biological signal if the data to be integrated and analysed together, which is becoming the case more and more often, therefore necessitating batch correction, lest the biological signal remains entangled with technical noise [385].

Now, speaking of the batch correction methods desinged for bulk RNA-seq. To author's knowledge, the most popular are limma and ComBat [386, 387]; both operate by fitting a linear regression for each gene with a coefficients for a batch structure. ComBat additionally employs empirical Bayessian shrinkage of batch coefficients by sharing the information between genes. Now, the nuance: bulk RNA-seq methods operate on the assumption that cell composition in different batches is identical, i.e. the systemctic differences in gene expression are attributed solely to technical variation, which we know is not the case [388]. A panolply of factors could affect the end-state of a scRNA-seq library, from the stochastic variation in RNA sampling to cell sorting, resulting in the batches of variable cell composition. Consequently, the batch coefficients in linear models would seize on a fraction of biological signal along with the technical variability producing biased results. Nevertheless, one cannot attribute this fallacy exclusively to bulk RNA-seq methods as far as every batch effect correction inevitably regresses out a fraction of biological signal, the question is how big of a fraction [385]. The answer to this question can only be given by rigorous benchmarking studies that cover tasks of simple batch-removal as well as the integration complex biological datasets [389]. Before that, however, let us consider methods designed with scRNA-seq data specifics in mind.

In 2018, three batch correction methods were published [304, 388, 390]; let us cover them in an order of popularity, from the lowest to the highest (according to citations). The first tool in a row, ZINB-WaVe, developed by Risso et al., presents a zero-inflated NB model that, along with modelling of the gene expression, can similarly include coefficients to account for batch effects, which, nevertheless, assumes the equality of cell composition as older regression-based tools [304]. The second popular method is based on mutual nearest neighbours (MNNs) and called accordingly. In the essence,

this approach operates on the notion that cells of the same type should be expressionally the most similar between batches. In details, consider two batches A and B that contain m and n cells accordingly. Then, for each cell i_A from batch A, the algorithm finds k nearest neighbours in batch B. Likewise, for each cell j_B in batch B the algorithm finds k nearest neighbours in batch A. When a pair of cells is contained within each others neighbourhood in both batches, these cells are labeled as MNNs [388]. MNNs are considered to be cells of the same type, therefore any systamtic differences in expression are attributed to technical variance and corrected. MNNs operate on a number of assumptions: 1) at least on cell type is shared between batches, 2) technical and biological signals are orthogonal, 3) technical variance is meagre compared to biological signal. While 1) and 3) are likely to hold true in most, if not all, cases, 2) raises concerns, insofar as orthogonality is not guaranteed. The third method is distributed within the Seurat framework and reigns supreme in popularity alongside its parent framework [390]. First iteration of the tool utilised Canonical Correlation Analysis (CCA), a method conceptually similar to PCA: whereas PCA constructs linear combinations of input features maximising variance, CCA does the same but maximises correlation of the features between datasest, to extract correlated features to harmonise the batches. The second iteration of this method that is currently present in Seurat v3 combines CCA with MNN: the algorithm first reduces the dimensions by CCA to a number of canonical variables and then looks for MNNs (called "anchors") within the correlated subspace [391]. Since the PCA was mentioned, it is worth to address another popular method that was published in 2019: Harmony [392]. This method operates fuzzy clustering within the subspace derived from PCA. It caclulates global centroids that, ideally, should cover similar cell types from both batches and then calculates correction factors relative to the centroid's centre for each batch. After the cells from different batches are corrected, the algorithm repeats itself until a convergence criterion is reached. Finally, one approach is of interest as it explicitly renounces the assumption that technical and biological signals should be orthogonal [393]. The algorithm, LIGER, builds upon the previously published integrative Negative Matrix Factorisation (iNMF) method honed to identify two sets of shared and dataset specific factors (called "metagenes" in the LIGER publication). Loadings of the sahred metagenes are used after to construct a neighbour graph to prevent spurious assignment of diveregent cell types to one cluster. Finally, more exotic methods leveraged deep neural networks to construct a probabilistic framework for a wide variety of tasks including data integration [394]. The main drawback of this approach, if it can be called this way, is its complixty, which can take time to build an intuition about.

In summary, going back to the posed question: how big of a fraction (of biological vairance is removed by integration)? With a bird's eye view

of the available tools, one cannot but notice that the initial question of biological variance recovery get complemented by others, more practical inquiries. For example, how the memory requirements scale with the sizes of the datasets to be integrated? Does the size of the datasets affect the integration perforamnce? Are some tools better in recovering cell states or trajectories? How strong the batch effect a method can correct? Luckily, the answer to all this inquiries is the same as to the founding question: benchmarking. A recent work by Luecken et al. shines as, to author's knowledge, the most comprehensive and up-to-date benchmarking study of scRNA-seq integration tools that covers all raised question [389]. For the author, it is only left to conlude that, alike to TI and clustering, there is no universal solution for batch effect removal, and one needs to pick a method based on situational demands of the task at hand.

2.15 Thesis scope

From observation of the contemporary trends in the single-cell field, one can soundly conclude that the expansion of number of cells per study is favoured among all other directions of improvement. In this regard, the exponential growth of the cell number makes the task of automatic cell annotation exponentially more critical [395]. This is particularly true for cancer studies, where the inherent heterogeneity of cancer cells places a vital requirement of robustness upon the annotation methodology. To exemplify, in the analysis of a scRNA-seq from a tumour sample, cancer cells often cluster separately, potentially due to the difference in their clonal origin and phenotypic constraints imposed by the environment, whereas the annotation should classify them into one group. In the deliberation on the cancer evolution, the author referenced temporal genomic profiling of ccRCC that exhaustively catalogued the potential evolutionary trajectories of this tumour type [234]. Six different tumour evolutionary trajectories were identified in that study, each with a unique clonal structure and mutational signature.

Nevertheless, all trajectories gravitated towards a singular gene pathway, hinting at the existence of a gene expression signature, from which an investigator can infer the degree of carcinogenic activation of that pathway. However, the gene signature design capable of capturing carcinogenic expression patterns across different tumour types is more complex. This complexity arises from the cancer's tissue of origin, whose expression patterns are still noticeable after malignant transformation [396]. This issue with tumour heterogeneity is further exacerbated by the variability of expression signatures due to the cell cycle and effects of the tumour microenvironment, e.g. hypoxic condition, immune response, etc. [397].

In consideration of these methodological impositions, the author and

his colleagues (the group) set out to develop Ikarus, an automated single-cell classifier that can robustly discern cancer cells from normal in different cancer types (Publication 1) [398]. The designed classifier machine operates in a multi-step framework that combines gene signature and graph-based methods [399]. Since the machine was conceptualised to work across cancer types, the algorithm to build up cancer gene signatures was developed intentionally versatile, i.e. the default gene signature identified by the group can be refined to better suit the analysis goals by incorporating new single-cell datasets as a reference.

Regarding gene signature, it is paramount to properly characterise and validate the latter to ensure that the obtained combination of genes is not composed of correlated features, consistent across different model systems and sequencing technologies, and not already defined. To this end, the qualities mentioned above of the gene set were ensured by an all-around validation devised by the author. Given the method's versatility, the group's efforts were steered to ascertain that the procedure could be extended to other multi-omics data, for which CNVs were chosen as a case study [400].

Earlier, the author touched upon translating our knowledge about cancer to the clinic and discussed the history and development of biological systems used to model cancer. A trend is evident among the studies in this field: the expansion of -omics profiling. This trend is omnipresent across all models, from old to new, including cell lines, tumour xenografts and organoids. Results from the publications accompanying releases of new multi-omic datasets univocally affirm the benefits of multi-omic profiles whereby novel drug sensitivities and in-sensitivities are discovered. Nevertheless, the speed with which transcriptomic and multi-omics to this matter permeate into clinical practice is surprisingly low. To this end, a standard approach to diagnostics has evolved in cancer centres. This approach relies on deep sequencing of a selection of transcripts whose widespread mutability is supported by decades of cancer genomics [192]. Dictated by reasons of economic feasibility and ethical needs to strike a delicate balance between errors of the first and second kind, the panel-based approach persists in cancer diagnostics. Despite that, the technology advances unyieldingly, and the costs of sequencing a transcriptome from a tumour biopsy are steadily approaching clinical affordability. In publication 2, the group set out to test if adding transcriptomic modality to panel mutation tests significantly increases the power to predict drug response, and if yes, to which extent and for which categories of drugs. To be sure that results generalise to other test systems, the analysis was repeated in cancer cell lines, xenografts, and ex vivo treated fresh tumour specimens [148, 163, 401].

3 Material and methods

3.1 The scraping of citation data from the Web of Science

Source data for Figures 1 and 2 were pulled manually from the Web of Science advanced search web page. For robustness, the queries were constructed to explicitly look for the technology in question and avoid counting unrelated references. The following queries were used to search for publications mentioning scRNA-seq:

(TS="single-cell" AND TS="RNA sequencing") OR (TS="single-cell" AND TS="RNA seq") OR (TS="single-cell" AND TS="transcriptomics") and DT=Article

The queries below were used to scan for the publications that reference RNA-seq:

TS="RNA seq" OR TS="RNA sequencing" and DT=Article

And microarray gene expression assays:

TS="oligonucleotide array" OR (TS="microarray" AND TS="gene expression") OR (TS="array" AND TS="gene expression") and DT=Article

4 Results

4.1 Publication I

Jan Dohmen*, *Artem Baranovskii**, Jonathan Ronen, Bora Uyar, Vedran Franke, and Altuna Akalin, Identifying tumor cells at the single-cell level using machine learning
* These authors contributed equally to the work

This study was published on 30th May, 2022
in Genome Biology volume 23, article number: 123
https://doi.org/10.1186/s13059-022-02683-1

Supplementary materials related to this publication could be found in Appendix I.

Dohmen *et al. Genome Biology* (2022) 23:123
https://doi.org/10.1186/s13059-022-02683-1

Genome Biology

METHOD

Open Access

Identifying tumor cells at the single-cell level using machine learning

Jan Dohmen[1], Artem Baranovskii[2,3], Jonathan Ronen[1], Bora Uyar[1], Vedran Franke[1*] and Altuna Akalin[1*]

*Correspondence: vedran.franke@mdc-berlin.de; altuna.akalin@mdc-berlin.de

[1] Bioinformatics and Omics Data Science Platform, Berlin Institute For Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Hannoversche Str.28, 10115, Berlin, Germany
Full list of author information is available at the end of the article

Abstract

Tumors are complex tissues of cancerous cells surrounded by a heterogeneous cellular microenvironment with which they interact. Single-cell sequencing enables molecular characterization of single cells within the tumor. However, cell annotation—the assignment of cell type or cell state to each sequenced cell—is a challenge, especially identifying tumor cells within single-cell or spatial sequencing experiments. Here, we propose ikarus, a machine learning pipeline aimed at distinguishing tumor cells from normal cells at the single-cell level. We test ikarus on multiple single-cell datasets, showing that it achieves high sensitivity and specificity in multiple experimental contexts.

Keywords: Single-cell genomics, Machine learning, Cell type classification, Cancer

Background

Cancer is a disease that stems from the disruption of cellular state. Through genetic perturbations, tumor cells attain cellular states that give them proliferative advantage over the surrounding normal tissue [1]. The inherent variability of this process has hampered efforts to find highly effective common therapies, thereby ushering the need for precision medicine [2]. The scale of single-cell experiments is poised to revolutionize personalized medicine by effective characterization of the complete heterogeneity within a tumor for each individual patient [3, 4].

Recent expansion of single-cell sequencing technologies has exponentially increased the scale of knowledge attainable through a single biological experiment [5]. The information contained within a single high-throughput single-cell experiment enables not only characterization of variable stable states (i.e., cell types, and cell states) but also functional annotation of individual cells, such as prediction of the differentiation potential, susceptibility to perturbations, and inference of cell–cell interactions [6].

As with all new technologies, high-throughput single-cell sequencing also created new computational challenges [7]. A problem in single-cell data analysis is cell annotation—assignment of a particular cell type or a cell state to each sequenced cell. The size of

the generated datasets made manual annotation approaches utterly unfeasible, while the peculiarities of data generation prompted the development of novel innovative classification methods [8–13]. This is especially apparent in datasets stemming from cancer tissues, where the variability in the transcriptomic states does not conform to classically defined cell types. An outstanding question is whether there exist transcriptomic commonalities between cancer cells originating from different cancers, and whether it is possible to create a model which would discriminate between cancer cells and the surrounding tissue across different cancer types, and datasets.

High-throughput single-cell technologies provide unprecedented precision of characterization of biological systems, with all the technical and biological influences being evident in the data. In cancer biology, this heterogeneity of data composition presents a particular problem, because it is very hard to enumerate, and correct for, all of the technical and biological variables which are giving rise to the measured variability [14]. For example, cell dissociation produces artifacts which mimic MAP kinase pathway activation [15], while it is impossible to know the exact environment influencing each cell. Cells might be in gradients of oxygen availability, under different physical constraints, or influenced by multiple varying signaling molecules. This variability presents a challenge for developing machine learning models, because the data coming from different conditions will have different underlying distributions, meaning that methods trained on one dataset will not generalize to other datasets [16].

Currently, there are three types of methods for mitigating distributional differences between single-cell RNA sequencing datasets: (1) manifold matching methods that try to find commonalities between low dimensional representations of multiple datasets and align them into one space [17]; (2) domain adaptation deep learning tools that try to model (explicitly or implicitly) the batch effects through the latent space embeddings [18–24]; and gene set based classifiers that use learned marker genes and robust statistics to transfer knowledge between datasets [8].

An issue recurrently arising in machine learning in biology is how to determine the generalization boundaries of trained models (i.e., on which datasets the model will fail). It is not evident whether the learned model will perform equally well on the data profiled using different sequencing technologies (i.e., Drop-Seq, 10X, CEL-seq or Fluidigm C1), produced by different laboratories, or originating from a different biological source (i.e., same cell types, coming from different human individuals). Because the sources of the heterogeneity are frequently unknown, the models need to be explicitly tested for robustness on datasets corresponding to different biological conditions and profiled using different technologies [25]. Therefore, special care needs to be taken that the learned associations really are between variables of interest and are not confounded by the properties of the data. Both manifold matching and domain adaptation methods follow a tradeoff between the removal of unwanted variance, while preserving biological heterogeneity [25]. Gene signature based methods lie on the opposite part of the tradeoff spectrum, whereby the gene lists represent a strong inductive bias about a biological property (cell type). If the gene lists are carefully tested, then the methods achieve a markedly low false positive rate.

We set out to answer a simple question: "Is it possible to make a classifier that would correctly differentiate tumor cells from normal cells in multiple cancer types?". We have

built ikarus, a stepwise machine learning pipeline for tumor cell classification. ikarus consists of two steps: (1) discovery of a comprehensive tumor cell signature in the form of a gene set by consolidation of multiple expertly annotated single-cell datasets and (2) training of a robust logistic regression classifier for stringent discrimination of tumor and normal cells followed by a network-based propagation of cell labels using a custom built cell–cell network [26]. With the goal of developing a robust, sensitive, and reproducible in silico tumor cell sorter, we have tested ikarus on multiple single-cell datasets of various cancer types, obtained using different sequencing technologies, to ascertain that it achieves high sensitivity and specificity in multiple experimental contexts. We have strictly adhered to machine learning best practices to avoid contamination of results by information leakage from training into testing process.

Results

Identification of a robust marker gene set

Cell type annotations in any particular experiment are inherently noisy. This is partly due to the properties of single-cell data, such as the different number of detected genes in each cell, the influence of sample processing, and our limited knowledge of biomarkers that are necessary for comprehensive annotation of cell types and cell states. We hypothesized that we can find robust markers of cellular states by comparing multiple independent annotated datasets from diverse origins.

We have employed a two-step procedure to find tumor-specific gene markers. First, using differential expression analysis, we selected genes that are either enriched or depleted in cancer cells per dataset (see 9). To obtain the final gene list, we took an intersection of the gene sets from each of the datasets (Fig. 1A). We have applied a standard cross validation approach for gene set selection, where datasets were either used as training, validation, and test sets. For cross validation, we have used the two lung cancer datasets from Laughney [27] and Lambrechts [28], a colorectal cancer from [29], neuroblastoma dataset from Kildisiute [30], and a head and neck cancer datasets from [31]. For each pair of datasets, we have selected the gene signature and trained the logistic classifier. The resulting classifier accuracy was validated on the datasets that were not used for training (Additional File 3: Cross validation results). As the performance metric, we have chosen a minimal balanced accuracy on the validation set (measuring the worst performance of the classifier on the validation set). The cross validation procedure showed that the gene signature selection using multiple datasets increases the generalization performance of the classifier (Fig. 1C). The best performing classifier combined the colorectal cancer dataset from Lee et al. [29] with the lung cancer from Laughney et al. [27], and achieved a minimal balanced accuracy of 0.97 on the validation data. The performance of the best performing gene set was tested on the hepatocellular carcinoma [32] (balanced accuracy of 0.93), and the lung carcinoid dataset [33] (balanced accuracy of 0.99).

The resulting tumor gene signature contained 162 genes that were significantly enriched in cancer cells across multiple datasets (Additional File 2: Gene Signatures). The resulting set of genes showed high specificity for cancer cells, from the head and neck cancer samples [31] (Additional File 1: Fig. S1A). This result indicates that the gene

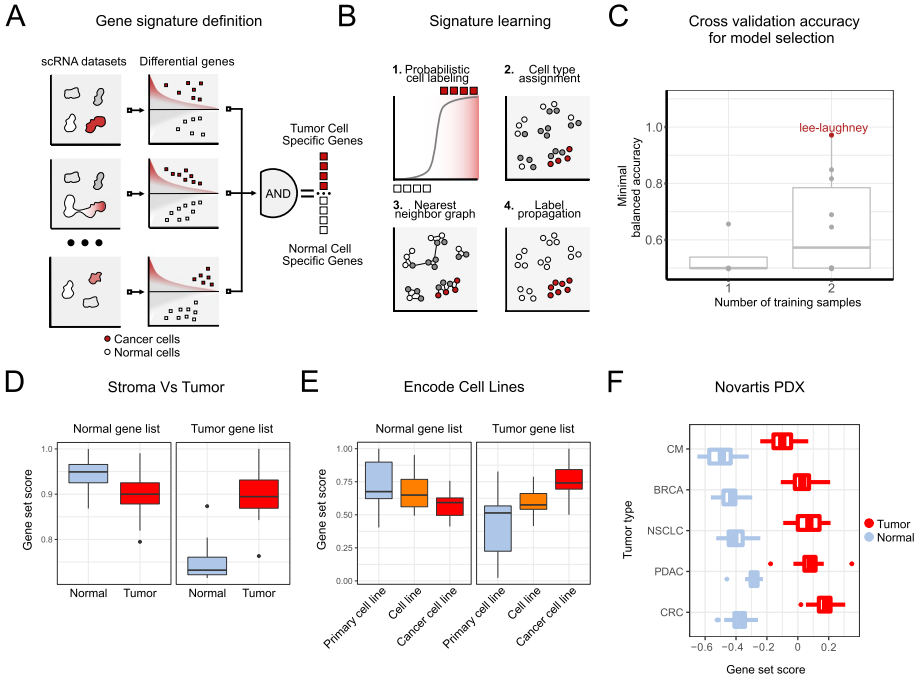


Fig. 1 Integration of multiple datasets enables robust extraction of informative gene sets. **A**, **B** ikarus workflow. ikarus is a two-step procedure for classifying cells. In the first step, integration of multiple expert labeled datasets enables the extraction of robust gene markers. The gene markers are then used in a composite classifier consisting of logistic regression and network propagation. **C** Comparison of cross validation accuracy for signature derivation and model selection. Minimal balanced accuracy on the validation set was chosen as the metric of choice (i.e., worse performance on the test set). Models trained on just one dataset achieved lower balanced accuracy than models trained on two datasets (p value given by the two sided Wilcoxon test is 0.063). The combination of colorectal cancer from Lee et al. [29] and lung cancer from Laughney et al. [27] achieved the highest minimal balanced accuracy of 0.97. **D** Comparison of gene signature scores in laser microdissected gastric cancer data [34]. The normal gene list shows lower signature scores in cancer samples (p value 0.052, $N=8$, Mood's median test), when compared to the cancer-associated normal tissue. The tumor gene signature is significantly higher for cancer samples than the normal tissue (p value 0.003, $N=8$, Mood's median test). **E** Primary cells and cancer cell lines have significantly different gene signature distributions. The normal-cell gene signature shows a gradual reduction in gene signature score distribution when compared in primary cells, cell lines, and tumor cell lines. The gene signature shows the complete opposite effect. Cancer cell lines have the higher gene signature score distribution, followed by cell lines, and primary cells. Distributions were compared using pairwise Wilcoxon tests with BH-FDR correction. All adjusted p values were lower than 0.01. **F** Patient-derived xenografts (PDX) show significantly higher tumor gene signature score, than the normal gene signature score. The same pattern is observed in multiple cancer types. Normal and tumor signature distributions were compared using Wilcoxon tests, for each cancer type, followed by BH-FDR correction. All adjusted p values were lower than 0.01

set contains information relevant for discriminating tumor cells from non-tumor cells in multiple different tumor types.

The same procedure was applied to the healthy cell types. We extracted genes enriched in each cell type, when compared to the tumor cells. The resulting gene set was then merged between multiple datasets. This "normal" cell gene signature contains both cell type specific markers and genes which are specifically depleted in the tumor cells (Additional File 1: Fig. S1B).

To validate the specificity of the novel tumor specific gene set, we have analyzed a gastric cancer dataset [34], where multiple areas of cancer and cancer-associated normal

cells were separated using laser-capture microdissection (LCM) and profiled using RNA sequencing. Using normal and tumor gene signatures that were identified by ikarus, we have scored the tumor and the associated normal cells. As expected, dissected sections coming from the cancerous lesions had significantly higher median tumor score than the surrounding normal tissue (Fig. 1D, right panel). In line with the latter, normal tissue scored higher than cancerous lesions when the normal gene signature was scored (Fig. 1D, left panel).

As another line of evidence, we have downloaded the expression data for primary, normal, and cancer cell lines from the ENCODE database [35, 36] (see 9). Tumor signature scores were on average highest in cancer cell lines, diminished in normal stable cell lines, reaching its lowest average in primary cells (Fig. 1E, left panel). When scoring using the normal (non-cancer) cell signature, an opposite trend was observed, i.e., score was highest in primary cells, intermediate in normal stable cells and the lowest in cancer cell lines (Fig. 1E, right panel).

Further, we tested the discriminatory power of the normal and tumor gene lists in multiple cancer types. To this end, we have used the patient-derived xenograft (PDX) samples from five cancer types provided by [37] and all of the cancer cell lines provided by the cancer cell line encyclopedia (CCLE) [38]. The tumor signature score was significantly higher than the normal signature score in all PDX cancer types (Fig. 1F) and all cancer cell lines screened in CCLE (Additional File 1: Fig. S1C). Surprisingly, the tumor signature list produced significantly reduced scores for cell lines stemming from blood-related cancers (LAML, CLL, LCML, MM, DLBC).

Accurate delineation of cancer cells

In the first step of classification, ikarus derives the tumor and normal gene set scores. The tumor and normal gene set scores are then used in a logistic regression classifier, to delineate cells with high probability of being tumorous or normal. The classification step is followed by the propagation of the cancer/normal label through a custom based cell–cell network (Fig. 1B). The cell–cell network is derived from the same gene sets that are used for robust scoring. By using only tumor or cell type specific genes, the resulting network separates communities that represent either tumor or normal cell states.

Figure 2A shows the performance of ikarus classification on all of the validation and test datasets. ikarus achieves an average balanced accuracy of 0.98 which is substantially higher than other classical machine learning methods. In addition to the standard machine learning methods (SVM, random forest, and logistic regression), we have compared ikarus to the top ranking tailored cell type classifiers, as evaluated in the recent comparison of methods for cell classification [8]: SingleCellNet [9], ACTINN [39], ItClust [10], scCaps [40], scHPL [12], CellAssign [41] from scvi-tools [42], and scMRMA [43]. We would like to emphasize that for the published methods, we have used the default hyperparameter settings from the corresponding descriptions or provided tutorials.

We have chosen balanced accuracy as a measure of performance because of the large imbalance of classes. The datasets contained, on average, 7 times more normal cells than annotated cancer cells (Additional File 2: Datasets). To give an unbiased view on the performance, Fig. 2B shows the area under the receiver operating characteristic (AUROC) distribution for the different datasets. ikarus also achieves a higher average AUROC

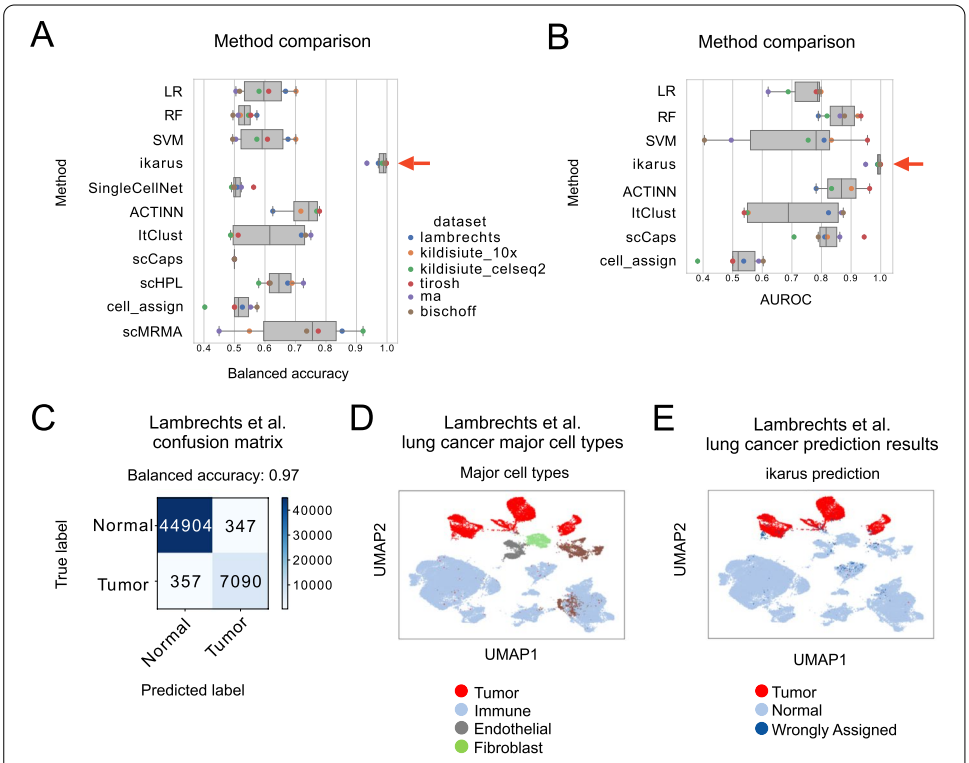


Fig. 2 ikarus accurately classifies tumor and normal cells. **A** Balanced accuracy for classification of tumor and normal cells, for each of the test datasets. Red arrow highlights performance of ikarus classifier. ikarus achieved a significantly higher performance than competing methods (AUC distributions were compared using ANOVA with post hoc pairwise comparison. *p* values were adjusted using BH-FDR. All adjusted *p* values were lower than 0.01). **B** Area under receiver operating characteristic (AUROC) performance for each classifier. AUROC could only be calculated for methods outputting class probabilities. **C** Confusion matrix showing the performance of ikarus classifier on the Lambrechts lung cancer dataset. **D** Cell type annotation of the Lambrechts et al. dataset. **E** Lambrechts dataset labeled by ikarus classifier

than other methods. Having a high AUROC value and low balanced accuracy is an indication of class imbalance. The classification error of the classical machine learning methods, having high AUROC and low balanced accuracy, is not uniformly distributed—they struggle with a high false positive rate.

We have additionally compared all methods with datasets subsampled to include an equal number of normal and tumor cells. ikarus showed a higher median balanced accuracy in discriminating tumor cells from normal cells (Additional File 1: Fig. S2A). During the comparison of subsampled datasets, we have noticed an increase in variance of ikarus results. This is because the subsampling reduces the connectivity of the cell–cell network which is used for network propagation.

We have also tested the performance of different classification methods by scaling down the input genes, from all profiled genes, to the tumor and normal gene signatures. The reduction of input to only normal and tumor gene signatures surprisingly increased the performance of all classifiers, indicating that the signatures contain information for proper discrimination between tumor and normal cells (Additional File 1: Fig. S2B).

Figure 2C shows the classification accuracy for the Lambrechts lung cancer dataset [28]. The Lambrechts dataset was not used for training nor gene signature definition. Figure 2D and E show the classification accuracy overlaid on UMAP [44] embeddings

of the Lambrechts lung cancer dataset [28]. ikarus correctly classifies normal cells, irrespective of the underlying cell types. The erroneous classifications are equally distributed between false positives and false negatives. (UMAPs for other validation and test datasets are reported in Additional File 1: Fig. S3 A-E).

In order to test the robustness of ikarus across different single-cell sequencing technologies, we applied ikarus on a dataset of neuroblastoma tumors sequenced by either 10X genomics or CEL-Seq2 protocols by Kildisiute et al. [30]. ikarus achieved a high classification accuracy (balanced accuracy of 0.98) on all datasets, irrespective of the profiling technology (Figures S3B and S3C). The false positive rate we observed in the test datasets (1–3%) can be partly attributable to occasional erroneous labeling of cells by the authors of the corresponding studies. The lack of a perfectly labeled single-cell tumor sequencing dataset makes it difficult to quantify the actual rate of false positive predictions by our method. One possible strategy to remedy this issue is to test our method on a dataset that is presumably free of tumor cells, in other words, a healthy tissue sample. To ascertain the actual false positive rate for tumor cell classification, we have tested ikarus on the single-cell data from peripheral blood of a healthy individual [45], where all cells are expected to be non-tumorous. ikarus labeled all cells as non-tumorous (Additional File 1: Fig. S3F).

Because the datasets used for training and testing consist predominantly of carcinomas, we decided to test ikarus performance on a synovial sarcoma sample [46]. On sarcoma samples, ikarus achieved a reduced balanced accuracy of 0.51, which was primarily driven by a high false negative rate—ikarus missed sarcoma tumor cells (Additional File 1: Fig. S3G).

Further, we were interested in how the accuracy of the classification changes in regards to the size and structure of the gene set. First, we have conducted an ablation study, where we removed from 20 to 80% of randomly selected genes from the list (Fig. 3A). The removal of up to 40% of the genes from the list leads to a ~12% (from 99 to 87%) drop in median accuracy. If 80% of the gene list is removed, the classification becomes random (median accuracy tends to ~50%).

Next, we explored how much the accuracy of the classification depends on individual genes. To test this, we sequentially removed each individual gene from the set and repeated the classification. For 160 out of the 162 genes, there was no observable change in the classification accuracy on the test datasets (Fig. 3B). The accuracy on the Lambrechts lung cancer [28] dataset was, however, particularly sensitive to the omittance of two genes: serum amyloid A (SAA1) and fibrinogen beta chain (FGB) (Fig. 3C). Each gene is a marker for a tumor specific cell cluster in the Lambrechts dataset (Fig. 3C, D), and their removal influences the classification probability of cells constituting that particular cluster. Such dependence was not observed for other test datasets (Additional File 2: Effects of SAA1 and FGB).

Properties of the tumor gene signature

Having observed the high accuracy performance of ikarus based on the detected tumor gene signature, we ventured forth to obtain a deeper characterization of the functional

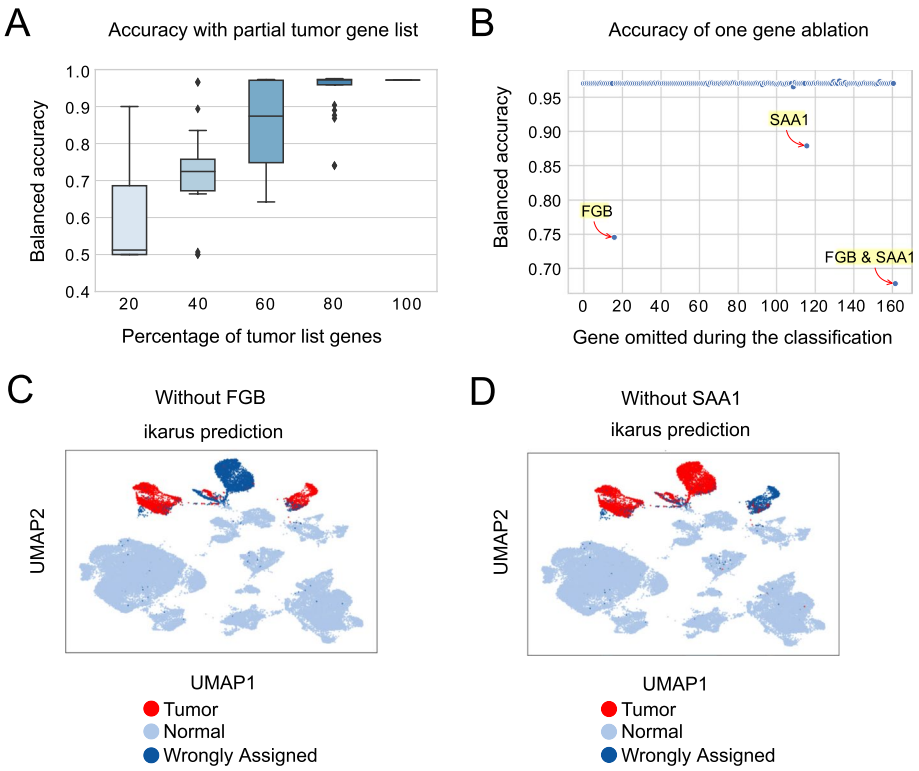


Fig. 3 ikarus performance under perturbation. **A** Performance of ikarus classifier, on the Lambrechts' validation set, when 20 to 80% of the input list is removed. ikarus performance is significantly reduced when the 60% of the gene list is removed (p value < 0.01, as determined by a one sample t-test, with BH-FDR correction). **B** Performance of ikarus classifier, on the Lambrechts' validation set, upon single gene removal from the gene list. One hundred sixty out of the 162 genes were inconsequential for the classification. FGB (fibrinogen) and SAA1 (serum amyloid alpha) ablation significantly reduced the classification accuracy, but only on the Lambrechts et al. lung cancer dataset. **C**, **D** FGB and SAA1 are respectively strong markers for one of the cell state clusters in the lung cancer dataset

content of these genes. Specifically, their involvement in the development of cancer and their roles in the prognosis for the patients.

Firstly, we were interested whether the genes within the tumor gene signature conform into expression modules, or whether their expression distribution is independent. We calculated the pairwise Pearson correlation between the genes from the signature for all datasets. To our surprise, the correlation between the genes was largely zero (Fig. 4A). We found only a single module (containing 34 genes) that was robustly present in all datasets (Additional File 1: Fig. S4A). Genes in this module are annotated as belonging to the cell cycle. We further inspected whether the classification accuracy depends on these cell cycle-related genes. The removal of the 34 cell cycle-related genes did not affect the classification accuracy (results not shown).

The tumor gene signature had, to our surprise, little overlap with established cancer-related gene sets. When compared with the gene sets annotated in the CancerSEA database of cancer functional states [47], our tumor gene list had zero or very few overlaps with most CancerSEA gene sets, except for the cell cycle genes, which shared only 9 genes with our tumor gene list (Fig. 4B). Co-expression analysis, using SEEK [48], again showed that the tumor gene signature is partially related to cell cycle and

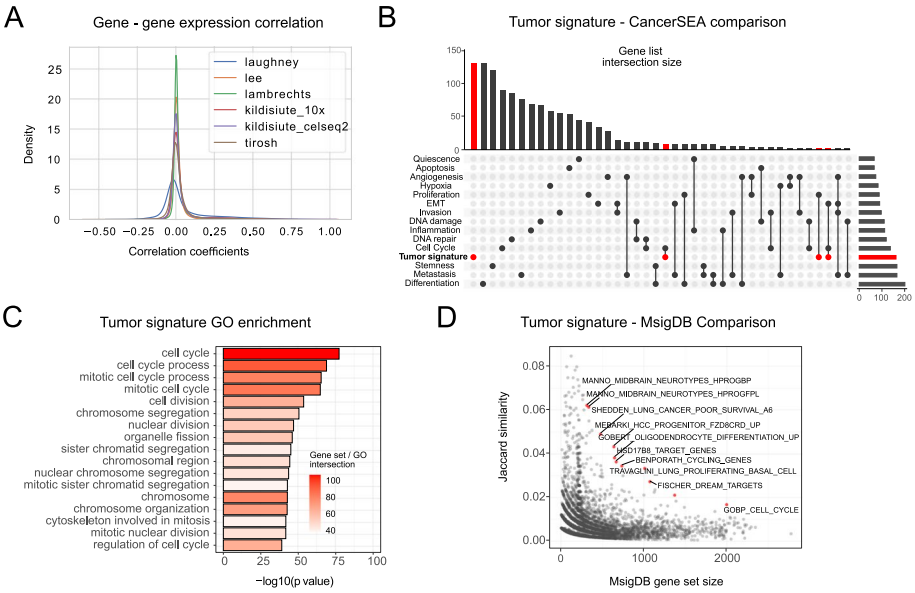


Fig. 4 Properties of the tumor gene signature. **A** Tumor gene signature co-expression analysis in tumor cells. Co-expression is measured as Pearson correlation between all pairs of genes from the gene list. **B** Tumor gene signature shows little overlap with known cancer-associated gene sets. The tumor gene signature was compared to the CancerSEA database. Out of the 162 genes, 134 showed no overlap with any of the gene lists. 9 genes overlapped with the cell cycle gene list. Only intersections of size 3 and more are shown. **C** Results of the gene co-expression analysis using SEEK. Top 150 co-expressed genes identified by SEEK were used as an input for GO enrichment analysis. **D** Tumor gene signature shows limited overlap with most of the gene sets from MsigDB. The most enriched gene set corresponds to the cell cycle

DNA replication (Fig. 4C). In addition, we saw no overlap with the cancer hallmarks from MSIGDB [49] (Additional File 1: Fig. S4B). When compared to the complete MSIGDB database, the tumor gene signature preferentially overlapped with the cell cycle hallmark (Fig. 4D). Gene ontology (GO) analysis using gprofiler2 [50] showed an enrichment of terms exclusively related to cell cycle and mitosis (Additional File 1: Fig. S4C). We have tested the GO and SEEK enrichment after the removal of the cell cycle module. The analysis did not result in any statistically enriched terms.

The enrichment of cell cycle and DNA replication-related functional terms in our tumor gene set (Additional File 2: Enrichment analysis of tumor gene signature) led us to hypothesize that the novel gene set differentiates promptly cycling cells. To test this hypothesis, we inspected the correlation of the tumor gene set scores with the growth rates detected in Patient Derived Xenograft (PDX) samples from [37] and the doubling times of the cancer cell lines from CCLE [38]. Unexpectedly, there was no correlation between the tumor signature score and the PDX growth rates in any of the reported cancer types (Additional File 1: Fig. S4C). Repetition of the analysis on the cell line doubling times from CCLE again revealed the same lack of correlation (Additional File 1: Fig. S4C).

We were interested in whether the tumor cell signature is predictive of survival outcomes in cancer. From the protein atlas database (http://www.proteinatlas.org) [51], we extracted genes predictive of survival in one or more cancers. The overlap of tumor gene signature with the extracted gene set showed that more than 75% of

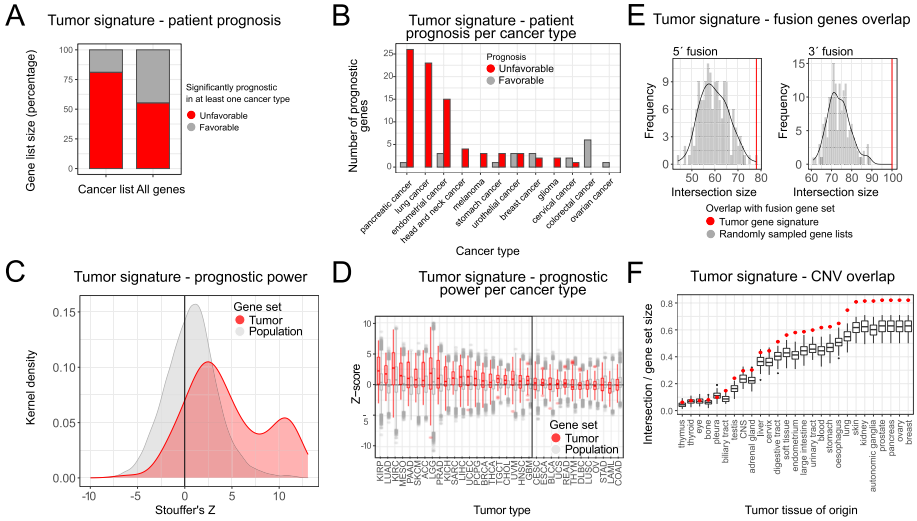


Fig. 5 Tumor gene signature is predictive of survival and associated with copy number variations. **A** Tumor gene signature genes are more likely to be significantly prognostic for patient survival in at least one cancer type (enrichment *p* value was lower than 0.01, as determined by a chi-square test). Data was extracted from the protein atlas. **B** The association of tumor gene signature genes with poor survival outcomes is evident in multiple cancer types. *p* values for within cancer comparisons are reported in Additional File 3: Statistics. **C** Data from [52] show that the tumor gene signature genes have much higher Stoufer's *z* (association with poor survival outcomes) than rest of protein coding genes (*p* value was lower than 0.01, based on the Wilcoxon rank-sum test). **D** Increased association of the tumor signature genes with negative survival is apparent in 23 out of 33 cancer types (the BH-FDR adjusted *p* value was lower than 0.01 as determined by Wilcoxon rank-sum test). **E** Genes from the tumor gene signature are more likely to participate in 3′ or 5′ fusions. When compared to sets of randomly drawn genes, the tumor signature genes had a significantly higher probability of participating in genomic fusions. The red vertical line depicts the overlap of the tumor signature list with the corresponding gene fusion list (the BH-FDR adjusted *p* value was lower than 0.01, as determined by the resampling test). **F** Genes from the tumor gene signature are frequently found in amplified or deleted genomic regions. We have measured the percentage of the gene list which overlaps with the known CNV regions for each cell line in the CCLE dataset. The tumor gene signature was found to be enriched in 22/27 cancer types (BH-FDR adjusted empirical *p* value was lower than 0.01 based on resampling tests). Background distributions were derived from expression matched randomly sampled gene lists

tumor signature genes are predictive of unfavorable prognosis in at least one cancer type (Fig. 5A). Interestingly, when stratified by cancer type, tumor signature genes reported to be unfavorable are overrepresented among 5 cancer types: liver, renal, pancreatic, lung, and endometrial cancers (Fig. 5B). An analogous analysis was done with data taken from [52], where the authors systematically calculated the risk predictive status for all genes in TCGA cancer types. The analysis showed that the cancer specific genes have a significantly higher Stoufer's Z value (a measure of how significantly the gene expression predicts the risk status in any cancer) than the rest of the annotated gene set (Fig. 5C). Furthermore, the same trend was observed in 21 out of 33 profiled cancer types (Welch two sample t-test, Bonferroni adjusted p value < 0.05) (Fig. 5D).

Next, we wanted to explore how often the genes from the tumor gene list participate in genomic rearrangements, particularly gene fusions, which are frequent drivers of oncogenic events in multiple cancer types. We downloaded the known cancer gene fusions from the ChiTaRS [53] database and inspected the overlap with the novel cancer defining gene list. To establish enrichment, we compared the overlap with a background

consisting of random gene sets. Genes from the tumor gene signature have a significantly higher probability of participating in both 3′ and 5′ fusions, than a random set of genes (Fig. 5E).

Gain and loss of DNA content is a ubiquitous property of tumor cells. Copy number variation (CNV) profiles that arise from genomic rearrangements create unique genomic signatures that can be used for characterization and discrimination of different tumor types [54]. We wondered how prevalent genes from the tumor gene signature are in the known CNV regions. To this end, we have compared the intersection of the tumor signature list with the CNV data from CCLE. We compared the tumor gene list intersection to a background distribution constructed by randomly sampling expression matched gene sets. The tumor gene list had a significantly higher overlap with the known CNV regions in the majority of profiled cancer types (Fig. 5F) irrespective of CNV frequency.

Multi-omics analysis reduces the false positive rate of classification

Characterization of biological systems from multiple viewpoints often produces synergistic insights into the underlying biology. We wondered whether the classification accuracy of ikarus algorithm can be improved by using multi-omics measurements. To this end, we have used inferCNV [55] to extract copy number variations (CNVs) from the single-cell RNA sequencing data. InferCNV is a Bayesian method, which agglomerates the expression signal of genomically adjointed genes to ascertain whether there is a gain or loss of a certain larger genomic segment. We have used inferCNV to call copy number variations in all samples used in the manuscript.

Firstly, we wondered whether the copy number variations could be used as universal markers for discriminating between normal and cancer cells. We trained random forest classifiers to discriminate between the expert labeled normal and tumor cells. One classifier was trained on each sample. Each of the classifiers was tested on all samples. As expected, when evaluated on the sample which was used for the training, each random forest classifier correctly discriminated between the cancer and tumor cells (Fig. 6A). The classifiers, however, did not generalize to other cancer types—they all suffered from a high false positive rate. We tried to improve the generalization of the classification by training on multiple datasets. We trained a random forest classifier on joint Lee et al. and Laughney et al. data and tested on all other datasets. Using multiple datasets for training did not improve the results of the classification on out of sample cells (Fig. 6B).

We then wondered whether the CNV calls could be used in conjunction with the gene expression data to improve ikarus classification of tumor and normal cells. We looked at the average CNV value and the variance of CNV values in cells, which were misclassified by ikarus in data from Lee et al. and Laughney et al. Both the average CNV value and the variance of CNVs were significantly higher in cancer cells, which were misclassified as normal cells (Fig. 6C). This indicated that by integrating the CNV scores with the gene expression classifier, we might increase the classification accuracy.

We have added an additional proofreading step into the classification procedure. We trained a logistic classifier on inferred CNVs, with ikarus predicted cell type labels as the dependent variable. Cells which obtained highly probable discordant class labels from the CNV classifier had their labels flipped. Using the proofreading step, the average balanced accuracy stayed the same for all of the samples. We have however noticed a

sudden drop in the false positive rate, with a marginal increase in the false negative rate (Fig. 6D, Additional File 2: Results).

Discussion

We have implemented a two-step approach for solving a problem that is perceived as simple: discriminating tumor cells from normal cells. In the first step, ikarus pipeline integrates multiple expert labeled datasets to extract gene sets which discriminate tumor cells from normal cells. In the second step, ikarus uses a robust gene set scoring along with adaptive network propagation for cell classification. By using robust gene set scoring and network propagation, we have mitigated two common problems in single-cell analysis: the influence of batch effects on sample comparison and parameter optimization during clustering.

The effect of technical differences between single-cell datasets is usually resolved using integration methods. Single-cell integration methods require extensive tuning of sets of parameters, most of which have non-intuitive effects on the results. Moreover, the accuracy of the resulting integrations cannot be trivially evaluated without extensive usage of biological priors. Gene set scoring methods are robust, because they use "within sample" rank based scores instead of direct comparison of measured expression values between different samples. The only technical variable that influences gene set scoring is the percentage of genes from the gene list which are detected in each cell. We have however extensively tested the influence of the number of genes on the classification accuracy.

A common step in single-cell analysis is aggregation of cells into clusters, which are then used for cell type annotation. Clustering is, however, a procedure with an inherently high number of parametric options. It is extremely hard, if not impossible, to choose a set of parameters that would produce the same level of accuracy (same cell types) on different datasets, which often necessitates manual intervention to deduce the best clustering resolution. Because cell types and cell states form highly connected modules within the cell–cell similarity graph, we have therefore opted to replace clustering with network propagation. Network propagation is a procedure where the uncertainty of cell annotation can be reduced by integrating the annotation score of each individual cell with the scores of its nearest neighbors. Network propagation represents a parameterless alternative to clustering, while retaining the same level of sensitivity for cell annotation.

By exploring a multi-omics approach, we have tried to increase the accuracy of the normal–tumor cell discrimination. Using inferred CNVs, we have shown that the information from copy number variations does not generalize across different cancer types. The inclusion of the copy number variation as a proofreading step reduced the false positive rate of the classifiers. It is still an open question, though, by how much would single-cell multi-omic measurements improve the classification (for example, by concurrently measuring mutations, CNVs, chromatin accessibility, and expression in the same single cells). Currently, such methods are either in their infancy, and the required data is not available, or have a very limited profiling range (profile only a handful of loci).

ikarus is currently constrained by the reliance on well annotated single-cell datasets. For both the gene set definition and testing, we rely on expert provided cell annotation. This requirement has limited our training and testing capabilities to the handful of profiled, and annotated cancer types. We have determined that ikarus produces accurate

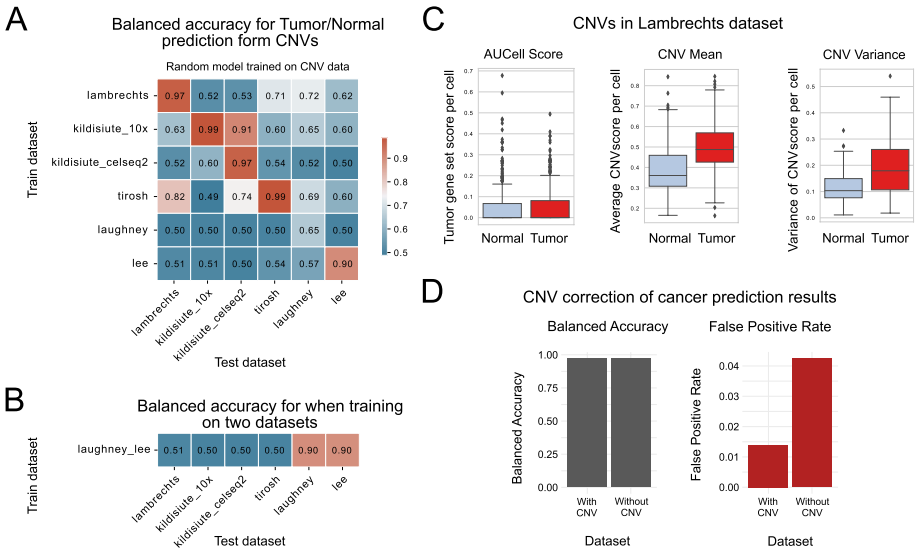


Fig. 6 Multi-omics decreases the false positive rate. **A** Balanced accuracy of random forest classifiers trained on each of the tumor types. The classifiers have excellent performance on the same samples they were trained on or on similar tumors (such as the Kildisiute et al. neuroblastoma sample), while they fail to generalize to other tumor types. **B** Training on multiple tumor types does not improve the generalization of the classifiers. **C** Cells misclassified by ikarus can be discriminated based on the average CNV and variance of CNV values. Tumor cells misclassified as normal cells have significantly higher values of both the average CNV and the variance of CNV, than the corresponding normal cells. **D** Integration of the CNV proofreading decreases the false positive rate from 4% to 1%, with the same average balanced accuracy

classifications in epithelial tumors, and the neuroblastoma; however, it showed reduced accuracy in classifying cells from synovial sarcoma. This implies that multiple trained models will be required for comprehensive discrimination of all cancer types. The exponentially increasing number of single-cell datasets will enable us to increase both the number of training datasets, as well as to test ikarus on currently unavailable tumor types, for instance, soft tissue sarcomas. Moreover, the increasing quality of single-cell datasets, most importantly, increasing gene coverage, will also increase the utility of ikarus as a gene set based classification method.

Conclusion

By integrating multiple datasets, we have derived a tumor signature gene list which is surprisingly refractory to annotation. The gene list contains a sub-module ($n = 34$) that encompasses genes involved in the cell cycle. All of the other genes, however, showed little modularity and a lack of enrichment in any single annotation category. Interestingly, the genes were highly expressed in all of the available PDX and CCLE cancer models. The ablation studies showed that the classifier was robust to the removal of any one of the genes. The low co-expression, combined with the lack of sensitivity to the gene removal indicates that the tumor signature genes provide mutually synergistic information towards the classification accuracy.

ikarus classifier, however, is not limited to tumor cell detection. It can be used to detect any cellular state, such as cell types. The only requirements are that the cellular state is present in at least two independent experiments, which are expertly annotated.

Automatic, parameterless discrimination between tumor cells and the surrounding tumor-associated tissue has multifactorial utility. Tumor cells can be streamlined into algorithms for neoepitope prediction, thereby enabling direct, clinically relevant insights. Furthermore, the increasing availability of multi-omics measurements would enable automatic genetic characterization of tumor subpopulations and the subpopulation-based recommendation of best therapeutic course of action. Application of automatic tumor classification on spatial sequencing datasets enables direct annotation of histological samples, thereby facilitating automated digital pathology.

The current scale of development in single-cell biology (on both the technological and computational levels) shows promise for quantitative characterization of the complete tumor heterogeneity, for each individual. However, before the personalized medicine approach can be readily adopted, every step in the data analysis needs to be completely automated, with robust performance guarantees. ikarus pipeline represents one step towards the implementation of personalized cancer therapy.

Methods

ikarus workflow

The presented ikarus pipeline consists of two major steps. In the first step, ikarus uses multiple expert labeled datasets to define gene signatures and builds a cell type specific classifier. In the second step, based on the constructed gene sets and classifier, unknown cells of interest are scored. The classifier's scores are then propagated through a custom cell-cell network, which eventually leads to the cell annotations. While the first step is optional, as users can provide their own gene lists, the latter steps are mandatory to make a prediction. For making predictions ikarus' API is modeled as the scikit-learn workflow, which means 1. load data, 2. initialize a model, 3. fit the model, and 4. make the actual predictions on unknown data. In general, annotated data objects are used as data format (AnnData, https://anndata.readthedocs.io). Each individual step is described in more detail in corresponding subsections below.

Defining gene signature lists

The count matrices of the input AnnData objects are expected to be normalized to the total number of reads per cell and log transformed with a base of 2 and a pseudocount of 1. In addition to that, each AnnData object must contain for each cell the corresponding cell type in the observation section, possibly in multiple columns for multiple hierarchical cell type levels (e.g., tumor and normal cells, or tumor, epithelial and immune cells).

It is important to take care that the input data is not scaled and that it contains the complete set of profiled genes and not a preselected set (such as highly variable genes).

Then, for each gene in the input dataset, a *t*-test with overestimated variance is used to compute an approximation of log 2-fold change between two cell type classes, one upregulated and one downregulated class. Those classes are provided by the user and should be chosen in accordance with the considered columns of the AnnData observation section. Users can either perform only one comparison (e.g., tumor vs. normal

cells) but also multiple (e.g., tumor vs. epithelial cells and tumor vs. immune cells). This is done independently for each dataset.

For each gene and for each pair-wise comparison, the associated log 2 fold changes (if $p_{\text{adj}} < 0.1$, neglected otherwise) of the different input datasets are averaged. According to these average values the genes are then sorted, highest to lowest and a user-defined number of top genes is selected (for our analyses, we used the top 300 genes). The final list of upregulated genes is derived by taking either the intersection or the union of selected genes across all of the comparisons.

The whole procedure is performed once for the case that the class of interest (e.g., tumor) is upregulated (here we take the intersection of selected genes across all of the comparisons) and once for the case that this class is downregulated (here we take the union of selected genes across all of the comparisons). That way, we obtain two final gene sets. One set representing genes enriched in the class of interest (i.e., enriched in tumor cells), and a set depleted in the class of interest (depleted in tumor cells).

A combination of Lee, Laughney, Lambrechts, Tirosh, and Kildisiute datasets was used to conduct a cross validation analysis. For each pair of datasets, gene signature selection was performed, followed by training of the logistic classifier. The resulting classifier accuracy was validated on the three datasets which were not used for training. The accuracy of the top performing classifiers was furthermore tested on the hepatocellular carcinoma (Ma) and carcinoid datasets (Bischoff). Cross validation results can be found in the Additional File 3: Cross Validation Results. As the performance metric, a minimal balanced accuracy on the validation set was chosen (i.e., what is the worst performance of the classifier on the validation set).

For comparison, classifiers were also trained on gene lists extracted from each of the datasets.

Cell scoring using gene sets

Both the tumor and normal gene sets were used to score each of the cells in each of the experiments using AUCell [56]. As input to AUCell, we provide the gene expression matrices that were normalized to the total number of sequenced reads per cell and subsequently transformed using the $\log2(x+1)$ function. AUCell requires that the dataset contains at least 80% of the genes from the input gene set.

We have noticed that the AUCell scores do not behave properly in some of the bulk sequencing datasets. Namely, samples which had similar transcriptomes sometimes had widely different AUCell scores. The user is encouraged to use different gene set scoring algorithms like ssGSEA.

Logistic classifier training

A logistic classifier was trained on the combined Lee et al. and Laughney et al. datasets. Scores of normal and tumor gene signatures were used as the input and the tumor/normal class assignment as the target variables.

Cell annotation using network propagation

ikarus implements the cell annotation as an iterative two-step process of cell type assignment and label propagation. In each iteration, we assign labels to cells with a decreasing

stringency threshold, which are then propagated to their nearest neighbors. Firstly, the labels are assigned to the most probable cells, based on a robust stringency threshold. Cells below the stringency threshold have their LR probabilities masked to zero. The stringency threshold is defined based on the order statistic of the gene set score difference between the two classes of interest. In the first iteration, it is the 90% percentile of the (tumor–normal) gene set score difference. The label propagation is then obtained by computing the dot product of neighborhood connectivities and LR class probability estimates. Annotations are derived from the propagated class probabilities. Within each iteration step, the stringency threshold is reduced using exponential decay:

$$N(t) = N_0 e^{-\lambda t}$$

where

N_0 is the starting stringency threshold;

t is an iteration step;

λ is an exponential decay constant.

The cell–cell graph, used for label propagation, is constructed using the normal and tumor gene signatures according to [57], as adopted in [58].

The label propagation procedure is repeated until less than 0.1% of cell annotations change.

CNV correction

Classification improvement using copy number variations

To improve the classification results, we used inferred CNV scores as an additional source of information. InferCNV [55] was used to compute CNV scores. A cutoff value of 0.1 was chosen for gene selection. CNV prediction was performed via HMM (Hidden Markov Models). For tumor sub-clustering the parameters were kept default (hclust = 'ward.D2', tumor_subcluster_pval = 0.05), though tumor_subcluster_partition_method was set to 'qnorm' as this is claimed to be reasonable faster than 'random_trees' No prior information on distinct clusters was provided.

In a self-supervised fashion, we used the current ikarus cell annotations as pseudo-labels to train an additional logistic regression model (LR). The LR takes as its input per cell inferred CNV values and predicts the cell annotations. The LR itself is trained on all cells from the validation dataset, e.g., Lambrechts et al., Kildisiute et al., Puram et al. This model was then used to make predictions on the same dataset assuming that logistic regression should not overfit on this task. The outcome is then considered as the final corrected ikarus prediction.

Gene set characterization

Gene set activity in cell lines and PDX models

Tumor and normal gene signature scores were calculated for bulk RNA sequencing data from laser microdissected data from gastric cancer, ENCODE cell lines, CCLE cell lines, and PDX data. Because AUCell was developed for single-cell RNA sequencing data, the signature scores were calculated using ssGSEA as implemented in the escape Bioconductor package [59]. The tumor gene signature scores were compared to the cell line doubling times and PDX growth rates that were provided as annotations to the datasets.

ENCODE cell lines dataset was further stratified into three groups: primary cell line, cell line, and cancer cell line. The stratification was done manually based on the annotation provided by ENCODE.

Comparison with published gene sets

The tumor signature gene set assembled in this study was characterized by comparison with publicly available gene sets provided by multiple public resources. We considered gene sets of various provenances, e.g., all Homo sapiens gene sets published by MsigDB [49], cancer-specific gene sets that represent distinct functional tumor states (CancerSEA) [47], novel gene lists covering previously unidentified members of cellular signaling pathways [60].

Gene sets as provided by MsigDB ($n = 31120$) were assessed via the interface provided by the R package msigdbr version 7.2.1. Next, intersections and unions of the cancer gene set (this study) with every human gene set from that release (version 7.4) of MsigDB, as well as the members of the intersections and original sizes of query gene sets, were computed.

CancerSEA resource provides a collection of functional cancer gene sets derived from a multitude of single-cell studies, thereby supplying a single-cell level scope to the cancer functional hallmarks. For characterization of the cancer gene set assembled in this study, we downloaded 14 gene sets $n_{min} = 66$, $n_{max} = 201$ from the CancerSEA resource (http://biocc.hrbmu.edu.cn/CancerSEA/goDownload) representative of distinct cancer functional states and intersected it with our gene set. The results of this analysis were presented with the UpSet plot framework [61] implemented in the ComplexHeatmap R package [62], mode intersect.

To account for recent advances in the annotation of cellular signaling pathways, we downloaded a novel collection of gene sets composed of genes previously unmapped to any signal transduction pathway. Namely, we acquired 11 gene sets of various sizes $n_{min} = 10$, $n_{max} = 164$) and intersected with the tumor signature gene set of this study. The visualization of this analysis was similar, i.e., the intersections were presented with an UpSet plot implemented in ComplexHeatmap R package, mode intersect.

Gene fusions

Data on human gene fusions were downloaded from the ChiTaRS resource as was provided on August 16, 2019 (http://chitars.md.biu.ac.il/index.html) [53]. First, we constructed a background distribution from randomly selected sets of genes that were expression-matched to the tumor signature gene set (this study). Every random gene set was intersected with fused genes from the database and the resulting intersection sizes were used to fill a background distribution. Lastly, the tumor signature gene set from this study was intersected with the list of fused genes to compare with the background distribution. This analysis was done separately on 5′- and 3′-fused genes.

Co-expression analysis

To investigate genes that are co-expressed with the tumor signature gene list across many datasets, we took advantage of web-based platform SEEK (https://seek.princeton.edu/seek/) [48]. We queried the tumor gene list to the SEEK search engine and downloaded a SEEK-generated ranked list of co-expressed genes. For further analyses we used top 150 genes from the ranked list.

Gene ontology (GO) analysis

The GO analyses throughout this study were done using the framework provided by gprofiler2 R package [50]. From the default run settings, p values threshold was changed to 10e−4 and correction_method option set to g_SCS.

Gene set sensitivity testing

To measure ikarus' robustness on the extracted tumor gene list, we performed the following analyses:

Gene set size

Using the Lambrechts et al. lung cancer dataset, we iteratively computed ikarus' balanced accuracies ablating a random section of the tumor gene list in a cumulative stepwise manner before scoring and prediction steps. Namely, we randomly removed 20, 40, 60, and 80% percent of the tumor gene list. Every ablation percentage was itself reiterated 25 times. The predictions were not CNV-corrected.

Single gene ablation

Further, we investigated the prediction value of individual genes in the gene list. We employed a similar procedure as before, but in contrast to ablating the whole sections of the list, we removed an individual gene from the list per iteration before computing ikarus' balanced accuracies.

Gene set prognostic power analysis

To infer the prognostic power of the generated gene set, we referred to prognostic data available in the TCGA. Namely, we downloaded a dataset of Cox-proportional hazard model z-scores that were generated for every gene expression feature across all available tumor types [43]. The distributions of the gene set z-scores were compared to the distribution of all gene expression z-scores (population) in every cancer type individually. Additionally, the cited research provided estimates of Stoufer's z-scores per gene expression feature. This metric represents a normalized prognostic average over all cancer types available in the dataset. Here, the same procedure was used; we compared the distribution of Stoufer's Zs in the gene set to the distribution of all gene expression features.

CNV analysis

To investigate the overrepresentation of copy number amplifications among the genes in the extracted tumor gene list, we referred to the COSMIC database. Namely, we downloaded a complete COSMIC collection of copy number alterations and stratified it by tumor tissue of origin ($n = 27$). Next, we iteratively intersected the tumor gene list from

this study with significantly amplified genes (denoted as "gain" in the COSMIC table) over tumor tissues of origin. As a random control, we prepared similarly sized random gene sets ($n = 162$) that were expression-matched to the original tumor gene list. Expression matching was done on Laughney et al., Lee et al., and Lambrechts et al. datasets independently. In total, 150 random gene sets were generated, 50 sets per expression dataset.

Comparison to reference methods

The methods used as reference for ikarus were installed and used to the best of our knowledge. We used default hyperparameter settings from the corresponding description or provided tutorials. We did not perform any kind of hyperparameter optimization. We would like to point out that the published classification methods have many tunable parameters, and tuning the parameters might significantly increase their performance. Both CellAssign and scMRMA assume a marker gene list for the target cell type prediction. We provided here the tumor and normal gene signatures generated with ikarus.

Dataset subsampling

To ascertain whether the classification methods accuracy can be improved by balancing the classes, we randomly subsampled 1000 tumor and 1000 normal cells 100 times for each of the following datasets used for validation, Lambrechts et al., Kildisiute et al., and Puram et al., and evaluated the classifier performance on the datasets.

Statistical testing

Statistical tests performed, groups in comparison, and sample sizes are summarized in Supplementary Additional File 3: Statistical Comparisons. In cases of multiple testing, p values were adjusted using Benjamini-Hochberg (FDR) method [63]. For situations where tests were not applicable, random background distributions were simulated against which the probabilities of observing an event under question were estimated. Testing approaches of such kind are reported as "empirical" in Supplementary Additional File 3: Statistical Comparisons. If the adjusted p value was lower than 0.01, it was reported as statistically significant.

For comparing the distribution of non-tumor cells from ikarus' misclassifications for the Lambrechts et al. lung cancer dataset with the actual distribution of cell types, we performed pairwise for each of the misclassified groups 2×2 fisher exact test (Additional File 2: Lambrechts misclassification).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02683-1.

Additional file 1. Supplementary figures 1–4.

Additional file 2. Tables with data description and results.

Additional file 3. Results of statistical analysis.

Additional file 4. Review history.

Acknowledgements

We would like to sincerely thank Florian Uhlitz for a very constructive discussion and critical reading of the manuscript.

Review history
Review history is available as Additional file 4.

Peer review information
Stephanie McClelland was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
AA conceptualized and planned the project. JD and AB jointly executed all of the computational analyses. BU, VF, and AA wrote the manuscript, with comments from JD and AB. JR critically edited the manuscript. BU, VF, and AA jointly supervised the work. VF led the day-to-day supervision and made sure the project was completed in a pre-defined timeline. All authors read, edited, and confirmed the final manuscript. AA acquired funding and supervised the project.

Funding
Open Access funding enabled and organized by Projekt DEAL. Akalin lab is supported by Berlin Institute of Health: Digital Health Accelerator in connection with this project. In addition, we acknowledge funding from the German Federal Ministry of Education and Research (BMBF) as part of the RNA Bioinformatics Center of the German Network for Bioinformatics Infrastructure (de.NBI) [031 A538C RBC (de.NBI)].

Availability of data and materials
Code
ikarus is a python package that can be found on the following link:
https://github.com/BIMSBbioinfo/ikarus
Code for reproducing the figures can be found on the following link:
https://github.com/BIMSBbioinfo/ikarus%2D%2D-auxiliary
Zenodo repository of the software libraries is available here [64].
Both repositories are available under the MIT license.
Datasets
Single-cell RNA-seq data:
Gene expression values for single-cell RNA-seq experiments are available through the corresponding publications. If not explicitly declared otherwise, the 10X genomics protocol was used for scRNA-seq.
Laughney et al. provide a lung adenocarcinoma (primary tumors and metastases) dataset that include 40505 cells coming from 17 patients. For our purpose, 1091 cells are considered tumorous, and 39,414 are normal.
63,689 cells from 23 colorectal cancer patients are coming from Lee et al. 16,248 cells are considered tumorous and 47,441 are normal. After our cross validation analysis, these two datasets serve as input for model building.
A non-small-cell lung cancer dataset is coming from Lambrechts et al. It considers 52,698 cells, of which 7447 are tumorous and 45,251 are normal. This dataset is used for model testing.
Puram et al. published 5578 single cells from 18 head and neck squamous cell carcinoma patients. They performed Fluorescence-activated cell sorting for scRNA-seq. 2215 cells are tumorous, 3363 cells are normal. We use this dataset for model testing.
Kildisiute et al. published a neuroblastoma cell atlas. We used 6442 cells (10X) from 5 patients (1766 tumorous, 4676 normal) and 13281 cells (CEL-seq2) from 16 patients (1630 tumorous, 11651 normal) as two distinct datasets for model testing. Ma et al. made a hepatocellular carcinoma dataset available with 17,164 tumor and 39,557 non-tumor cells. It was used as another test set. We also used a lung carcinoid dataset by Bischoff et al. for testing. It includes 8097 tumor and 55,230 non-tumor cells. Jerby-Arnon et al. published a synovial sarcoma dataset that we used for testing. It includes 8323 tumor and 851 non-tumor cells. A comprehensive description of the datasets can be found in the Additional File 2: Datasets.
For both input datasets, Laughney et al. lung adenocarcinoma and Lee et al. colorectal cancer, we considered a refined annotation for tumorous cells. Based on gene sets from MSigDB (v7.1) [49] hallmark collection HALLMARK_E2F_TARGETS, HALLMARK_G2M_CHECKPOINT, HALLMARK_MYC_TARGETS_V1, HALLMARK_MYC_TARGETS_V2, HALLMARK_P53_PATHWAY, HALLMARK_MITOTIC_SPINDLE, HALLMARK_HYPOXIA, HALLMARK_ANGIOGENESIS, and HALLMARK_GLYCOLYSIS, we scored each cell. If the average over all considered hallmark gene list scores (in the range 0–1) exceeds a reasonable threshold (0.45 for Laughney et al., 0.35 for Lee et al.), the cell is considered tumorous. Thresholds are chosen to minimize the amount of false positives with respect to the initial annotation of normal and tumor cell sources. The distribution of normal and tumor cell sources obtained from the initial annotation is provided in Additional File 2: Datasets.
ENCODE cell line dataset:
Gene expression values for primary cells, cell lines, and cancer cell lines were downloaded in batch from the ENCODE portal with the following query: Assay title: "polyA plus RNA-seq"; Status: "released"; Perturbation: "not perturbed"; Organism: "Homo sapiens"; Biosample classification: "cell line", "primary cell"; Genome assembly: "GRCh38". Identifiers corresponding to the acquired data totaling 860 files are provided in the supplementary materials (Additional File 2: Encode IDs). Downloaded expression tables were merged and standardized by a custom R script prepare.data_encode.R to a combined gene expression matrix that includes all input data, HGNC symbol gene annotation and cell annotations. For gene expression quantification log2(TPM) with a pseudocount of 1 was used. Based on those components, an AnnData object is created which is then provided as an input to the ikarus package. Cancer cell line annotation was done manually and is provided in Additional File 2: Enrichment analysis of tumor gene signature.
Microdissection dataset:
The gastric cancer microdissection dataset comprises laser capture microdissected (LCM) stromal and cancer regions collected from a patient cohort ($n = 8$) totaling 16 samples. Microdissected tissue for each sample was pooled together before library preparation to account for the absence of replicates. Gene expression quantification of stromal and cancer samples, as provided by the authors of the study [34] in the form of raw counts, was first standardized to ikarus format and then used as an input to ikarus pipeline.
Databases:

- Human protein atlas (https://www.proteinatlas.org/humanproteome/pathology) [51]
- Prognostic genes [52]
- Gene fusion (ChiTaRs) [53]
- SEEK (co-expression database) (https://seek.princeton.edu/seek/) [48]
- g:Profiler (https://biit.cs.ut.ee/gprofiler/) [50]
- CancerSEA (http://biocc.hrbmu.edu.cn/CancerSEA/home.jsp) [47]
- MsigDB (GO, Hallmark gene sets) [49]
- Atlas of co-essential modules [60]
- DepMap Achilles scores (https://depmap.org/portal/download/) [65]
- COSMIC (cancer.sanger.ac.uk) [66]

Declarations

Ethics approval and consent to participate

Ethics approval is not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Bioinformatics and Omics Data Science Platform, Berlin Institute For Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Hannoversche Str.28, 10115, Berlin, Germany. [2]Non-coding RNAs and Mechanisms of Cytoplasmic Gene Regulation Lab, Berlin Institute for Medical Systems Biology, Hannoversche Str. 28, 10115 Berlin, Germany. [3]Free University Berlin, Kaiserswerther Str. 16-18, 14195 Berlin, Germany.

Received: 23 November 2021 Accepted: 6 May 2022
Published online: 30 May 2022

References

1. Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. Nat Rev Genet. 2019;20:404–16.
2. Moscow JA, Fojo T, Schilsky RL. The evidence framework for precision cancer medicine. Nat Rev Clin Oncol. 2018;15:183–92.
3. Bassiouni R, Gibbs LD, Craig DW, Carpten JD, McEachron TA. Applicability of spatial transcriptional profiling to cancer research. Mol Cell. 2021;81:1631–9.
4. Nath A, Bild AH. Leveraging Single-cell approaches in cancer precision medicine. Trends Cancer Res. 2021;7:359–72.
5. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13:599–604.
6. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15:e8746.
7. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21:31.
8. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20:194.
9. Tan Y, Cahan P. SingleCellNet: A computational tool to classify single cell RNA-Seq data across platforms and across species. Cell Syst. 2019;9:207–213.e2.
10. Hu J, Li X, Hu G, Lyu Y, Susztak K, Li M. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. Nat Mach Intell. 2020;2:607–18.
11. Andreatta M, Corria-Osorio J, Müller S, Cubas R, Coukos G, Carmona SJ. Interpretation of T cell states from single-cell transcriptomics data using reference atlases. Nat Commun. 2021;12:2965.
12. Michielsen L, Reinders MJT, Mahfouz A. Hierarchical progressive learning of cell identities in single-cell data. Nat Commun. 2021;12:2799.
13. Ranjan B, Schmidt F, Sun W, Park J, Honardoost MA, Tan J, et al. scConsensus: combining supervised and unsupervised clustering for cell type identification in single-cell RNA sequencing data. BMC Bioinformatics. 2021;22:186.
14. Grabski IN, Irizarry RA. A probabilistic gene expression barcode for annotation of cell-types from single cell RNA-seq data. bioRxiv. 2020:2020.01.05.895441. https://doi.org/10.1101/2020.01.05.895441.
15. van den Brink SC, Sage F, Vértesy Á, Spanjaard B, Peterson-Maduro J, Baron CS, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods. 2017;14:935–6.
16. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21:12.
17. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nat Biotechnol. 2021. https://doi.org/10.1038/s41587-021-00895-7.
18. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053–8.
19. Brbić M, Zitnik M, Wang S, Pisco AO, Altman RB, Darmanis S, et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. Nat Methods. 2020;17:1200–6.
20. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11:2338.

21. Zhou X, Chai H, Zeng Y, Zhao H, Yang Y. scAdapt: virtual adversarial domain adaptation network for single cell RNA-seq data classification across platforms and species. Brief Bioinform. 2021;22. https://doi.org/10.1093/bib/bbab281.
22. Ge S, Wang H, Alavi A, Xing E, Bar-Joseph Z. Supervised adversarial alignment of single-cell RNA-seq data. J Comput Biol. 2021;28:501–13.
23. Chen L, He Q, Zhai Y, Deng M. Single-cell RNA-seq data semi-supervised clustering and annotation via structural regularized domain adaptation. Bioinformatics. 2021;37:775–84.
24. Kimmel JC, Kelley DR. Semisupervised adversarial neural networks for single-cell classification. Genome Res. 2021;31:1781–93.
25. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19:41–50.
26. Ronen J, Akalin A. netSmooth: Network-smoothing based imputation for single cell RNA-seq. F1000Res. 2018;7:8.
27. Laughney AM, Hu J, Campbell NR, Bakhoum SF, Setty M, Lavallée V-P, et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. Nat Med. 2020;26:259–69.
28. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med. 2018;24:1277–89.
29. Lee H-O, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nat Genet. 2020;52:594–603.
30. Kildisiute G, Kholosy WM, Young MD, Roberts K, Elmentaite R, van Hooff SR, et al. Tumor to normal single-cell mRNA comparisons reveal a pan-neuroblastoma cancer cell. Sci Adv. 2021;7. https://doi.org/10.1126/sciadv.abd3311.
31. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171:1611–1624.e24.
32. Ma L, Wang L, Khatib SA, Chang C-W, Heinrich S, Dominguez DA, et al. Single-cell atlas of tumor cell evolution in response to therapy in hepatocellular carcinoma and intrahepatic cholangiocarcinoma. J Hepatol. 2021;75:1397–408.
33. Bischoff P, Trinks A, Wiederspahn J, Obermayer B, Pett JP, Jurmeister P, et al. The single-cell transcriptional landscape of lung carcinoid tumors. Int J Cancer. 2022. https://doi.org/10.1002/ijc.33995.
34. Grunberg N, Pevsner-Fischer M, Goshen-Lago T, Diment J, Stein Y, Lavon H, et al. Cancer-associated fibroblasts promote aggressive gastric cancer phenotypes via heat shock factor 1-mediated secretion of extracellular vesicles. Cancer Res. 2021;81:1639–53.
35. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 2013;489:57–74.
36. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.
37. Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. Nat Med. 2015;21:1318–25.
38. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature. 2019;569:503–8.
39. Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics. 2020;36:533–8.
40. Wang L, Nie R, Yu Z, Xin R, Zheng C, Zhang Z, et al. An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell RNA-sequencing data. Nat Mach Intell. 2020;2:693–703.
41. Zhang AW, O'Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. Nat Methods. 2019;16:1007–15.
42. Gayoso A, Lopez R, Xing G, Boyeau P, Wu K, Jayasuriya M, et al. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. bioRxiv. 2021:2021.04.28.441833. https://doi.org/10.1101/2021.04.28.441833.
43. Li J, Sheng Q, Shyr Y, Liu Q. scMRMA: single cell multiresolution marker-based annotation. Nucleic Acids Res. 2022;50:e7.
44. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv [stat.ML]. 2020; Available: http://arxiv.org/abs/1802.03426.
45. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8:14049.
46. Jerby-Arnon L, Neftel C, Shore ME, Weisman HR, Mathewson ND, McBride MJ, et al. Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. Nat Med. 2021;27:289–300.
47. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res. 2019;47:D900–8.
48. Zhu Q, Wong AK, Krishnan A, Aure MR, Tadych A, Zhang R, et al. Targeted exploration and analysis of large cross-platform human transcriptomic compendia. Nat Methods. 2015;12:211–4 3 p following 214.
49. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.
50. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:W191–8.
51. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357. https://doi.org/10.1126/science.aan2507.
52. Smith JC, Sheltzer JM. Genome-wide identification and analysis of prognostic features in human cancers. bioRxiv. 2021:2021.06.01.446243. https://doi.org/10.1101/2021.06.01.446243.
53. Gorohovski A, Tagore S, Palande V, Malka A, Raviv-Shay D, Frenkel-Morgenstern M. ChiTaRS-3.1-the enhanced chimeric transcripts and RNA-seq database matched with protein-protein interactions. Nucleic Acids Res. 2017;45:D790–5.
54. Shlien A, Malkin D. Copy number variations and cancer. Genome Med. 2009;1:62.

55. Tickle T, Tirosh I, Georgescu C, Brown M, Haas B. inferCNV of the Trinity CTAT Project. Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA. USA. Available: https://github.com/broadinstitute/infercnv.
56. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14:1083–6.
57. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci U S A. 2005;102:7426–31.
58. Haghverdi L, Büttner M, Alexander Wolf F, Buettner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods. 2016:845–8. https://doi.org/10.1038/nmeth.3971.
59. Borcherding N, Andrews J. escape: Easy single cell analysis platform for enrichment. 2021.
60. Wainberg M, Kamber RA, Balsubramani A, Meyers RM, Sinnott-Armstrong N, Hornburg D, et al. A genome-wide atlas of co-essential modules assigns function to uncharacterized genes. Nat Genet. 2021;53:638–49.
61. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: visualization of intersecting sets. IEEE Trans Vis Comput Graph. 2014;20:1983–92.
62. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32:2847–9.
63. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc. 1995;57:289–300.
64. Dohmen J, Baranovskii A, Ronen J, Uyar B, Franke V, Akalin A. Tumor cell classification at the single cell level. Zenodo. 2022. https://doi.org/10.1101/2021.10.15.463909.
65. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. Cell. 2017;170:564–576.e16.
66. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47:D941–7.

Publisher's Note

4.2 Publication II

*Artem Baranovskii**, Irem Gündüz*, Vedran Franke, Bora Uyar & Altuna Akalin, Multi-Omics Alleviates the Limitations of Panel Sequencing for Cancer Drug Response Prediction
* These authors contributed equally to the work

This study was published on 15th November, 2022
in MDPI Cancers 2022, 14(22), 5604
https://doi.org/10.3390/cancers14225604

Supplementary materials related to this publication could be found in Appendix II.

cancers

MDPI

Communication

Multi-Omics Alleviates the Limitations of Panel Sequencing for Cancer Drug Response Prediction

Artem Baranovskii [1,†], Irem B. Gündüz [2,†], Vedran Franke [3], Bora Uyar [3,*] and Altuna Akalin [3,*]

[1] Non-Coding RNAs and Mechanisms of Cytoplasmic Gene Regulation Lab, Berlin Institute for Medical Systems Biology, Max Delbrück Center (MDC) for Molecular Medicine, Hannoversche Str. 28, 10115 Berlin, Germany
[2] Integrative Cellular Biology & Bioinformatics Lab, Saarland University, 66123 Saarbrücken, Germany
[3] Max Delbrück Center (MDC) for Molecular Medicine, Bioinformatics and Omics Data Science Platform, The Berlin Institute for Medical Systems Biology, Hannoversche Str. 28, 10115 Berlin, Germany
* Correspondence: bora.uyar@mdc-berlin.de (B.U.); altuna.akalin@mdc-berlin.de (A.A.)
† These authors contributed equally to this work.

Simple Summary: Cancer is a complex, heterogeneous collection of diseases with hundred of different subtypes. Genomic aberrations that are primarily thought to be the root causes of different cancers have been clinically used as evidence for both the diagnosis and also matching individual patients to proper treatment options. However, the complexity of cancer manifests itself differently in each patient when inspected at the molecular level. Even patients with the same cancer type rarely have identical root causes for the same disease. Without an extensive molecular profile of a patient, it has been challenging to match the patients to the best treatment options. To remedy this, comprehensive genomic profiling panels have been developed to monitor hundreds of genes for a given patient, which has helped broaden the treatment options for patients. However, genomic aberrations detected in such panels still do not reflect the full complexity of how a tumour responds to cancer drugs. In this study, we demonstrate that using an additional layer of molecular information (called the transcriptome) on top of genomic aberrations that can be detected with cancer gene panels can provide significant improvements in predicting the cancer drug response in pre-clinical cancer models. Thus, this study serves as a push towards incorporating the transcriptome measurements more routinely in (pre-)clinical practice.

Abstract: Comprehensive genomic profiling using cancer gene panels has been shown to improve treatment options for a variety of cancer types. However, genomic aberrations detected via such gene panels do not necessarily serve as strong predictors of drug sensitivity. In this study, using pharmacogenomics datasets of cell lines, patient-derived xenografts, and ex vivo treated fresh tumor specimens, we demonstrate that utilizing the transcriptome on top of gene panel features substantially improves drug response prediction performance in cancer.

Keywords: multi-omics; cancer; drug response prediction; pharmacogenomics; panel sequencing

Citation: Baranovskii, A.; Gündüz, I.B.; Franke, V.; Uyar, B.; Akalin, A. Multi-Omics Alleviates the Limitations of Panel Sequencing for Cancer Drug Response Prediction. *Cancers* **2022**, *14*, 5604. https://doi.org/10.3390/cancers14225604

Academic Editor: Hubert Hackl

Received: 19 October 2022
Accepted: 12 November 2022
Published: 15 November 2022

1. Introduction

Cancer is a collection of diseases characterized by abnormal cellular growth and the invasion of other body parts. It affected 19 million people in 2020 and was the cause of 9.5 million deaths that year alone [1]. Cancer has been primarily considered to be a disease of the genome, where the accumulation of alterations is the underlying cause of the transformation of normal cells into malignant cancerous cells with survival and proliferation advantages [2]. Genetic alterations of this kind have been studied to understand the mechanisms of cancer and to develop targeted therapies. The latter and companion diagnostic tools have transformed oncology [3], promising more precise treatments tailored to tumors' genetic profiles. Various targeted therapies have been successfully developed to counteract

the defects in the molecular machinery borne out of such oncogenic mutations [4–6]. To this date, most of the markers approved for targeted therapy decisions are single-gene markers [7]. It has thus become crucial to develop accurate, sensitive, and high-throughput genomic assays to accommodate the increasingly genotype-based therapeutic approaches. Commercial companies, such as Foundation Medicine, as well as large cancer research centers, such as Memorial Sloan Kettering and Dana Farber, have produced their panel sequencing assays to guide therapy for cancer patients [8]. These techniques examine genes that are frequently mutated in cancer to assess mutations and copy number variations. Especially for diagnostics, the approved methods for targeted drugs are usually the presence or absence of the mutations. Therefore, the assay developers focus primarily on mutation calling accuracy as a metric of the usefulness and accuracy of the assay [8].

Although comprehensive genomic profiling using cancer gene panels has demonstrated value in broadening the treatment options for patients based on matching a patient's genomic lesions to cancer driver gene aberrations associated with FDA-approved treatment indications [8,9], the presence/absence of mutations in such genes does not necessarily translate into improved predictive power for estimating the patient's response to the potential treatments. While for some drugs the variation in drug response can be explained by a very specific mutation (for instance, BRAF V600E mutation is a strong predictor for response to Vemurafenib in metastatic melanoma [4]), for many drugs the knowledge of the mechanism of action is missing. This is because many drugs are discovered via phenotypic screening of model systems rather than target-based approaches [10]. Of note, such single mutation markers for a given cancer type are not necessarily good markers for other cancer types. For instance, BRAF V600E, while a good predictor for metastatic melanoma, is a poor predictor of response in metastatic colorectal cancer [11]. More importantly, the latest compilation of the hallmarks of cancer recognized in the field includes factors such as the non-mutational epigenetic aberrations, the involvement of the immune system in the tumor microenvironment, and the composition of the microbiome along with genomic defects [12]. These layers of information cannot be sufficiently captured by focusing on the restricted set of genomic alterations and necessitate other data modalities. Among those, transcriptome profiling—besides being a cheap and accessible option in terms of logistics—has been shown to yield strong predictors of drug response [13–15].

Here, we set out to quantify the extent to which the usage of the transcriptome as an additional data modality improves the drug response prediction performance compared to only the genetic features restricted to the cancer gene panels (such as mutations and copy number variations). We leveraged publicly available pharmacogenomics datasets, including genomic and transcriptomic profiles and drug sensitivity measurements in three types of datasets: cancer cell lines (using the CCLE database [16] and PRISM project [17]), ex vivo treated fresh tumor specimens from Acute Myeloid Leukemia patients (BeatAML) [18], and patient-derived xenografts (PDX) [19]. These datasets span three vastly different model systems to anti-cancer drug efficacy testing, each of which varies in biological complexity and comes with unique challenges and advantages. Testing across these datasets should deliver an exhaustive assessment of the importance of transcriptomic features.

2. Results

In all three settings, with an application of out-of-the-box machine learning techniques (see Section 4), we modelled drug responses for all available drugs in two reported data modalities, using only panel gene features (panel (PS)) or using the transcriptomic features on top of panel features (multi-omics (MO)). While achieving only moderate predictive power (CCLE mean R-square ~10%, n = 396; BeatAML mean R-square ~12%, n = 106), the MO modalities of the CCLE and BeatAML datasets showed an overall increase in predictive power (up to a 5-fold improvement for certain drugs) in comparison to PS data (Figure 1A and Figure S1a,b, Table S1). Of note, we observed a significant positive correlation ($r = 0.4$, $p < 0.0001$) between the percentage of gene expression features among the top 100 features and an increase in MO's predictive power over PS (Figure 1B). Modelling in

xenografts generally conferred similar results. Ten out of twelve drugs showed a significant increase in MO's predictive power of PS (Wilcoxon's $p < 0.05$) (Figure 1C,D, Table S1). These results were obtained by building random forest regression models; however, we have also reproduced similar findings, where transcriptome features added substantial predictive power on top of panel features, using both Elastic Net (GLMnet) and Support Vector Machines (with radial kernels) (Figure S1 and Table S1—see Section 4).

The improvement of MO over PS across all datasets is nearly univocal, yet heterogeneous. The most extreme improvement we have observed was for Venetoclax in beatAML dataset, where using the panel features yielded an R^2 value of 0.03, and the MO features yielded an R^2 value of 0.49. Hence, the top predictive features for Venetoclax consisted solely of cell type and cancer hallmark signatures (Figure S2c). For some drugs, a 4- to 5-fold improvement in drug response prediction was recorded. For others, the improvement was modest (e.g., ~0.025 change in R^2 between MO and PS). To an extent, this difference could be explained by a drug's mechanism of action (MOA), as some perturb larger shares of cellular machinery than others. We selected the CCLE dataset as the most representative to test this (n drugs = 396) (Figures 1D and S2a,b). Among the drugs for which we observed the most improvement, there are histone deacetylase (HDAC) inhibitors, a relatively novel class of anti-cancer drugs that interferes with epigenetic regulation [20]. This MOA likely affects the transcriptome on a broad scale via secondary effects that arise from altered transcription. Likewise, topoisomerase inhibitors and bromodomain inhibitors drive wide transcriptional changes across the genome. The former inhibits the action of DNA topoisomerases that lead to the activation of the DNA damage response cascade [21]. Bromodomain inhibitors compete for the bromodomains of the respective proteins and prevent binding of the latter to acetylated histones and transcription factors [22]. Among the drugs with a defined target pathway, the effects of MO improvement are more modest. Altogether, the scale of off-target transcriptional perturbation seems to be beneficial for MO modelling, as it produces a signal outside of the defined panel's reach.

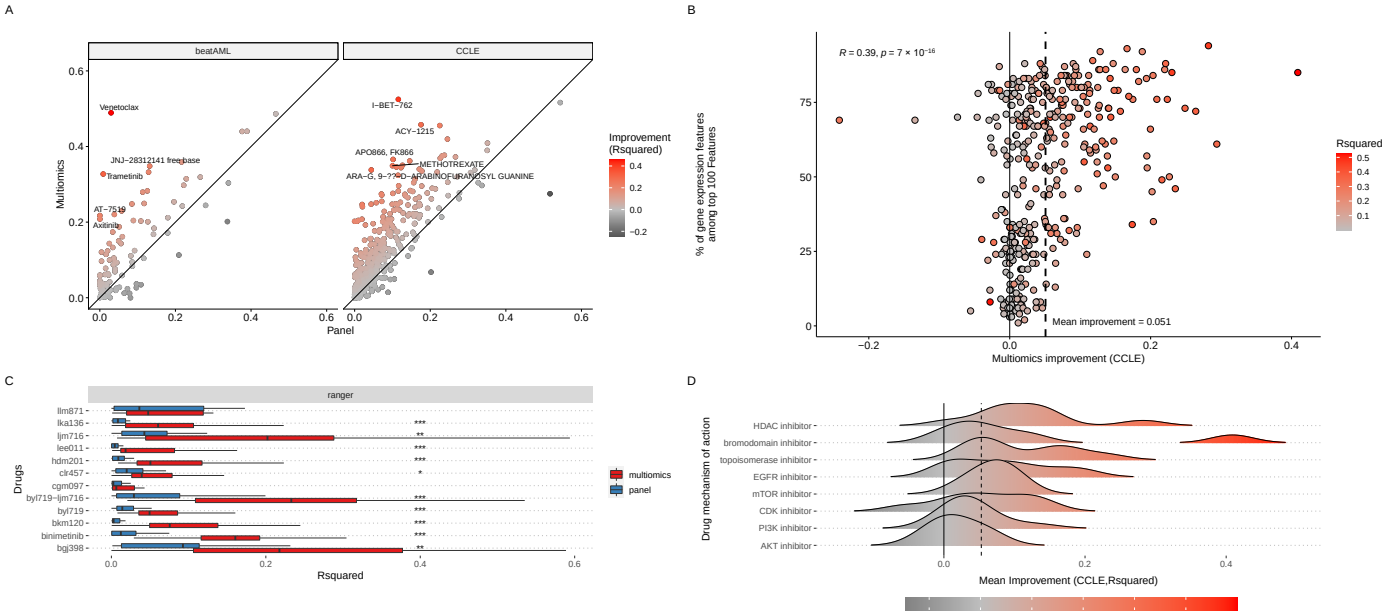


Figure 1. Evaluation of the performance of drug response prediction when using only panel-seq features (mutations and/or copy number variations) or using transcriptome features in combination with panel-seq features (multi-omics). (**A**) Improvement of multi-omics (as in R-square metric) in comparison to panel-seq features for the testing portion of BeatAML (**left** panel) and CCLE (**right** panel) datasets for 106 and 396 drugs, respectively. (**B**) Correlation of the prediction performance improvement (multi-omics vs. panel-seq) with respect to the proportion of transcriptome features among the top 100 most important predictors of drug response for CCLE datasets. (**C**) Improvement of multi-omics (in red) (as in R-square metric) in comparison to panel-seq features (in blue) for the testing portion of the PDX dataset for 12 drugs. Stars above the boxplots represent significance levels: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$. (**D**) Drug classes with loose pathway specificity show higher average improvement in MO over PS. Drug classes (y-axis) are ordered by average improvement in MO and filtered to keep only those that have a minimum of five drugs in a class. The dashed line corresponds to the global average improvement in MO (0.051) as reported in figure panel (**C**).

3. Discussion

In this study, we set out to demonstrate two points. First, genomic features derived with comprehensive genomic profiling methods (panel sequencing) and used during clinical/pre-clinical drug development often have limited predictive power for drug response in cancer pre-clinical models. Second, we sought to elucidate how the drug response prediction could be improved using additional transcriptomic features. We showed that using cell type and cancer hallmark gene signature scores derived from the transcriptome on top of panel-derived features improves predictive power across different pre-clinical models (cell lines, patient-derived xenografts, ex vivo treated human samples), irrespective of the modelling method used. Our main aim in this study was not to argue that panel sequencing should be replaced by transcriptome profiling, but rather to demonstrate that using the transcriptomic features could have added benefit for treatment response prediction. The logic we follow is that if modelling drug responses using only panel features has limited power in pre-clinical models, then it would be even more limited for actual patients who would receive such treatments based on few marker genotypes, as the pre-clinical models cannot perfectly represent the complexity of the tumor or its microenvironment. Our second aim in this study was to look for general trends across drugs and datasets using off-the-shelf methods with a fair comparison of feature sets used in the modelling. We are mainly interested in the general trends; therefore, we refrain from making strong conclusions about individual drugs. However, we have quantified the predictive importance of all the studied features for each drug as supplementary material. It is important to note that the top markers we report would be correlative in nature. One would need to use causal inference methods to figure out drug-specific causal biomarkers.

Although the pre-clinical models we studied here do not perfectly reflect the complexity of the actual tumor microenvironment in a human patient, this kind of a large-scale data analysis of drug responses would not be feasible to carry out on actual patients due to logistical and ethical limitations. While replicating this kind of a large-scale analysis could not be extended to actual human samples, follow-up studies could address some questions that we have not addressed here. First of all, there could be many more alternatives in regard to how to preprocess the genomic and transcriptomic features in terms of converting the input data into less noisy and more information-dense latent features, optionally with added prior knowledge. For instance, mutation data could be converted into cancer mutational signatures, or transcriptome data could be integrated with prior knowledge networks to derive causal subnetwork features. Moreover, different layers of omics data modalities could be integrated using multi-omics integration methods. More sophisticated deep learning-based drug response modelling tools could be used, including pre-training and transfer learning approaches. Finally, the robustness and generalization power of the prediction models could be compared in a cross-dataset setting, where the models are built and validated in independently acquired resources or cross-tissue settings, where the models could be evaluated on cell lines derived from a tissue of interest unrelated to the tissues that are considered during the training procedure. However, our purpose here was not to benchmark the potential data processing or modelling algorithms, but rather to have a fair comparison of distinct feature sets in terms of their predictive power for drug response.

4. Methods

4.1. Data/Code Availability

In this study, the following publicly available pharmacogenomics datasets were used:

Cancer Cell Line Encyclopedia (CCLE) [16] downloaded from https://depmap.org/portal/download/, accessed on 11 April 2022. The drug response measurements from the PRISM project [17] were used for the corresponding CCLE samples.

Patient-Derived Xenografts (PDX) [19].

BeatAML: ex vivo drug sensitivity screening of acute myeloid leukemia patient tumor specimens [18] downloaded and processed using the PharmacoGx R package [23].

Annotation of drug classes was downloaded from Drug Repurposing Hub project [24].

All codes to download, process, analyze the datasets, and reproduce the figures in this manuscript can be found here: https://github.com/BIMSBbioinfo/multiomics_vs_panelseq, accessed on 11 April 2022.

4.2. Data Processing

Mutations: The mutation data were converted into a matrix of mutation counts per gene per sample. The resulting matrix was further filtered to only keep mutation data for genes in the OncoKB cancer gene list (https://www.oncokb.org/cancerGenes, accessed on 15 April 2022.). Mutation data was available in all three datasets.

Copy Number Variations: The copy number variation data was used as downloaded from the respective resources. Copy number variation data was also filtered to only keep genes found in the OncoKB cancer gene list. Both the CCLE and PDX datasets contained CNV data available, but it was not available for the BeatAML dataset.

Gene Expression (Transcriptome): To reduce the dimensionality and obtain less noisy features, the gene expression datasets were converted into gene-set activity scores using single-sample gene set scoring (singscore R package [25]). The gene sets utilized in this study were the Cancer Hallmarks gene signatures (50 gene sets) from the MSIGDB database [26] and tumor microenvironment-related gene sets (64 gene sets) curated in the xCell R package [27]. Gene expression data was available for all three datasets.

Drug Sensitivity Measures: For the CCLE and PDX datasets, AUC (area under the curve) scores derived from dose–response curves as published in the respective resources were used in the prediction models. For the BeatAML dataset, recomputed AAC (area above the curve) scores were used as downloaded via the PharmacoGx R package [23].

4.3. Drug Response Modelling with Machine Learning

For drug response modelling, we considered two main scenarios based on the availability of datasets. In the first set, we considered mutation and/or copy number variation data for genes found in the OncoKB cancer gene list, which aims to simulate panel-sequencing. In the second set, we considered the features as in the first set along with the whole transcriptome profiling as an additional data modality, which is further converted into gene-set activity scores. The second set represents the multi-omics condition, in which the panel features are concatenated with the transcriptome features (gene-set scores).

4.4. For Both Settings, All Three Datasets (CCLE, PDX, BeatAML) Were Analyzed with the Following Protocol

Only drugs that were treated on at least 100 samples were considered.

For each drug, the samples were split into training (70% of samples) and testing groups (30% of samples). See Tables S1–S3 for the specific sample counts used for each drug.

Caret R package [28] was used to build random forest regression models (using either ranger [29], logistic regression models (glmnet) [30], and support vector machines (svmRadial)) on the training data, where the genomic/transcriptomic features were used as predictors and the drug response values were used as the outcome variable. We used 3-times repeated 5-fold cross-validation for hyperparameter tuning to find the best model parameters based on the training data. Near-zero-variation filtering, scaling, and centering were applied as data processing steps. Applying principal component analysis (PCA) as a processing step led to poorer prediction results for both multi-omics and panel-seq features. However, the overall trend was the same, where MO-based models yielded better results than PS-based models (Figure S1); therefore, we excluded PCA processing when reporting the main results. For the PDX samples, this step was repeated 20 times by resampling the training/testing portions. This was only applied for the PDX samples, due to the small number of drugs (N = 12) treated on at least 100 samples.

The final model performance was evaluated on the testing data. Spearman rank correlation, R-square, and root mean squared error (RMSE) metrics were computed, and the R-square metric was used to report the results in the final figures.

Feature importance rankings were obtained using 'caret::varImp' function and are provided in Table S2 for each modelling method and for each drug.

Drugs were summarized according to their mechanism of action and were evaluated by mean multi-omics improvement.

5. Conclusions

We believe that in order to better understand cancer and develop better drugs and diagnostics, we need to make use of all the molecular features by integrating different omics datasets. In this manuscript, using multi-omics and machine learning techniques, we showed that multi-omics has indeed superior performance for drug response prediction in cancer.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/cancers14225604/s1: Figure S1. Performance comparisons of different drug response prediction models trained by using only panel-seq features (mutations and/or copy number variations) or using transcriptome features in combination with panel-seq features (multi-omics) using different pre-processing options (with PCA in A) and without PCA (in B) and using different machine learning methods: random forests, elastic nets, and support vector machines. (A) Comparisons of different drug response models trained with preprocessed panel-seq and multi-omics features of beatAML and CCLE datasets using three different methods with scaling, entering, and near-zero variance filtering. (B) Comparisons of drug response models, trained with preprocessed (scaled/centered/filtered for near-zero-variation) and dimensionally reduced (using PCA) panel-seq and multi-omics features of beatAML and CCLE datasets. (C) Multi-omics (red) improvements (in terms of R-squared metric) compared to panel-seq features (blue) of the test section of the 12-drug PDX dataset, using the elastic net regression (glmnet) model. Stars above the boxplots represent significance levels: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$. (D) Multi-omics (red) improvements (in terms of R-squared metric) compared to panel-seq features (blue) of the test section of the 12-drug PDX dataset, using the radial support vector machine (svmRadial) model. Stars above the boxplots represent significance levels: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$. Figure S2. (A) Classes of drugs based on the average improvement in multi-omics over panel-seq when the logistic regression (glmnet) model was used for drug response prediction. (B) Classes of drugs based on the average improvement in multi-omics over panel-seq when the radial support vector machine (svmRadial) model was used for drug response prediction. Mean improvement on overall drugs marked with dashes. (C) Top 20 cell type and cancer hallmark gene signatures associated with Venetoclax response prediction for beatAML samples using a random forest model. Table S1: Drug response prediction performance metrics for each machine learning method and pharmacogenomics dataset. Table S2: Feature importance metrics derived from each machine learning model built for each drug in each pharmacogenomics dataset. Table S3: Improvement in prediction performance across drug classes per dataset (except for PDX dataset) per built model.

Author Contributions: Conceptualization and initial planning: A.A.; Data collection and formal analysis: A.B., I.B.G., V.F. and B.U.; Writing and editing of the manuscript: all authors; Supervision: B.U. and A.A.; Funding acquisition: A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research partly funded by Berlin Institute of Health, Digital Health Accelerator grant number and The APC was funded by Helmholtz Association.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data analysed in this study have been previously published (See Section 4.1). The raw and processed data along with the code that was used to process, analyse, and visualise the findings reported in this study can be found at our GitHub repository: https://github.com/BIMSBbioinfo/multiomics_vs_panelseq (accessed on 11 April 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef] [PubMed]
2. Weinberg, R.A.; Hanahan, D. The Hallmarks of Cancer. *Cell* **2000**, *100*, 57–70.
3. Bedard, P.L.; Hyman, D.M.; Davids, M.S.; Siu, L.L. Small molecules, big impact: 20 years of targeted therapy in oncology. *Lancet* **2020**, *395*, 1078–1088. [CrossRef]
4. Chapman, P.B.; Hauschild, A.; Robert, C.; Haanen, J.B.; Ascierto, P.; Larkin, J.; Dummer, R.; Garbe, C.; Testori, A.; Maio, M.; et al. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. *N. Engl. J. Med.* **2011**, *364*, 2507–2516. [CrossRef] [PubMed]
5. Shaw, A.T.; Kim, D.-W.; Nakagawa, K.; Seto, T.; Crinó, L.; Ahn, M.-J.; De Pas, T.; Besse, B.; Solomon, B.J.; Blackhall, F.; et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N. Engl. J. Med.* **2013**, *368*, 2385–2394. [CrossRef]
6. Shaw, A.T.; Kim, D.-W.; Mehra, R.; Tan, D.S.W.; Felip, E.; Chow, L.Q.M.; Camidge, D.R.; Vansteenkiste, J.; Sharma, S.; de Pas, T.; et al. Ceritinib in ALK-rearranged non-small-cell lung cancer. *N. Engl. J. Med.* **2014**, *370*, 1189–1197. [CrossRef]
7. Chakravarty, D.; Gao, J.; Phillips, S.; Kundra, R.; Zhang, H.; Wang, J.; Rudolph, J.E.; Yaeger, R.; Soumerai, T.; Nissan, M.H.; et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis. Oncol.* **2017**. [CrossRef]
8. Cheng, D.T.; Mitchell, T.N.; Zehir, A.; Shah, R.H.; Benayed, R.; Syed, A.; Chandramohan, R.; Liu, Z.Y.; Won, H.H.; Scott, S.N.; et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J. Mol. Diagn.* **2015**, *17*, 251–264. [CrossRef]
9. Karol, D.; McKinnon, M.; Mukhtar, L.; Awan, A.; Lo, B.; Wheatley-Price, P. The Impact of Foundation Medicine Testing on Cancer Patients: A Single Academic Centre Experience. *Front. Oncol.* **2021**, *11*, 687730. [CrossRef]
10. Swinney, D.C.; Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discov.* **2011**, *10*, 507–519. [CrossRef]
11. Tabernero, J.; Ros, J.; Élez, E. The Evolving Treatment Landscape in BRAF-V600E-Mutated Metastatic Colorectal Cancer. *Am. Soc. Clin. Oncol. Educ. Book* **2022**, *42*, 254–263. [CrossRef] [PubMed]
12. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **2022**, *12*, 31–46. [CrossRef] [PubMed]
13. Chen, J.; Zhang, L. A survey and systematic assessment of computational methods for drug response prediction. *Brief. Bioinform.* **2021**, *22*, 232–246. [CrossRef] [PubMed]
14. Sharifi-Noghabi, H.; Jahangiri-Tazehkand, S.; Smirnov, P.; Hon, C.; Mammoliti, A.; Nair, S.K.; Mer, A.S.; Ester, M.; Haibe-Kains, B. Drug sensitivity prediction from cell line-based pharmacogenomics data: Guidelines for developing machine learning models. *Brief. Bioinform.* **2021**, *22*, bbab294. [CrossRef]
15. Rodon, J.; Soria, J.-C.; Berger, R.; Miller, W.H.; Rubin, E.; Kugel, A.; Tsimberidou, A.; Saintigny, P.; Ackerstein, A.; Braña, I.; et al. Genomic and transcriptomic profiling expands precision cancer medicine: The WINTHER trial. *Nat. Med.* **2019**, *25*, 751–758. [CrossRef]
16. Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A.A.; Kim, S.; Wilson, C.J.; Lehar, J.; Kryukov, G.V.; Sonkin, D.; et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **2012**, *483*, 603–607. [CrossRef] [PubMed]
17. Corsello, S.M.; Nagari, R.T.; Spangler, R.D.; Rossen, J.; Kocak, M.; Bryan, J.G.; Humeidi, R.; Peck, D.; Wu, X.; Tang, A.A.; et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer* **2020**, *1*, 235–248. [CrossRef]
18. Tyner, J.W.; Tognon, C.E.; Bottomly, D.; Wilmot, B.; Kurtz, S.E.; Savage, S.L.; Long, N.; Schultz, A.R.; Traer, E.; Abel, M.; et al. Functional genomic landscape of acute myeloid leukaemia. *Nature* **2018**, *562*, 526–531. [CrossRef]
19. Gao, H.; Korn, J.M.; Ferretti, S.; Monahan, J.E.; Wang, Y.; Singh, M.; Zhang, C.; Schnell, C.; Yang, G.; Zhang, Y.; et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **2015**, *21*, 1318–1325. [CrossRef]
20. Kim, H.-J.; Bae, S.-C. Histone deacetylase inhibitors: Molecular mechanisms of action and clinical trials as anti-cancer drugs. *Am. J. Transl. Res.* **2011**, *3*, 166–179.
21. Pommier, Y. Topoisomerase I inhibitors: Camptothecins and beyond. *Nat. Rev. Cancer* **2006**, *6*, 789–802. [CrossRef] [PubMed]
22. Pérez-Salvia, M.; Esteller, M. Bromodomain inhibitors and cancer therapy: From structures to applications. *Epigenetics* **2017**, *12*, 323–339. [CrossRef] [PubMed]
23. Smirnov, P.; Safikhani, Z.; El-Hachem, N.; Wang, D.; She, A.; Olsen, C.; Freeman, M.; Selby, H.; Gendoo, D.; Grossmann, P.; et al. PharmacoGx: An R package for analysis of large pharmacogenomic datasets. *Bioinformatics* **2016**, *32*, 1244–1246. [CrossRef]
24. Corsello, S.M.; Bittker, J.A.; Liu, Z.; Gould, J.; McCarren, P.; Hirschman, J.E.; Johnston, S.E.; Vrcic, A.; Wong, B.; Khan, M.; et al. The Drug Repurposing Hub: A next-generation drug library and information resource. *Nat. Med.* **2017**, *23*, 405–408. [CrossRef] [PubMed]
25. Foroutan, M.; Bhuva, D.D.; Lyu, R.; Horan, K.; Cursons, J.; Davis, M.J. Single sample scoring of molecular phenotypes. *BMC Bioinform.* **2018**, *19*, 404. [CrossRef]

26. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef]
27. Aran, D.; Hu, Z.; Butte, A.J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **2017**, *18*, 220. [CrossRef]
28. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
29. Wright, M.N.; Ziegler, A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **2017**, *77*, 1–17. [CrossRef]
30. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]

5 Discussion

This thesis comprises two studies with three unifying themes: machine learning, multi-omics, and cancer biology. First, each work showcases an application of machine learning techniques, be it out-of-the-box methods, or custom-built machines, to conceptually different problems. Second, both studies gauge the virtue of the transition from mono- to multi-omic analysis regarding a trained model's predictive ability. Finally, within the scope of cancer biology, these studies' immediate goals and results do not hold much in common; therefore, each work entailed unique conceptual obstacles that required not less unique solutions. The development of Ikarus , the pipeline to segregate single-cell profiles into tumour and non-tumour, demanded resolutions to the engineering problems encountered during the algorithm's optimisation alongside the questions arising from the conceptualisation of validation procedures. On the other hand, the second work was focused on a multi-omic approach to analysis in cancer models and didn't concern itself with the method development.

5.1 Robust annotation of cancer cells in scRNA-seq data

The intention behind developing Ikarus was to tackle a seemingly elementary challenge: to distinguish between neoplastic and healthy cells within a single-cell dataset. Our methodology constitutes two stages, both implemented within the Ikarus pipeline.

At the outset, Ikarus amalgamates a user-provided array of expert-annotated scRNA-seq datasets and constructs cancer cell-specific gene signatures including up- and down-regulated genes. Subsequently, Ikarus scores the constructed gene sets within single cells and passes on the scores to the neighbouring cells via versatile network propagation. These mechanisms allow Ikarus to circumvent two prevailing difficulties in a single-cell annotation: batch effects and unstable clustering.

Conventionally, batch effects stemming from technical disparities between single-cell datasets are addressed by integration methodologies; these are numerous but univocally demand extensive parameter tuning. In turn, that often leads to non-intuitive effects on the integration solution. Further, the precise accuracy of the resultant integrations can only be effortlessly assessed with a reliance on comprehensive biological priors, i.e. reference datasets. In this regard, gene set scoring offers a robust and computationally feasible alternative: single-sample rank-based evaluation eliminates the need for integration to compare disparate samples. Nevertheless, gene set scoring is not perfect as with any other method. The primary technical variable to affect the technique is the proportion of the gene set present in the given samples, inasmuch as strong disbalance in the representation of gene set in

different samples produces unstable scores. This aspect has been thoroughly interrogated in our work with random simulations of scorings with incomplete gene sets.

A routine single-cell analysis necessarily involves aggregating cells into clusters that are subsequently annotated and compared to a high-fidelity reference, e.g. cell atlas. However, irrespective of the chosen methodology, clustering parametrisation is exceptionally challenging to yield stable solutions across different datasets. Naturally, this problem renders automatic clustering unfeasible, if not outright impossible, and often necessitates manual tuning. Therefore, in our pipeline, we supplanted clustering with graph-based network propagation, a parameterless alternative to clustering that retains similar sensitivity to annotation. This method reduces the uncertainty in cell classes by integrating the scores of densely connected neighbours.

In an extension of the pipeline, we engaged a multi-omic strategy and aimed to improve the classification accuracy by including a data modality of CNVs inferred from scRNA-seq. The main advantage of this type of multi-omics is that additional data modalities are directly inferred from the same mono-omic experiment. In this regard, a posteriori integration of CNVs in a validation step increased the precision of a classifier, i.e. decreased the number of false positives. However, this effect was not universal across the diverse cancer types we analysed. Theoretically, appending more multi-omic classification stages would further increase the machine's performance. For instance, tools for inference of transposable elements have been recently published and were able to delineate differentiating stem cells. These data could be utilised to implement a cancer-cell de-differentiation inference, whereas transposon expression can serve as a proxy for the latter.

Nevertheless, the inquiry regarding the full potential of multi-omic integration in single-cell analysis remains unresolved. For instance, scWGS and single-cell epigenetic profiling are hindered by the low and uneven sequence coverage, while di-omic profiling of transcriptome and proteome in single cells is limited by the used array of antibodies that often range a hundred.

The expansion of single-cell omics concurs with the development of spatial transcriptomics. New methodologies, like 10x Visium, ¬that utilise a barcoded grid to capture spatial information from a tissue slide, allow interrogation of tissue sections in a spatial context with near single-cell precision, e.g. two to ten cells per grid unit. In the case of cancer, the usual annotation of tissue sections is done based on histological evaluation by a pathologist: a laborious and tedious process. In this regard, Ikarus can help to streamline the tissue section annotation process, albeit the potential caveats arising from the uneven aggregation of cells per grid unit need to be

explored.

Ikarus's main methodological limitation is its reliance on expert-annotated reference single-cell datasets. The latter is used in the identification of gene signatures and validation of a classifier, the two steps of the Ikarus pipeline. Therefore, the availability of the reference constrains the applicability of Ikarus to a shallow pool of faithfully annotated datasets. So far, we have ascertained Ikarus's performance within the domain of epithelial tumours and neuroblastoma, whereas the classification of single cells in synovial sarcoma was inaccurate. Therein, the Ikarus needs to accommodate multiple models to comprehensively classify different cancer types, which necessitates the expansion of a catalogue of high-fidelity references. In this respect, the rapidly expanding volume of single-cell datasets will yield new references and potentially include hitherto unexamined cancer types as soft tissue tumours.

5.2 Multi-omics fare better in the prediction of drug response in cancer models

Genomic profiling techniques commonly utilised in pharmaceutical practice revolve around in-depth deep profiling of a specific set of well-conserved and validated genomic features, i.e. panel sequencing.

The pharmaceutical practice utilises genomic profiling techniques in clinical and pre-clinical practice. In this regard, panel sequencing, i.e. in-depth profiling of a specific panel of well-conserved and validated genomic features, is the most widespread approach. In our work, we investigated two points; First, the set of genomic features assayed by panel sequencing has a constrained capacity for drug response prediction in pre-clinical cancer models. Second, the predictive power can be improved by expanding the panel of transcriptomic features.

We show that adding gene expression-derived rank-based signature scores to the panel features drastically enhances the predictive ability across various cancer models, including cell lines, patient-derived xenografts and ex-vivo human samples, regardless of the chosen machine learning algorithm. The improvements in prediction are most pronounced in drugs whose MOA induces collateral transcriptomic changes, such as HDAC, bromodomain, and DNA topoisomerase inhibitors. In our rationale, we adhered to the workflow of pharmaceutical practice, whereby drugs are tested in pre-clinical cancer models and then prescribed to patients based on several selected genetic markers elucidated from pre-clinical models. If panel sequencing fares poorly in cancer models, its prediction efficacy in patients would be exacerbated even further, as cancer models are imperfect in capturing the tumour microenvironment and heterogeneity. Notably, our work does not

propose a complete departure from panel sequencing in favour of a multi-omic approach. Rather, it showcases that the complementary transcriptomic features generally magnify the predictive power and where this magnification is the most pronounced.

In our study, we examined pre-clinical cancer models spanning a wide range of biological complexity, albeit the majority of the samples in our analysis came from the cancer cell lines, which is the simplest model and likely the farthest from mirroring the biological complexity of an actual tumour within a patient. Although the large-scale pharmacogenomic analysis cannot be replicated in the actual human samples due to ethical and logistical constraints, considering modern trends to expand and improve next-generation cancer models like organoids and xenografts, the subsequent studies could leverage their potential in the near future. In this regard, it would be valuable to test the generalisation capacity of the constructed models. For instance, how the models perform in cross-tissue settings and if the predictive power remains consistent when models are trained on one cancer model system and tested on another, e.g. train the model in organoids and test in xenografts, and vice versa.

As far as we were concerned with general trends, we utilised standard data pre-processing and out-of-the-shelf machine learning methods. In this regard, alternative data pre-processing could be employed to de-noise the input: both transcriptomic and genomic features could be transformed into information-dense latent features; transcriptomic data could be integrated into gene networks; mutations could be transformed into cancer mutational signatures based on the a priori knowledge from large scale cancer studies. Finally, different -omics data modalities could be integrated alongside transcriptomic and genomic features, potentially yielding a potent substrate for sophisticated deep learning approaches.

6 Bibliography

[1] Kuska B. "Beer, Bethesda, and biology: How genomics came into being". In: *J Natl Cancer Inst* 90.2 (1998), p. 93.

[2] Charles A. Janeway et al. "Hypergammaglobulinemia associated with severe recurrent and chronic nonspecific infection". In: *Ama american journal of diseases of children* 88.3 (1954), p. 515.

[3] et al. Holmes Beulah. "Fatal granulomatous disease of childhood. An inborn abnormality of phagocytic function". In: *Lancet* (1966), pp. 1225–8.

[4] Robert L. Baehner and David G. Nathan. "Leukocyte oxidase: defective activity in chronic granulomatous disease". In: *Science* 155.3764 (1967), pp. 835–836.

[5] Brigitte Royer-Pokora et al. "Cloning the gene for an inherited human disorder chronic granulomatous disease on the basis of its chromosomal location". In: *Nature* 322.6074 (1986), pp. 32–38.

[6] Dorothy H. Andersen. "Cystic fibrosis of the pancreas and its relation to celiac disease: a clinical and pathologic study". In: *American journal of Diseases of Children* 56.2 (1938), pp. 344–399.

[7] Francis S. Collins. "The human genome project and the future of medicine". In: *Annals of the New York Academy of Sciences* 882.1 (1999), pp. 42–55.

[8] John R. Riordan et al. "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA". In: *Science* 245.4922 (1989), pp. 1066–1073.

[9] David Botstein et al. "Construction of a genetic linkage map in man using restriction fragment length polymorphisms". In: *American journal of human genetics* 32 (1980), p. 3.

[10] U. S. Doe Joint Genome Institute et al. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (2001), pp. 860–921.

[11] J. Craig Venter et al. "The sequence of the human genome". In: *Science* 291.5507 (2001), pp. 1304–1351.

[12] International SnpMap Working Group. "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms". In: *Nature* 1.42 (2001), pp. 928–933.

[13] Richard A. Gibbs et al. "The international HapMap project". In: 2003.

[14] Victor E. Velculescu et al. "Serial analysis of gene expression". In: *Science* 270.5235 (1995), pp. 484–487.

[15] Victor E. Velculescu et al. "Characterization of the yeast transcriptome". In: *Cell* 88.2 (1997), pp. 243–251.

[16] Andre Goffeau et al. "Life with 6000 genes". In: *Science* 274.5287 (1996), pp. 546–567.

[17] Sydney Brenner et al. “Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays”. In: *Nature biotechnology* 18.6 (2000), pp. 630–634.

[18] Toshiyuki Shiraki et al. “Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage”. In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15776–15781.

[19] The Fantom Consortium, the RIKEN PMI, and CLST (DGT). “A promoter-level mammalian expression atlas”. In: *Nature* 507 (2014), pp. 462–470. URL: `https://doi.org/10.1038/nature13182`.

[20] CC. Hon, J. Ramilowski, J. Harshbarger, et al. “An atlas of human long non-coding RNAs with accurate 5’ ends”. In: *Nature* 543 (2017), pp. 199–204. URL: `https://doi.org/10.1038/nature21374`.

[21] David J. Lockhart et al. “Expression monitoring by hybridization to high-density oligonucleotide arrays”. In: *Nature biotechnology* 14.13 (1996), pp. 1675–1680.

[22] Mark Schena et al. “Quantitative monitoring of gene expression patterns with a complementary DNA microarray”. In: *Science* 270.5235 (1995), pp. 467–470.

[23] Genevieve Pietu et al. “Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array”. In: *Genome Research* 6.6 (1996), pp. 492–503.

[24] Mark D. Adams et al. “Complementary DNA sequencing: expressed sequence tags and human genome project”. In: *Science* 252.5013 (1991), pp. 1651–1656.

[25] Mark D. Adams et al. “Sequence identification of 2,375 human brain genes”. In: *Nature* 355.6361 (1992), pp. 632–634.

[26] Robert J. Lipshutz et al. “High density synthetic oligonucleotide arrays”. In: *Nature genetics* 21.1 (1999), pp. 20–24.

[27] Rafael A. Irizarry et al. “Exploration, normalization, and summaries of high density oligonucleotide array probe level data”. In: *Biostatistics* 4.2 (2003), pp. 249–264.

[28] Andrew I. Su et al. “A gene atlas of the mouse and human protein-encoding transcriptomes”. In: vol. 101. 16. 2004, pp. 6062–6067.

[29] Michael Ashburner et al. “Gene ontology: tool for the unification of biology”. In: *Nature genetics* 25.1 (2000), pp. 25–29.

[30] C. elegans Sequencing Consortium. “Genome sequence of the nematode C. elegans: a platform for investigating biology”. In: *Science* 282.5396 (1998), pp. 2012–2018.

[31] Mark D. Adams et al. “The genome sequence of Drosophila melanogaster”. In: *Science* 287.5461 (2000), pp. 2185–2195.

[32] European Bioinformatics Institute. “Initial sequencing and comparative analysis of the mouse genome”. In: *Nature* 420.6915 (2002), pp. 520–562.

[33] Gerald M. Rubin et al. “Comparative genomics of the eukaryotes”. In: *Science* 287.5461 (2000), pp. 2204–2215.

[34] Serafim Batzoglou et al. “Human and mouse gene structure: comparative analysis and application to exon prediction”. In: *Proceedings of the fourth annual international conference on Computational molecular biology*. 2000.

[35] Abel Ureta-Vidal, Laurence Ettwiller, and Ewan Birney. “Comparative genomics: genome-wide analysis in metazoan eukaryotes”. In: *Nature Reviews Genetics* 4.4 (2003), pp. 251–262.

[36] Amos Bairoch and Rolf Apweiler. “The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000”. In: *Nucleic acids research* 28.1 (2000), pp. 45–48.

[37] Wendy Baker et al. “The EMBL nucleotide sequence database”. In: *Nucleic Acids Research* 28.1 (2000), pp. 19–23.

[38] Alex Bateman et al. “The Pfam protein families database”. In: *Nucleic acids research* 28.1 (2000), pp. 263–266.

[39] Dennis A. Benson et al. “GenBank”. In: *Nucleic acids research* 28.1 (2000), pp. 15–18.

[40] Minoru Kanehisa et al. “The KEGG databases at GenomeNet”. In: *Nucleic acids research* 30.1 (2002), pp. 42–46.

[41] Joseph L. DeRisi, Vishwanath R. Iyer, and Patrick O. Brown. “Exploring the metabolic and genetic control of gene expression on a genomic scale”. In: *Science* 278.5338 (1997), pp. 680–686.

[42] Kam D. Dahlquist et al. “GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways”. In: *Nature genetics* 31.1 (2002), pp. 19–20.

[43] Todd R. Golub et al. “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”. In: *Science* 286.5439 (1999), pp. 531–537.

[44] S. Kropf and J. Lauter. “Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data”. In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 44.7 (2002), pp. 789–800.

[45] Scott W. Doniger et al. “MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data”. In: *Genome biology* 4.1 (2003), pp. 1–12.

[46] Vamsi K. Mootha et al. “PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes”. In: *Nature genetics* 34.3 (2003), pp. 267–273.

[47] Aravind Subramanian et al. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles”. In: vol. 102. 43. 2005, pp. 15545–15550.

[48] Douglas Hanahan and Robert A. Weinberg. “The hallmarks of cancer”. In: *Cell* 100.1 (2000), pp. 57–70.

[49] Douglas Hanahan. “The hallmarks of cancer”. In: *Hallmarks of cancer: new dimensions.* 12.1 (2022), pp. 31–46.
[50] Eran Segal et al. “Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data”. In: *Nature genetics* 34.2 (2003), pp. 166–176.
[51] Sridhar Ramaswamy et al. “A molecular signature of metastasis in primary solid tumors”. In: *Nature genetics* 33.1 (2003), pp. 49–54.
[52] Jan Ihmels et al. “Revealing modular organization in the yeast transcriptional network”. In: *Nature genetics* 31.4 (2002), pp. 370–377.
[53] Amos Tanay et al. “Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data”. In: *Proceedings of the National Academy of Sciences* 101.9 (2004), pp. 2981–2986.
[54] Eran Segal et al. “A module map showing conditional activity of expression modules in cancer”. In: *Nature genetics* 36.10 (2004), pp. 1090–1098.
[55] Richard O. Hynes. “Metastatic potential: generic predisposition of the primary tumor or rare, metastatic variants—or both?” In: *Cell* 113.7 (2003), pp. 821–823.
[56] Rene Bernards and Robert A. Weinberg. “Metastasis genes: a progression puzzle”. In: *Nature* 418.6900 (2002), pp. 823–823.
[57] Nicholas Wade. “Cloning Gold Rush Turns Basic Biology into Big Business: Cloning a gene can help raise 50 million dollars for your company. Will the laboratory suffer?” In: *Science* 208.4445 (1980), pp. 688–692.
[58] Yoshio Miki et al. “A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1”. In: *Science* 266.5182 (1994), pp. 66–71.
[59] Richard Wooster et al. “Identification of the breast cancer susceptibility gene BRCA2”. In: *Nature* 378.6559 (1995), pp. 789–792.
[60] Pieter Vos et al. “AFLP: a new technique for DNA fingerprinting”. In: *Nucleic acids research* 23.21 (1995), pp. 4407–4414.
[61] Giulia C. Kennedy et al. “Large-scale genotyping of complex DNA”. In: *Nature biotechnology* 21.10 (2003), pp. 1233–1237.
[62] Hajime Matsuzaki et al. “Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays”. In: *Nature Methods* 1.2 (2004), pp. 109–111.
[63] Joel N. Hirschhorn and Mark J. Daly. “Genome-wide association studies for common diseases and complex traits”. In: *Nature reviews genetics* 6.2 (2005), pp. 95–108.
[64] Hana Lango Allen. “Hundreds of variants clustered in genomic loci and biological pathways affect human height”. In: *Nature* 467.7317 (2010), pp. 832–838.

[65] Elizabeth K. Speliotes et al. “Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index”. In: *Nature genetics* 42.11 (2010), pp. 937–948.
[66] Douglas F. Easton et al. “Genome-wide association study identifies novel breast cancer susceptibility loci”. In: *Nature* 447.7148 (2007), pp. 1087–1093.
[67] Douglas F. Easton. “How many more breast cancer predisposition genes are there?” In: *Breast Cancer Research* 1.1 (1999), pp. 1–4.
[68] Brent W. Zanke et al. “Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24”. In: *Nature genetics* 39.8 (2007), pp. 989–994.
[69] Rosalind A. Eeles et al. “Multiple newly identified loci associated with prostate cancer susceptibility”. In: *Nature genetics* 40.3 (2008), pp. 316–321.
[70] Christopher I. Amos et al. “Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25. 1”. In: *Nature genetics* 40.5 (2008), pp. 616–622.
[71] Michael S. Lawrence et al. “Mutational heterogeneity in cancer and the search for new cancer-associated genes”. In: *Nature* 499.7457 (2013), pp. 214–218.
[72] Theodosius Dobzhansky. “Nothing in biology makes sense except in the light of evolution”. In: *The american biology teacher* 75.2 (2013), pp. 87–91.
[73] Frederick Sanger. “Sequences, sequences, and sequences”. In: *Annual review of biochemistry* 57.1 (1988), pp. 1–29.
[74] Frederick Sanger et al. “Nucleotide sequence of bacteriophage phiX174 DNA”. In: *Nature* 265.5596 (1977), pp. 687–695.
[75] Lloyd M. Smith et al. “Fluorescence detection in automated DNA sequence analysis”. In: *Nature* 321.6071 (1986), pp. 674–679.
[76] Olena Morozova and Marco A. Marra. “Applications of next-generation sequencing technologies in functional genomics”. In: *Genomics* 92.5 (2008), pp. 255–264.
[77] Pettersson Nyren, Bertil Pettersson, and Mathias Uhlen. “Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay”. In: *Analytical biochemistry* 208.1 (1993), pp. 171–175.
[78] Mostafa Ronaghi, Mathias Uhlen, and Paal Nyren. “A sequencing method based on real-time pyrophosphate”. In: *Science* 281.5375 (1998), pp. 363–365.
[79] Marcel Margulies et al. “Genome sequencing in microfabricated high-density picolitre reactors”. In: *Nature* 437.7057 (2005), pp. 376–380.
[80] Jay Shendure et al. “Accurate multiplex polony sequencing of an evolved bacterial genome”. In: *Science* 309.5741 (2005), pp. 1728–1732.

[81] S. Balasubramanian and D. R. Bentley. “Polynucleotide arrays and their use in sequencing”. In: *Patent WO* 1 (2001).

[82] Simon T. Bennett et al. “Toward the $1000 human genome”. In: (2005), pp. 373–382.

[83] Hutchison Iii and A. Clyde. “DNA sequencing: bench to bedside and beyond”. In: *Nucleic acids research* 35.18 (2007), pp. 6227–6237.

[84] Elliot J. Lefkowitz et al. “Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)”. In: *Nucleic acids research* 46 (2018), pp. 708–717.

[85] Zhong Wang, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[86] Thomas E. Royce, Joel S. Rozowsky, and Mark B. Gerstein. “Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification”. In: *Nucleic acids research* 35 (2007), p. 15.

[87] Ali Mortazavi et al. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature methods* 5.7 (2008), pp. 621–628.

[88] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. “TopHat: discovering splice junctions with RNA-Seq”. In: *Bioinformatics* 25.9 (2009), pp. 1105–1111.

[89] Robert Piskol, Gokul Ramaswami, and Jin Billy Li. “Reliable identification of genomic variants from RNA-seq data”. In: *The American Journal of Human Genetics* 93.4 (2013), pp. 641–651.

[90] Ugrappa Nagalakshmi et al. “The transcriptional landscape of the yeast genome defined by RNA sequencing”. In: *Science* 320.5881 (2008), pp. 1344–1349.

[91] Brian T. Wilhelm et al. “Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution”. In: *Nature* 453.7199 (2008), pp. 1239–1243.

[92] Nicole Cloonan et al. “Stem cell transcriptome profiling via massive-scale mRNA sequencing”. In: *Nature methods* 5.7 (2008), pp. 613–619.

[93] Heng Li et al. “The sequence alignment/map format and SAMtools”. In: *bioinformatics* 25.16 (2009), pp. 2078–2079.

[94] Stephen F. Altschul et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. In: *Nucleic acids research* 25.17 (1997), pp. 3389–3402.

[95] W. James Kent. “BLAT—the BLAST-like alignment tool”. In: *Genome research* 12.4 (2002), pp. 656–664.

[96] Zemin Ning, Anthony J. Cox, and James C. Mullikin. “SSAHA: a fast search method for large DNA databases”. In: *Genome research* 11.10 (2001), pp. 1725–1729.

[97] Ruiqiang Li et al. “SOAP: short oligonucleotide alignment program”. In: *Bioinformatics* 24.5 (2008), pp. 713–714.

[98] Heng Li, Jue Ruan, and Richard Durbin. “Mapping short DNA sequencing reads and calling variants using mapping quality scores”. In: *Genome research* 18.11 (2008), pp. 1851–1858.

[99] John C. Marioni et al. “RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays”. In: *Genome research* 18.9 (2008), pp. 1509–1517.

[100] Ben Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome biology* 10.3 (2009), pp. 1–10.

[101] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[102] Helen Davies et al. “Mutations of the BRAF gene in human cancer”. In: *Nature* 417.6892 (2002), pp. 949–954.

[103] P. Andrew Futreal et al. “A census of human cancer genes”. In: *Nature reviews cancer* 4.3 (2004), pp. 177–183.

[104] Peter C. Nowell. “The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression”. In: *Science* 194.4260 (1976), pp. 23–28.

[105] Peter C. Nowell. *Tumor progression: a brief historical perspective*. Vol. 12. Seminars in cancer biology. No. 4: Academic Press, 2002.

[106] Carlo C. Maley et al. “Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett’s esophagus”. In: *Cancer research* 64.10 (2004), pp. 3414–3427.

[107] Laura D. Wood et al. “The genomic landscapes of human breast and colorectal cancers”. In: *Science* 318.5853 (2007), pp. 1108–1113.

[108] Aya Sasaki et al. “Filamin associates with Smads and regulates transforming growth factor-beta signaling”. In: *Journal of Biological Chemistry* 276.21 (2001), pp. 17871–17877.

[109] Sohrab P. Shah et al. “Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution”. In: *Nature* 461.7265 (2009), pp. 809–813.

[110] Philip J. Stephens et al. “Complex landscapes of somatic rearrangement in human breast cancer genomes”. In: *Nature* 462.7276 (2009), pp. 1005–1010.

[111] Timothy J. Ley et al. “DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome”. In: *Nature* 456.7218 (2008), pp. 66–72.

[112] Erin D. Pleasance et al. “A small-cell lung cancer genome with complex signatures of tobacco exposure”. In: *Nature* 463.7278 (2010), pp. 184–190.

[113] David Dickson. “Wellcome funds cancer database”. In: *Nature* 401.6755 (1999), pp. 729–729.

[114] Nazneen Rahman et al. “PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene”. In: *Nature genetics* 39.2 (2007), pp. 165–167.

[115] Peter J. Campbell et al. “Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing”. In: *Proceedings of the National Academy of Sciences* 105.35 (2008), pp. 13081–13086.

[116] Ultan McDermott et al. “Genomic alterations of anaplastic lymphoma kinase may sensitize tumors to anaplastic lymphoma kinase inhibitors”. In: *Cancer research* 68.9 (2008), pp. 3389–3395.

[117] Simon A. Forbes et al. “COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer”. In: *Nucleic acids research* 38 (2010), pp. D652–D657.

[118] ICGC. “International network of cancer genome projects”. In: *Nature* 464.7291 (2010), pp. 993–998.

[119] Kun Zhang et al. “Sequencing genomes from single cells by polymerase cloning”. In: *Nature biotechnology* 24.6 (2006), pp. 680–686.

[120] cancer.gov. *TCGA timeline*. accessed on 22.05.2023. 2018. URL: `https://www.cancer.gov/ccg/research/genome-sequencing/tcga/history/timeline-milestones`.

[121] Francis S. Collins and Anna D. Barker. “Mapping the cancer genome”. In: *Scientific American* 296.3 (2007), pp. 50–57.

[122] Cancer Genome Atlas Research Network. “Comprehensive genomic characterization defines human glioblastoma genes and core pathways”. In: *Nature* 455 (2008), p. 7216.

[123] Cancer Genome Atlas Research Network. “Integrated genomic analyses of ovarian carcinoma”. In: *Nature* 474 (2011), p. 7353.

[124] Cancer Genome Atlas Network. “Comprehensive molecular characterization of human colon and rectal cancer”. In: *Nature* 487 (2012), p. 7407.

[125] Cancer Genome Atlas Research Network. “Comprehensive genomic characterization of squamous cell lung cancers”. In: *Nature* 489 (2012), p. 7417.

[126] Brigham, Women’s Hospital, and Harvard Medical. “Comprehensive molecular portraits of human breast tumours”. In: *Nature* 490.7418 (2012), pp. 61–70.

[127] Cancer Genome Atlas Research Network. “Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia”. In: *New England Journal of Medicine* 368.22 (2013), pp. 2059–2074.

[128] Douglas A. Levine et al. “Integrated genomic characterization of endometrial carcinoma”. In: *Nature* 497.7447 (2013), pp. 67–73.

[129] Cancer Genome Atlas Research Network Analysis working group. "Comprehensive molecular characterization of clear cell renal cell carcinoma". In: *Nature* 499.7456 (2013), pp. 43–49.

[130] Levi A. Garraway and William R. Sellers. "Lineage dependency and lineage-survival oncogenes in human cancer". In: *Nature Reviews Cancer* 6.8 (2006), pp. 593–602.

[131] Data Coordinating Center. "The cancer genome atlas pan-cancer analysis project". In: *Nature genetics* 45.10 (2013), pp. 1113–1120.

[132] Katherine A. Hoadley et al. "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin". In: *Cell* 158.4 (2014), pp. 929–944.

[133] S. et al. Monti. "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene". In: *Machine Learning* 1.52 (2003), pp. 91–118.

[134] Matthew D. Wilkerson and D. Neil Hayes. "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking". In: *Bioinformatics* 26.12 (2010), pp. 1572–1573.

[135] Dvir Aran, Marina Sirota, and Atul J. Butte. "Systematic pan-cancer analysis of tumour purity". In: *Nature communications* 6 (2015), p. 1.

[136] Katherine A. Hoadley et al. "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer". In: *Cell* 173.2 (2018), pp. 291–304.

[137] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis". In: *Bioinformatics* 25.22 (2009), pp. 2906–2912.

[138] Francisco Sanchez-Vega et al. "Oncogenic signaling pathways in the cancer genome atlas". In: *Cell* 173.2 (2018), pp. 321–337.

[139] Web of Science. *Query: GP=" Canc Genome Atlas Res Network"*. Accessed on 25.05.2023. URL: `https://www.webofscience.com/wos/woscc/advanced-searc`.

[140] Robert A. Weinberg. "Coming full circle—from endless complexity to simplicity and back again". In: *Cell* 157.1 (2014), pp. 267–271.

[141] Amin Hojat et al. "Procurement and storage of surgical biospecimens". In: *Biobanking: Methods and Protocols :* (2019), pp. 65–76.

[142] Maryam Shabihkhani et al. "The procurement, storage, and quality assurance of frozen blood and tissue biospecimens in pathology, biorepository, and biobank settings". In: *Clinical biochemistry* 47.4-5 (2014), pp. 258–266.

[143] Levi A. Garraway et al. "Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma". In: *Nature* 436.7047 (2005), pp. 117–122.

[144] David B. Solit et al. "BRAF mutation predicts sensitivity to MEK inhibition". In: *Nature* 439.7074 (2006), pp. 358–362.

[145] William M. Lin et al. “Modeling genomic diversity and tumor dependency in malignant melanoma”. In: *Cancer research* 68.3 (2008), pp. 664–673.

[146] R. H. Shoemaker et al. “Development of human tumor cell line panels for use in disease-oriented drug screening”. In: *Progress in clinical and biological research* 276 (1988), pp. 265–286.

[147] John N. Weinstein et al. “An information-intensive approach to the molecular pharmacology of cancer”. In: *Science* 275.5298 (1997), pp. 343–349.

[148] Jordi Barretina et al. “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. In: *Nature* 483.7391 (2012), pp. 603–607.

[149] Mathew J. Garnett et al. “Systematic identification of genomic markers of drug sensitivity in cancer cells”. In: *Nature* 483.7391 (2012), pp. 570–575.

[150] Wanjuan Yang et al. “Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells”. In: *Nucleic acids research* 41 (2012), pp. D955–D961.

[151] Channing Yu et al. “High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines”. In: *Nature biotechnology* 34.4 (2016), pp. 419–423.

[152] Steven M. Corsello et al. “Discovering the anticancer potential of non-oncology drugs by systematic viability profiling”. In: *Nature cancer* 1.2 (2020), pp. 235–248.

[153] Aviad Tsherniak et al. “Defining a cancer dependency map”. In: *Cell* 170.3 (2017), pp. 564–576.

[154] broadinstitute.org. *DepMap Timeline*. 2019. URL: `https://www.broadinstitute.org/news/broad-institute-launches-academic-industrial-consortium-cancer-dependency-studies`.

[155] Jesse S. Boehm and Todd R. Golub. “An ecosystem of cancer cell line factories to support a cancer dependency map”. In: *Nature Reviews Genetics* 16.7 (2015), pp. 373–374.

[156] Andrew L. Hong et al. “Integrated genetic and pharmacologic interrogation of rare cancers”. In: *Nature communications* 7 (2016), p. 1.

[157] Jose V. Castell and Maria Jose Gomez-Lechon. “Liver cell culture techniques”. In: *Hepatocyte Transplantation: Methods and Protocols :* (2009), pp. 35–46.

[158] J. Miki and J. S. Rhim. “Prostate cell cultures as in vitro models for the study of normal stem cells and cancer stem cells”. In: *Prostate cancer and prostatic diseases* 11.1 (2008), pp. 32–39.

[159] Xuefeng Liu et al. “Cell-restricted immortalization by human papillomavirus correlates with telomerase activation and engagement of

the hTERT promoter by Myc". In: *Journal of virology* 82.23 (2008), pp. 11568–11576.

[160] Xuefeng Liu et al. "ROCK inhibitor and feeder cells induce the conditional reprogramming of epithelial cells". In: *The American journal of pathology* 180.2 (2012), pp. 599–607.

[161] Aliya Fatehullah, Si Hui Tan, and Nick Barker. "Organoids as an in vitro model of human development and disease". In: *Nature cell biology* 18.3 (2016), pp. 246–254.

[162] Marc Van de Wetering et al. "Prospective derivation of a living organoid biobank of colorectal cancer patients". In: *Cell* 161.4 (2015), pp. 933–945.

[163] Megan Cully. "Xenograft encyclopaedia identifies drug combination opportunities". In: *Nature Reviews Drug Discovery* 14.12 (2015), pp. 819–819.

[164] John J. Tentler et al. "Patient-derived tumour xenografts as models for oncology drug development". In: *Nature reviews Clinical oncology* 9.6 (2012), pp. 338–350.

[165] Andrea Bertotti et al. "A Molecularly Annotated Platform of Patient-Derived Xenografts ("Xenopatients") Identifies HER2 2 as an Effective Therapeutic Target in Cetuximab-Resistant Colorectal CancerHER2 Amplification and Cetuximab Resistance in Colon Cancer". In: *Cancer discovery* 1.6 (2011), pp. 508–523.

[166] Hui Gao et al. "High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response". In: *Nature medicine* 21.11 (2015), pp. 1318–1325.

[167] Mahmoud Ghandi et al. "Next-generation characterization of the cancer cell line encyclopedia". In: *Nature* 569.7757 (2019), pp. 503–508.

[168] David P. Nusinow et al. "Quantitative proteomics of the cancer cell line encyclopedia". In: *Cell* 180.2 (2020), pp. 387–402.

[169] Justin Lamb et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease". In: *Science* 313.5795 (2006), pp. 1929–1935.

[170] Timothy R. Hughes et al. "Functional discovery via a compendium of expression profiles". In: *Cell* 102.1 (2000), pp. 109–126.

[171] Huck CP Gargiulo AV Burns GM. "Dyclonine hydrochloride–a topical agent for managing pain". In: *Ill Dent J.* 61.4 (1992), pp. 303–4.

[172] Ramesh Kekuda et al. "Cloning and functional expression of the human type 1 sigma receptor (hSigmaR1)". In: *Biochemical and biophysical research communications* 229.2 (1996), pp. 553–558.

[173] Russell A. Wilke et al. "K+ channel modulation in rodent neurohypophysial nerve terminals by sigma receptors and not by dopamine receptors". In: *The Journal of physiology* 517.2 (1999), pp. 391–406.

[174] K. B. Glaser. "Gene expression histone deacetylase inhibitors in T24 and MDA cell lines". In: *Cancer Therapy* 2.2 (2003), pp. 151–63.
[175] Walton JD. "HC-toxin". In: *Phytochemistry* 67.14 (2006).
[176] et al. Göttlicher M Minucci S. "Valproic acid defines a novel class of HDAC inhibitors inducing differentiation of transformed cells". In: *EMBO J.* 20.24 (2001), pp. 6969–78. URL: 10.1093/emboj/20.24.6969.
[177] Andrea M. Brum et al. "Connectivity Map-based discovery of parbendazole reveals targetable human osteogenic pathway". In: *Proceedings of the National Academy of Sciences* 112.41 (2015), pp. 12711–12716.
[178] Junli Liu et al. "Treatment of obesity with celastrol". In: *Cell* 161.5 (2015), pp. 999–1011.
[179] Michelle L. Churchman et al. "Efficacy of retinoids in IKZF1-mutated BCR-ABL1 acute lymphoblastic leukemia". In: *Cancer cell* 28.3 (2015), pp. 343–356.
[180] Alok R. Singh et al. "PI-3K inhibitors preferentially target CD15+ cancer stem cell population in SHH driven medulloblastoma". In: *PloS one* 11 (2016), p. 3.
[181] Aravind Subramanian et al. "A next generation connectivity map: L1000 platform and the first 1,000,000 profiles". In: *Cell* 171.6 (2017), pp. 1437–1452.
[182] George C. Tseng and Wing H. Wong. "Tight clustering: a resampling-based approach for identifying stable and tight patterns in data". In: *Biometrics* 61.1 (2005), pp. 10–16.
[183] GTEx Consortium et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans". In: *Science* 348.6235 (2015), pp. 648–660.
[184] Livnat Jerby-Arnon et al. "A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade". In: *Cell* 175.4 (2018), pp. 984–997.
[185] Uri Ben-David et al. "Genetic and transcriptional evolution alters cancer cell line drug response". In: *Nature* 560.7718 (2018), pp. 325–330.
[186] Tathiane M. Malta et al. "Machine learning identifies stemness features associated with oncogenic dedifferentiation". In: *Cell* 173.2 (2018), pp. 338–354.
[187] Marcin Pilarczyk et al. "Connecting omics signatures and revealing biological mechanisms with iLINCS". In: *Nature Communications* 13 (2022), p. 1.
[188] Nathaniel Lim and Paul Pavlidis. "Evaluation of connectivity map shows limited reproducibility in drug repositioning". In: *Scientific reports* 11 (2021), p. 1.
[189] Minji Jeon et al. "Transforming L1000 profiles to RNA-seq-like profiles with deep learning". In: *BMC bioinformatics* 23 (2022), p. 1.

[190] Daniel A. Haber and Victor E. Velculescu. "Blood-Based Analyses of Cancer: Circulating Tumor Cells and Circulating Tumor DNABlood-Based Analysis of Cancer". In: *Cancer discovery* 4.6 (2014), pp. 650–661.

[191] Yuxuan Wang et al. "Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas". In: *Science translational medicine* 7.293 (2015), pp. 104–293.

[192] Joshua D. Cohen et al. "Detection and localization of surgically resectable cancers with a multi-analyte blood test". In: *Science* 359.6378 (2018), pp. 926–930.

[193] Tony S. Mok et al. "Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma". In: *New England Journal of Medicine* 361.10 (2009), pp. 947–957.

[194] Brian J. Druker et al. "Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia". In: *New England Journal of Medicine* 355.23 (2006), pp. 2408–2417.

[195] Michael C. Heinrich et al. "Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor". In: *Journal of clinical oncology* 21.23 (2003), pp. 4342–4349.

[196] Jessica J. Lin, Gregory J. Riely, and Alice T. Shaw. "Targeting ALK: Precision Medicine Takes on Drug ResistanceOvercoming Resistance to ALK Inhibitors". In: *Cancer discovery* 7.2 (2017), pp. 137–155.

[197] Arnaud Boyer et al. "Drug repurposing in malignant pleural mesothelioma: a breath of fresh air?" In: *European Respiratory Review* 27 (2018), p. 147.

[198] Don Benjamin et al. "Rapamycin passes the torch: a new generation of mTOR inhibitors". In: *Nature reviews Drug discovery* 10.11 (2011), pp. 868–880.

[199] Hope S. Rugo et al. "A randomized phase II trial of ridaforolimus, dalotuzumab, and exemestane compared with ridaforolimus and exemestane in patients with advanced breast cancer". In: *Breast cancer research and treatment* 165 (2017), pp. 601–609.

[200] Diana Cirstea et al. "Dual inhibition of Akt/mammalian target of rapamycin pathway by nanoparticle albumin-bound–rapamycin and perifosine induces antitumor activity in multiple myeloma". In: *Molecular cancer therapeutics* 9.4 (2010), pp. 963–975.

[201] seer.cancer.gov. *Seer*. accessed on 30.05.23. URL: `https://seer.cancer.gov/statfacts/html/all.html;`.

[202] Jens G. Lohr et al. "Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy". In: *Cancer cell* 25.1 (2014), pp. 91–101.

[203] Samra Turajlic et al. "Resolving genetic heterogeneity in cancer". In: *Nature Reviews Genetics* 20.7 (2019), pp. 404–416.

[204] Martincorena et al. “Universal patterns of selection in cancer and somatic tissues”. In: *Cell* 171.5 (2017), pp. 1029–1041.

[205] Rachel Marty Pyke et al. “Evolutionary pressure against MHC class II binding cancer mutations”. In: *Cell* 175.2 (2018), pp. 416–428.

[206] Trevor A. Graham and Andrea Sottoriva. “Measuring cancer evolution from the genome”. In: *The Journal of pathology* 241.2 (2017), pp. 183–191.

[207] Nicolai J. Birkbak et al. “Paradoxical Relationship between Chromosomal Instability and Survival Outcome in CancerChromosomal Instability and Cancer Survival”. In: *Cancer research* 71.10 (2011), pp. 3447–3452.

[208] Ruchira et al. S. Datta. “Modelling the evolution of genetic instability during tumour progression”. In: *Evolutionary applications* 6.1 (2013), pp. 20–33.

[209] Lawrence A. Loeb. “Mutator phenotype in cancer: origin and consequences”. In: *Seminars in cancer biology* 20 (2010).

[210] Austin L. Hughes. “Near neutrality: leading edge of the neutral theory of molecular evolution”. In: *Annals of the New York Academy of Sciences* 1133.1 (2008), pp. 162–179.

[211] Martincorena et al. “High burden and pervasive positive selection of somatic mutations in normal human skin”. In: *Science* 348.6237 (2015), pp. 880–886.

[212] Jessica Okosun et al. “Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma”. In: *Nature genetics* 46.2 (2014), pp. 176–181.

[213] Ruli Gao et al. “Punctuated copy number evolution and clonal stasis in triple-negative breast cancer”. In: *Nature genetics* 48.10 (2016), pp. 1119–1130.

[214] Niles Eldredge, Stephen Jay Gould, and Thomas J. M. Schopf. “Punctuated equilibria: an alternative to phyletic gradualism”. In: *Models in paleobiology* (1972), pp. 82–115.

[215] Philip J. Stephens et al. “Massive genomic rearrangement acquired in a single catastrophic event during cancer development”. In: *Cell* 144.1 (2011), pp. 27–40.

[216] Sylvan C. Baca et al. “Punctuated evolution of prostate cancer genomes”. In: *Cell* 153.3 (2013), pp. 666–677.

[217] Richard Goldschmidt. *The material basis of evolution*. Vol. 28. Yale University Press, 1982.

[218] James Hicks et al. “Novel patterns of genome rearrangement and their association with survival in breast cancer”. In: *Genome research* 16.12 (2006), pp. 1465–1479.

[219] Sylvan C. Baca et al. “Punctuated evolution of prostate cancer genomes”. In: *Cell* 153.3 (2013), pp. 666–677.

[220] Philipp M. Altrock, Lin L. Liu, and Franziska Michor. "The mathematics of cancer: integrating quantitative models". In: *Nature Reviews Cancer* 15.12 (2015), pp. 730–745.

[221] Ignacio A. Rodriguez-Brenes, Natalia L. Komarova, and Dominik Wodarz. "Tumor growth dynamics: insights into evolutionary processes". In: *Trends in ecology & evolution* 28.10 (2013), pp. 597–604.

[222] Diaz Jr, A. Luis, et al. "The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers". In: *Nature* 486.7404 (2012), pp. 537–540.

[223] J. Randolph Hecht et al. "Lack of correlation between epidermal growth factor receptor status and response to panitumumab monotherapy in metastatic colorectal cancer". In: *Clinical Cancer Research* 16.7 (2010), pp. 2205–2213.

[224] Bartlomiej Waclaw et al. "A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity". In: *Nature* 525.7568 (2015), pp. 261–264.

[225] Jordan M. Winter et al. "Absence of E-cadherin expression distinguishes noncohesive from cohesive pancreatic cancer". In: *Clinical cancer research* 14.2 (2008), pp. 412–418.

[226] Andrea Sottoriva et al. "A Big Bang model of human colorectal tumor growth". In: *Nature genetics* 47.3 (2015), pp. 209–216.

[227] Kimberly D. Siegmund et al. "Many colorectal cancers are "flat" clonal expansions". In: *Cell Cycle* 8.14 (2009), pp. 2187–2193.

[228] Adam Humphries et al. "Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution". In: *Proceedings of the National Academy of Sciences* 110.27 (2013), pp. 2490–2499.

[229] Sridhar Ramaswamy et al. "A molecular signature of metastasis in primary solid tumors". In: *Nature genetics* 33.1 (2003), pp. 49–54.

[230] Rene Bernards and Robert A. Weinberg. "Metastasis genes: a progression puzzle". In: *Nature* 418.6900 (2002), pp. 823–823.

[231] Saife N. Lone et al. "Liquid biopsy: A step closer to transform diagnosis, prognosis and future of cancer treatments". In: *Molecular cancer* 21.1 (2022), pp. 1–22.

[232] Samuel W. Brady, Alexander M. Gout, and Jinghui Zhang. "Therapeutic and prognostic insights from the analysis of cancer mutational signatures". In: *Trends in Genetics* 38.2 (2022), pp. 194–208.

[233] tracerx.co.uk. *TracerX study*. accessed on 01.06.2023. URL: `http://tracerx.co.uk/`.

[234] Samra Turajlic et al. "Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal". In: *Cell* 173.3 (2018), pp. 595–610.

[235] Mariam Jamal-Hanjani et al. “Tracking the evolution of non–small-cell lung cancer”. In: *New England Journal of Medicine* 376.22 (2017), pp. 2109–2121.

[236] Nicholas McGranahan et al. “Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade”. In: *Science* 351.6280 (2016), pp. 1463–1469.

[237] Santiago Zelenay et al. “Cyclooxygenase-dependent tumor growth through evasion of immunity”. In: *Cell* 162.6 (2015), pp. 1257–1270.

[238] Ian J. Frew and Holger Moch. “A clearer view of the molecular complexity of clear cell renal cell carcinoma”. In: *Annual Review of Pathology: Mechanisms of Disease* 10 (2015), pp. 263–289.

[239] Thomas J. Mitchell et al. “Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx renal”. In: *Cell* 173.3 (2018), pp. 611–623.

[240] Sarah M. Nielsen et al. *Von Hippel-Lindau disease: genetics and role of genetic counseling in a multiple neoplasia syndrome*. American Society of Clinical Oncology, 2016.

[241] Samra Turajlic et al. “Tracking cancer evolution reveals constrained routes to metastases: TRACERx renal”. In: *Cell* 173.3 (2018), pp. 581–594.

[242] M. Bianchi et al. “Distribution of metastatic sites in renal cell carcinoma: a population-based analysis”. In: *Annals of Oncology* 23.4 (2012), pp. 973–980.

[243] Rana R. McKay et al. “Impact of bone and liver metastases on patients with renal cell carcinoma treated with targeted therapy”. In: *European urology* 65.3 (2014), pp. 577–584.

[244] A. Bex et al. “Immediate versus deferred cytoreductive nephrectomy (CN) in patients with synchronous metastatic renal cell carcinoma (mRCC) receiving sunitinib (EORTC 30073 SURTIME)”. In: *Annals of Oncology* 28 (2017).

[245] Lucy R. Yates et al. “Genomic evolution of breast cancer metastasis and relapse”. In: *Cancer cell* 32.2 (2017), pp. 169–184.

[246] Michael C. Haffner et al. “Tracking the clonal origin of lethal prostate cancer”. In: *The Journal of clinical investigation* 123.11 (2013), pp. 4918–4922.

[247] Johannes G. Reiter et al. “Lymph node metastases develop through a wider evolutionary bottleneck than distant metastases”. In: *Nature genetics* 52.7 (2020), pp. 692–700.

[248] Faiyaz Notta et al. “A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns”. In: *Nature* 538.7625 (2016), pp. 378–382.

[249] Matthew G. Field et al. “Punctuated evolution of canonical genomic aberrations in uveal melanoma”. In: *Nature communications* 9 (2018).

[250] David W. Cescon et al. “Circulating tumor DNA and liquid biopsy in oncology”. In: *Nature Cancer* 1.3 (2020), pp. 276–290.
[251] Patricia J. Brooks et al. “Isolation of salivary cell-free DNA for cancer detection”. In: *Plos one* 18 (2023), p. 5.
[252] Mikkel H. Christensen et al. “DREAMS: deep read-level error model for sequencing data applied to low-frequency variant calling and circulating tumor DNA detection”. In: *Genome Biology* 24 (2023), p. 1.
[253] Jin H. Bae et al. “Single duplex DNA sequencing with CODEC detects mutations with high sensitivity”. In: *Nature Genetics* (2023), pp. 1–9.
[254] Moritz Gerstung et al. “The evolutionary history of 2,658 cancers”. In: *Nature* 578.7793 (2020), pp. 122–128.
[255] Verena Korber et al. “Neuroblastoma arises in early fetal development and its evolutionary duration predicts outcome”. In: *Nature Genetics* (2023), pp. 1–12.
[256] cancerresearchuk.org. *PEACE clinical trial*. Accessed on 02.06.2023. 2022. URL: `https://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/a-study-looking-at-blood-and-tissue-samples-to-learn-more-about-advanced-cancer-peace;`.
[257] Iain C. Macaulay et al. “G&T-seq: parallel sequencing of single-cell genomes and transcriptomes”. In: *Nature methods* 12.6 (2015), pp. 519–522.
[258] Stefanie Barthel et al. “Single-cell profiling to explore pancreatic cancer heterogeneity, plasticity and response to therapy”. In: *Nature Cancer* (2023), pp. 1–14.
[259] James Eberwine et al. “Analysis of gene expression in single live neurons”. In: *Proceedings of the National Academy of Sciences* 89.7 (1992), pp. 3010–3014.
[260] Antonio Peixoto et al. “Quantification of multiple gene expression in individual cells”. In: *Genome research* 14.10 (2004), pp. 1938–1947.
[261] Kazuki Kurimoto et al. “Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis”. In: *Nature protocols* 2.3 (2007), pp. 739–752.
[262] Daniel Ramskold et al. “Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells”. In: *Nature biotechnology* 30.8 (2012), pp. 777–782.
[263] Saiful Islam et al. “Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq”. In: *Genome research* 21.7 (2011), pp. 1160–1167.
[264] Sanja Vickovic et al. “Massive and parallel expression profiling using microarrayed single-cell sequencing”. In: *Nature communications* 7 (2016), p. 1.

[265] Mauro J. Muraro et al. “A single-cell transcriptome atlas of the human pancreas”. In: *Cell systems* 3.4 (2016), pp. 385–394.

[266] Amit Zeisel et al. “Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq”. In: *Science* 347.6226 (2015), pp. 1138–1142.

[267] Linas Mazutis et al. “Single-cell analysis and sorting using droplet-based microfluidics”. In: *Nature protocols* 8.5 (2013), pp. 870–891.

[268] H. Christina Fan, Glenn K. Fu, and Stephen P. A. Fodor. “Combinatorial labeling of single cells for gene expression cytometry”. In: *Science* 347 (2015), p. 6222.

[269] Pachter lab. *scRNA-seq intro*. Accessed on 04.06.2023. URL: `https://pachterlab.github.io/kallistobustools/tutorials/scRNA-seq_intro/python/scRNA-seq_intro/;`.

[270] Valentine Svensson, Roser Vento-Tormo, and Sarah A. Teichmann. “Exponential scaling of single-cell RNA-seq in the past decade”. In: *Nature protocols* 13.4 (2018), pp. 599–604.

[271] Charles Gawad, Winston Koh, and Stephen R. Quake. “Single-cell genome sequencing: current state of the science”. In: *Nature Reviews Genetics* 17.3 (2016), pp. 175–188.

[272] Claudia Spits et al. “Whole-genome multiple displacement amplification from single cells”. In: *Nature protocols* 1.4 (2006), pp. 1965–1970.

[273] Frank B. Dean et al. “Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification”. In: *Genome research* 11.6 (2001), pp. 1095–1099.

[274] P. Burlet et al. “Multiple displacement amplification improves PGD for fragile X syndrome”. In: *Molecular human reproduction* 12.10 (2006), pp. 647–652.

[275] Roger S. Lasken and Timothy B. Stockwell. “Mechanism of chimera formation during the Multiple Displacement Amplification reaction”. In: *BMC biotechnology* 7 (2007), pp. 1–11.

[276] H. K. Telenius, N. P. Carter, and C. E. Rebb. “Degenerate oligonucleotide-primed PCR: General amplification of target DNA by single degenerate primer”. In: *Genomics* 13 (1992), pp. 718–725.

[277] John P. Langmore. “Rubicon Genomics, Inc”. In: *Pharmacogenomics* 3.4 (2002), pp. 557–560.

[278] Chenghang Zong et al. “Genome-wide detection of single-nucleotide and copy-number variations of a single human cell”. In: *Science* 338.6114 (2012), pp. 1622–1626.

[279] Yong Hou et al. “Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing”. In: *Gigascience* 4.1 (2015), pp. 13742–015.

[280] Lei Huang et al. “Single-cell whole-genome amplification and sequencing: methodology and applications”. In: *Annual review of genomics and human genetics* 16 (2015), pp. 79–102.

[281] David R. Bentley et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *Nature* 456.7218 (2008), pp. 53–59.

[282] Nicholas Navin et al. “Tumour evolution inferred by single-cell sequencing”. In: *Nature* 472.7341 (2011), pp. 90–94.

[283] Joao M. Alves et al. “Clonality and timing of relapsing colorectal cancer metastasis revealed through whole-genome single-cell sequencing”. In: *Cancer letters* 543 (2022).

[284] Jens G. Lohr et al. “Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer”. In: *Nature biotechnology* 32.5 (2014), pp. 479–484.

[285] Emily S. Park et al. “Isolation and genome sequencing of individual circulating tumor cells using hydrogel encapsulation and laser capture microdissection”. In: *Lab on a Chip* 18.12 (2018), pp. 1736–1749.

[286] Ermin Hodzic et al. “Identification of conserved evolutionary trajectories in tumors”. In: *Bioinformatics* 36 (2020).

[287] Valentin Svensson. *Variance stabilising transformation*. Accessed on 05.06.2023. 2017. URL: `https://www.nxn.se/valent/2017/10/15/variance-stabilizing-scrna-seq-counts;`.

[288] Wouter Saelens et al. “A comparison of single-cell trajectory inference methods”. In: *Nature biotechnology* 37.5 (2019), pp. 547–554.

[289] Constantin Ahlmann-Eltze and Wolfgang Huber. “Comparison of transformations for single-cell RNA-seq data”. In: *Nature Methods* (2023), pp. 1–8.

[290] Francis J. Anscombe. “The transformation of Poisson, binomial and negative-binomial data”. In: *Biometrika* 35.3 (1948), pp. 246–254.

[291] David I. Warton. “Why you cannot transform your way out of trouble for small counts”. In: *Biometrics* 74.1 (2018), pp. 362–368.

[292] Christoph Hafemeister and Rahul Satija. “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”. In: *Genome biology* 20 (2019), p. 1.

[293] Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13 (2021), pp. 3573–3587.

[294] 10x Genomics. *How are the UMI counts are normalised*. Accessed on 05.06.2023. 2018. URL: `https://kb.10xgenomics.com/hc/en-us/articles/115004583806-How-are-the-UMI-counts-normalized-before-PCA-and-differential-expression-;`.

[295] Leo Breiman. “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”. In: *Statistical science* 16.3 (2001), pp. 199–231.

[296] Peter V. Kharchenko, Lev Silberstein, and David T. Scadden. "Bayesian approach to single-cell differential expression analysis". In: *Nature methods* 11.7 (2014), pp. 740–742.

[297] Valentine Svensson. "Droplet scRNA-seq is not zero-inflated". In: *Nature Biotechnology* 38.2 (2020), pp. 147–150.

[298] Huipeng Li et al. "Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors". In: *Nature genetics* 49.5 (2017), pp. 708–718.

[299] Mo Huang et al. "SAVER: gene expression recovery for single-cell RNA sequencing". In: *Nature methods* 15.7 (2018), pp. 539–542.

[300] Lingxue Zhu et al. "A unified statistical framework for single cell and bulk RNA sequencing data". In: *The annals of applied statistics* 12 (2018), p. 1.

[301] Wei-Chao Lin and Chih-Fong Tsai. "Missing value imputation: a review and analysis of the literature (2006–2017)". In: *Artificial Intelligence Review* 53 (2020), pp. 1487–1509.

[302] Peijie Lin, Michael Troup, and Joshua W. K. Ho. "CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data". In: *Genome biology* 18.1 (2017), pp. 1–11.

[303] Gokcen Eraslan et al. "Single-cell RNA-seq denoising using a deep count autoencoder". In: *Nature communications* 10 (2019), p. 1.

[304] Davide Risso et al. "A general and flexible method for signal extraction from single-cell RNA-seq data". In: *Nature communications* 9 (2018), p. 1.

[305] Beate Vieth et al. "powsimR: power analysis for bulk and single cell RNA-seq experiments". In: *Bioinformatics* 33.21 (2017), pp. 3486–3488.

[306] Wenhao Tang et al. "bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data". In: *Bioinformatics* 36.4 (2020), pp. 1174–1181.

[307] F. William Townes et al. "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model". In: *Genome biology* 20 (2019), pp. 1–16.

[308] Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. "Demystifying "drop-outs" in single-cell UMI data". In: *Genome biology* 21 (2020), p. 1.

[309] Abhishek Sarkar and Matthew Stephens. "Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis". In: *Nature genetics* 53.6 (2021), pp. 770–777.

[310] Christoph Hafemeister and Rahul Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome biology* 20 (2019), p. 1.

[311] Jan Lause, Philipp Berens, and Dmitry Kobak. "Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data". In: *Genome biology* 22.1 (2021), pp. 1–20.

[312] Saket Choudhary and Rahul Satija. “Comparison and evaluation of statistical error models for scRNA-seq”. In: *Genome biology* 23 (2022), p. 1.

[313] Jeremie Breda, Mihaela Zavolan, and Erik van Nimwegen. “Bayesian inference of gene expression states from single-cell RNA-seq data”. In: *Nature Biotechnology* 39.8 (2021), pp. 1008–1016.

[314] Mukund Thattai. “Universal Poisson statistics of mRNAs with complex decay pathways”. In: *Biophysical journal* 110.2 (2016), pp. 301–305.

[315] Changlin Wan et al. “LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data”. In: *Nucleic acids research* 47 (2019), p. 18.

[316] Xizhi Luo et al. “BISC: accurate inference of transcriptional bursting kinetics from single-cell transcriptomic data”. In: *Briefings in Bioinformatics* 23 (2022), p. 6.

[317] Yuchao Jiang, Nancy R. Zhang, and Mingyao Li. “SCALE: modeling allele-specific gene expression by single-cell RNA sequencing”. In: *Genome biology* 18.1 (2017), pp. 1–15.

[318] Philip Brennecke et al. “Accounting for technical noise in single-cell RNA-seq experiments”. In: *Nature methods* 10.11 (2013), pp. 1093–1095.

[319] et al. Satija Rahul. “Spatial reconstruction of single-cell gene expression data”. In: *Nature biotechnology* 5.33 (2015), pp. 495–502.

[320] Barbara Treutlein et al. “Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq”. In: *Nature* 509.7500 (2014), pp. 371–375.

[321] Alex M Ascension et al. “Triku: a feature selection method based on nearest neighbors for single-cell data”. In: *GigaScience* 11 (2022), giac017.

[322] Bobby Ranjan et al. “DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data”. In: *Nature Communications* 12.1 (2021), p. 5849.

[323] Peter V. Kharchenko. “The triumphs and limitations of computational methods for scRNA-seq”. In: *Nature Methods* 18.7 (2021), pp. 723–732.

[324] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9 (2008), p. 11.

[325] Etienne Becht et al. “Dimensionality reduction for visualizing single-cell data using UMAP”. In: *Nature biotechnology* 37.1 (2019), pp. 38–44.

[326] Dmitry Kobak and Philipp Berens. “The art of using t-SNE for single-cell transcriptomics”. In: *Nature communications* 10 (2019), p. 1.

[327] Dmitry Kobak and George C. Linderman. “Initialization is critical for preserving global data structure in both t-SNE and UMAP”. In: *Nature biotechnology* 39.2 (2021), pp. 156–157.

[328] John A. Lee, Diego H. Peluffo-Ordonez, and Michel Verleysen. “Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure”. In: *Neurocomputing* 169 (2015), pp. 246–261.

[329] Anna C. Belkina et al. “Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets”. In: *Nature communications* 10 (2019), p. 1.

[330] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome biology* 19 (2018), pp. 1–5.

[331] L. McInnes, J. Healy, and J. Melville. “Uniform manifold approximation and projection for dimension reduction”. arXiv preprint.

[332] George C. Linderman et al. “Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data”. In: *Nature methods* 16.3 (2019), pp. 243–245.

[333] Jan Niklas Bohm, Philipp Berens, and Dmitry Kobak. “Attraction-repulsion spectrum in neighbor embeddings”. In: *Journal of Machine Learning Research* 23.95 (2022), pp. 1–32.

[334] Mathieu Jacomy et al. “Forceatlas2, a continuous graph layout algorithm for handy network visualization”. In: *Medialab center of research* 560 (2011), p. 4.

[335] F. Alexander Wolf et al. “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells”. In: *Genome biology* 20 (2019), pp. 1–9.

[336] Kelly Street et al. “Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics”. In: *BMC genomics* 19 (2018), pp. 1–16.

[337] Geoffrey Schiebinger et al. “Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming”. In: *Cell* 176.4 (2019), pp. 928–943.

[338] Andreas Schlitzer et al. “Identification of cDC1-and cDC2-committed DC progenitors reveals early lineage priming at the common DC progenitor stage in the bone marrow”. In: *Nature immunology* 16.7 (2015), pp. 718–728.

[339] Laleh Haghverdi et al. “Diffusion pseudotime robustly reconstructs lineage branching”. In: *Nature methods* 13.10 (2016), pp. 845–848.

[340] Manu Setty et al. “Wishbone identifies bifurcating developmental trajectories from single-cell data”. In: *Nature biotechnology* 34.6 (2016), pp. 637–645.

[341] Yusen Ye, Lin Gao, and Shihua Zhang. “Circular Trajectory Reconstruction Uncovers Cell-Cycle Progression and Regulatory Dynamics from Single-Cell Hi-C Maps”. In: *Advanced Science* 6 (2019), p. 23.

[342] Luca Albergante et al. “Robust and scalable learning of complex intrinsic dataset geometry via ElPiGraph”. In: *Entropy* 22 (2020), p. 3.

[343] Helena Todorov et al. “TinGa: fast and flexible trajectory inference with Growing Neural Gas”. In: *Bioinformatics* 36 (2020).

[344] Xiaojie Qiu et al. “Reversed graph embedding resolves complex single-cell trajectories”. In: *Nature methods* 14.10 (2017), pp. 979–982.

[345] et al. duVerle David A. “CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data”. In: *BMC bioinformatics* 17 (2016), pp. 1–17.

[346] Tapio L”onnberg et al. “Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria”. In: *Science immunology* 2 (2017), p. 9.

[347] Wouter Saelens et al. “A comparison of single-cell trajectory inference methods”. In: *Nature biotechnology* 37.5 (2019), pp. 547–554.

[348] Wouter Saelens. “Dynverse”. Accessed on 08.06.2023. 2019. URL: `https://dynverse.org/;`.

[349] Anna Klimovskaia et al. “Poincare maps for analyzing complex hierarchies in single-cell data”. In: *Nature communications* 11 (2020), p. 1.

[350] Manu Setty et al. “Characterization of cell fate probabilities in single-cell data with Palantir”. In: *Nature biotechnology* 37.4 (2019), pp. 451–460.

[351] Shuxiong Wang et al. “Cell lineage and communication network inference via optimization for single-cell transcriptomics”. In: *Nucleic acids research* 47 (2019), p. 11.

[352] Jiangyong Wei et al. “SCOUT: A new algorithm for the inference of pseudo-time trajectory using single-cell data”. In: *Computational Biology and Chemistry* 80 (2019), pp. 111–120.

[353] Aden Forrow and Geoffrey Schiebinger. “LineageOT is a unified framework for lineage tracing and trajectory inference”. In: *Nature communications* 12 (2021), p. 1.

[354] Louise Deconinck et al. “Recent advances in trajectory inference from single-cell omics data”. In: *Current Opinion in Systems Biology* 27 (2021).

[355] Vladimir Yu Kiselev et al. “SC3: consensus clustering of single-cell RNA-seq data”. In: *Nature methods* 14.5 (2017), pp. 483–486.

[356] Dominic Grun et al. “Single-cell messenger RNA sequencing reveals rare intestinal cell types”. In: *Nature* 525.7568 (2015), pp. 251–255.

[357] Bosiljka Tasic et al. “Adult mouse cortical cell taxonomy revealed by single cell transcriptomics”. In: *Nature neuroscience* 19.2 (2016), pp. 335–346.

[358] Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. “Overlapping community detection in networks: The state-of-the-art and comparative study”. In: *Acm computing surveys* 45.4 (2013), pp. 1–35.

[359] Evan Z. Macosko et al. “Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets”. In: *Cell* 161.5 (2015), pp. 1202–1214.

[360] Santo Fortunato and Marc Barthelemy. “Resolution limit in community detection”. In: *Proceedings of the national academy of sciences 104*. 1, 2007, pp. 36–41.

[361] Giovanni Pasquini et al. “Automated methods for cell type annotation on scRNA-seq data”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 961–969.

[362] Aviv Regev et al. “The human cell atlas”. In: *elife* 6 (2017).

[363] Muzlifah Haniffa et al. “A roadmap for the human developmental cell atlas”. In: *Nature* 597.7875 (2021), pp. 196–205.

[364] Nicholas Schaum et al. “Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium”. In: *Nature* 562 (2018).

[365] Xiaoping Han et al. “Mapping the mouse cell atlas by microwell-seq”. In: *Cell* 172.5 (2018), pp. 1091–1107.

[366] Dongsheng Chen et al. “Single cell atlas for 11 non-model mammals, reptiles and birds”. In: *Nature Communications* 12 (2021), p. 1.

[367] Xinxin Zhang et al. “CellMarker: a manually curated resource of cell markers in human and mouse”. In: *Nucleic acids research* 47 (2019), pp. 721–728.

[368] Oscar Franzen, Li-Ming Gan, and Johan L. M. Bjorkegren. “PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data”. In: *Database* 2019 (2019).

[369] Huating Yuan et al. “CancerSEA: a cancer single-cell state atlas”. In: *Nucleic acids research* 47 (2019), pp. 900–908.

[370] Xin Shao et al. “scCATCH: automatic annotation on cell types of clusters from single-cell RNA sequencing data”. In: *Iscience* 23 (2020), p. 3.

[371] Ziwei Wang, Hui Ding, and Quan Zou. “Identifying cell types to interpret scRNA-seq data: how, why and more possibilities”. In: *Briefings in functional genomics* 19.4 (2020), pp. 286–291.

[372] H. Atakan Ekiz et al. “CIPR: a web-based R/shiny app and R package to annotate cell clusters in single cell RNA sequencing experiments”. In: *BMC bioinformatics* 21.1 (2020), pp. 1–15.

[373] Rui Fu et al. “clustifyr: an R package for automated single-cell RNA sequencing cluster classification”. In: *F1000Research* 9 (2020).

[374] Tallulah S. Andrews and Martin Hemberg. “M3Drop: dropout-based feature selection for scRNASeq”. In: *Bioinformatics* 35.16 (2019), pp. 2865–2867.

[375] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. "scmap: projection of single-cell RNA-seq data across data sets". In: *Nature methods* 15.5 (2018), pp. 359–362.

[376] Dvir Aran et al. "Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage". In: *Nature immunology* 20.2 (2019), pp. 163–172.

[377] Jurrian K. De Kanter et al. "CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing". In: *Nucleic acids research* 47 (2019), p. 16.

[378] Yuqi Tan and Patrick Cahan. "SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species". In: *Cell systems* 9.2 (2019), pp. 207–213.

[379] Yingxin Lin et al. "scClassify: sample size estimation and multiscale classification of cells using single and multiple reference". In: *Molecular systems biology* 16 (2020).

[380] Jose Alquicira-Hernandez et al. "scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data". In: *Genome biology* 20.1 (2019), pp. 1–17.

[381] Feiyang Ma and Matteo Pellegrini. "ACTINN: automated identification of cell types in single cell RNA sequencing". In: *Bioinformatics* 36.2 (2020), pp. 533–538.

[382] Peng Xie et al. "SuperCT: a supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles". In: *Nucleic acids research* 47 (2019), p. 8.

[383] Tamim Abdelaal et al. "A comparison of automatic cell identification methods for single-cell RNA sequencing data". In: *Genome biology* 20 (2019), pp. 1–19.

[384] Xinlei Zhao et al. "Evaluation of single-cell classifiers for single-cell RNA sequencing data sets". In: *Briefings in bioinformatics* 21.5 (2020), pp. 1581–1595.

[385] Stephanie C Hicks et al. "Missing data and technical variability in single-cell RNA-sequencing experiments". In: *Biostatistics* 19.4 (2018), pp. 562–578.

[386] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[387] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127.

[388] Laleh Haghverdi et al. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". In: *Nature biotechnology* 36.5 (2018), pp. 421–427.

[389] Malte D Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". In: *Nature methods* 19.1 (2022), pp. 41–50.

[390] Andrew Butler et al. “Integrating single-cell transcriptomic data across different conditions, technologies, and species”. In: *Nature biotechnology* 36.5 (2018), pp. 411–420.

[391] Tim Stuart et al. “Comprehensive integration of single-cell data”. In: *Cell* 177.7 (2019), pp. 1888–1902.

[392] Ilya Korsunsky et al. “Fast, sensitive and accurate integration of single-cell data with Harmony”. In: *Nature methods* 16.12 (2019), pp. 1289–1296.

[393] Joshua D Welch et al. “Single-cell multi-omic integration compares and contrasts features of brain cell identity”. In: *Cell* 177.7 (2019), pp. 1873–1887.

[394] Romain Lopez et al. “Deep generative modeling for single-cell transcriptomics”. In: *Nature methods* 15.12 (2018), pp. 1053–1058.

[395] Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. “A curated database reveals trends in single-cell transcriptomics”. In: *Database* (2020).

[396] Levi A. Garraway and William R. Sellers. “Lineage dependency and lineage-survival oncogenes in human cancer”. In: *Nature Reviews Cancer* 6.8 (2006), pp. 593–602.

[397] Siel Olbrecht et al. “High-grade serous tubo-ovarian cancer refined with single-cell RNA sequencing: specific cell subtypes influence survival and determine molecular subtype classification”. In: *Genome Medicine* 13 (2021), pp. 1–30.

[398] Jan Dohmen et al. “Identifying tumor cells at the single-cell level using machine learning”. In: *Genome Biology* 23 (2022), p. 1.

[399] Jonathan Ronen and Altuna Akalin. “netSmooth: Network-smoothing based imputation for single cell RNA-seq”. In: *F1000Research* (2017), pp. 7–8.

[400] Trinity CTAT Project. *inferCNV*. Accessed on 11.06.2023. 2023. URL: `https://github.com/broadinstitute/inferCNV;`.

[401] Jeffrey W. Tyner et al. “Functional genomic landscape of acute myeloid leukaemia”. In: *Nature* 562.7728 (2018), pp. 526–531.

7 Publication list and contributions

Publication I (Chapter 4.2)

Jan Dohmen*, *Artem Baranovskii**, Jonathan Ronen, Bora Uyar, Vedran Franke, and Altuna Akalin, Identifying tumor cells at the single-cell level using machine learning, Genome Biology 2022, 23, 123, https://doi.org/10.1186/s13059-022-02683-1

* These authors contributed equally to the work

Own contributions:

1. Devised analyses to characterise a gene signature set reported in the paper. (Figure 1 D, E; Figure 4 B, C, D; Figure 5; Figure S1, Figure S4 B).
2. Development of the gene set scoring functionality within Ikarus package.
3. Testing and debugging of the Ikarus package. Post release support on github (https://github.com/BIMSBbioinfo/ikarus).
4. Parallel analysis of the single-cell datasets with JD. Results reported on Figure 2 C, D, E; Figure S3 A, B, C, E, F.
5. Statistical testing of the reported findings.
6. Writing of the result sections for gene set characterisation, related methods, and proof-reading of the manuscript.

Publication II (Chapter 4.2)

*Artem Baranovskii**, Irem Gündüz*, Vedran Franke, Bora Uyar & Altuna Akalin, Multi-Omics Alleviates the Limitations of Panel Sequencing for Cancer Drug Response Prediction, Cancers 2022, 14(22), 5604,
https://doi.org/10.3390/cancers14225604
* These authors contributed equally to the work

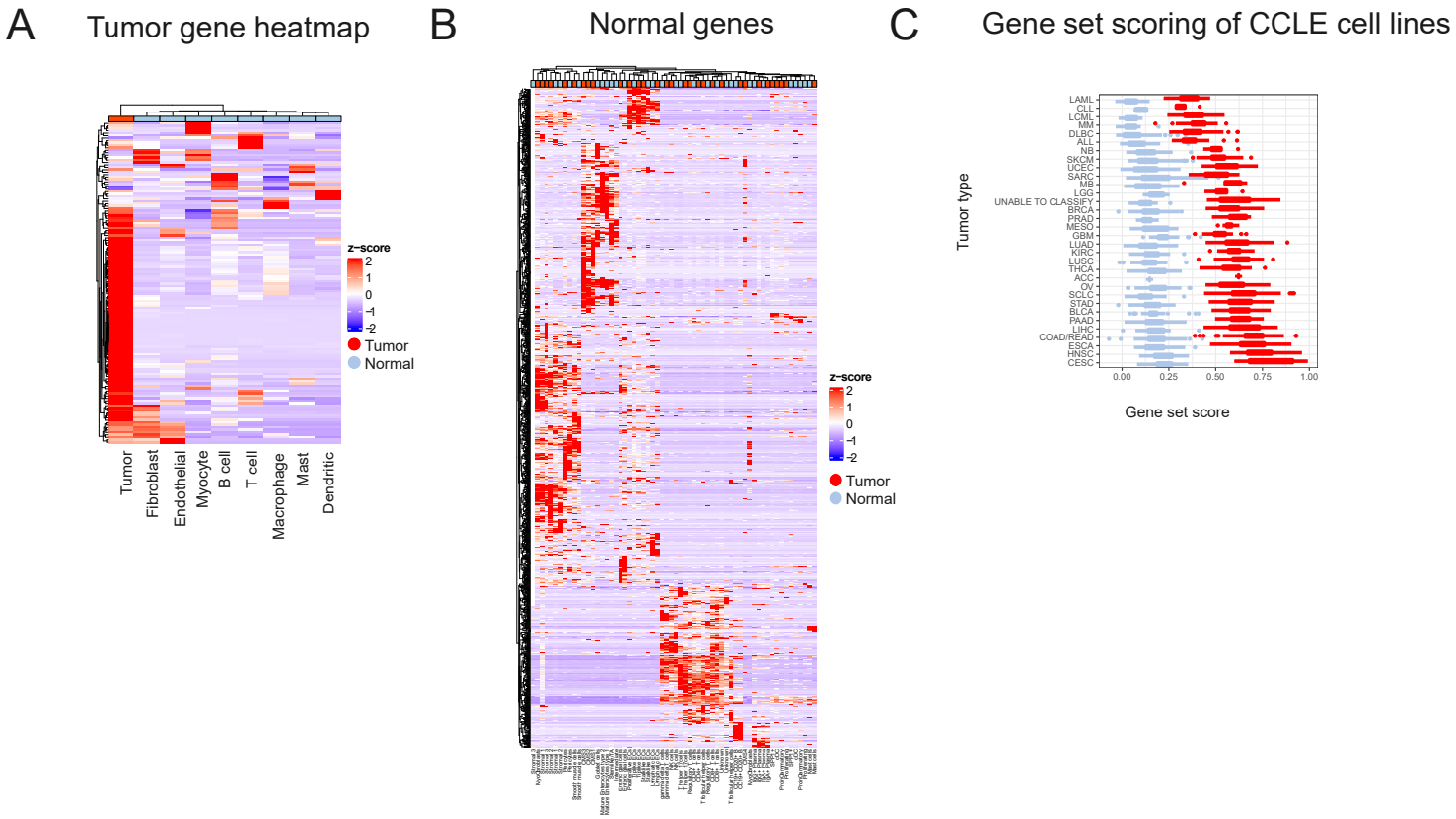
Own contributions:

1. Data collection and preparation (CCLE and PDX datasets).
2. Prediction of drug responses in CCLE and PDX datasets in Multi-omics and Panel sequencing modalities (Figure 1 A, B, C, Figure S1).
3. Analysis of effects drug classes have on drug response prediction improvement in Multi-omics (Figure 1 D; Figure S2).
4. Joint writing of the manuscript (Simple Summary, Introduction, Results, Discussion and Methods sections).

8 Appendix - Extended data

8.1 Appendix I - Extended data for Publication I

Fig S1.



Tumor and normal gene signature characterization

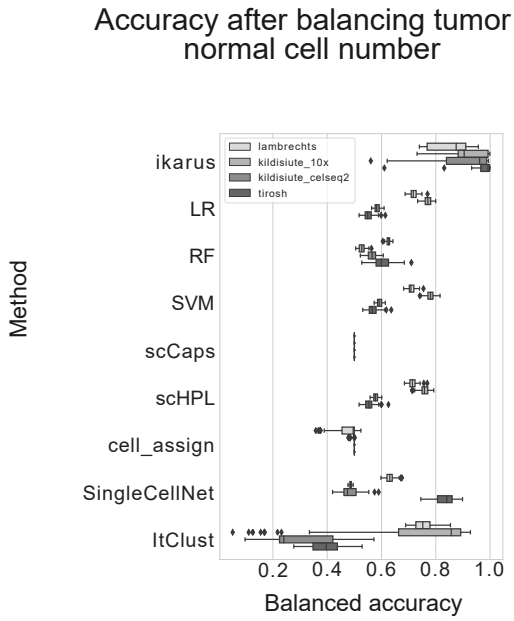
A) The heat map depicts the expression of the tumor specific genes in the Tirosh head and neck cancer dataset [31], which was not used for the signature definition. Out of the 162 genes from the signature, 132 were found to be expressed in the Tirosh dataset. The tumor gene signature contains two sets of genes: genes that are highly enriched in tumor cells compared to all other cells and genes that are highly enriched in tumor cells compared to each individual cell type.

B) The normal gene signature list visualized on the Tirosh dataset (which was not used in the gene list definition). Normal gene signature captures mostly cell type specific gene expression. The tumor/normal classification designates whether the cells originated from the tumor or the healthy dataset.

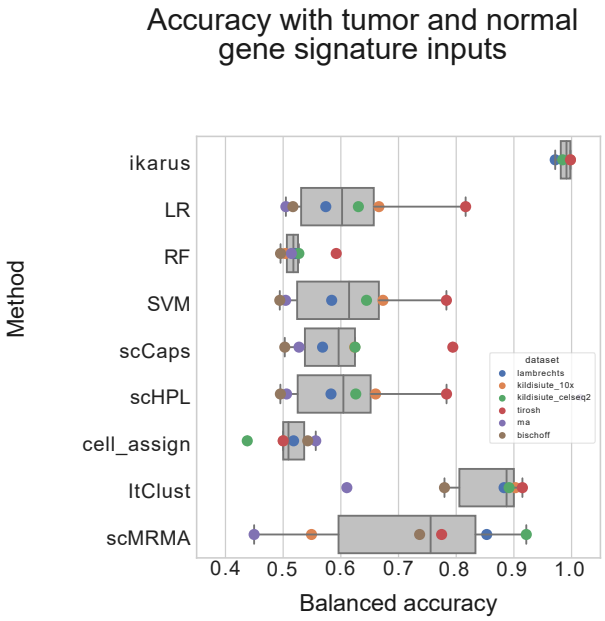
C) Tumor and normal gene signature scores of the cancer cell line encyclopedia (CCLE) data. Tumor gene signature shows a significantly higher score distribution in all cancer types present in the dataset. Normal and tumor signature distributions were compared using Wilcoxon tests, for each cancer type, followed by BH-FDR correction. All adjusted p values were lower than 0.01.

Fig S2.

A



B



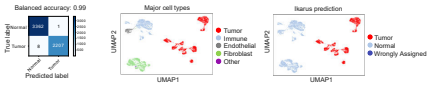
Additional tests of ikarus performance

A) Performance of ikarus and competing classifiers on datasets where the tumor and normal classes have been balanced by sampling. The sampling procedure was repeated 100 times (Distributions of results were compared using ANOVA with post hoc pairwise comparison. P values were adjusted using BH-FDR. All adjusted p values were lower than 0.01).

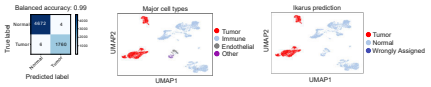
B) Performance of ikarus and competing classifiers when using tumor and normal gene signatures as inputs, instead of all genes. (Distributions of results were compared using ANOVA with post hoc pairwise comparison. P values were adjusted using BH-FDR. All adjusted p values were lower than 0.05).

Fig S3.

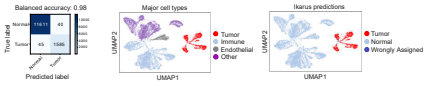
A Tirosh et al. Head and Neck Cancer



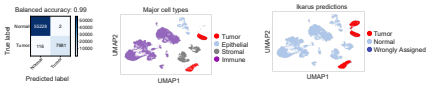
B Kildisiute et al. Neuroblastoma 10x



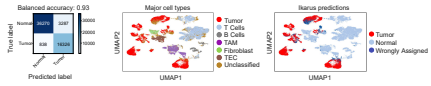
C Kildisiute et al. Neuroblastoma celseq2



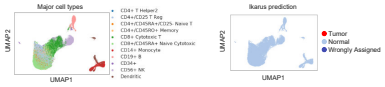
D Bischoff et al. Lung carcinoid 10x



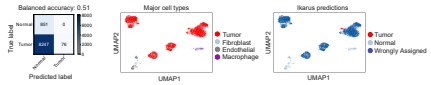
E Ma et al. Hepatocellular carcinoma 10x



F Zheng et al. PBMC 10x



G Livnat et al. Sarcoma 10x



ikarus performance on multiple test datasets

A) Performance of ikarus classifier on the Puram et al. Head and Neck Cancer dataset.

B) Performance of ikarus classifier on the Kildisiute et al. Neuroblastoma dataset sequenced with 10X.

C) Performance of ikarus classifier on the Kildisiute et al. Neuroblastoma sequenced with CEL-Seq2.

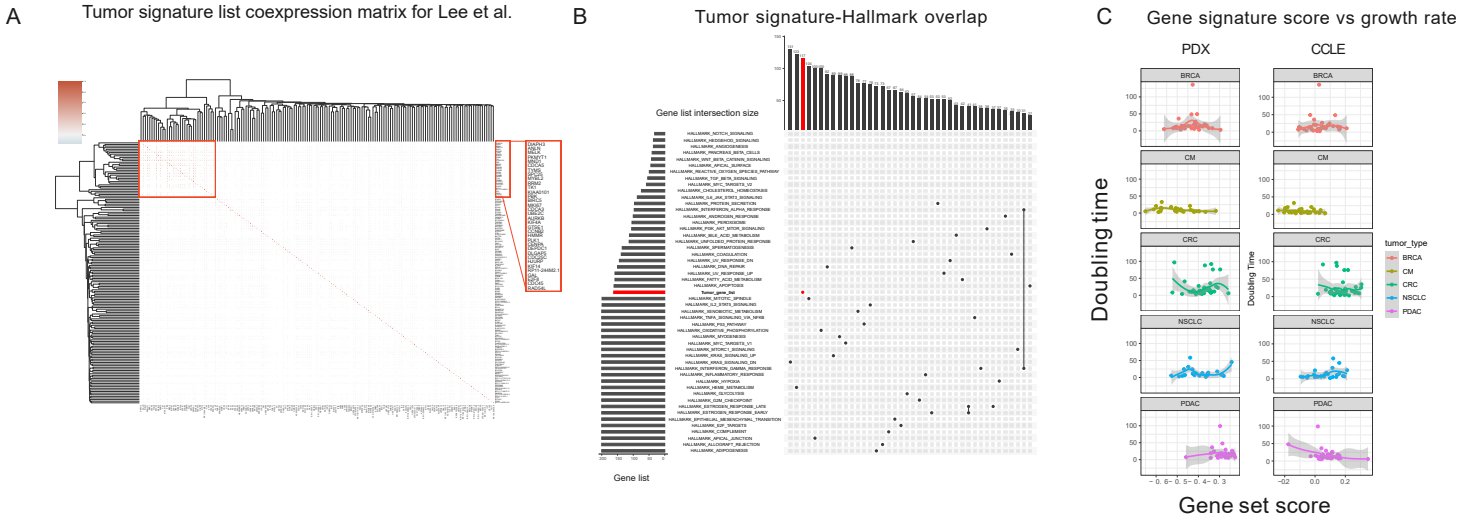
D) Performance of ikarus classifier on the Bischoff et al. Lung carcinoid sequenced with 10Xx.

E) Performance of ikarus classifier on the Ma et al. Hepatocellular carcinoma sequenced with 10Xx.

F) ikarus correctly recognizes all cells from a healthy peripheral blood as non-tumorous.

G) ikarus shows a reduction in sensitivity when discrimination tumor form normal cell in a sarcoma sample [46].

Fig S4.



Additional annotation of the tumor gene signature

A) Correlation analysis of the tumor gene signature from Lee et al. Genes belonging to the cell cycle module are marked with a red rectangle. The comprehensive list of gene names can be found in the Table S1 - Gene signatures.

B) Tumor gene signature shows limited overlap with most of the hallmark gene sets. Only intersections of size 27 and more are shown.

C) Relationship between the tumor gene signature scores and the growth rate of PDX and CCLE models. There is limited correlation between the growth rate and the gene signatures.

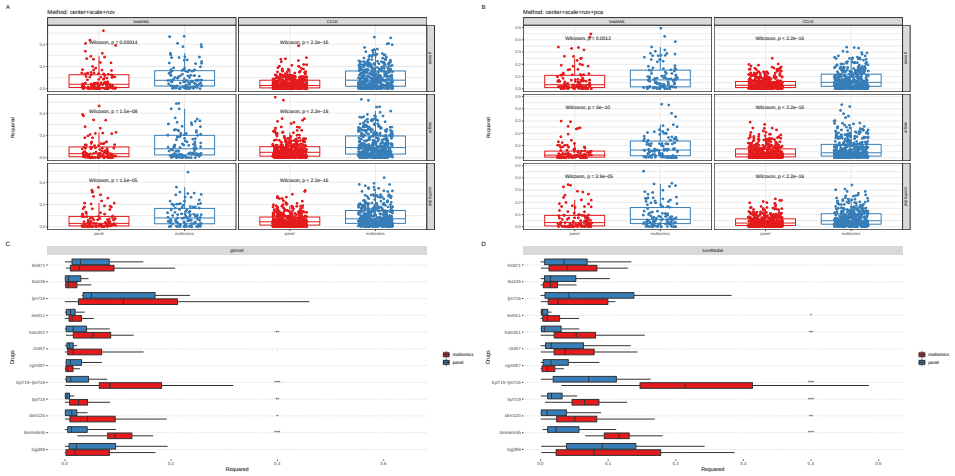


Figure 3: Figure S1. Performance comparisons of different drug response prediction models trained by using only panel-seq features (mutations and/or copy number variations) or using transcriptome features in combination with panel-seq features (multi-omics) using different pre-processing options (with PCA in A) and without PCA (in B) and using different machine learning methods: random forests, elastic nets, and support vector machines. (A) Comparisons of different drug response models trained with preprocessed panel-seq and multi-omics features of beatAML and CCLE datasets using three different methods with scaling, entering, and near-zero variance filtering. (B) Comparisons of drug response models, trained with preprocessed (scaled/centered/filtered for near-zero-variation) and dimensionally reduced (using PCA) panel-seq and multi-omics features of beatAML and CCLE datasets. (C) Multi-omics (red) improvements (in terms of R-squared metric) compared to panel-seq features (blue) of the test section of the 12-drug PDX dataset, using the elastic net regression (glmnet) model. Stars above the boxplots represent significance levels: *** for p ¡ 0.001, ** for p ¡ 0.01, * for p ¡ 0.05. (D) Multi-omics (red) improvements (in terms of R-squared metric) compared to panel-seq features (blue) of the test section of the 12-drug PDX dataset, using the radial support vector machine (svmRadial) model. Stars above the boxplots represent significance levels: *** for p ¡ 0.001, ** for p ¡ 0.01, * for p ¡ 0.05.

8.2 Appendix II - Extended data for Publication II

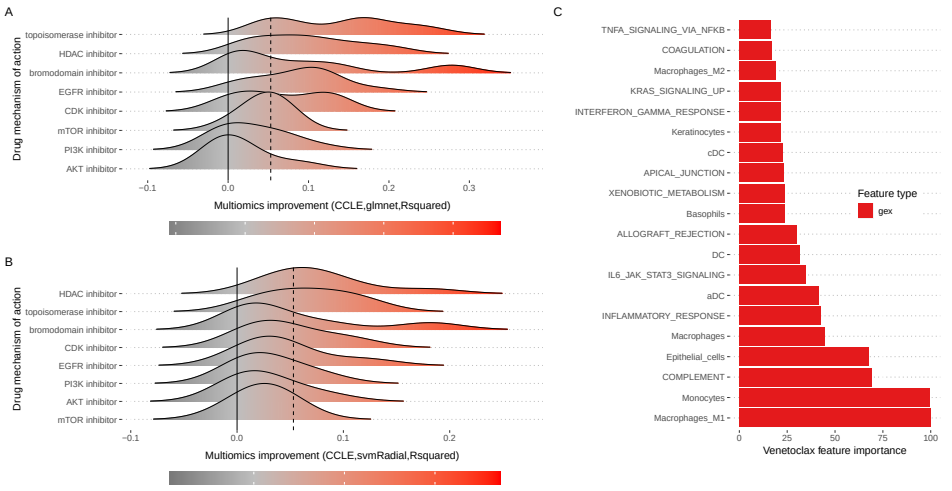


Figure 4: Figure S2. (A) Classes of drugs based on the average improvement in multi-omics over panel-seq when the logistic regression (glmnet) model was used for drug response prediction. (B) Classes of drugs based on the average improvement in multi-omics over panel-seq when the radial support vector machine (svmRadial) model was used for drug response prediction. Mean improvement on overall drugs marked with dashes. (C) Top 20 cell type and cancer hallmark gene signatures associated with Venetoclax response prediction for beatAML samples using a random forest model. Table S1: Drug response prediction performance metrics for each machine learning method and pharmacogenomics dataset. Table S2: Feature importance metrics derived from each machine learning model built for each drug in each pharmacogenomics dataset.