CHAPTER

3

Machine learning applications in cancer genomics

Omar El-Charif[a], *Russell Schwartz*[b], *Ye Yuan*[c] and *Yifeng Tao*[d]

[a]Department of Medicine, Section of Hematology/Oncology, The University of Chicago, Chicago, IL, United States; [b]Computational Biology Department and Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, United States; [c]Department of Radiation Oncology, NYU Langone, New York, NY, United States; [d]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States

KEY POINTS

- Common genomic technologies include microarrays, RNAseq, DNAseq, and sequencing technologies for characterizing epigenetic and regulatory status.
- Principal uses of machine learning (ML) in cancer genomics to date have included tumor subtyping, driver gene/mutation discovery, biomarker identification, and pharmacogenomics.
- Major challenges for applying ML to cancer genomics include difficulties of data acquisition, data sparsity, inter- and intratumor heterogeneity, and validation of ML-derived biomarkers.
- Important current issues for the field include development of methods for clinical application of whole-genome sequencing and single-cell sequencing, interpretable ML for oncology, and increased training in genomics and ML for healthcare professionals.

Introduction

Thirty-two years separated the development of the first DNA sequencing technology using primer extension in 1971 and the sequencing of all 3 billion nucleotides in the human genome

Machine Learning and Artificial Intelligence in Radiation Oncology
https://doi.org/10.1016/B978-0-12-822000-9.00006-9

as part of the Human Genome Project in 2001 (Lander et al., 2001; Venter et al., 2001), completed at an estimated cost of 2.7 billion US dollars. Less than 20 years later, the estimated cost of whole-genome sequencing (WGS) crossed the milestone of <$1000 per genome (Nakagawa & Fujita, 2018), with some startups offering direct-to-consumer WGS for $299 in 2020 (https://nebula.org/whole-genome-sequencing/). The rate of decline in cost has been driven by tremendous advances in next-generation sequencing (NGS) technologies, some of which have been repurposed to interrogate genome-wide gene expression (transcriptomics), methylation and chromatin status (epigenomics), splice variants, and more (Lappalainen et al., 2019). The increased availability of sequencing technologies has led to a plethora of multi-omic data in clinical research. The complexity of these data is compounded by high intra-tumor (McGranahan & Swanton, 2015) and inter-tumor (Kornelia, 2007) heterogeneity requiring advanced computational analysis and data mining tools to identify reproducible and clinically actionable patterns in the presence of multiple hypothesis testing on the order of 10^6–10^9 per phenotype. This challenge is being addressed by bringing together biological/clinical science with the computational sciences. Advances in machine learning (ML) have shown particular promise to address many of the limitations of more conventional statistical analyses by facilitating the identification of sparse signals in large, noisy data and predicting outcomes potentially with no a priori hypotheses or assumptions about data distribution (Leung et al., 2016).

Oncology has thus far been at the forefront of the genomic revolution, with large-scale projects identifying both germline and somatic variants and transcriptomic signatures to predict cancer risk (van't Veer et al., 2002, 2003), subtype histologically similar but clinically distinct cancers (Dawson et al., 2013; Nielsen et al., 2010; Parker et al., 2009), predict tumor response to therapy (Sicklick et al., 2019), identify driving mutations (Lawrence et al., 2014) and pathways (Vandin et al., 2012; Vogelstein et al., 2013), and nominate novel therapeutic targets (Goldman & Melo, 2003; Paez et al., 2004). This effort has been facilitated by large-scale community efforts at data generation, including The Cancer Genome Atlas (Weinstein et al., 2013) and the International Cancer Genome Consortium (Hudson et al., 2010) among others. Aside from tumor classification and characteristics, oncology has also been a driving force in the field of pharmacogenomics (Evans & Relling, 1999; Relling & Evans, 2015), aimed at explaining and exploiting inter-patient variability in drug response by interrogating, for example, variants influencing the function of enzymes and transporters in pharmacokinetic pathways. Around 40% of the ~400 drug-gene pairs on FDA labels are for oncological drugs, most of which were discovered using candidate gene approaches and conventional statistical methods prior to genome-wide interrogation and ML techniques. Furthermore, most are rarely used routinely in clinical settings, highlighting what is arguably the most significant challenge facing the field — translating cancer research into clinical implementation.

The promise of integrating ML with large scale multi-omic databases is only starting to come to fruition as costs of data collection decline and the baton is passed to data analytics. Today, the cost of sequencing, at least of bulk tumors, is no longer a significant obstacle to making precision genomic medicine a routine part of cancer treatment, but the computational tools and expertise to make use of these data in the clinic are still lacking. In this chapter, we will give a brief overview of sequencing technologies and discuss the promises of ML in deriving clinical utility from these data. We will also review some current applications and successful examples of clinical biomarkers and conclude by addressing some of the hurdles that remain as well as future directions.

Overview of genomic technologies

The study of the molecular basis of cancers has been revolutionized by the rise of genomic technologies, which has made it possible to observe genetic and genomic variations that cause cancer in unprecedented scale and detail (Lappalainen et al., 2019). From early methods that first made it possible to study bulk tumors at a whole-genome scale (DeRisi et al., 1996) to modern innovations for profiling large numbers of single cells (Navin et al., 2011; Suvà & Tirosh, 2019), new genomic technologies have found some of their most prominent uses in cancer research. Changes in genomic technologies have gone hand-in-hand with advances in our understanding of cancer and our increasing appreciation for the important roles of intratumor heterogeneity, biological systems and networks, and complex somatic mutation processes in understanding how cancer develops and progresses and how it might better be treated and monitored. In this section, we survey some of the genomic technologies that have been most influential in driving new directions in cancer treatment and research.

Microarrays

One of the basic data types for characterizing tumor genomics is gene expression data, which quantifies the number of transcripts present for different genes or gene isoforms in a sample. Microarrays proved the first technology for profiling gene expression on a scale of whole-genomes and large patient cohorts. While sequencing-based methods for gene expression profiling preceded microarrays (Adams et al., 1991; Velculescu et al., 1995), they were initially far too costly to adapt at large scale. Microarrays were developed as a platform enabling more efficient profiling of activities for potentially large numbers of genes or transcripts simultaneously (DeRisi et al., 1996). The basic technology consists of a slide decorated with a series of "spots" each defined by a DNA sequence and used to capture complementary sequences from a genomic sample. By measuring relative intensities of fluorescent probes attached to these complementary sequences, usually in comparison between a sample and a control, a microarray makes it possible to estimate fold changes in gene expression levels. While most commonly used for protein-coding genes, microarrays could also be used to profile microRNA expression or expression of other non-coding sequences (Nelson et al., 2004). However, this technology had some important disadvantages, such as a high noise level, an ability to profile only previously known sequences, and a limitation to measuring relative rather than absolute expression levels (Tinker et al., 2006). Nonetheless, its excellent scalability, eventually allowing on the order of a million parallel assays per microarray at low cost, made it a crucial technology for bringing whole-genome analysis into widespread use.

Microarrays found numerous important applications in early cancer genomic studies. Microarrays were crucial to early whole-genome expression profiling of tumors, and in the process led to the development of expression-signature-based definitions of tumor subtypes (Perou et al., 2000) that helped to reveal how clinically similar tumors might arise from very different molecular mechanisms and how this recognition could lead to improved prognostic prediction (van't Veer et al., 2002, 2003). They also facilitated systems biology approaches to identifying networks and pathways implicated in particular tumors and thus identify potential therapeutic targets. While subtyping is continually being refined, more sophisticated

variants of the subtyping arising from the early microarray-derived models remain in common use today.

While microarrays were most widely employed as a technology for expression profiling, they also found important use as a means of profiling DNA variations. "SNP chips" provided an early way to efficiently test for many known genetic variants in parallel (Gunderson et al., 2005). SNP arrays can also be used to perform a coarse-grained form of copy number alteration (CNA) analysis (Dumur et al., 2003). Copy number arrays of this kind were also influential in developing some of the first efforts at profiling subclonal variation in cancers through regional profiling of bulk copy number variation (Navin et al., 2010).

While microarrays are still used for RNA expression profiling, SNP genotyping, and DNA copy number analysis, they have been largely displaced in practice by sequencing technologies described below, as well as by more specialized alternatives not covered here, such as the Nanostring platform (Kulkarni, 2011), that may allow different tradeoffs of accuracy, scale, and cost.

RNA-seq

Expression microarrays have been largely superseded in practice by RNA sequencing (RNA-seq) as a technology for profiling gene expression. In RNA-seq, one isolates RNA from a sample and derives the sequences of large numbers of putatively randomly-selected fragments from the resulting RNA transcripts, which can then be mapped back to a genome to infer expression levels at gene, transcript, exon, or even single-base resolutions (Mortazavi et al., 2008). Variants of RNA-seq have been available essentially since automated Sanger sequencing technology first came into in widespread use (Hunkapiller et al., 1991), for example in the use of expressed sequence tags (ESTs) in many early gene expression studies (Adams et al., 1991), but they were long cost-prohibitive for widespread use, particularly for large cohorts. "Next-generation" sequencing technologies (Metzker, 2010) combined with new computer algorithms for interpreting much larger volumes of genomic data (Li & Durbin, 2009) greatly reduced the cost of RNA-seq, and these costs have continued to fall as technologies have advanced. While the data from RNA-seq is typically used similarly to that from microarrays, RNA-seq offers a number of technical advantages that led to its widespread adoption once costs became manageable. These advantages include greater accuracy, absolute rather than relative quantification, finer resolution, and ability to perform novel transcript discovery.

RNA-seq has been widely adopted for cancer genomics, notably as one of the standard technologies for some of the largest community sequencing efforts to date, such as The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and the International Cancer Genome Sequencing Consortium (ICGC) (Hudson et al., 2010).

More recently, cancer genomics has been greatly influenced by the advent of single-cell RNA-sequencing platforms. Similar to bulk RNA-seq, single-cell RNA-seq (scRNA-seq) was quickly appreciated for its value to cancer genomics and drove seminal studies into subclonal heterogeneity in expression programs (Patel et al., 2014), mechanisms of expression variation implicated in metastasis (Tirosh et al., 2016), immune infiltration (Chung et al., 2017; Zheng et al., 2017), and other forms of stromal contamination. As technologies have matured, its advantages over bulk sequencing have led to widespread adoption. While large

public repositories of single-cell cancer genomic data are not yet available on the scales of the TCGA or ICGC, scRNA-seq has become a core technology for other large efforts to profile cell-to-cell expression variation more systematically (HuBMAP Consortium, 2019). Other innovative uses of scRNA-seq with a particular value for cancer genomics include various technologies for spatial transcriptomics, in which one profiles expression mapped to spatial regions in tumor or stroma in situ (Ståhl et al., 2016).

DNA-seq

DNA sequencing (DNA-seq) has also become a key tool for cancer genomic studies. Genome sequencing has been a crucial enabling technology for all modern work in cancer genomics, making it possible to identify cancer-associated genes from genome-wide studies and map them reliably to specific genomic regions. This has facilitated discovery and typing of genetic variations genome-wide, including structural variations implicated in many cancers, and provided reference genomes typically used in interpreting RNA-seq and other genomic data sources. As DNA-seq became cost-effective to perform at large scale, it has proven particularly valuable for cancer genomics, where it has enabled the systematic profiling of the highly idiosyncratic somatic mutation burden characteristic of most cancers. DNA-seq can be used to identify and quantify many variant types, including single nucleotide variants (SNVs), copy number alterations (CNAs), and various kinds of structural variations (SVs), making it far more versatile than early array methods for typing somatic variation. DNA-seq, has therefore become a standard part of large cancer genome sequencing efforts (Hudson et al., 2010; Weinstein et al., 2013) and is now finding important uses in clinical practice (Gagan & Van Allen, 2015).

DNA-seq comes in many variants, often with tradeoffs in practice. Most cancer genomic studies and clinical uses of DNA-seq are still limited to whole-exome sequence (WES) (Weinstein et al., 2013), or to more limited targeted sequencing (Bybee et al., 2011). Recent studies using whole-genome sequencing (WGS) have, however, provided a great deal of insight into structural and non-coding variations unavailable to WES (Li et al., 2020; The International Cancer Genome Consortium et al., 2020), suggesting the value of transitioning to WGS as it becomes more cost-effective. DNA-seq often incorporates mate paired or paired-end sequence, which allows for a more complete discovery of SVs. Another alternative is long-read technologies that produce potentially much longer sequences from single chromosomes than prior next-generation sequencing methods. Prominent long read technologies include platforms offered by Pacific Biosciences and Oxford Nanopore (English et al., 2012; Jain et al., 2018), which likewise have advantages for structural variant discovery and for better "phasing" mutations to discover how different clones are related to one another. Single-cell DNA-seq (scDNA-seq) has also come into widespread use in basic research into cancer genomics due to its value in studying clonal heterogeneity and for reconstructing clonal evolution (Navin et al., 2011). scDNA-seq remains considerably more costly and technically challenging than scRNA-seq, however, and is thus in more limited use. Another class of special sequencing technology of particular relevance to cancers are so-called "liquid biopsy" methods, which provide ways to diagnose solid tumors and track their progression from non-invasive blood draws (Crowley et al., 2013). A variety of related technologies for liquid biopsy currently exist, including techniques based on the isolation of circulating tumor cells

(CTCs) shed by tumors (Alix-Panabières & Pantel, 2013), small fragments of circulating free DNA (cfDNA) released typically during tumor cell death (Diaz & Bardelli, 2014), and isolating DNA from extracellular vesicles (EVs) that may be released by living tumor cells in situ (Contreras-Naranjo et al., 2017), with each offering some unique advantages (Zhang et al., 2017). Liquid biopsy is now also finding clinical use through commercially available platforms such as Guardant360 (Guardant Health, Inc.) and PlasmaSELECT (Personal Genome Diagnostics, Inc.).

Epigenetic and regulatory sequencing

Sequencing technologies have also found widespread use in cancer genomics as a way to profile epigenetic markers and protein-DNA interactions that influence gene activity and can underlie functional variation in cancers. DNA methylation has proven an important marker of genetic regulation that is frequently perturbed in cancers and can provide markers of tumor progression and prognostic power for outcomes (Das & Singal, 2004). A next-generation technology, bisulfite sequencing, first made it practical to profile methylation at a whole-genome scale (Meissner et al., 2005). The Oxford Nanopore technology has recently brought methylation typing to long-read sequencing (Simpson et al., 2017). Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) provides a different form of epigenetic information, allowing genome-wide profiling of accessible chromatin (Buenrostro et al., 2015), which likewise can identify regulatory phenotypes frequently perturbed in cancers. Chromatin immunoprecipitation sequencing (ChIP-seq), a sequence-based variant of the earlier microarray-based "ChIP-chip" method, provides yet another view of regulatory interactions by identifying regions of DNA to which proteins are bound (Axel et al., 2009; Ren et al., 2000), which again often show variation in cancers. Chromatin conformation capture (CCC) (Dekker et al., 2002), a technique for profiling three-dimensional chromatin structure, has given rise to a diverse array of related technologies that have come into use in cancer genomics because chromatin structure is frequently perturbed in cancers and may suggest likely sites of structural variation (Fudenberg et al., 2011).

Applications of genomics in oncology

Tumor subtyping

Somatic variants are playing an ever-growing role in guiding cancer management, due to more granular tumor characterization beyond TNM staging and chromosomal analysis, but also due to the success of mutation-targeting therapies. Among the earliest prototypes of targeted therapies is the anti-EGFR drug gefitinib, designed after years of in vitro and in vivo experiments demonstrating the role of EGFR in oncogenesis (Arteaga, 2003; Mendelsohn & Baselga, 2000; Woodburn, 1999). Only after FDA approval, an increased awareness developed around heterogeneity in response to gefitinib in NSCLC patients, leading to retrospective (Lynch et al., 2004) followed by prospective (Fukui et al., 2008) analyses identifying common mutations in EGFR underlying good response. This initiated a paradigm shift in clinical trial methodology whereby mutational analysis guides study design

(Russo et al., 2015). We discuss this example because it provides two lessons: the well-established lesson of the promise of genomic characterization of tumors, and more importantly, the tediousness of this decades-long process from bench to bedside. More recently, Big Data initiatives have introduced a new paradigm, shifting oncological research toward high throughput, promising to create cheaper and more productive methods to reach personalized oncology. One such example noted above is The Cancer Genome Atlas (TCGA), which characterized genomic and transcriptomic profiles for thousands of tumors along with clinical data and pathology reports (Hutter & Zenklusen, 2018; Weinstein et al., 2013).

With millions of data points in each sample, the need to reduce dimensionality, identify relevant signals, and build understanding beyond single genes and pathways emerged. ML has been utilized for this purpose at various stages in the TCGA research pipeline. For example, to facilitate somatic variant identification, D.E. Wood et al. simulated tumor exomes by sequencing normal peripheral samples and engineering mutations to train Cerebro, an extremely randomized trees classifier, a variant of random forest classifiers (Geurts et al., 2006). They subsequently established the sensitivities and specificities of Cerebro and multiple variant calling algorithms in this simulated data. These investigators then applied their trained mutation-calling algorithm to TCGA and compared the mutation-calling accuracy to the established standard, the PanCanAtlas consensus calling method. They showed high levels of concordance and a statistically significant association between quality of PanCancer-Atlas calls in TCGA and concordance rates with Cerebro.

Aside from variant calling, ML has been used to identify shared latent characteristics of tumors using genome-wide data. Work by Malta et al. (2018) trained a penalized one-class logistic regression ML algorithm to identify patterns consistent with "stemness" using RNA-seq and genome-wide methylation data from cell lines ranging from high stemness (embryonic stem cell lines or induced pluripotent stem cells) to low stemness (their differentiated endodermal, mesodermal, or ectodermal progenies). The trained algorithm was then used to identify this "stemness" characteristic in the TCGA tumor RNAseq and methylation data. As expected, it was found that the stemness index was more prominent in metastatic cancer, but this research also unexpectedly discovered an associated immune checkpoint expression phenotype, possibly suggesting a stronger immune response to de-differentiated phenotypes.

Dawson et al. (2013) integrated the molecular profiles of both copy number variations and RNA expressions to group breast cancer patients into 10 distinct clusters, with different clinical prognoses and providing insights on the potential drivers.

While the problems illustrated by these examples continue to be at the forefront of computational research in oncology, they at least provide proof-of-concept demonstrations for the potential that Big Data and ML offer to interpret the complete multi-omic profile of a patient's tumor in the context of databases characterizing clinically pertinent features such as similarity to tumors from other sites and probability of treatment response.

Driver mutation discovery

Although a single tumor may contain hundreds to thousands of somatic alterations, there is a general consensus that not all mutations contribute equally to the initialization and progression of the tumor. Only a small amount of the somatic alterations arising either early or at

critical time points during tumor development contribute to tumor progression and selective advantages over the neighboring cell clones. These are named "driver" mutations/genes (Futreal et al., 2004). Meanwhile, "passenger" mutations happen along with the driver mutations but do not contribute to the tumor development. Although initial research focused on finding "driver genes" (Greenman et al., 2007; Sjöblom et al., 2006), more recent research has focused on the non-synonymous "driver mutations". Researchers identify cancer driver mutations through various strategies: based on mutation frequency and location, through causal inference or ensembled methods of multiple tools.

Mutation-rate-based methods are built upon the assumption that driver mutations should have higher recurrences than the background mutation rate. MuSiC is a package suite that contains seven modules, where the core module estimates the background mutation rate and utilizes three statistical tests to distinguish the significantly mutated genes from others (Dees et al., 2012). MutSigCV aims to reduce the false positive significant genes, such as the olfactory receptor genes and muscle protein titin genes, by accurately estimating the mutational heterogeneity in cancer (Lawrence et al., 2013). It stratifies the statistical significance by taking into account the heterogeneity across patients within a specific cancer type, mutational signatures of tumors, and regional differences across the genome. However, low recurrence mutated drivers are prone to be neglected in the mutation rate based methods.

Mutation location-based approaches assume that mutations present in conserved locations of corresponding proteins are more likely to lead to defective proteins that cannot perform their biological functions properly. CHASM used a random forest classifier to distinguish driver mutations from missense passengers based on a set of features, including contextual information of protein structure (Carter et al., 2009). Reva et al. introduced a functional impact score (FIS) to identify the amino acid residue alterations that happened in the evolutionary conserved regions (Reva et al., 2011). Similarly, Gonzalez-Perez and Lopez-Bigas proposed the metric functional impact (FI) (Gonzalez-Perez & Lopez-Bigas, 2012) based on a few tools estimating the functional effect of amino acid substitutions, such as PolyPhen-2 (Adzhubei et al., 2013) and SIFT (Kumar et al., 2009). HotSpot3D further discovered the functional mutations with the guide of 3D protein structure (Niu et al., 2016).

Instead of inference of drivers in an unsupervised way, researchers have proposed methods based on causal inference, i.e., based on the analysis of a large pool of cancer samples, to infer which somatic mutations contribute to the change of downstream phenotypes, such as transcriptome or protein expression (Cai et al., 2019; Wang et al., 2018).

There exist tools that combine an ensemble of various tools or pipelines to systematically identify the potential drivers followed by potential manual curation of the driver list. For example, IntOGen combined results from both mutation-rate-based and mutation-location-based algorithms (Abel Gonzalez-Perez et al., 2013). Bailey et al. conducted a comprehensive aggregation of cancer drivers through the integration of 26 driver discovery tools (Bailey et al., 2018).

Biomarkers of outcome in clinical practice

Treatment strategies in modern oncology have evolved from one size fits all regimens to complex multimodality and increasingly personalized therapies. The adoption of breast conserving therapy over radical mastectomy in early-stage breast cancer is a classic example

of how the use of clinical risk factors led to more personalized treatment regimens that decreased overtreatment and treatment-associated morbidity (Kroman et al., 2004). In the decades since, we have developed a better understanding of the biological pathways involved in cancer progression leading to the integration of histopathologic and molecular biomarkers in clinical risk-stratification. The clinical utilization of molecular biomarkers is poised to increase, especially as the development of next-generation high-throughput sequencing technologies now allows not only the ability to interrogate the expression levels of tens of thousands of genes but also to uncover novel driver mutations and epigenetic changes. From a clinician's viewpoint, there is a pressing need for better prognostic, or better yet, predictive/prescriptive biomarkers that can help patients and clinicians make more personalized treatment decisions. Several validated genomic biomarkers, including Oncotype DX, DCISionRT, MammaPrint, and PAM50 are currently being used in the clinic as decision support tools. In this section, we will review the development, training, validation, and interpretation of several of the most commonly used biomarker tests.

Breast cancer is a biologically heterogeneous disease and risk-stratification of patients using clinical factors is insufficient. Expression profiling can separate breast cancer into four major intrinsic molecular subtypes (luminal A, luminal B, basal-like/triple negative, and HER2-enriched) that have differential clinical outcomes. Integrating these intrinsic subtypes with clinical risk factors led to the development of the PAM50 biomarker. The PAM50 test was developed to both classify patient breast cancer samples as one of the four major intrinsic subtypes and to stratify patients by their risk of breast cancer relapse utilizing the expression levels of 50 selected genes and 5 control genes (Parker et al., 2009). The PAM50 intrinsic subtype classifier was trained on a dataset of 189 breast cancer and 29 normal samples taken from a heterogeneous cohort of patients with node-negative breast cancer who received no adjuvant chemotherapy or endocrine therapy. Gene expression was derived from a mix of qRT-PCR and microarrays. For subtype classification, the authors performed hierarchical clustering of the expression levels of 1906 genes and were able to identify clusters representative of the five intrinsic breast cancer subtypes in the majority of the breast cancer samples. These 1906 genes were further minimized to a 50 gene set that represented the top 10 most significant genes per subtype. Using this reduced geneset, the authors then trained a Ridge-penalized multivariable Cox model on a cohort of untreated node-negative patients to obtain a Risk of Recurrence score (ROR) that estimates the probability of relapse within 5 years. The actual numeric ROR score for each patient is derived from the sum of the coefficients of the Cox model. In order to stratify patients into discrete risk groups using the continuous ROR score, thresholds were chosen that resulted in no patients from the training set with luminal A subtypes in the high-risk group and no basal-like subtypes in the low-risk group.

The prognostic performance of the PAM50 ROR score in estimating distant recurrence risk has been validated in secondary analyses of large, randomized prospective trials. The first validation study was performed in a cohort of 786 women with non-metastatic breast cancer treated in British Columbia (Nielsen et al., 2010). Of note, in contrast to the patient characteristics of the studies used to develop PAM50, the majority (71%) of patients in this validation study had node-positive disease and all patients received adjuvant endocrine therapy with 5 years of tamoxifen. Despite these differences, both classifications by intrinsic subtype and the PAM50 ROR score provided additional prognostic information regarding 5-year disease-specific survival when compared to risk stratification by clinicopathologic variables alone.

Another more recent validation study of the prognostic performance of the PAM50 test was accomplished for a cohort of 1620 hormone-receptor positive post-menopausal breast cancer patients treated with tamoxifen alone or tamoxifen and anastrozole enrolled on the ABCSG-8 trial (Gnant et al., 2014). The authors found that the addition of both the continuous ROR score or the discrete risk groups significantly improved the prognostic performance of risk stratification by clinicopathological variables alone. One study compared the PAM50 ROR score to the Oncotype DX recurrence score in estimating the distant recurrence risk of postmenopausal women with hormone-positive breast cancer treated with either tamoxifen or anastrozole alone (Dowsett et al., 2013). While both genomic tests assigned similar numbers, but non-overlapping groups of patients to the low-risk group, the PAM50 ROR score stratified more patients to the high-risk group than Oncotype DX. Interestingly, the study found that in terms of prognostic performance, that the PAM50 ROR outperformed Oncotype DX in estimating distant recurrence risk.

Another clinically relevant genomic biomarker that is available to breast patients and clinicians is the MammaPrint test. The development cohort for MammaPrint consisted of 117 women with early-stage, node-negative breast cancer with the majority of patients presenting with sporadic invasive breast cancer and 20 patients harboring BRCA1/2 germline mutations (van't Veer et al., 2002). The expression levels of 25,000 genes were measured with microarrays and this feature space was then reduced by selecting for genes with greater than a two-fold difference in expression levels resulting in roughly 5000 selected genes. Unsupervised hierarchical clustering of the expression of these approximately 5000 significantly regulated genes and 98 tumors revealed two distinct clusters with disparate risk for distant metastases following therapy. The authors also went on to develop a prognostic signature for distant metastasis risk using supervised machine learning on a subset of 78 tumors from patients with sporadic breast cancers of whom 34 developed metastases within 5 years post-treatment. Training of this prognostic signature involved a separate feature selection step that reduced the full 25,000 gene microarray dataset to around 5000 highly regulated genes. Another feature selection step was implemented by calculating the Pearson correlation coefficient for each gene in the reduced gene set and the risk of distant metastasis, the clinical outcome of interest, which resulted in the selection of 231 outcome-associated genes. Finally, these 231 genes were ranked by the magnitude of their correlation coefficient and further refined using forward feature selection with leave-one-out cross-validation to obtain an optimal set of 70 genes that yielded a cross-validation accuracy of 83%. The classification performance of this final 70-gene prognostic signature was evaluated on an unseen test set of 19 patients with sporadic node-negative breast cancer and achieved an accuracy of approximately 90% across different thresholds optimized for accuracy or sensitivity. A larger test set from the same institution (van de Vijver et al., 2002) confirmed that the 70-gene signature could outperform existing clinicopathologic risk stratification schemes. An independent validation of this signature was performed by the TRANSBIG consortium (Buyse et al., 2006) in Europe, which collected breast cancer samples from 307 patients and stratified patients by their risk for distant metastases within 5-year using either the 70-gene signature or a risk stratification tools based on clinicopathologic factors (Adjuvant! online software (Ravdin et al., 2001)). It was determined in this study that risk stratification using the 70-gene signature was significantly better in predicting both 5-year distant metastasis risk and 10-year overall survival than clinicopathologic factors alone.

The MINDACT trial was the first prospective, randomized validation study of the MammaPrint 70-gene signature and was initially designed to include women with early-stage, node-negative breast cancers, although this was later expanded to also include women with up to three positive nodes. Patients were assigned into both clinical and genomic risk groups based on clinicopathologic variables or the 70 gene signature, respectively, and those with discordant clinical and genomic risk scores were then randomized to receive adjuvant chemotherapy. However, there were no statistically significant differences in outcomes with or without adjuvant chemotherapy within the discordant risk groups in terms of the primary endpoint of distant metastatic risk or the secondary endpoints of disease-free and overall survival. Based on the findings of MINDACT, there is currently not sufficient evidence to support the use of the MammaPrint 70-gene signature to predict the benefit of adjuvant chemotherapy (Cardoso et al., 2016; Markopoulos et al., 2020).

The only prospectively validated predictive biomarker to date is the 21-gene Oncotype DX recurrence score, which is currently used to determine the benefit of the addition of adjuvant chemotherapy in patients with hormone receptor positive, node-negative breast cancer (Soonmyung Paik et al., 2004; Sparano et al., 2015, 2018, 2020; Sparano & Paik, 2008). The Oncotype DX recurrence score was originally developed from a curated set of 250 genes that were collated from prior studies of dysregulated genes in breast cancer. The authors performed additional feature selection by identifying genes that were highly correlated with breast cancer recurrence in three published studies of gene expression profiling in breast cancer. The three studies used were intentionally heterogeneous in order to select for robust gene sets associated with recurrence risk (Cobleigh et al., 2003; Esteban et al., 2003; Paik et al., 2003, p. 82). Five reference genes and 16 tumor-related genes were eventually selected for the final recurrence score model. Clustering and PCA of the 16 tumor-related genes showed that they broadly defined a few functional groups such as genes associated with proliferation, invasion, and HER2 which increased the recurrence score or genes associated with estrogen receptor signaling that decreased the recurrence score. Patients were then risk-stratified by their continuous recurrence score into low, intermediate, and high-risk groups using cutoff points derived from the results of the NSABP B-20 trial. The authors evaluated the recurrence score on 668 patients with ER-positive, node-negative breast cancer treated with tamoxifen enrolled on the NSABP B-14 trial and found on multivariate Cox models that the recurrence score significantly improved prognostic information beyond standard clinicopathologic factors such as tumor grade, size, age, and receptor status.

The use of the Oncotype DX recurrence score as a prognostic and predictive biomarker was recently prospectively validated in the randomized, phase III TAILORx trial (Sparano et al., 2018). The trial enrolled over 10,000 women with hormone receptor positive, HER2-negative, node-negative breast cancer who met the criteria for consideration of adjuvant chemotherapy. Enrolled patients were then risk-stratified by their Oncotype DX recurrence scores with all high-risk patients assigned to receive hormone therapy and chemotherapy and all low-risk patients assigned to hormone therapy alone. Patients who had an intermediate recurrence score, defined as 11−25 in TAILORx, were randomized to hormone therapy alone or hormone therapy and chemotherapy. The primary endpoint was the noninferiority of hormone therapy alone in the study population of patients with an intermediate recurrence score. At 9 years, there was no significant difference in invasive disease-free survival, distant or locoregional failure, or overall survival thereby confirming the utility of Oncotype DX as a

biomarker for chemotherapy use in patients with intermediate (11−25) recurrence scores. Given these results, the 21-gene Oncotype DX recurrence score is now the preferred genomic biomarker test to determine the need for adjuvant chemotherapy for patients with hormone receptor-positive, node-negative breast cancer in the NCCN guidelines (*Breast Cancer (Version 6.2020)*, 2020). The application of Oncotype DX to predict chemotherapy benefit for node-positive breast cancer patients is currently undergoing prospective validation in the ongoing RxPONDER trial (Jasem et al., 2017).

While the biomarkers discussed thus far are only applicable to women with invasive breast cancer, a potentially useful biomarker, called DCISionRT, was recently developed to guide recommendations for adjuvant radiation following surgery for ductal carcinoma in situ (DCIS) (Bremer et al., 2018). The DCISionRT test utilizes both molecular markers as well as a subset of clinicopathologic factors in the final model. The test was developed on archived tumor samples from 526 women with DCIS treated at Uppsala University Hospital in Sweden and at University of Massachusetts Hospital in the United States. An initial set of molecular and clinicopathologic features were selected from previously published reports and expert communications based on their association with DCIS recurrence or disease progression. Further feature selection was performed using forward and backward selection with cross-validation. It should be noted that in contrast with the previously discussed biomarkers that assess transcript-level gene expression, the molecular features included in DCISionRT are derived from immunohistochemical staining of tissue. Ultimately, seven molecular markers (PR, FOXA1, COX-2, SIAH2, HER2, Ki-67, P16/INK4A) and four clinicopathologic features (age, tumor size, margin status, and palpability) were selected for the final predictive model. Internal validation of the DCISionRT Decision Score demonstrated improved prognostic performance over clinicopathologic risk-stratification. External validation of the DCISIONionRT DS was recently published and supported the prognostic utility of the DS in 455 women with DCIS who underwent lumpectomy with or without radiation (Weinmann et al., 2020). However, the validity of the DCISionRT DS as a predictive biomarker for adjuvant radiation in DCIS remains to be proven in a prospective randomized clinical trial.

The biomarkers reviewed here are a small subset of the molecular signatures that are currently in development or undergoing validation studies. As next-generation sequencing techniques continue to improve and become more integrated into the clinical process of care, the opportunities to apply biological insights in oncology treatment decisions will continue to expand. Patients are also becoming more aware and in certain cases proactively requesting molecular tests, especially as these biomarkers become more commercially visible and accessible. This is both an opportunity and a challenge. Clinicians will be challenged to understand how to apply these molecular biomarkers in the correct clinical scenarios as well as how to interpret the results. It is tempting to base treatment decisions on molecular tests that have validated prognostic but unclear predictive utility. The vast majority of currently available biomarker tests are prognostic. A predictive or prescriptive biomarker must be validated in a prospective trial that randomizes patients to a treatment strategy by their biomarker derived risk. Lastly, it is also important for clinicians to understand at a high level how these biomarker tests are developed and in particular the unique challenges of choosing the appropriate patient cohorts and genes of interest as these factors can determine which patients may benefit from a particular molecular signature. These and other potential hurdles will be addressed in the section to follow.

Pharmacogenomics

The ability to optimally recommend existing anti-cancer drugs, and discover novel agents, with the facilitation of genomic profiles is a critical task in chemotherapy and pharmacogenomics for precision medicine and personalized treatment of cancer patients (Moffat et al., 2014). Since the molecular mechanisms of patients with the same cancer type can be distinct, drugs effective in one patient may be ineffective in others (Evans & Relling, 1999; Relling & Evans, 2015). In clinical practice, clinicians identify the differences across patients through cancer subtypes, and provide the best treatment based on the molecular targets and cancer subtypes (Gu et al., 2016). In some subtypes, the targeted pathways or mechanism of action (MoA) are well understood. For example, since the PI3K/AKT/mTOR pathway plays a crucial role in some ER-positive breast cancers, the mTOR inhibitor everolimus can be an effective treatment in subsets of patients (Yardley et al., 2013). Similarly, HER2-targeted antibodies such as trastuzumab can significantly improve the recurrence and survival of early-stage HER2 positive breast cancer patients (Goutsouliak et al., 2020; Slamon et al., 2011).

The extensive screening assays of cancer cell lines have made it possible to test the drug resistance of cancer cells using a panel of potential anti-cancer drugs. Several popular cancer cell line drug sensitivity datasets are publicly accessible, including the NCI-60 (Shoemaker, 2006), CCLE (Barretina et al., 2012), and GDSC (Yang et al., 2013) datasets. Most of these datasets consist of both the response data of cell lines to drugs in the form of IC50 or activity area, as well as the molecular profiles of the cell lines, including somatic mutations, CNAs, transcriptomic expressions, and methylation levels. Computational methods, such as ML models trained over these data, can suggest potentially effective drugs for a cell line if genomic information is provided. For example, a Bayesian multitask multiple kernel learning (MKL) model utilizing Bayesian inference, multitask learning, and kernelized regression was able to achieve the best prediction performance in the NCI-DREAM drug sensitivity prediction challenge (Costello et al., 2014). The prediction accuracy of models can be further improved by taking into account the external knowledge such as cell line or drug similarities (Wei et al., 2019; Zhang et al., 2015). More recently, researchers showed that the models utilizing the features generated by an autoencoder neural network pretrained on the genomic and transcriptomic features of TCGA data can be effective competitors with classical machine learning models trained on the CCLE and GDSC data alone (Chiu et al., 2019; Ding et al., 2018).

There are limitations to making inferences from the cancer cell lines discussed above. For instance, the data from these specific cell line assays cannot uncover additional MoAs that may be therapeutic targets for new drugs or molecules. Computational researchers have proposed a few solutions for this, including drug discovery (Moffat et al., 2014) and drug repurposing/repositioning (Zhang et al., 2020). Another drawback with utilizing cell line data is the huge gap between in vitro cell lines and in vivo real tumors that consist of a mixed population of cancer cells, normal stromal cells, and infiltrating immune cells. One potential solution is leveraging transfer learning to apply models fitted on in vitro cell lines to in vivo tumor models, such as patient-derived xenografts (Bhattacharyya et al., 2020; Sharifi-Noghabi et al., 2019).

Common hurdles of machine learning in genomics

ML can provide a powerful and versatile set of tools, as the preceding discussion demonstrates. When bringing them to new applications or data sets, though, they can be easily applied in ways that are unsound or less than optimal. This section considers some of the common challenges to applying ML successfully in genomics work, particularly for cancer genomics, and some strategies by which they may be approached.

Challenges in data acquisition

Obtaining sufficient quantity and quality of data plays an essential role in successfully applying ML in any context, and cancer genomics is no exception. With advances in sequencing techniques and the rapid expansion in the scale of genomic data, a few recurring issues related to data acquisition and sharing have emerged. These include challenges in the availability of datasets, cleaning and summarizing raw datasets, and performing ML analysis while assuring the privacy and security of potentially sensitive genomic data.

There have been several unified efforts internationally in collecting comprehensive genomic and pathological data of cancer samples in past years, such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) noted previously. In addition to the original datasets, which can contain petabytes of genomic data of various types (Hutter & Zenklusen, 2018), researchers and clinicians may also rely on summary statistics or aggregated results such as information on inferred driver mutations or specific subtypes of cancer data. Efforts to provide such data resources therefore typically involve considerable downstream processing and analysis to supplement them with informative derived data, associate them with relevant metadata, and make the raw data easier for the general research community to use. Table 3.1 provides a few examples in the form of a list, not meant to be comprehensive, of sources of cancer genomic data and associated derived data and metadata, most of which are accessible through easy to use web user interface. Reference data resources are TCGA/GDC (Jensen et al., 2017), dbGaP (Tryka et al., 2014), ICGC (Hudson et al., 2010), EGA (Lappalainen et al., 2015), METABRIC (Pereira et al., 2016), NCI-60 (Shoemaker, 2006), CCLE (Barretina et al., 2012), GDSC (Yang et al., 2013), and CancerSEA (Huating Yuan et al., 2019).

Apart from the original large scale and heterogeneity of the data, general researchers may face challenges in accessing genomic data due to the needs for privacy and security. This is especially an issue in the clinical area or potentially personally identifiable data, both of which are issues for cancer genomics. Researchers traditionally used de-identification techniques to directly remove sensitive Protected Health Information (PHI) such as name or address, which is defined in the Health Insurance Portability and Accountability Act of 1996 (HIPAA) (Cohen & Mello, 2018). However, this practice can not fully protect the privacy of patients or research subjects. For example, data may be subject to a linkage attack, which matches de-identified datasets with external databases to expose the identity of the individuals in a seemingly anonymous dataset (Sweeney, 2002).

Another traditional strategy is to grant data access only to qualified researchers or groups. For example, both TCGA and ICGC provide different access tiers to researchers based on

TABLE 3.1 Examples of important cancer genomic data sets of value to ML applications in cancer genomics.

Databases	Pan-cancer	Tumor or cell line	URL	Comment
TCGA/GDC (Jensen et al., 2017)/ dbGaP (Tryka et al., 2014)	Y	Tumor	https://portal.gdc.cancer.gov/	TCGA data were hosted through dbGaP before 2016, but they are now hosted through GDC.
ICGC (Hudson et al., 2010)	Y	Tumor	https://icgc.org/	
EGA (Lappalainen et al., 2015)	Y	Tumor	https://ega-archive.org/	
METABRIC (Pereira et al., 2016)	N (breast cancer)	Tumor	https://www.cbioportal.org/study/summary?id=brca_metabric	Large scale breast cancer dataset
NCI-60 (Shoemaker, 2006)	Y	Cell line	https://dtp.cancer.gov/discovery_development/nci-60/	Drug sensitivity data
CCLE (Barretina et al., 2012)	Y	Cell line	https://portals.broadinstitute.org/ccle	Drug sensitivity data
GDSC (Yang et al., 2013)	Y	Cell line	https://www.cancerrxgene.org/	Drug sensitivity data

need and demonstrated capacity to protect sensitive data (Hudson et al., 2010; Weinstein et al., 2013). Researchers or general audiences without specific training can only access the less sensitive public aggregated part of the data. However, more sensitive data, such as the original bam/vcf files, and germline information are only available to selected research groups that are reviewed by a committee. Those researchers judged to be qualified will typically also be required to take additional training and conduct experiments or analysis under restrictions, such as the use of a protected computing environment that provides additional guarantees for the privacy of research subjects.

More recently, methods for privacy-preserving data analysis have come into use for sharing data in more limited ways that inherently protect privacy and security. For example, researchers in the area of cryptography and ML have proposed a paradigm of differential privacy (Dwork, 2008) to cope with the challenge of providing access to a set of samples while not leaking information on individual patient samples. Roughly speaking, differential privacy works by adding carefully designed random noise to data in ways that make it provably impossible under certain assumptions to extract information about the data for any specific individual. Recently, this technology has been applied to the area of genomics (Berger & Cho, 2019; Cho et al., 2018). Under such a framework, researchers are still able to apply potentially sophisticated ML inference algorithms to the data but can only access the final output of analysis instead of the raw genomic data.

Data sparsity

Although there has been a boom in genomic analysis of cancer cohorts in recent years, the public cancer cohorts usually include tens of thousands of patients, or, for private ones, up to hundreds of thousands (Chalmers et al., 2017). For more specialized kinds of questions, cohort sizes may range from just tens of samples to thousands. In contrast, the ImageNet dataset in computer vision contains more than 14 million images (Deng et al., 2009) while the Yelp reviews dataset used in natural language processing (NLP) studies contains more than 5 million reviews (McAuley & Leskovec, 2013). At the same time, genomic data normally have high dimensional feature sets. For human beings, we have more than 20,000 genes and 324,000,000 known variants in total. The limited samples, high dimensions, and large noise characterizing the genomic data can lead to fragile machine learning models, e.g., overfitting and lack of interpretability. Researchers have tackled high dimensional data with various approaches, including feature selection and feature transformation.

Small sample sizes relative to the feature set present a challenge for almost any kind of ML. In some cases, dealing with small sample sizes may mean favoring different ML methods that are less data intensive, e.g., using support vector machines (SVMs) instead of popular deep learning approaches (Brown et al., 2000; Guyon et al., 2002). Extra attention to protecting models from overfitting is also warranted. ML offers a variety of common methods for protecting a model from overfitting during learning, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or other regularization strategies that bias toward simple models (Kuha, 2004). It is also important to validate robustness to data subsampling and to independent data sets post-hoc. The use of prior knowledge to bias a model to reflect expected outcomes can also help mitigate the effects of limited training data.

Another way to mitigate the model complexity is leveraging the effectiveness of other related data available, e.g., through transfer learning (Weiss et al., 2016), multitask learning (Caruana, 1998), and semi-supervised learning (Hady & Schwenker, 2013). These methods utilize knowledge from other applications, with the hypothesis that the data entails specific structure and can improve performance when the number of samples is limited by capturing that structure. In transfer learning, instead of working on the current task, the model is first pretrained on a related task, and then part of the pretrained model parameters are transferred to the current task to boost the performance. For example, in order to predict the differentially expressed RNAs from the somatic mutations in a tumor, researchers first pretrained "gene embeddings" using a Mut2Vec model in an unsupervised way (Kim et al., 2018), and then transferred the gene embeddings to improve the RNA expression prediction task (Tao, Cai, et al., 2020). In multitask learning, it is thought that multiple prediction tasks with limited sample sizes are closely related to each other, and therefore it is proposed that sharing model parameters across these tasks can improve the performance of individual tasks. Yuan et al. formulated the resistance of cell lines to multiple anti-cancer drugs as a multitask problem and found that the collaboration of these individual tasks facilitates the overall prediction (Yuan et al., 2016). With semi-supervised learning, we seek to overcome the problem of having access only to limited amounts of labeled data for which a property of interest is known, but also with access to many more unlabeled samples. By integrating the prior assumption that similar samples are more likely to share the same label, semi-supervised methods can improve the prediction of labels by utilizing a large number of unlabeled samples. For

example, Bair and Tibshirani utilized gene expression data to predict breast cancer patient survival, and found it outperformed both supervised learning or clustering methods alone (Bair & Tibshirani, 2004).

One particularly important class of methods for dealing with data sparsity is dimensionality reduction, which refers to a set of strategies for shrinking the set of features from which we seek to learn. One simple version of data sparsity is straightforward: since many features are redundant in a dataset, it is possible to just select a subset of essential features that are important to the task. A few major categories of methods have been developed for this purpose (Saeys et al., 2007), including filter (Xing et al., 2001), wrapper (Kohavi & John, 1997), and embedded methods (Robert Tibshirani, 1996). Taking the cancer type classification task through microarray expression profiles of tumors as an example (Zhao & Wu, 2016), the training process of a machine learning model is equivalent to optimizing an objective function. In the case of wrapper methods, the subset of genes is selected that can achieve the best performance on the validation dataset. In practice, it is computationally infeasible to find the optimal subset of around 20,000 features, but many heuristic algorithms have been proposed to select suboptimal solutions, e.g., stepwise forward selection. In embedded methods, however, additional regularization terms are added on the model parameters to the original objective/loss function. A widely used version of this is the L1-regularization of the parameters, or Lasso (Tibshirani, 1997). The L1 regularization is equivalent to having a Laplacian prior to the model parameters, therefore enforcing a sparse solution, where most of the coefficients are zeros, i.e., not selected. Wrapper methods are in general computationally expensive and prone to overfit. Therefore it is necessary to have a proper split of the dataset during tuning and evaluation, e.g., nested cross-validation (Cawley & Talbot, 2010). Embedded methods are faster and easier to implement in practices (Cawley & Talbot, 2010).

Apart from the data-driven dimension reduction methods, computational biologists may also incorporate a biological database for dimension reduction through a knowledge-driven approach. In cancer genomics, a common approach is to reduce the gene-level expressions into pathway-level expressions (Drier et al., 2013; Park et al., 2009; Tao et al., 2019; Tao, Lei, et al., 2020). A pathway is defined as a set of closely related genes that are interact in the genome or participate in the same or similar molecular processes, and therefore are likely to show correlated expression. A few knowledge bases can be utilized, e.g., the DAVID database (Dennis et al., 2003, p. P3), KEGG pathway database (Kanehisa & Goto, 2000), and Gene Ontology (Mi et al., 2013). This kind of knowledge-driven dimension reduction method can especially be effective when the sample number is limited.

Unsupervised learning methods are also useful ways of reducing noise and reducing the dimension of input features to the model. Principal components analysis (PCA) can identify and utilize the correlation across features from the data to represent a complex dataset with a reduced set of derived features. PCA rotates the coordinates of the original features, such that the variance of samples is largest to the first axis of the new coordinate, and is second largest to the second axis, etc. For noise reduction purposes, a method will select the top k dimensions that are able to explain at least (1-ε) of the total variance of the samples, where ε is typically set to be 0.05 or 0.01. In other cases, one might choose the top 50 or 100 or 200 dimensions. PCA is also often used as a preprocessing step for nonlinear visualization tools such as t-SNE, as introduced below. More recently, some researchers have utilized the

"autoencoder" to extract the nonlinear hidden states of the original input (Alavi et al., 2018; Hinton & Salakhutdinov, 2006). The autoencoder is a type of neural network with the output the same as the input, and the central layer of the autoencoder is used as the compact representation of input features, which can be used for recovering the dense high-dimensional input signals. It can be proved that PCA is equivalent to an autoencoder with one hidden layer under specific conditions (Bourlard & Kamp, 1988).

PCA has the advantage of a linear transformation. It tries to find a good viewpoint to present the data instead of changing the data structure. PCA is not the best option for visualization in many cases because the differences of samples may exist in other axes instead of the first two. Therefore, nonlinear transformations to the two-dimensional space is necessary. t-SNE (van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2018) are widely used methods for visualization, which are based on the intuition that the near samples in the original feature space should be close to each other in the reduced 2-D space as well. The two methods and improved variants are widely used in the analysis of bulk and single-cell molecular profiles (Abdelmoula et al., 2016; Linderman et al., 2019). UMAP in general yields smoother boundaries than t-SNE. Another advance of UMAP is its flexibility of choosing kernel functions, which defines similarity or distance between samples, making it suitable for applications of single-cell RNA data (Wang et al., 2017). Researchers also use nonlinear transformations such as autoencoder (introduced in the previous subsection) or VAE for dimension reduction of genomic data. However, these only capture local information instead of global, thus may not be a suitable way for feature engineering. More advanced methods from the field of manifold learning may be employed to decompose data into lower-dimensional models even if they form complex substructures in the full feature space.

Inter-tumor heterogeneity

Another frequently encountered challenge of cancer genomic data in machine learning is high inter-tumor heterogeneity. Depending on the tumor type, there may be a number of subtypes with substantially different molecular mechanisms, confounding many forms of genomic analysis (Parker et al., 2009). Where a subtyping is well defined and understood, pre-partitioning data by subtype (Hofree et al., 2013) may avoid some of the challenges of confounding data at the cost of reducing cohort sizes and making it hard to infer cross-subtype effects. Even if a subtyping is not fully understood, unsupervised clustering approaches (e.g., *k*-means clustering) can be used as a preprocessing step to partition data into more homogeneous subgroups before further analysis (Liu et al., 2006). Subtyping remains an area of active study, however, and meaningful subtypes continue to be discovered and refined. Furthermore, even within a defined subtype, most cancers commonly exhibit somatic hypermutability (Lawrence et al., 2013), leading typically to large amounts of idiosyncratic, functionally irrelevant passenger mutations that can expand data dimensionality, further confounding analyses and limiting the effectiveness of simple strategies for dimensionality reduction. General strategies for dealing with data sparsity and large feature sets discussed above may mitigate such problems. Specialized methods, such as training models on mutation burdens at the gene or pathway level rather than individual variants, can also reduce, but not eliminate, these challenges.

Intra-tumor heterogeneity

Another recurring challenge of cancer genomic data is intra-tumor heterogeneity (ITH), i.e., cell-to-cell variability within single tumors (Jamal-Hanjani et al., 2015; Swanton, 2012). ITH is a common feature of cancers arising from the process of clonal evolution within a tumor and from the somatic hypermutability processes frequently active in tumors during this process (Loeb, 1991). As a result, cancer genomic data often must be interpreted as mixtures of cell types (clones) that may exhibit different mutations, epigenetic markers, or patterns of gene activity. This clonal heterogeneity is further exacerbated by contributions from infiltrating immune cells or other stromal contamination (Tao et al., 2019; Zhu et al., 2019). ITH is a confounding factor for many common ML analyses, as genetic or genomic signals that underlie tumor function are obscured by clones that lack those signals. The problem is particularly vexing for prognostic prediction because progression processes in cancers, such as metastasis or the development of drug resistance, frequently proceed from relatively rare clones within a tumor (Heselmeyer-Haddad et al., 2012) and thus prediction based on the dominant genomic features of a tumor may poorly predict the behavior of the tumor as a whole.

While ITH is problematic for ML in cancer genomics, it can be dealt with by computational or experimental approaches. Computational strategies for working with ITH typically involve the use of genomic deconvolution, a strategy for computationally separating mixed genomic signals to infer likely signals of specific clones within a tumor that can then be examined separately during machine learning (Venet et al., 2001). While such approaches were initially developed specifically for resolving tumor impurity, by separating tumor and stromal contributions (Etzioni et al., 2005), the idea was later extended to resolve distinct clones within single tumors. Numerous deconvolution methods exist today, some relying on multiple samples from a single tumor to resolve clonal mixtures and others on comparison across tumors to resolve common features of progression across a cohort. Clonal deconvolution is often combined with tumor phylogenetics (Schwartz & Schäffer, 2017), i.e., inference of trees describing the evolution of clonal states in a tumor, to better resolve substructure (Beerenwinkel et al., 2005; Schwartz & Shackney, 2010).

As single-cell genomics has become more practical and widespread, it has increasingly displaced deconvolution methods for resolving clonal mixtures computationally from bulk sequencing data. The peculiar error characteristics of different single-cell technologies — which may include high error rates, high rates of allelic dropout or other missing data, and aberrations such as doublet sequences — create distinct challenges for ML from single-cell data that generally must be resolved to adapt ML analysis methods from bulk to single-cell data (Suvà & Tirosh, 2019). A handful of methods now exist as well for combining bulk and single-cell data in common analyses to achieve some advantages for each method type (Lei et al., 2020; Malikic, Jahn, et al., 2019; Malikic, Mehrabadi, et al., 2019).

Other common data issues

Another issue that can plague many genomic applications is imbalanced data. Imbalanced data refers to datasets that are skewed to some possible outcomes over others. An example would be the challenge of predicting a rare complication or atypical progression outcome (Diz et al., 2016), e.g., predicting those patients with poor outcomes in cancers that are rarely

fatal, because there are few examples from which to train a model. General methods for dealing with data sparsity — such as reliance on prior knowledge, model regularization, or strategies such as transfer learning — may also help address challenges of imbalanced data. In addition, specialized strategies including the upsampling of minor categories via the generation of artificial "decoy" data (Blagus & Lusa, 2013), or the downsampling of major categories, can mitigate the problem of imbalance. When coping with imbalanced data, it is crucial to choose proper evaluation metrics. Common intuitive assessments such as accuracy and the receiver operating characteristic (ROC) curve do not describe the minor class properly. Instead, other measures such as the F1 score and precision-recall curve are in general better choices in these cases (Davis & Goadrich, 2006).

Missing and/or inconsistent annotation of data are likewise common problems of cancer genomic data (de Souto et al., 2015). While data quality has generally improved across genomic technologies over time, noisy assays and complex biology can lead to missing fields in subsets of data points, posing yet another problem for ML. This may be particularly a concern for clinical data, where standards for expert annotation may still be less precise and consistent than is ideal for automated inference. Furthermore, large consortium efforts, for all their value to the scientific community, can introduce problems of standardization across partner sites. While changes in practice concurrent with the broader adoption of electronic health records (EHRs) may help, ML methods must be able to cope with all of these issues to make use of them. Cleaning data of poorly annotated data points or data fields can resolve some such problems (Jianfang Liu et al., 2018). ML may also rely on imputation, i.e., using simpler learning models or other heuristics to infer likely values for missing data, or be engineered directly to allow for unknown or uncertain values in data (Beretta & Santaniello, 2016).

This remains far from an exhaustive list about the challenges of genomic data and cancer genomic data in particular. It is intended, however, to highlight some of the common issues and provide an overview of typical ML strategies for dealing with these or similar problems. When other difficulties arise, some of the same strategies discussed above may prove useful in achieving good performance in less-than-ideal conditions for machine learning inference.

Future directions

The idea of personalized and precision oncology is no longer new, but the field continues to rapidly advance through a combination of dramatic improvements in genomic methods and other synergistic biotechnologies. This, combined with a burgeoning field of computational cancer biology using innovative machine learning methods, is realizing the promise of translating masses of genomic data into actionable information for clinicians. There is no indication that these trends are slowing down and indeed they seem likely to continue to accelerate. It is thus fair to consider how the field might continue to evolve in the coming years and how future clinicians and clinical cancer researchers might take advantage of innovations yet to come.

Sequencing technologies continue to become more versatile and affordable, and one can fairly speculate on how that will impact future cancer treatment. While targeted or whole-exome sequencing remain the standard in clinical practice, costs of whole-genome sequencing are now almost negligible compared to the cost of cancer treatment. We can thus anticipate

WGS becoming routinely available as part of cancer treatment and other medical care, along with other emerging genomic technologies, enabling a host of new downstream computational analyses to become routine in cancer care. Other genomic technologies that are still maturing, such as "liquid biopsy" — i.e., the genomic characterization of tumor cells using peripheral blood — can likewise be expected to become more routinely available as part of cancer treatment or even routine public health screening (Mattox et al., 2019).

Although single-cell sequencing technologies appeared late in the last century (Eberwine et al., 1992) and gained a great deal of attraction in the past decade in academia (Hwang et al., 2018), as far as we know, they have not yet become routine in clinical practice. However, single-cell sequencing has a potentially wide application in oncology, for example, identifying the chemo-resistant tumor clones or malignant cell populations from the tumor tissue (Haque et al., 2017). One common feature of many such technologies is that once a biotechnology becomes widely available for deriving genomic data, new computational advances can be added with minimal cost and no additional burden to the patient. Also, improvements in collecting high-quality clinical data will help solve some of the hurdles associated with data sparsity and lead to a more efficient pipeline from biomarker development to validation. We will likely not know for some time what specific directions will have clinical impact, but the enormous growth in computational cancer biology suggests many possibilities for reconstructing tumor evolution, progression, and cell migration and applying ML to project its future trajectory.

One further prediction is that an emerging era of big data medicine and computationally augmented reasoning will usher in a need for rethinking standards of training for clinicians and other healthcare professionals, who will not necessarily be inventing such technologies, but will all need to understand and use them (Welch et al., 2014). Standard medical education today provides aspiring physicians only limited training as to their use and how to critically evaluate these tools. Increasing effort at developing interpretable ML models may help bridge the gap between the state-of-the-art of ML research and their use by non-experts, but cannot eliminate the need for physicians to understand and determine when ML tools and analyses are appropriate for them and how to weigh results of such analyses that may critically depend on model assumptions and the available data. We can suggest that more advanced training in data science, computational thinking, and statistical reasoning will be key to preparing future generations of oncologists to practice medicine in a climate where it is necessary to work with and evaluate computational inferences alongside their own training and judgment.

References

Abdelmoula, W. M., Balluff, B., Englert, S., Dijkstra, J., Reinders, M. J. T., Walch, A., McDonnell, L. A., & Lelieveldt, B. P. F. (2016). Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of Mass spectrometry imaging data. *Proceedings of the National Academy of Sciences of the United States of America, 113*(43), 12244–12249. https://doi.org/10.1073/pnas.1510227113

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M. A. K., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., & Venter, J. C. (1991). Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science, 252*(5013), 1651–1656. https://doi.org/10.1126/science.2047873

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics, 76*(1), 7–20. https://doi.org/10.1002/0471142905.hg0720s76

Alavi, A., Ruffalo, M., Parvangada, A., Huang, Z., & Bar-Joseph, J. (2018). A web server for comparative analysis of single-cell RNA-seq data. *Nature Communications, 9*(1).

Alix-Panabières, C., & Pantel, K. (2013). Circulating tumor cells: Liquid biopsy of cancer. *Clinical Chemistry, 59*(1), 110–118. https://doi.org/10.1373/clinchem.2012.194258

Arteaga, C. L. (2003). ErbB-targeted therapeutic approaches in human cancer. *Experimental Cell Research, 284*(1), 122–130. https://doi.org/10.1016/S0014-4827(02)00104-0

Axel, V., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., & Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature, 457*(7231), 854–858. https://doi.org/10.1038/nature07730

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sohini, S., Denis, B., Amila, W., Antonio, C., Wendl, M. C., Jaegil, K., Brendan, R., Kwok-Shing, N. P., Jin, J. K., Song, C., Zixing, W., Jianjiong, G., Qingsong, G., Fang, W., Minwei, L. E., Loris, M., … Mariamidze, A. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell, 173*(2), 371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060

Bair, E., & Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology, 2*(4), e108. https://doi.org/10.1371/journal.pbio.0020108

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., … Garraway, L. A. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature, 483*(7391), 603–607. https://doi.org/10.1038/nature11003

Beerenwinkel, N., Rahnenführer, J., Kaiser, R., Hoffman, D., Selbig, J., & Lengauer, T. (2005). Mtreemix: A software package for learning and using mixture models of mutagenetic trees. *Bioinformatics, 21*(9), 2106–2107. https://doi.org/10.1093/bioinformatics/bti274

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making, 16*(Suppl. 3), 74. https://doi.org/10.1186/s12911-016-0318-z

Berger, B., & Cho, H. (2019). Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biology, 20*, 128. https://doi.org/10.1186/s13059-019-1741-0

Bhattacharyya, R., Ha, M. J., Liu, Q., Akbani, R., Liang, H., & Baladandayuthapani, V. (2020). Personalized network modeling of the pan-cancer patient and cell line interactome. *JCO Clinical Cancer Informatics, 4*, 399–411. https://doi.org/10.1200/cci.19.00140

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics, 14*, 106. https://doi.org/10.1186/1471-2105-14-106

Bourlard, H., & Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics, 59*(4–5), 291–294. https://doi.org/10.1007/BF00332918

Breast cancer (Version 6.2020). (2020). National Comprehensive Cancer Network. https://www.nccn.org/professionals/physician_gls/pdf/breast.pdf

Bremer, T., Whitworth, P. W., Patel, R., Savala, J., Barry, T., Lyle, S., Leesman, G., Linke, S. P., Jirström, K., Zhou, W., Amini, R.-M., & Wärnberg, F. (2018). A biological signature for breast ductal carcinoma in situ to predict radiotherapy benefit and assess recurrence risk. *Clinical Cancer Research, 24*(23), 5895–5901. https://doi.org/10.1158/1078-0432.ccr-18-0842

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America, 97*(1), 262–267. https://doi.org/10.1073/pnas.97.1.262

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology, 109*(1), 21–29. https://doi.org/10.1002/0471142727.mb2129s109

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., … Straehle, C. (2006). Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute, 98*(17), 1183–1192. https://doi.org/10.1093/jnci/djj329

Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., Hermansen, R. A., Byers, R. L., Clement, M. J., Udall, J. A., Wilcox, E. R., & Crandall, K. A. (2011). Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution, 3*(1), 1312–1323. https://doi.org/10.1093/gbe/evr106

Cai, C., Cooper, G. F., Lu, K. N., Ma, X., Xu, S., Zhao, Z., Chen, X., Xue, Y., Lee, A. V., Clark, N., Chen, V., Lu, S., Chen, L., Yu, L., Hochheiser, H. S., Jiang, X., Wang, Q. J., & Lu, X. (2019). Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference. *PLoS Computational Biology, 15*(7), e1007088. https://doi.org/10.1371/journal.pcbi.1007088

Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J.-Y., Brain, E., Causeret, S., DeLorenzi, M., Glas, A. M., Golfinopoulos, V., Goulioti, T., Knox, S., Matos, E., Meulemans, B., Neijenhuis, P. A., Nitz, U., Passalacqua, R., … Piccart, M. (2016). 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. *New England Journal of Medicine, 375*(8), 717–729. https://doi.org/10.1056/NEJMoa1602253

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: Computational prediction of driver missense mutations. *Cancer Research, 69*(16), 6660–6667. https://doi.org/10.1158/0008-5472.CAN-09-1133

Caruana, R. (1998). In S. Thrun, & L. Pratt (Eds.), *Multitask learning* (pp. 95–133). Springer US.

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research: JMLR, 11*, 2079–2107.

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., Huang, F., He, Y., Sun, J., Tabori, U., Kennedy, M., Lieber, D. S., Roels, S., White, J., Otto, G. A., … Frampton, G. M. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine, 9*(1), 34. https://doi.org/10.1186/s13073-017-0424-2

Chiu, Y.-C., Chen, H.-I. H., Zhang, T., Zhang, S., Gorthi, A., Wang, L.-J., Huang, Y., & Chen, Y. (2019). Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics, 12*(1), 18. https://doi.org/10.1186/s12920-018-0460-9

Cho, H., Wu, D. J., & Berger, B. (2018). Secure genome-wide association analysis using multiparty computation. *Nature Biotechnology, 36*(6), 547–551. https://doi.org/10.1038/nbt.4108

Chung, W., Eum, H. H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., Ryu, H. S., Kim, S., Lee, J. E., Park, Y. H., Kan, Z., Han, W., & Park, W.-Y. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications, 8*(1), 1–12. https://doi.org/10.1038/ncomms15081

Cobleigh, M. A., Bitterman, P., Baker, J., Cronin, M., Liu, M. L., Borchik, R., Tabesh, B., Mosquera, J. M., Walker, M. G., & Shak, S. (2003). Tumor gene expression predicts distant disease-free survival (DDFS) in breast cancer patients with 10 or more positive nodes: High throughput RT-PCR assay of Paraffin-embedded tumor tissues. *Program Proceedings – American Society of Clinical Oncology, 22*.

Cohen, I. G., & Mello, M. M. (2018). HIPAA and protecting health information in the 21st century. *JAMA, the Journal of the American Medical Association, 320*(3), 231–232. https://doi.org/10.1001/jama.2018.5630

Contreras-Naranjo, J. C., Wu, H.-J., & Ugaz, V. M. (2017). Microfluidics for exosome isolation and analysis: Enabling liquid biopsy for personalized medicine. *Lab on a Chip, 17*(21), 3558–3577. https://doi.org/10.1039/C7LC00592J

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S. A., Mpindi, J.-P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Collins, J. J., Gallahan, D., Singer, D., Saez-Rodriguez, J., … Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology, 32*(12), 1202–1212. https://doi.org/10.1038/nbt.2877

Crowley, E., Di Nicolantonio, F., Loupakis, F., & Bardelli, A. (2013). Liquid biopsy: Monitoring cancer-genetics in the blood. *Nature Reviews Clinical Oncology, 10*(8), 472–484. https://doi.org/10.1038/nrclinonc.2013.110

Das, P. M., & Singal, R. (2004). DNA methylation and cancer. *Journal of Clinical Oncology, 22*(22), 4632–4642. https://doi.org/10.1200/jco.2004.07.151

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd international conference on machine learning*.

Dawson, S.-J., Rueda, O. M., Aparicio, S., & Caldas, C. (2013). A new genome-driven integrated classification of breast cancer and its implications. *The EMBO Journal, 32*(5), 617–628. https://doi.org/10.1038/emboj.2013.19

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K., & Ding, L. (2012). MuSiC: Identifying mutational significance in cancer genomes. *Genome Research, 22*(8), 1589–1598. https://doi.org/10.1101/gr.134635.111

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science, 295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biology, 4*(5), P3.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., & Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics, 14*(4), 457–460. https://doi.org/10.1038/ng1296-457

Diaz, L. A., & Bardelli, A. (2014). Liquid biopsies: Genotyping circulating tumor DNA. *Journal of Clinical Oncology, 32*(6), 579–586. https://doi.org/10.1200/jco.2012.45.2011

Ding, M. Q., Chen, L., Cooper, G. F., Young, J. D., & Lu, X. (2018). Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Molecular Cancer Research, 16*(2), 269–278. https://doi.org/10.1158/1541-7786.MCR-17-0378

Diz, J., Marreiros, G., & Freitas, A. (2016). Applying data mining techniques to improve breast cancer diagnosis. *Journal of Medical Systems, 40*(9), 203. https://doi.org/10.1007/s10916-016-0561-y

Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Dunbier, A. K., Cowens, J. W., Ferree, S., Storhoff, J., Schaper, C., & Cuzick, J. (2013). Comparison of PAM50 risk of recurrence score with Onco type DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *Journal of Clinical Oncology, 31*(22), 2783–2790. https://doi.org/10.1200/JCO.2012.46.1558

Drier, Y., Sheffer, M., & Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences, 110*(16), 6388–6393. https://doi.org/10.1073/pnas.1219651110

Dumur, C. I., Dechsukhum, C., Ware, J. L., Cofield, S. S., Best, A. M., Wilkinson, D. S., Garrett, C. T., & Ferreira-Gonzalez, A. (2003). Genome-wide detection of LOH in prostate cancer using human SNP microarray technology. *Genomics, 81*(3), 260–269. https://doi.org/10.1016/S0888-7543(03)00020-X

Dwork, C. (2008). Differential privacy: A survey of results. In *Theory and applications of models of computation* (pp. 1–19).

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., & Coleman, P. (1992). Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences of the United States of America, 89*(7), 3010–3014. https://doi.org/10.1073/pnas.89.7.3010

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., Gibbs, R. A., & Liu, Z. (2012). Mind the gap: Upgrading genomes with Pacific biosciences RS long-read sequencing technology. *PLoS One, 7*(11), e47768. https://doi.org/10.1371/journal.pone.0047768

Esteban, J., Baker, J., Cronin, M., Liu, M. L., Llamas, M. G., Walker, M. G., Mena, R., & Shak, S. (2003). Tumor gene expression and prognosis in breast cancer: Multi-gene RT-PCR assay of Paraffin-embedded tissue. *Proceedings of American Society of Clinical Oncology, 22*.

Etzioni, R., Hawley, S., Billheimer, D., True, L. D., & Knudsen, B. (2005). Analyzing patterns of staining in immunohistochemical studies: Application to a study of prostate cancer recurrence. *Cancer Epidemiology, Biomarkers & Prevention, 14*(5), 1040–1046. https://doi.org/10.1158/1055-9965.EPI-04-0584

Evans, W. E., & Relling, M. V. (1999). Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science, 286*(5439), 487–491. https://doi.org/10.1126/science.286.5439.487

Fudenberg, G., Getz, G., Meyerson, M., & Mirny, L. A. (2011). High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nature Biotechnology, 29*(12), 1109–1113. https://doi.org/10.1038/nbt.2049

Fukui, T., Ohe, Y., Tsuta, K., Furuta, K., Sakamoto, H., Takano, T., Nokihara, H., Yamamoto, N., Sekine, I., Kunitoh, H., Asamura, H., Tsuchida, T., Kaneko, M., Kusumoto, M., Yamamoto, S., Yoshida, T., & Tamura, T. (2008). Prospective study of the accuracy of EGFR mutational analysis by high-resolution melting analysis in small samples obtained from patients with non-small cell lung cancer. *Clinical Cancer Research, 14*(15), 4751–4757. https://doi.org/10.1158/1078-0432.CCR-07-5207

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., & Stratton, M. R. (2004). A census of human cancer genes. *Nature Reviews Cancer, 4*(3), 177–183. https://doi.org/10.1038/nrc1299

Gagan, J., & Van Allen, E. M. (2015). Next-generation sequencing to guide cancer therapy. *Genome Medicine, 7*(1). https://doi.org/10.1186/s13073-015-0203-x

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gnant, M., Filipits, M., Greil, R., Stoeger, H., Rudas, M., Bago-Horvath, Z., Mlineritsch, B., Kwasny, W., Knauer, M., Singer, C., Jakesz, R., Dubsky, P., Fitzal, F., Bartsch, R., Steger, G., Balic, M., Ressler, S., Cowens, J. W., Storhoff, J., … Nielsen, T. O. (2014). Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: Using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone. *Annals of Oncology, 25*(2), 339–345. https://doi.org/10.1093/annonc/mdt494

Goldman, J. M., & Melo, J. V. (2003). Chronic Myeloid Leukemia — Advances in biology and new approaches to treatment. *New England Journal of Medicine, 349*(15), 1451–1464. https://doi.org/10.1056/NEJMra020777

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A., & Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods, 10*(11), 1081–1082. https://doi.org/10.1038/nmeth.2642

Gonzalez-Perez, A., & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Research, 40*(21), e169. https://doi.org/10.1093/nar/gks743

Goutsouliak, K., Veeraraghavan, J., Sethunath, V., De Angelis, C., Osborne, C. K., Rimawi, M. F., & Schiff, R. (2020). Towards personalized treatment for early stage HER2-positive breast cancer. *Nature Reviews Clinical Oncology, 17*(4), 233–250. https://doi.org/10.1038/s41571-019-0299-9

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., … Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature, 446*(7132), 153–158. https://doi.org/10.1038/nature05610

Gu, G., Dustin, D., & Fuqua, S. A. W. (2016). Targeted therapy for breast cancer and molecular mechanisms of resistance to treatment. *Current Opinion in Pharmacology, 31*, 97–103. https://doi.org/10.1016/j.coph.2016.11.005

Gunderson, K. L., Steemers, F. J., Lee, G., Mendoza, L. G., & Chee, M. S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics, 37*(5), 549–554. https://doi.org/10.1038/ng1547

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797

Hady, M. F. A., & Schwenker, F. (2013). Semi-supervised learning. In M. Bianchini, M. Maggini, & L. Jain (Eds.), *Handbook on neural information processing* (pp. 215–239). Springer Science and Business Media LLC. https://doi.org/10.1007/978-3-642-36657-4_7

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine, 9*(1).

Heselmeyer-Haddad, K., Berroa Garcia, L. Y., Bradley, A., Ortiz-Melendez, C., Lee, W. J., Christensen, R., Prindiville, S. A., Calzone, K. A., Soballe, P. W., Hu, Y., Chowdhury, S. A., Schwartz, R., Schäffer, A. A., & Ried, T. (2012). Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of MYC during progression. *American Journal Of Pathology, 181*(5), 1807–1822. https://doi.org/10.1016/j.ajpath.2012.07.012

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature Methods, 10*(11), 1108–1115. https://doi.org/10.1038/nmeth.2651

HuBMAP Consortium. (2019). The human body at cellular resolution: The NIH human biomolecular atlas program. *Nature, 574*(7777), 187–192.

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., … Yang, H. (2010). International network of cancer genome projects. *Nature, 464*(7291), 993–998. https://doi.org/10.1038/nature08987

Hunkapiller, T., Kaiser, R. J., Koop, B. F., & Hood, L. (1991). Large-scale and automated DNA sequence determination. *Science, 254*(5028), 59–67. https://doi.org/10.1126/science.1925562

Hutter, C., & Zenklusen, J. C. (2018). The cancer genome atlas: Creating lasting value beyond its data. *Cell, 173*(2), 283–285. https://doi.org/10.1016/j.cell.2018.03.042

Hwang, B., Ji, H. L., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine, 50*(8), 1–14. https://doi.org/10.1038/s12276-018-0071-8

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology, 36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Jamal-Hanjani, M., Quezada, S. A., Larkin, J., & Swanton, C. (2015). Translational implications of tumor heterogeneity. *Clinical Cancer Research, 21*(6), 1258–1266. https://doi.org/10.1158/1078-0432.CCR-14-1429

Jasem, J., Fisher, C. M., Amini, A., Shagisultanova, E., Rabinovitch, R., Borges, V. F., Elias, A., & Kabos, P. (2017). The 21-gene recurrence score assay for node-positive, early-stage breast cancer and impact of RxPONDER trial on chemotherapy decision-making: Have clinicians already decided? *Journal of the National Comprehensive Cancer Network, 15*(4), 494–503.

Jensen, M. A., Ferretti, V., Grossman, R. L., & Staudt, L. M. (2017). The NCI genomic data commons as an engine for precision medicine. *Blood, 130*(4), 453–459. https://doi.org/10.1182/blood-2017-03-735654

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Kim, S., Lee, H., Kim, K., & Kang, J. (2018). Mut2Vec: Distributed representation of cancerous mutations. *BMC Medical Genomics, 11*(2), 33. https://doi.org/10.1186/s12920-018-0349-7

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324. https://doi.org/10.1016/s0004-3702(97)00043-x

Kornelia, P. (2007). Breast cancer: Origins and evolution. *Journal of Clinical Investigation*, 3155–3163. https://doi.org/10.1172/jci33295

Kroman, N., Holtveg, H., Wohlfahrt, J., Jensen, M.-B., Mouridsen, H. T., Blichert-Toft, M., & Melbye, M. (2004). Effect of breast-conserving therapy versus radical mastectomy on prognosis for young women with breast carcinoma. *Cancer, 100*(4), 688–693. https://doi.org/10.1002/cncr.20022

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research, 33*(2), 188–229. https://doi.org/10.1177/0049124103262065

Kulkarni, M. M. (2011). Digital multiplexed Gene Expression analysis using the NanoString nCounter system. *Current Protocols in Molecular Biology, 2011*. https://doi.org/10.1002/0471142727.mb25b10s94

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols, 4*(7), 1073–1081. https://doi.org/10.1038/nprot.2009.86

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860–921. https://doi.org/10.1038/35057062

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., Vaughan, B., Laurent, T., Rowland, F., Marin-Garcia, P., Barker, J., Jokinen, P., Carreño Torres, A., Rambla de Argila, J., Martinez Llobet, O., … Flicek, P. (2015). The European genome-phenome archive of human data consented for biomedical research. *Nature Genetics, 47*(7), 692–695. https://doi.org/10.1038/ng.3312

Lappalainen, T., Scott, A. J., Brandt, M., & Hall, I. M. (2019). Genomic analysis in the age of human genome sequencing. *Cell, 177*(1), 70–84. https://doi.org/10.1016/j.cell.2019.02.032

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., & Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature, 505*(7484), 495–501. https://doi.org/10.1038/nature12912

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature, 499*(7457), 214–218. https://doi.org/10.1038/nature12213

Lei, H., Lyu, B., Gertz, E. M., Schäffer, A. A., Shi, X., Wu, K., Li, G., Xu, L., Hou, Y., Dean, M., & Schwartz, R. (2020). Tumor copy number deconvolution integrating bulk and single-cell sequencing data. *Journal of Computational Biology, 27*(4), 565–598. https://doi.org/10.1089/cmb.2019.0302

Leung, M. K. K., Delong, A., Alipanahi, B., & Frey, B. J. (2016). Machine learning in genomic medicine: A review of computational problems and data sets. *Proceedings of the IEEE, 104*(1), 176–197. https://doi.org/10.1109/jproc.2015.2494198

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., & Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods, 16*(3), 243–245. https://doi.org/10.1038/s41592-018-0308-4

Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., Kumar, K., Khurana, E., Waszak, S., Korbel, J. O., Haber, J. E., Imielinski, M., Akdemir, K. C., Alvarez, E. G., Baez-Ortega, A., Beroukhim, R., Boutros, P. C., Bowtell, D. D. L., Brors, B., Burns, K. H., … Campbell, P. J. (2020). Patterns of somatic structural variation in human cancer genomes. *Nature, 578*(7793), 112–121. https://doi.org/10.1038/s41586-019-1913-9

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., & Hu, H. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell, 173*(2), 400–416.e11. https://doi.org/10.1016/j.cell.2018.02.052

Liu, J., Mohammed, J., Carter, J., Ranka, S., Kahveci, T., & Baudis, M. (2006). Distance-based clustering of CGH data. *Bioinformatics, 22*(16), 1971–1978. https://doi.org/10.1093/bioinformatics/btl185

Loeb, L. A. (1991). Mutator phenotype may be required for multistage carcinogenesis. *Cancer Research, 51*(12), 3075–3079.

Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., Harris, P. L., Haserlat, S. M., Supko, J. G., Haluska, F. G., Louis, D. N., Christiani, D. C., Settleman, J., & Haber, D. A. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *New England Journal of Medicine, 350*(21), 2129–2139. https://doi.org/10.1056/NEJMoa040938

van der Maaten, L., & Hinton, G. (2008). Visualizing data using T-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C., & Beerenwinkel, N. (2019). Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature Communications, 10*(1), 1–12. https://doi.org/10.1038/s41467-019-10737-5

Malikic, S., Mehrabadi, F. R., Ciccolella, S., Rahman, Md K., Ricketts, C., Haghshenas, E., Seidman, D., Hach, F., Hajirasouliha, I., & Sahinalp, S. C. (2019). PhISCS: A combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Research, 29*(11), 1860–1877. https://doi.org/10.1101/gr.234435.118

Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., Kamińska, B., Huelsken, J., Omberg, L., Gevaert, O., Colaprico, A., Czerwińska, P., Mazurek, S., Mishra, L., Heyn, H., Krasnitz, A., Godwin, A. K., Lazar, A. J., Stuart, J. M., … Wiznerowicz, M. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell, 173*(2), 338–354.e15. https://doi.org/10.1016/j.cell.2018.03.034

Markopoulos, C., Hyams, D. M., Gomez, H. L., Harries, M., Nakamura, S., Traina, T., & Katz, A. (2020). Multigene assays in early breast cancer: Insights from recent phase 3 studies. *European Journal of Surgical Oncology, 46*(4), 656–666. https://doi.org/10.1016/j.ejso.2019.10.019

Mattox, A. K., Bettegowda, C., Zhou, S., Papadopoulos, N., Kinzler, K. W., & Vogelstein, B. (2019). Applications of liquid biopsies for cancer. *Science Translational Medicine, 11*(507), eaay1984. https://doi.org/10.1126/scitranslmed.aay1984

McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. In *RecSys 2013 — Proceedings of the 7th ACM conference on recommender systems* (pp. 165–172). https://doi.org/10.1145/2507157.2507163

McGranahan, N., & Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell, 27*(1), 15–26. https://doi.org/10.1016/j.ccell.2014.12.001

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software, 3*(29), 861. https://doi.org/10.21105/joss.00861

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research, 33*(18), 5868–5877. https://doi.org/10.1093/nar/gki901

Mendelsohn, J., & Baselga, J. (2000). The EGF receptor family as targets for cancer therapy. *Oncogene, 19*(56), 6550–6565. https://doi.org/10.1038/sj.onc.1204082

Metzker, M. L. (2010). Sequencing technologies — The next generation. *Nature Reviews Genetics, 11*(1), 31–46. https://doi.org/10.1038/nrg2626

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the panther classification system. *Nature Protocols, 8*(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092

Moffat, J. G., Rudolph, J., & Bailey, D. (2014). Phenotypic screening in cancer drug discovery — Past, present and future. *Nature Reviews Drug Discovery, 13*(8), 588–602. https://doi.org/10.1038/nrd4366

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods, 5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Nakagawa, H., & Fujita, M. (2018). Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Science, 109*(3), 513–522. https://doi.org/10.1111/cas.13505

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., & Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature, 472*(7341), 90–94. https://doi.org/10.1038/nature09807

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Månér, S., Zetterberg, A., Hicks, J., & Wigler, M. (2010). Inferring tumor progression from genomic heterogeneity. *Genome Research, 20*(1), 68–80. https://doi.org/10.1101/gr.099622.109

Nelson, P. T., Baldwin, D. A., Scearce, L. M., Oberholtzer, J. C., Tobias, J. W., & Mourelatos, Z. (2004). Microarray-based, high-throughput gene expression profiling of microRNAs. *Nature Methods, 1*(2), 155–161. https://doi.org/10.1038/nmeth717

Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J., Reed, J., Cheang, M. C. U., Mardis, E. R., Perou, C. M., Bernard, P. S., & Ellis, M. J. (2010). A comparison of PAM50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor–positive breast cancer. *Clinical Cancer Research, 16*(21), 5222–5232. https://doi.org/10.1158/1078-0432.ccr-10-1282

Niu, B., Scott, A. D., Sengupta, S., Bailey, M. H., Batra, P., Ning, J., Wyczalkowski, M. A., Liang, W. W., Zhang, Q., McLellan, M. D., Sun, S. Q., Tripathi, P., Lou, C., Ye, K., Jay Mashl, R., Wallis, J., Wendl, M. C., Chen, F., & Ding, L. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nature Genetics, 48*(8), 827–837. https://doi.org/10.1038/ng.3586

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, T. J., & Naoki, K. (2004). EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy. *Science, 304*(5676), 1497–1500. https://doi.org/10.1126/science.1099314

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, R., Walker, M., Watson, D., Park, T., & Bryant, J. (2003). *Multi-gene RT-PCR assay for predicting recurrence in node negative breast cancer patients-NSABP studies B-20 and B-14* (p. 82). Breast Cancer Research and Treatment.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J., & Wolmark, N. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine, 351*(27), 2817–2826. https://doi.org/10.1056/nejmoa041588

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., & Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology, 27*(8), 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370

Park, Y., Shackney, S., & Schwartz, R. (2009). Network-based inference of cancer progression from microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6*(2), 200–212. https://doi.org/10.1109/TCBB.2008.126

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suva, M. L., Regev, A., & Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science, 344*(6190), 1396–1401. https://doi.org/10.1126/science.1254257

Pereira, B., Chin, S. F., Rueda, O. M., Vollan, H. K. M., Provenzano, E., Bardwell, H. A., Pugh, M., Jones, L., Russell, R., Sammut, S. J., Tsui, D. W. Y., Liu, B., Dawson, S. J., Abraham, J., Northen, H., Peden, J. F., Mukherjee, A., Turashvili, G., Green, A. R., … Caldas, C. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications, 7*. https://doi.org/10.1038/ncomms11479

Perou, C. M., Sørile, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Ress, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergammenschlkov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747–752. https://doi.org/10.1038/35021093

Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hewlett, J., Gerson, N., & Parker, H. L. (2001). Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *Journal of Clinical Oncology, 19*(4), 980–991. https://doi.org/10.1200/JCO.2001.19.4.980

Relling, M. V., & Evans, W. E. (2015). Pharmacogenomics in the clinic. *Nature, 526*(7573), 343–350. https://doi.org/10.1038/nature15817

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., & Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science (New York, N.Y.), 290*(5500), 2306–2309. https://doi.org/10.1126/science.290.5500.2306

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research, 39*(17), e118. https://doi.org/10.1093/nar/gkr407

Russo, A., Franchina, T., Ricciardi, G. R. R., Picone, A., Ferraro, G., Zanghì, M., Toscano, G., Giordano, A., & Adamo, V. (2015). A decade of EGFR inhibition in EGFR-mutated non small cell lung cancer (NSCLC): Old successes and future perspectives. *Oncotarget, 6*(29), 26814–26825. https://doi.org/10.18632/oncotarget.4254

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Schwartz, R., & Schäffer, A. A. (2017). The evolution of tumour phylogenetics: Principles and practice. *Nature Reviews Genetics, 18*(4), 213–229. https://doi.org/10.1038/nrg.2016.170

Schwartz, R., & Shackney, S. E. (2010). Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics, 11*. https://doi.org/10.1186/1471-2105-11-42

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., & Ester, M. (2019). MOLI: Multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics, 35*(14), i501–i509. https://doi.org/10.1093/bioinformatics/btz318. Oxford University Press.

Shoemaker, R. H. (2006). The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer, 6*(10), 813–823. https://doi.org/10.1038/nrc1951

Sicklick, J. K., Kato, S., Okamura, R., Schwaederle, M., Hahn, M. E., Williams, C. B., De, P., Krie, A., Piccioni, D. E., Miller, V. A., Ross, J. S., Benson, A., Webster, J., Stephens, P. J., Lee, J. J., Fanta, P. T., Lippman, S. M., Leyland-Jones, B., & Kurzrock, R. (2019). Molecular profiling of cancer patients enables personalized combination therapy: The I-PREDICT study. *Nature Medicine, 25*(5), 744–750. https://doi.org/10.1038/s41591-019-0407-5

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods, 14*(4), 407–410. https://doi.org/10.1038/nmeth.4184

Sjöblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S. D., Willis, J., Dawson, D., Willson, J. K. V., Gazdar, A. F., … Velculescu, V. E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science, 314*(5797), 268–274. https://doi.org/10.1126/science.1133427

Slamon, D., Eiermann, W., Robert, N., Pienkowski, T., Martin, M., Press, M., Mackey, J., Glaspy, J., Chan, A., Pawlicki, M., Pinter, T., Valero, V., Liu, M.-C., Sauter, G., von Minckwitz, G., Visco, F., Bee, V., Buyse, M., Bendahmane, B., … Crown, J. (2011). Adjuvant trastuzumab in HER2-positive breast cancer. *New England Journal of Medicine, 365*(14), 1273–1283. https://doi.org/10.1056/NEJMoa0910383

de Souto, M. C. P., Jaskowiak, P. A., & Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. *BMC Bioinformatics, 16*(1). https://doi.org/10.1186/s12859-015-0494-3

Sparano, J. A., Gray, R. J., Makower, D. F., Albain, K. S., Saphner, T. J., Badve, S. S., Wagner, L. I., Kaklamani, V. G., Keane, M. M., Gomez, H. L., Reddy, P. S., Goggins, T. F., Mayer, I. A., Toppmeyer, D. L., Brufsky, A. M., Goetz, M. P., Berenberg, J. L., Mahalcioiu, C., Desbiens, C., … Sledge, G. W. (2020). Clinical outcomes in early breast cancer with a high 21-gene recurrence score of 26 to 100 assigned to adjuvant chemotherapy plus endocrine therapy: A secondary analysis of the TAILORx randomized clinical trial. *JAMA Oncology, 6*(3), 367–374. https://doi.org/10.1001/jamaoncol.2019.4794

Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., Geyer, C. E., Dees, E. C., Goetz, M. P., Olson, J. A., Lively, T., Badve, S. S., Saphner, T. J., Wagner, L. I., Whelan, T. J., Ellis, M. J., Paik, S., Wood, W. C., Ravdin, P. M., … Sledge, G. W. (2018). Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine, 379*(2), 111–121. https://doi.org/10.1056/NEJMoa1804710

Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., Geyer, C. E., Dees, E. C., Perez, E. A., Olson, J. A., Zujewski, J. A., Lively, T., Badve, S. S., Saphner, T. J., Wagner, L. I., Whelan, T. J., Ellis, M. J., Paik, S., Wood, W. C., … Sledge, G. W. (2015). Prospective validation of a 21-gene expression assay in breast cancer. *New England Journal of Medicine, 373*(21), 2005–2014. https://doi.org/10.1056/NEJMoa1510764

Sparano, J. A., & Paik, S. (2008). Development of the 21-gene assay and its application in clinical practice and clinical trials. *Journal of Clinical Oncology, 26*(5), 721–728. https://doi.org/10.1200/jco.2007.15.1068

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Fernández Navarro, J., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., … Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science, 353*(6294), 78–82. https://doi.org/10.1126/science.aaf2403

Suvà, M. L., & Tirosh, I. (2019). Single-cell RNA sequencing in cancer: Lessons learned and emerging challenges. *Molecular Cell, 75*(1), 7–12. https://doi.org/10.1016/j.molcel.2019.05.003

Swanton, C. (2012). Intratumor heterogeneity: Evolution through space and time. *Cancer Research, 72*(19), 4875–4882. https://doi.org/10.1158/0008-5472.CAN-12-2217

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570. https://doi.org/10.1142/S0218488502001648

Tao, Y., Cai, C., Cohen, W. W., & Lu, X. (2020). From genome to phenome: Predicting multiple cancer phenotypes based on somatic genomic alterations via the genomic impact transformer. *Pacific Symposium on Biocomputing, 25*(2020), 79–90. World Scientific Publishing Co. Pte Ltd http://psb.stanford.edu/.

Tao, Y., Lei, H., Fu, X., Lee, A. V., Ma, J., & Schwartz, R. (2020). Robust and accurate deconvolution of tumor populations uncovers evolutionary mechanisms of breast cancer metastasis. *Bioinformatics, 36*(S1), i407–i416. https://doi.org/10.1093/bioinformatics/btaa396

Tao, Y., Lei, H., Lee, A. V., Ma, J., & Schwartz, R. (2019). Phylogenies derived from matched transcriptome reveal the evolution of cell populations and temporal order of perturbed pathways in breast cancer brain metastases. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 11826, pp. 3–28). Springer. https://doi.org/10.1007/978-3-030-35210-3_1

The International Cancer Genome Consortium, The Cancer Genome Atlas Consortium, & The Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. *Nature, 578*(7793), 82–93.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine, 16*(4), 385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3

Tinker, A. V., Boussioutas, A., & Bowtell, D. D. L. (2006). The challenges of gene expression microarrays for the study of human cancer. *Cancer Cell, 9*(5), 333–339. https://doi.org/10.1016/j.ccr.2006.05.001

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J. R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., … Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science, 352*(6282), 189–196. https://doi.org/10.1126/science.aad0501

Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Zhen, Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M., & Feolo, M. (2014). NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Research, 42*(D1), D975–D979.

Vandin, F., Upfal, E., & Raphael, B. J. (2012). De novo discovery of mutated driver pathways in cancer. *Genome Research, 22*(2), 375–385. https://doi.org/10.1101/gr.120477.111

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Bernards, R., & Friend, S. H. (2003). Expression profilling predicts outcome in breast cancer. *Breast Cancer Research, 5*(1), 57–58. https://doi.org/10.1186/bcr562

van't Veer, L. J., Dai, H., Van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530–536. https://doi.org/10.1038/415530a

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science, 270*(5235), 484–487. https://doi.org/10.1126/science.270.5235.484

Venet, D., Pecasse, F., Maenhaut, C., & Bersini, H. (2001). Separation of samples into their constituents using gene expression data. *Bioinformatics, 17*(1), S279–S287. https://doi.org/10.1093/bioinformatics/17.suppl_1.S279

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The sequence of the human genome. *Science (New York, N.Y.), 291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schrieber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., & Parrish, M. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine, 347*(25), 1999–2009.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science, 340*(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Wang, Z., Ng, K.-S., Chen, T., Kim, T.-B., Wang, F., Shaw, K., Scott, K. L., Meric-Bernstam, F., Mills, G. B., Chen, K., & Bader, J. S. (2018). Cancer driver mutation prediction through Bayesian integration of multi-omic data. *PLoS One, 13*(5), e0196939. https://doi.org/10.1371/journal.pone.0196939

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., & Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods, 14*(4), 414–416. https://doi.org/10.1038/nmeth.4207

Wei, D., Liu, C., Zheng, X., & Li, Y. (2019). Comprehensive anticancer drug response prediction based on a simple cell line-drug complex network model. *BMC Bioinformatics, 20*(1), 44. https://doi.org/10.1186/s12859-019-2608-9

Weinmann, S., Leo, M. C., Francisco, M., Jenkins, C. L., Barry, T., Leesman, G., Linke, S. P., Whitworth, P. W., Patel, R., & Pellicane, J. (2020). Validation of a ductal carcinoma in situ biomarker profile for risk of recurrence after breast-conserving surgery with and without radiotherapy. *Clinical Cancer Research, 26*(15), 4054–4063. https://clincancerres.aacrjournals.org/content/26/15/4054.full.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chu, A., … Kling, T. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics, 45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*(1).

Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., & Schneider, M. V. (2014). Bioinformatics curriculum guidelines: Toward a definition of core competencies. *PLoS Computational Biology, 10*(3), e1003496. https://doi.org/10.1371/journal.pcbi.1003496

Woodburn, J. R. (1999). The epidermal growth factor receptor and its inhibition in cancer therapy. *Pharmacology & Therapeutics, 82*(2–3), 241–250. https://doi.org/10.1016/s0163-7258(98)00045-x

Xing, E. P., Jordan, M. I., & Karp, R. M. (2001). Feature selection for high-dimensional genomic microarray data. In *Proceedings of the eighteenth international conference on machine learning*.

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C., McDermott, U., & Garnett, M. J. (2013). Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research, 41*(1), D955–D961. https://doi.org/10.1093/nar/gks1111

Yardley, D. A., Noguchi, S., Pritchard, K. I., Burris, H. A., Baselga, J., Gnant, M., Hortobagyi, G. N., Campone, M., Pistilli, B., Piccart, M., Melichar, B., Petrakova, K., Arena, F. P., Erdkamp, F., Harb, W. A., Feng, W., Cahana, A., Taran, T., Lebwohl, D., & Rugo, H. S. (2013). Everolimus plus exemestane in postmenopausal patients with HR+ breast cancer: BOLERO-2 final progression-free survival analysis. *Advances in Therapy, 30*(10), 870–884. https://doi.org/10.1007/s12325-013-0060-1

Yuan, H., Paskov, I., Paskov, H., González, A. J., & Leslie, C. S. (2016). Multitask learning improves prediction of cancer drug sensitivity. *Scientific Reports, 6*. https://doi.org/10.1038/srep31619

Yuan, H., Yan, M., Zhang, G., Liu, W., Deng, C., Liao, G., Xu, L., Luo, T., Yan, H., Long, Z., Shi, A., Zhao, T., Xiao, Y., & Li, X. (2019). CancerSEA: A cancer single-cell state atlas. *Nucleic Acids Research, 47*(D1), D900–D908. https://doi.org/10.1093/nar/gky939

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., Liu, X. S., & Leslie, C. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Computational Biology, 11*(9), e1004498. https://doi.org/10.1371/journal.pcbi.1004498

Zhang, W., Xia, W., Lv, Z., Ni, C., Xin, Y., & Yang, L. (2017). Liquid biopsy for cancer: Circulating tumor cells, circulating free DNA or exosomes? *Cellular Physiology and Biochemistry, 41*(2), 755–768. https://doi.org/10.1159/000458736

Zhang, Z., Zhou, L., Xie, N., Nice, E. C., Zhang, T., Cui, Y., & Huang, C. (2020). Overcoming cancer therapeutic bottleneck by drug repurposing. *Signal Transduction and Targeted Therapy, 5*(1), 1–25. https://doi.org/10.1038/s41392-020-00213-8

Zhao, G., & Wu, Y. (2016). Feature subset selection for cancer classification using weight local modularity. *Scientific Reports, 6*, 34759. https://doi.org/10.1038/srep34759

Zheng, C., Zheng, L., Yoo, J.-K., Guo, H., Zhang, Y., Guo, X., Kang, B., Hu, R., Huang, J. Y., Zhang, Q., Liu, Z., Dong, M., Hu, X., Ouyang, W., Peng, J., & Zhang, Z. (2017). Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell, 169*(7), 1342–1356.e16. https://doi.org/10.1016/j.cell.2017.05.035

Zhu, L., Narloch, J. L., Onkar, S., Joy, M., Broadwater, G., Luedke, C., Hall, A., Kim, R., Pogue-Geile, K., Sammons, S., Nayyar, N., Chukwueke, U., Brastianos, P. K., Anders, C. K., Soloff, A. C., Vignali, D. A. A., Tseng, G. C., Emens, L. A., Lucas, P. C., … Lee, A. V. (2019). Metastatic breast cancers have reduced immune cell recruitment but harbor increased macrophages relative to their matched primary tumors. *Journal for Immuno Therapy of Cancer, 7*, 265. https://doi.org/10.1186/s40425-019-0755-1