PeerJ Computer Science

Machine learning analysis of TCGA cancer data

Jose Liñares-Blanco[1,2], Alejandro Pazos[1,2,3] and Carlos Fernandez-Lozano[1,2,3]

[1] CITIC-Research Center of Information and Communication Technologies, University of A Coruna, A Coruña, Spain
[2] Department of Computer Science and Information Technologies, Faculty of Computer Science, University of A Coruna, A Coruña, Spain
[3] Grupo de Redes de Neuronas Artificiales y Sistemas Adaptativos. Imagen Médica y Diagnóstico Radiológico (RNASA-IMEDIR). Complexo Hospitalario Universitario de A Coruña (CHUAC), SERGAS, Universidade da Coruña, Instituto de Investigación Biomédica de A Coruña (INIBIC), A Coruña, Spain

ABSTRACT

In recent years, machine learning (ML) researchers have changed their focus towards biological problems that are difficult to analyse with standard approaches. Large initiatives such as The Cancer Genome Atlas (TCGA) have allowed the use of omic data for the training of these algorithms. In order to study the state of the art, this review is provided to cover the main works that have used ML with TCGA data. Firstly, the principal discoveries made by the TCGA consortium are presented. Once these bases have been established, we begin with the main objective of this study, the identification and discussion of those works that have used the TCGA data for the training of different ML approaches. After a review of more than 100 different papers, it has been possible to make a classification according to following three pillars: the type of tumour, the type of algorithm and the predicted biological problem. One of the conclusions drawn in this work shows a high density of studies based on two major algorithms: Random Forest and Support Vector Machines. We also observe the rise in the use of deep artificial neural networks. It is worth emphasizing, the increase of integrative models of multi-omic data analysis. The different biological conditions are a consequence of molecular homeostasis, driven by both protein coding regions, regulatory elements and the surrounding environment. It is notable that a large number of works make use of genetic expression data, which has been found to be the preferred method by researchers when training the different models. The biological problems addressed have been classified into five types: prognosis prediction, tumour subtypes, microsatellite instability (MSI), immunological aspects and certain pathways of interest. A clear trend was detected in the prediction of these conditions according to the type of tumour. That is the reason for which a greater number of works have focused on the BRCA cohort, while specific works for survival, for example, were centred on the GBM cohort, due to its large number of events. Throughout this review, it will be possible to go in depth into the works and the methodologies used to study TCGA cancer data. Finally, it is intended that this work will serve as a basis for future research in this field of study.

Submitted 3 March 2021
Accepted 17 May 2021
Published 12 July 2021

Corresponding author
Carlos Fernandez-Lozano,
carlos.fernandez@udc.es

Academic editor
Li Zhang

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj-cs.584**

Subjects Bioinformatics, Computational Biology, Artificial Intelligence, Data Mining and Machine Learning
Keywords Cancer, Machine learning, TCGA, Multi-omics, Data integration, BRCA, Random Forest, Support Vector Machines

INTRODUCTION

The appearance of the carcinogenic phenotype is the consequence of an alteration of one or more genes. In addition, the appearance of subtypes occurs in different ways in individuals of a population. Hence, a major problem that arises in cancer is the difficulty in its genetic diagnosis. Similar to Mendelian diseases, where the disease develops due to the alteration in the function of a single gene, the development of cancer is a consequence of epistatic behaviour of genes. There is already an extremely large search space in the identification of alterations in a single gene, including exonic and intronic mutations, single nucleotide polymorphisms (SNPs), copy number variants, indels, post-transcriptional alterations, post-translational alterations, three-dimensional assembly of the protein, epigenetic modifications, etc. Thus, the search space for alterations when we encounter a subgroup of 40 genes is immense. When we do not know exactly which genes are involved, we have to search among the more than 20,000 coding regions or even in whole genome sequence. In these cases the search space grows to incalculable levels. All this complexity is the result of intermolecular communications in and among cells, a phenomenon that constitutes an environment of molecular communication that is extremely complicated to understand and identify.

In order to lay the foundation and achieve great advances in the prevention, early detection, stratification and success in the treatment of cancer, it is necessary to identify the complete changes generated by each type of cancer in its genome. Further, researchers must understand how these changes interact with the cancer microenvironment, intra- and intercellularly, to manifest itself. Hence, the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) of the United States established The Cancer Genome Atlas (TCGA), with the aim of obtaining comprehensive multidimensional genomic maps of all key changes in several types and subtypes of cancer (*The Cancer Genome Atlas Research Network, 2008*). An initial pilot project in 2006 confirmed that an atlas of these changes could be specifically created for different types of cancer. Subsequently, TCGA has collected tissues from more than 11,000 cancer and healthy patients, an endeavour that allows the study of more than 33 types and subtypes of cancer, including 10 rare cancers. The most interesting aspect of this initiative is that all the information is free and accessible to any researcher who wants to focus their efforts on the disease. The different types of data presented by the TCGA project are summarised in Table 1 and Fig. 1 shows, for each cancer type, the percentage that each data type represents in the subtype's total. Data are provided open access to the community, a factor that facilitates the generation of novel models without requiring an initial financial investment to obtain the data. Therefore, there are increasingly specific models for the analysis of omics data. In particular, the rise and success of machine learning (ML) techniques to process a large amount of data is revolutionising bioinformatics and

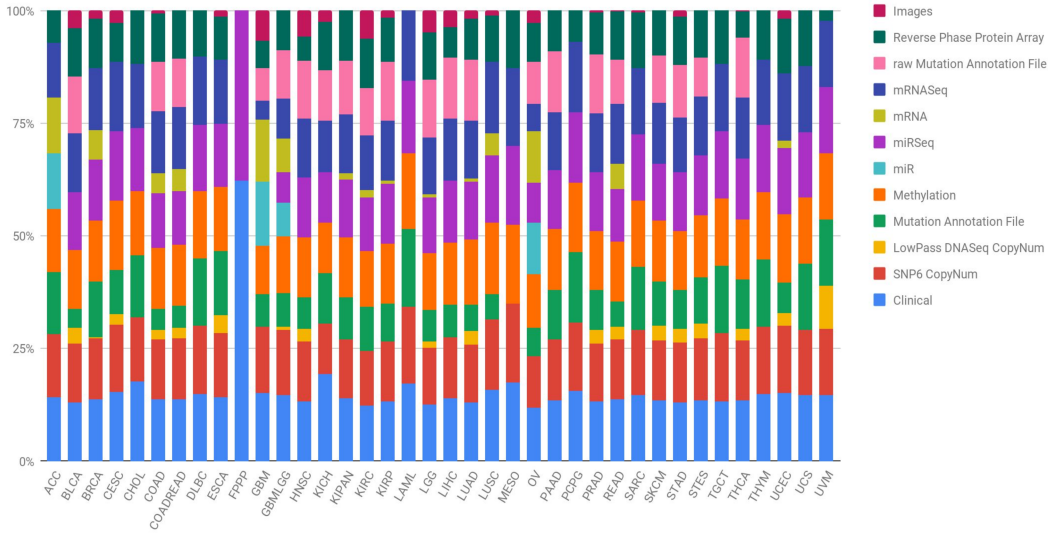


Figure 1 Quantification of the number of samples in the TCGA repository, classified by type of tumour and type of biotechnological analysis. Clin, Clinical; SNP6, SNP6 CopyNum; DNAseq, Low-Pass DNASeq CopyNum; Mutat, Mutation Annotation File; Met, Methylation; rawMut, rawMutation Annotation File; Prot, Reverse Phase Protein Array. Full-size DOI: 10.7717/peerj-cs.584/fig-1

conventional forms of genetic diagnosis. These methods have focused on making predictions by using general learning algorithms to find patterns in complex, larger and hard-to-handle problems. In addition, these ML methods work really well with very large datasets, even when the number of variables in each observation is much greater than the total number of observations ($n << p$).

This survey presents the state-of-the-art research on TCGA analysis using machine learning. Efforts have involved both supervised and unsupervised learning problems, as well as survival analysis, disease prognosis, cancer staging and pathways analysis to analyse different types of data ranging from multi-omics human cancer data to imaging. Therefore, review articles are needed to show an overview of machine learning-based analysis of TCGA data to highlight the findings and to discuss future research lines so that the obtained knowledge is useful and can be translated to clinical practice.

There are few published review articles on machine learning for biomedical genomic analysis (*Leung et al., 2015*; *Karczewski & Snyder, 2018*). These review articles are before 2018 and do not present a discussion on TCGA data nor a discussion on machine learning results neither present a multi-omic and imaging point of view for different biological questions. To the best of our knowledge, no survey has been conducted on Machine Learning analysis of TCGA using multi-level cancer data. Thus, this survey aims to present a comprehensive summary of the previous machine learning approaches applied to TCGA during the span of 2008-2020. The contributions of this review are:

- This review includes exhaustive review of the main results obtained by the TCGA consortium using conventional approaches in order to understand if machine learning is increasing the knowledge in the area.
- This review includes machine learning results by the TCGA consortium.

Table 1 Different types of data present in the TCGA repository.

DNA Sequencing	Whole genome sequences
	Whole exome sequences
	Sequences traces
	Mutations, including coding, splice site, germline and noncoding somatic variants
RNA sequencing	mRNA sequences (calculated expression per gene, exon, splice junction and isoform)
	miRNA sequences (calculated expression per miNRA and isoform)
	Total RNA sequences (calculated expression per gene, exon, splice junction and isoform)
	Expression signals per gene, exon, splice junction, miRNA and isoform
Copy number	Arrays (raw, unnormalized, normalized)
	Low-pass DNA sequencing (whole genomes sequences, variants and coverage)
Array-based expression	Gene expression (raw, normalized and calls)
	Exon expression (raw, normalized and calls)
	miRNA expression (raw, normalized and calls)
DNA methylation	Array-based methylation (raw signal intensity, calculated beta values)
Other	Protein expression (high-resolution images of protein arrays, raw signals, normalized expression and mass spectrometry protein)
	Microsatelite instability (markers and classification)
	ATAC-seq (chromatine accesibility)
Metadata	Clinical information about patients (e.g., sex, race, ethnicity, drugs taken, metastasis status and response to treatment)
	Information about samples (e.g., the weight of a sample portion, days to collect and time of freezing)
	Images of the tumors

- A classification of supervised, unsupervised and clustering methods that may point researchers to new approaches or new problems.
- Identification of data types mostly used in machine learning research of TCGA.
- A comprehensive discussion on biological questions solved by machine learning algorithms: prognosis, immunological phenotype, pathways, MSI status, and subtype prediction.
- A deeper examination of the most used TCGA cohort: Breast Cancer Adenocarcinoma (BRCA).
- We point data integration approaches as the future trend in TCGA analysis using machine learning.

We believe that researchers in machine learning, bioinformatics, biology, computational biology and data integration would benefit from the findings of this exhaustive and comprehensive review.

This manuscript is organised as follows. "Survey Methodology" explains the methodology used in this survey. "TCGA Consortium" presents the main results obtained by the TCGA consortium. In "Machine Learning as a Source of New Knowledge", we review the TCGA efforts with those algorithms as well as we present the most used

algorithms on supervised, unsupervised and clustering approaches for external researchers. Special attention with a subsection on medical imaging analysis using deep learning approaches in recent years. "Biological Questions Solved by Machine Learning Algorithms" discusses the capability of those algorithms to solve the biological problem with the highest performance score and find that the predictions are biologically of relevance. To this aim we divide and study five biological problems: prognosis, immunological phenotype, pathways, MSI and subtypes prediction. We finish with special emphasis on the analysis of the BRCA cohort. Finally, we conclude the review in 'Conclusions'.

SURVEY METHODOLOGY

This work is based on a literature review in machine learning-based analysis of TCGA cancer data. We searched for the main findings of the TCGA consortium using classical statistical approaches and works using machine learning and classify them into supervised, unsupervised and clustering methods. Furthermore, we considered of relevance to answer to the intitial biological question with sense, not only with a higher performance score. The search keywords, data sources and on criteria are discussed.

Search keywords

We initially reviewed the original TCGA consortim publication in order to carefully select the search keywords. The keywords used for the survey included the following terms to find the relevant papers: 'machine learning', 'TCGA'. We used the 'AND' and 'OR' Boolean operators to combine terms. After the initial subset of papers we refined the search keywords according with the most used machine learning models, type of problems and biological question: 'clustering', 'computer vision', 'deep learning', 'random forest', 'support vector machines', 'linear model', 'survival', 'MSI', 'prognosis', 'pathway', 'subtypes' or 'phenotype'.

Data sources

The papers included in this survey were retrieved from prominent journals indexed in diverse quality databases: Pubmed and Scopus.

Article inclusion/exclusion criteria

We decided which articles are eligible for the survey under the following inclusion/exclusion criteria:

- Inclusion criteria:
  - manuscripts written in the English language and published by indexed journals in Pubmed to ensure the health science specialization and Scopus using TCGA as the main source of data
- Exclusion criteria:
  - manuscripts using machine learning marginally or without solid biological conclusions
  - manuscripts in preprint without peer review

Article selection

The TCGA consortium papers were identified in the website and were included. Initially 345 papers were identified in Pubmed and Scopus using the search keywords. Of these, we filtered by the inclusion/exclusion criteria. In addition, duplicated papers retrieved from multiple sources were removed. Finally, more than 150 articles were included.

TCGA CONSORTIUM

TCGA began as a pilot project for 3 years, with a focus on the characterisation of three types of human cancer: glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC) and ovarian cancer (OV). TCGA currently presents data from a total of 38 different cohorts. Four of them (COADREAD, GBMLGG, KIPAN and STES) are not original—they are combinations of other cohorts. Among the remaining 34 cancer cohorts are tumours of different tissue types, as can be seen in Table 2. To date, TCGA has characterised and published about 33 different types of tumours in leading international journals. Table 2 provides greater depth for each of the publications that TCGA has made in each recruited cohort.

In 2018, a series of works were published in Cell editorial, where they were exhaustively analysed the samples recruited throughout the project. These studies led to the identification and examination of mechanisms that underlie all types of tumours. These findings allow researchers to draw conclusions about tumour origins, molecular biology and subtyping. In this series of publications—and in order to understand the molecular biology underlying cancer—the TCGA consortium cross-checked general molecular aspects in all tumour types. To this end, they exhaustively studied, in the more than 10,000 samples stored in their repository, the process of alternative splicing (*Kahles et al., 2018*) and they identified the specific variants (*Huang et al., 2018*) and driver genes (*Bailey et al., 2018*) that generate greater predisposition to tumour development. They also analysed the effect of enhancer activation on different tumour types (*Chen et al., 2018a*) and the effect of aneuploidy (*Taylor et al., 2018*). They also catalogued the variants of the 10 pathways that are most frequently altered in most tumours (*Sanchez-Vega et al., 2018*), in addition to alterations in genes related to the ubiquitin (*Ge et al., 2018*), DNA damage repair (*Knijnenburg et al., 2018*) and the MYC pathways (*Schaub et al., 2018*).

The consortium also features a strong technology component; they published an integrated pancancerous clinical data resource from TCGA with the aim of driving the analysis of high-quality survival results (*Liu et al., 2018a*). In addition, they conducted studies where they used ML and deep learning algorithms to identify stemness features in tumour cells (*Malta et al., 2018*), the prediction of Ras pathway activation (*Way et al., 2018*) and the detection of tumour infiltrating lymphocytes using images (*Saltz et al., 2018*). In *Ellrott et al. (2018)* they described the Multi-Center Mutation Calling project, which aims to generate a complete encyclopaedia of somatic mutations from TCGA data that allows a robust analysis for different tumour types. They performed different studies that proposed new classifications among tumours. For example, they identified new immune tumour types across the 33 types of cancer that differ by somatic aberrations,

Table 2 Enumeration of the different cohorts presented by the TCGA repository, classified according tothe tissue of origin of the tumour. In addition, the original paper published by the TCGA consortium is cited.

Cancer type	Acronym	Tissue	Citation
Breast Ductal/Lobular Carcinoma	BRCA	Breast	(*The Cancer Genome Atlas Network, 2012b*; *Ciriello et al., 2015*)
Glioblastoma Multiforme	GBM	Central Nervous System	(*The Cancer Genome Atlas Research Network, 2008*; *Verhaak et al., 2010*; *Noushmehr et al., 2010*; *Brennan et al., 2013*; *The Cancer Genome Atlas Research Network, 2015*, *Ceccarelli et al., 2016*)
Lower Grade Glioma	LGG	Central Nervous System	(*The Cancer Genome Atlas Research Network, 2015*)
Adrenocortical Carcinoma	ACC	Endocrine	(*Zheng et al., 2016*)
Papillary Thyroid Carcionma	THCA	Endocrine	(*Agrawal et al., 2014*)
Paraganglioma & Pheochromocytoma	PCPG	Endocrine	(*Fishbein et al., 2017*)
Cholangiocarcinoma	CHOL	Gastrointestinal	(*Farshidfar et al., 2017*)
Colon Adenocarcinoma	COAD	Gastrointestinal	(*The Cancer Genome Atlas Network, 2012a*)
Rectal Adenocarcinoma	READ	Gastrointestinal	(*The Cancer Genome Atlas Network, 2012a*)
Esophageal Cancer	ESCA	Gastrointestinal	(*The Cancer Genome Atlas Research Network, 2017c*)
Liver Hepatocellular Carcionoma	LIHC	Gastrointestinal	(*Ally et al., 2017*)
Pancreatic Ductal Adenocarcinoma	PAAD	Gastrointestinal	(*Raphael et al., 2017*)
Stomach Cancer	STAD	Gastrointestinal	(*The Cancer Genome Atlas Research Network, 2014a*)
Cervical Cancer	CESC	Gynecologic	(*The Cancer Genome Atlas Research Network, 2017b*)
Ovarian Serous Cystadenocarcinoma	OV	Gynecologic	(*The Cancer Genome Atlas Research Network, 2011*)
Uterine Carcinosarcoma	UCS	Gynecologic	(*Cherniack et al., 2017*)
Uterine Corpus Endometrial Carcinoma	UCEC	Gynecologic	(*Levine, 2013*)
Head and Neck Squamous Cell Carcinoma	HNSC	Head and Neck	(*The Cancer Genome Atlas Network, 2015*)
Uveal Melanoma	UVM	Head and Neck	(*Robertson et al., 2017b*)
Acute Myeloid Leukemia	AML	Hematologic	(*The Cancer Genome Atlas Research Network, 2013*)
Thymoma	THYM	Hematologic	(*Radovich et al., 2018*)
Cutaneous Melanoma	SKCM	Skin	(*Akbani et al., 2015*)
Sarcoma	SARC	Soft Tissue	(*The Cancer Genome Atlas Research Network, 2017a*)
Lung Adenocarcinoma	LUAD	Thoracic	(*The Cancer Genome Atlas Research Network, 2014c*; *Campbell et al., 2016*)
Lung Squamous Cell Carcinoma	LUSC	Thoracic	(*The Cancer Genome Atlas Research Network, 2012c*; *Campbell et al., 2016*)
Mesothelioma	MESO	Thoracic	(*Hmeljak et al., 2018*)
Chromophobe Renal Cell Carcinoma	KICH	Urologic	(*Davis et al., 2014*)
Clear Cell Kidney Carcinoma	KIRC	Urologic	(*The Cancer Genome Atlas Research Network, 2013*)

(Continued)

Table 2 (continued)

Cancer type	Acronym	Tissue	Citation
Papillary Kidney Carcinoma	KIRP	Urologic	(*The Cancer Genome Atlas Research Network, 2016*)
Prostate Adenocarcinoma	PRAD	Urologic	(*Abeshouse et al., 2015*)
Testicular Germ Cell Cancer	TGCT	Urologic	(*Shen et al., 2018*)
Urothelial Bladder Carcinoma	BLCA	Urologic	(*The Cancer Genome Atlas Research Network, 2014b*; *Robertson et al., 2017a*)
Diffuse Large B-cell Lymphoma	DLBC	Lymphatic tissue	

microenvironment and survival (*Thorsson et al., 2018*). Furthermore, they classified tumours based on metabolic expression and subsequently proposed different subtypes that were not previously contemplated (*Peng et al., 2018*). In addition, they carried out exhaustive studies on groupings of tumours according to their origin in order to elucidate new therapeutic targets that might be useful for gastrointestinal adenocarcinomas (*Liu et al., 2018b*), gynaecological tumours and breast cancers (*Berger et al., 2018*) and squamous carcinomas (*Campbell et al., 2018*). In these papers, they performed clustering techniques to subtype patients into new groups for treatment or diagnosis. Finally, they studied tumours by cell (*Hoadley et al., 2018*) and tissue (*Hoadley et al., 2014*) of origins.

There are many results reported by TCGA that have had a very important impact on oncology. The results obtained by the consortium show a roadmap to follow and open countless avenues in this field where new research groups, until now unable to carry out their research globally, will be able to report important results in this field.

MACHINE LEARNING AS A SOURCE OF NEW KNOWLEDGE

ML is the process by which machines acquire the ability to learn an action or behaviour. These processes are defined by different algorithms that enable the computer to learn a behaviour (classify, identify, etc.) and extract patterns from the data. These patterns are ultimately inherent knowledge of the problem to be analysed that the algorithms can extract and learn to identify. Subsequently, given a new case, these techniques can evaluate and predict to which group it is most likely to belong, always in accordance with prior knowledge. It is therefore critical that such techniques are applied with careful experimental design (*Fernandez-Lozano et al., 2016*) and that the data are as accurate as possible to define the problem. These techniques will learn and maximally exploit the intrinsic knowledge that underlies the data.

Depending on how this information extraction process is performed, we can speak of different approaches: supervised and unsupervised learning. Although in practice there are more types of learning, we will only focus on these two, mainly because these approaches have been the most widely used in biomedicine.

The TCGA consortium and ML

The TCGA consortium has analysed cancer based on ML algorithms, sometimes with novel approaches specifically designed for the TCGA data. TCGA researchers recently presented a new ML that can predict the differentiation of certain tumour tissues (*Malta et al., 2018*). In this case, using data from non-differentiated stem cells and their differentiated progenitors (data obtained from public repositories), they constructed two classes of indicators that reflect epigenetic and genetic expression traits of the cells. Once they constructed these descriptors, they used a variant of one-class logistic regression to classify the different TCGA samples according to their degree of differentiation, a crucial characteristic for the development of the tumour and its invasive potential.

Another study (*Way et al., 2018*) used three types of omics platforms (expression, copy number and mutation) to predict the activation of the Ras pathway, which has been widely studied throughout oncological research. This model predicted whether this pathway was activated using RNAseq expression data. From the copy number and mutation data, the researchers were able to label the patients to design a supervised learning problem. Therefore, it was observed that certain omic patterns could be predicted from different omic data. This enables the prediction of a significant number of characteristics in tumours. This approach was also performed in another study by modifying the target in order to predict the activation of the TP53 pathway (*Knijnenburg et al., 2018*).

In other study, deep learning based on convolutional neural networks (CNN) mapped tumour infiltrating lymphocytes (TIL) based on haematoxylin and eosin (H&E) images. In this case, 13 types of TCGA tumours exhibited almost perfect performance when differentiating these cell types (*Saltz et al., 2018*). In this work, the TCGA consortium highlighted the importance of the images it stores and questions their relatively limited use by different researchers in comparison with omics platforms. The images in the TCGA repository will be discussed in the following sections.

Popular ML models with TCGA data

The TCGA consortium has relied on both supervised and unsupervised ML techniques to extract new knowledge from its data. However, it is interesting to identify the work developed by other researchers who have used TCGA data. The approaches taken and the results obtained from the various published works will be discussed below.

Figure 2 shows the proportion of published papers according to the type of algorithm and the type of omic data used. We reviewed more than 100 papers that have used ML approaches with TCGA data. For each one, we identified: the algorithm and data type/s. Almost half of the identified works used variants of the support vector machine (SVM) or tree-based algorithms, followed by linear models as can be seen in Fig. 2A. On the other hand, Fig. 2B clearly shows that gene expression data is most abundant data type used in ML research. Other data types such as images, methylation, miRNA and copy number have been used, but majority in a combination with gene expression data.

The findings of this review highlight the low variability of reported research and analytical methods. It is true that the mostly used algorithms, Random Forest (RF) and SVM, as well as the types of omic data (expression) have reported promising results in

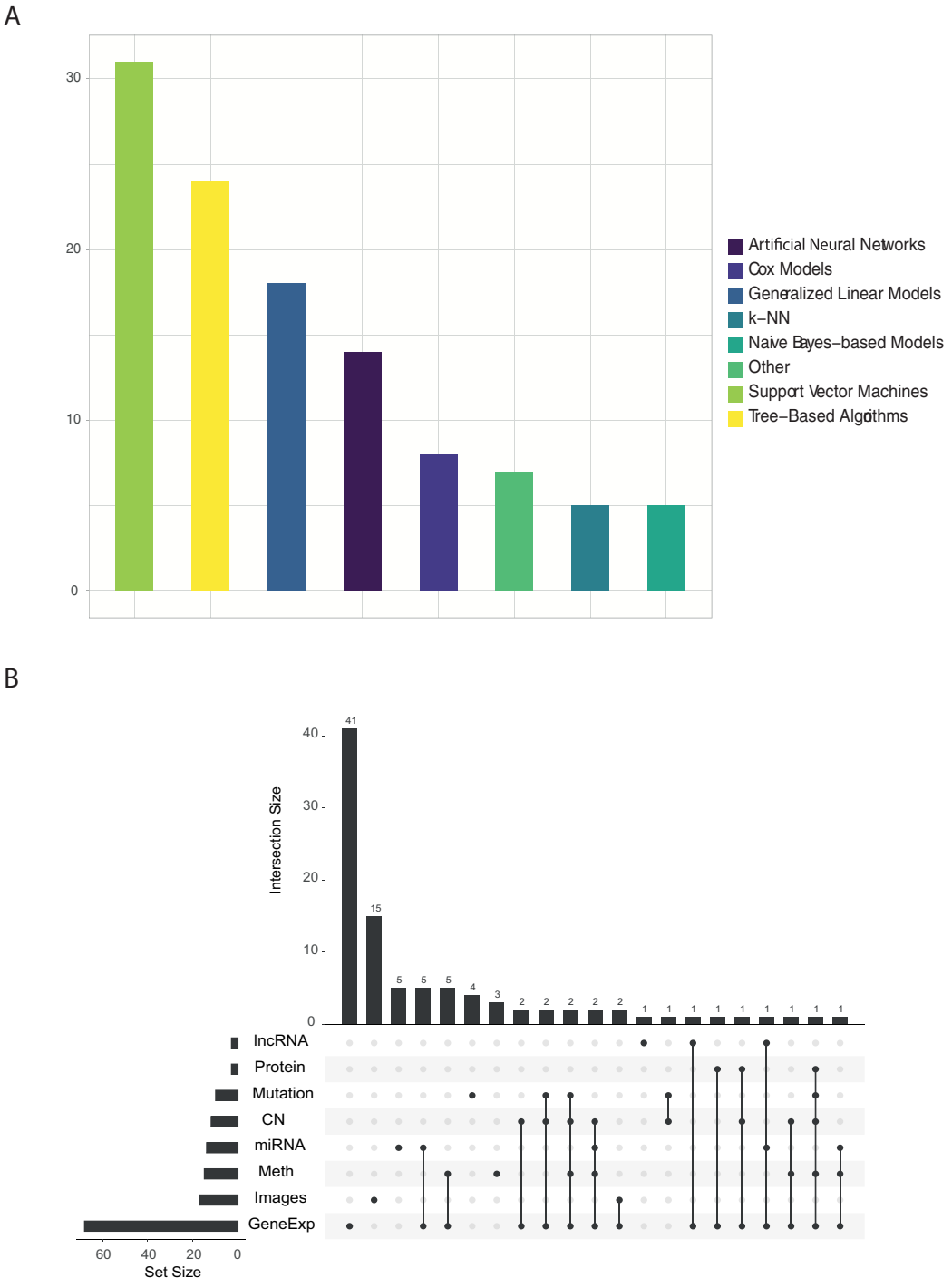


Figure 2 (A) Number of papers that used each type of algorithm, and (B) relations between omics data used in each work. Full-size DOI: 10.7717/peerj-cs.584/fig-2

the biomedical field during the last years. We believe that the low variability in the approaches established by researchers is mainly due to two reasons. First, the intrinsic characteristics of biomedical data, and specifically the omic data, present a much greater number of characteristics than observations. This fact is generally not idyllic for the

training of ML algorithms. In this sense, the use of algorithms is mainly determined by the use of which type of omic data is being analyzed. In the context mentioned above, certain algorithms are able to handle some characteristics of the data better than others. For example, neural networks are more sensitive to the lack of observations than in this case RF, SVM or linear models. Given that the vast majority of works identified have used expression data, it is logical to observe a high density of works that have used RF or SVM type algorithms. On the other hand, those works that have used image data are more likely to use neuron networks. Secondly, there is no doubt that the possibilities in the exploitation of these data by ML algorithms are yet to be discovered. A break in the arrival of ML-based applications in the field of biomedicine has been detected. This is partly due to the complexity of the omic data, and the need for specialists in this field for its modelling and good practical use. Possible applications that could revolutionize the field of biomedicine could be the use of NLP (Natural Language Processing) algorithms for the analysis of Whole Genome Sequencing (WGS) data.

After all, if there is something to highlight in the results observed in the Fig. 2B is the trend towards more and more work integrating different omic data. Even so, this trend is not reflected in Fig. 2A, in which a variety of algorithms and/or new known and standardized methodologies that can solve this problem are not observed. This is the great challenge in the coming years presented by biomedicine, which could generate very useful predictions for tackling complex diseases, such as cancer.

A general perspective of unsupervised learning with TCGA data

In oncology, clustering methods are extremely useful for subtyping or reclassification of patients in a particular cohort. Over the years, the classic clustering methods have been most widely used, including partitioning clustering or hierarchical clustering. Even today, they are widely used with their respective variations. For example, the TCGA consortium has used them to subtype different tumours (see publications in Table 2). The problem with these algorithms is that they can only model a single set of data and the concatenation of different types of data does not perform adequately. The complexity of the tumour is manifested at distinct biological levels; hence, methods that can accept different types of data are preferable. Thus, researchers developed a new integrative clustering method based on a joint latent variable model (iCluster) (*Shen, Olshen & Ladanyi, 2009*) and used it with TCGA data (*Shen et al., 2012*). *iCluster* fits a regularized latent variable model based clustering that generates an integrated cluster assigment based on joint inference across data types. In addition, the implementation in several programming languages is very intuitive. On other hand, an extended version (*iClusterPlus*) was also developed (*Mo et al., 2013*). One of the most important works using this method was (*Curtis et al., 2012*), identifying 12 different breast tumour subtypes.

In addition to beforementioned works, there are a huge examples of iCluster use with TCGA. For instance, in *Xie et al. (2019)* an integrative analysis was carried out with iCluster through RNAseq and proteomics data to analyse the OV subtype. The results showed two clusters with different survival rates; the method identified 18 mRNAs and 38 proteins as distinct molecules among subtypes. Another study proposed a modified

iCluster model to discover key processes in the tumour collection through unsupervised integration of multiple types of molecular data and functional annotations (*Bismeijer, Canisius & Wessels, 2018*). Further, *Mo et al. (2017)* described a novel modification (iClusterBayes) capable of jointly modelling omics data of continuous and discrete data types for the identification of tumour subtypes and relevant omics characteristics. In the work of *Kim et al. (2017)*, they modified this procedure to subtype patients using sequential double regularisation. Another pathway-based variant incorporates pathway data to group patients into cancer subtypes (*Mallavarapu et al., 2019*). Additionally, in *Jean-Quartier et al. (2021)* clustered GBM patients into several age subgroups with different age-related biomarkers. Finally, a work developed in *Nguyen et al. (2017)*, named PINS, allows omics data integration and molecular patient stratification automatically.

With the above, the trend in genome research is evident. An increasing number of works are attempting to integrate the greater amount of information provided by the different omic data into their models. Due to the complexity of cancer, stratifying patients according to a single source of information is becoming obsolete. Therefore, it is vitally important to improve models that are capable of multi-omic integration, as is the case with iCluster. Moreover, there is a need of novel approaches to automated medical decision pipelines building on machine learning, information fusion and explainability (*Holzinger et al., 2021*; *Barredo Arrieta et al., 2020*).

Medical imaging as a data source for ML algorithms

An important event occurred in 2012 during the celebration of the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (*Russakovsky et al., 2015*). A deep learning model (specifically, a CNN) halved the second best error rate in the image classification task. The goal of this challenge was the detection of objects and the classification of images using a large-scale database. Furthermore, deep learning algorithms can automatically find the best subset of features that describe the nuances of images. In addition, transfer learning was borne: an attempt to reuse the representation of the learning characteristic of one problem to solve another.

Deep learning techniques are on the rise in cancer research, namely for object detection and image classification. Initiatives such as TCGA offer the possibility of training deep learning models by making a large quantity of biomedical images available for research. Specifically, TCGA provides two types of images: tissue slide and digital imaging and communications in medicine (DICOM) images. DICOM images such as X-rays or computed tomography (CT), are used to extract quantitative characteristics from the images. Algorithms are trained to identify those characteristics. Histopathological images are used for direct image processing.

As discussed in previous sections, the TCGA consortium has used deep learning methods (*Saltz et al., 2018*). Specifically, they used CNN to detect tumour-infiltrating lymphocytes (TILs) based on H&E images in 13 tumour types. They reported a local spatial structure in the TIL patterns and their correlation with overall survival. These data modify densities and spatial structure among tumour types, immune subtypes and

molecular tumour subtypes. Spatial infiltration of lymphocytes might reflect particular aberration states of tumour cells.

Based on these findings, several studies have used this and other repositories to create their own models. It is important to distinguish among data types. On the one hand, there are works that have used radiological images for the classification of stages of gliomas (*Park et al., 2019*). In this work, they did not use the radiological images directly; rather, they extracted 250 characteristics from them to train their models, obtaining an area under the receiver operating characteristic curve (AUROC) of 72%. Notably, this model, which was validated with very heterogeneous cohorts such as TCGA, considerably reduced the performance. These results indicate that manual extraction of characteristics does not provide sufficient generalisation.

Sun et al. (2018b) utilised contrast-enhanced CT images and RNASeq data to assess CD8 cell infiltration in tumour biopsies. They first extracted features from both types of data to ultimately keep eight features and train an elastic-net regularised regression method. They used this signature to predict the response to anti-programmed cell death protein 1 (PD1) or anti-programmed death-ligand 1 (PDL1) treatments. Magnetic resonance imaging (MRI) was used in to predict the status of MGMT, a promoter of methylation that has been related to better outcomes on GBM patients integrated with expression data (accuracy of 73%; (*Kanas et al., 2017*)).

Fischer et al. (2018) reported a new method for histopathological image analysis—sparse coding—using a dictionary optimised for biomedical images. They stated that they generally obtained better performance rates compared to transfer learning. In *Yu et al. (2016)*, they predicted the prognosis of non-small cell lung tumours. Using the CellProfiler software, they extracted 9,879 quantitative characteristics and trained different algorithms, such as SVM or random forest. Finally, with a variant of the SVM algorithm, they achieved an AUROC of 81%. Besides, they developed a low-complexity method for classification and disease grading in histopathological images. This method—discriminative feature-oriented dictionary learning (DFDL)—learns from specific class dictionaries in such a way that under a dispersion restriction, the learned dictionaries allows it to represent a new image in a simplified way. However, it is unable to represent samples from other classes. *Coudray et al. (2018)* used histopathology images of lung cancers to classify squamous cell carcinomas, adenocarcinomas and normal samples with a 97% of AUROC. In the work of *Cheerla & Gevaert (2019)*, they were able to extract information from several datasets and obtain a model capable to predict patient prognosis. *Ertosun & Rubin (2015)* subtyped gliomas with CNN algorithms by using raw images for this task; there was more than 90% accuracy for glioma classification and almost 80% for glioma grade identification. *Rendleman et al. (2019)* used a CNN to evaluate distinct histological tumour growth patterns such as solid, micropapillary, acinar and cribriform (84% accuracy). An important work was developed in *Janowczyk et al. (2019)*. They developed an unsupervised encoder to compress four data modalities, including whole slide images (WSIs), into a single feature vector for each patient. The model was trained with TCGA data and predict single cancer overall survival, achieving a C-index of 0.78 overall.

It is important to highlight the need to pre-process the histopathological images before their analysis. This step is crucial to achieve great performances in the models. The images housed in TCGA are not homogeneous in size, shape and brightness. Therefore, it is necessary to use a pre-processing stage in order to standardise all the images before the analysis. Open source tools as HistoQC (*Janowczyk et al., 2019*) are relevant in the extraction of knowledge and the good use of images in research.

Biological questions solved by ML algorithms

In addition to all the existing omics data in TCGA, the inclusion of the clinical information from each patient increases the ability to generate analytical models. The dependent variable in supervised learning problems can potentially be any of the 100 clinical outcomes offered by TCGA, depending on the biological response to be answered. For classification problems, researchers have information on the anatomical division of the neoplasm, the clinical stage of the patient, TNM status, MSI status, ethnicity, age and gender, survival and/or relapse of the tumor among others. Thus, we can infer whether we can predict the anatomical division of the tumour or its clinical stage from the methylation marks of the patients (among other possibilities). For regression, we can use the initial age at diagnosis or the prognosis of the patient by means of the Karnofsky Performance Status Scale. Also, independently of clinical data, classificacion and regression models could be created to determinate other omics outcomes. For instance, sobreexpression of driver genes, methylation status or mutation types. In addition, the data storaged in TCGA repository allows any potential researcher to study survival in the cohort: it presents data on life status and the days that have elapsed between events, such as the death of the patient or other events of interest (relapse or disease-free survival).

The quantification of the number of papers published for each cancer subtype is shown in Fig. 3. As shown in this figure, the most used are those with the highest number of samples: BRCA, LUAD and OV. The great number of dimensions and observations, together with the large number of available clinical variables (pathological state, TNM classification status, drug effect, treatment response, etc.) generates an ideal data analysis environment for the use of both supervised and unsupervised ML techniques. For supervised learning problems, contingent on the dependent variable to be predicted, these problems may be regression (patient survival time, expression of a specific gene or individual age) or classification (classification of patients according to some driver gene status, disease or metastasis stages, etc.) problems. In terms of unsupervised learning problems, most work focuses on finding new subtypes of the disease. As for the other tumours, there is a significant decrease in the number of publications, mainly due to the number of samples collected. This fact is due to the intrinsic functioning of the ML algorithms, which, because they work on the basis of examples, are able to generalise more as the number of observations in their training phase increases. We can observe in the Fig. 3 also how there are several works that use different cohorts in the same analysis. After reviewing these papers, two trends have been observed in this type of article. Firstly, there are those that train models to predict cross-sectional and/or basic conditions of tumours. For example, in *Fischer et al. (2018)* they predicts MSI status from

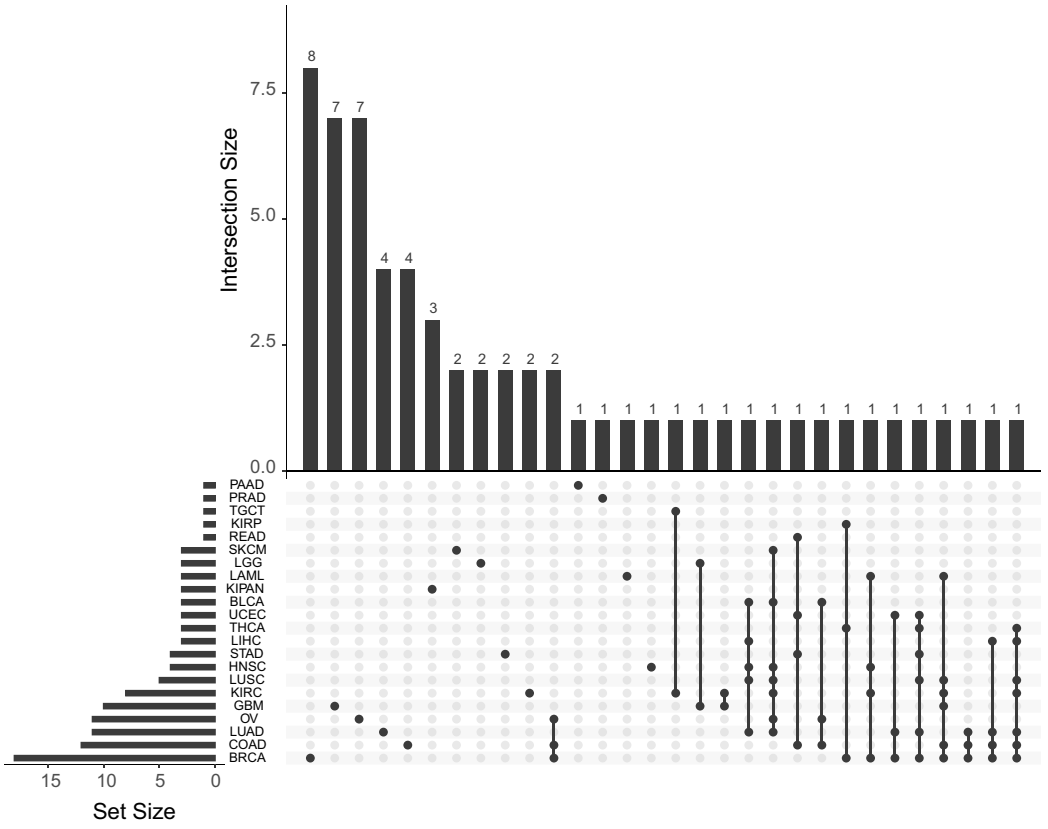


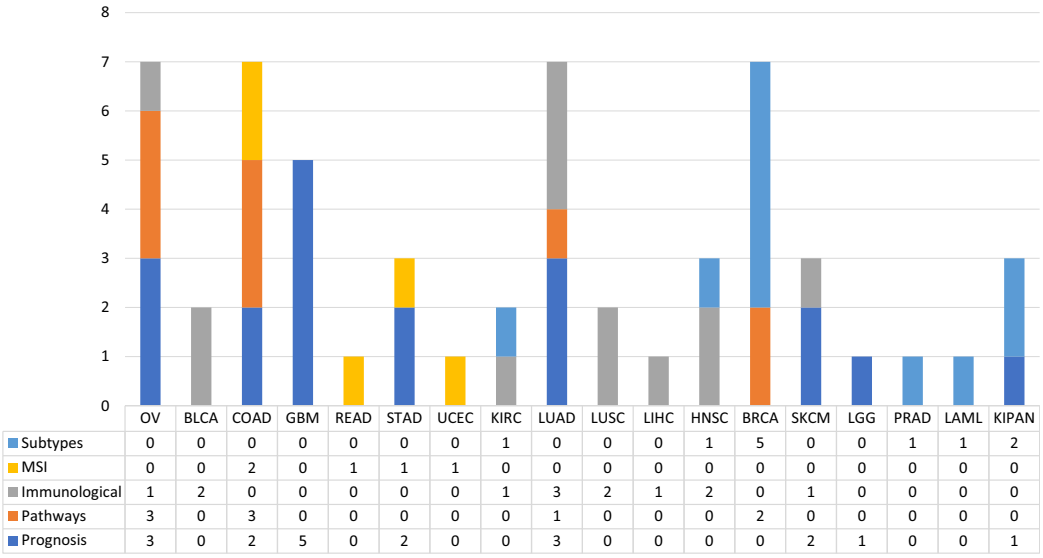
Figure 3 Number of papers published with each of the TCGA cohorts. Upset plot showing the number of works published with each tumor type and their combinations.

Full-size DOI: 10.7717/peerj-cs.584/fig-3

histopathological images. In this case, the different TCGA cohorts are treated together for the training of the models. On the other hand, other works have been identified in which the cohorts are used independently. These works are mainly based on model improvements or development of new technologies that are then tested with each cohort. This is the example of *Chen et al. (2018b)*, where they develop a new model of autoencoders for the search of new genetic signatures. This model is later validated in each of the available TCGA cohorts. Another example is *Cheerla & Gevaert (2017)*, where they obtain a model that recommends the type of treatment from miRNA expression data. This recommendation is validated in the different TCGA cohorts. Therefore, there are many approaches that can be used by researchers to use this type of data.

In this review, we classify the identified works on five major groups according biological problem solved. Although there are more than 100 variables in the TCGA clinical database, there is very little variability observed in the type of analysed problems.

In order to observe the distribution of publications according to this type of classification along the different types of tumours, pay attention to the Fig. 4. Figure 4 shows the distribution of the published papers according to the different types of tumors and the type of biological problem. The different biological problems show a different distribution according to tumor type. It can be seen how prognosis prediction is more



	OV	BLCA	COAD	GBM	READ	STAD	UCEC	KIRC	LUAD	LUSC	LIHC	HNSC	BRCA	SKCM	LGG	PRAD	LAML	KIPAN
Subtypes	0	0	0	0	0	0	0	1	0	0	0	1	5	0	0	1	1	2
MSI	0	0	2	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
Immunological	1	2	0	0	0	0	0	1	3	2	1	2	0	1	0	0	0	0
Pathways	3	0	3	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0
Prognosis	3	0	2	5	0	2	0	0	3	0	0	0	0	2	1	0	0	1

Figure 4 The proportion of published works with ML techniques according to the type of biological problem. Full-size DOI: 10.7717/peerj-cs.584/fig-4

common in GBM cohorts. In this case GBM is a type of tumour with high mortality rates, so it is a cohort where there are numerous events with which robust ML models can be created. Following GBM, OV and LUAD cohorts were the most used. Furthermore, it is observed how this type of problem is addressed in different cohorts. This is not the case for MSI prediction, as few tumours are defined by MSI status. The most common ones in this case are COAD, READ, STAD and UCEC. Paying attention to the prediction of subtypes, we see that the BRCA cohort is the most used. Regarding the immunological phenotype, the works have used cohorts mainly of solid tumours, which are the ones that present the best response to treatments with immunological therapies. Finally, few tumours have been addressed in the prediction of pathways. The works identified used the OV, COAD, LUAD and BRCA cohorts. The following sections are a review of the works according to the five classes identified.

Prognosis prediction

The prognosis in the different types of cancer varies greatly due to their heterogeneity, their environment and their unique behaviour in each patient. It is therefore crucial to be able to predict the events that will develop in the patient and have a direct effect on the prognosis of the cancer. These events can be deaths, recurrence and/or relapse events, metastases or the classification of patients into specific stages. Numerous studies have been identified that have addressed this field of study with ML-based analyses.

Within this category, many of the papers identified have aimed to predict events related to patient survival time. Furthermore, it has been observed that expression data are the most used in this type of problem, due to their better performance in predictions, together with methylation data (*Stephen & Lewis, 2013*). In *Wong, Rostomily & Wong (2019)* they use them as input from a deep learning network, while in *Fatai & Gamieldien (2018)* they use the SVM algorithm. In both problems they obtained gene signatures that were

highly correlated with the survival events of the patients. Other works have addressed this type of problem by integrating expression data with other data sets. For example, in *Yasser et al. (2018)*, using FS techniques, they obtained subgroups of features from sets of ANC, methylation and expression. In *Zhang et al. (2016)*, they add a layer of complexity, adding to the integration of miRNA data by means of multiple kernel and FS techniques. This technique was also used in *Srivastava et al. (2013)* for the integration of expression and miRNA data. On the other hand, one paper has used only lncRNA data capable of predicting survival events over 19 months (*Cheng, 2018*). Works were also identified that have addressed this problem from histopathological images. In these cases, they extract characteristics from the images in order to train different types of ML algorithms and be able to predict survival times and/or events (*Ing et al., 2017*; *Yu et al., 2016*; *Powell et al., 2017*).

In addition to survival events and times, there are other events that are interesting to predict for clinical practice. In this case, the events of tumour recurrence, which are when the tumour is detected again after a treatment process. Knowing, therefore, the probabilities of a cancer relapse in a given patient is interesting for clinicians. Using the ML approach, this problem has been addressed mainly with transcriptomical expression data. For example, in *Wang et al. (2018)*, from miRNA data, lncRNA and mRNA identified 36 features capable of classifying with 91% accuracy whether a tumour will recur or not. In *Zhou et al. (2018)*, *Sun et al. (2018a)*, *Xu et al. (2017)*, similar performances were obtained with RNASeq data, while in *Wei (2018)* from RNASeq data predict metastasis processes. On the other hand, in *Feng et al. (2018)* they predicted recurrence based on data from the tumour microenvironment.

Usually the different types of tumours are classified in different stages which correlate with different prognoses. Therefore, this has been another problem addressed by the researchers, in which the ML has been able to offer a solution. Again, the RNASeq data were the most used to address this problem. In *Fan et al. (2018)*, they obtained a signature of 12 genes capable of distinguishing patients with lung cancer with different risks, while in *Chen et al. (2017)* they identified pathways of interest capable of classifying the different stages of lung adenocarcinoma. For example, in *Yang, Xu & Zeng (2018)*, they used only the features corresponding to lncRNA, obtaining a signature of six lncRNA capable of classifying patients with melanoma according to their stages.

Immunological phenotype prediction

Currently, one of the most successful and promising therapies against cancer are drugs that act against immune checkpoint inhibitors (ICI). These drugs block the proteins produced by certain immune cells to prevent immune responses from becoming too strong. The activation of these checkpoints can cause the cells of our immune system not to be able to kill the cancer cells. The treatment of most types of tumours is helped by this type of therapy, although there are some that do not respond in the same way. This is the case of HGSOC tumours. In *Dai et al. (2018)*, they analysed genomic data from HGSOC patients to predict their immune phenotype of the tumour microenvironment. After a comparison with the analysis of other solid tumours, such as BLCA, SKCM, KIRC, LUSC

and LUAD, they identified ten dominant factors that determine the immunogenicity of HGSOCs. Using the ML they were able to classify tumours with high and low cytolytic activity, noting also that mutations in BRCA1 may be a good predictive biomarker for guiding ICI therapies of HGSOC patients.

Moreover, they developed and independently validated an eight-feature signature based on CD8 cell radiomic imaging for the response to (PD)-1 and (PD-L1). This imaging predictor provides a promising way to predict the immunologic phenotype of tumors and infer clinical outcomes for cancer patients who had been treated with anti-PD-1 and PD-L1.

Pathways prediction

Some of the genetic drivers specific to each tumour are well known, as well as certain pathways that influence the process of tumour development. Although the identification of status is a complex issue, it holds a great deal of information in the diagnosis and treatment of patients. This is why researchers have addressed this problem using ML techniques. After the review carried out in this work, works have been detected that were able to model this problem. Most of them are based on RNASeq data, with which they infer the status of different cancer driver pathways (*Rykunov et al., 2016*), damaged pathways (*Klein, Stern & Zhao, 2017*) and level of apoptosis (*Salvucci et al., 2017*). In *Chen et al. (2012)*, RNASeq data and copy number data are used to detect pathways capable of differentiating expression patterns between different phenotypes. In the case of *Ou-Yang et al. (2017)*, they developed a cross-platform method for the identification of new molecular pathways related to tumour types.

MSI status prediction

Microsatellite instability is the mutation predisposition of certain tumours due to defects in the DNA mismatch repair machinery. It is of great importance to identify MSI status in certain tumours as it is a great predictor and marker for diagnosis and treatment. In this review two papers were identified that have addressed this problem with MSI techniques. The first of these, called *Wang & Liang (2018)*, classified the different MSI subtypes based on mutational annotation data. They used an SVM algorithm and obtained a total accuracy of 0.91 for the COAD, READ, STAD and UCEC cohorts. They used a total of 22 features for the classification, such as the count of SNPs, indels, total mutations, missence mutations or the ratio between mutations and SNPs. On the other hand, in *Chen et al. (2018c)* they made a classification from the expression data. Using ML algorithms and FS techniques they obtained a classifier capable of discerning the different subtypes.

Subtypes prediction

Finally, another problem that has been addressed by researchers and where ML techniques can contribute significantly is the prediction of the different subtypes of the disease. It is interesting to recognise which are the different omic data sets that hold enough information to build a classification system robust enough to obtain the appropriate yields. As usual, RNASeq was the technique par excellence from which the data were obtained

to train the models (*Yang et al., 2014*; *Graudenzi et al., 2017*; *Gao et al., 2017*). In addition, the expression data were combined with other sets such as miRNA (*Wilop et al., 2016*; *Nair et al., 2015*), methylation (*List et al., 2014*) or miRNA and methylation (*Nguyen et al., 2017*).

In addition to expression data, two papers have used exclusively image data to classify subtypes of the disease. Firstly from MRI images (*Sutton et al., 2017*) and with qCT-TA data (*Kocak et al., 2018*). Other work, for example, used mutation data (*Vural, Wang & Guda, 2016*) and miRNA data (*Muhamed Ali et al., 2018*).

It is logical to think that the ML algorithms now attempt to analyse the most studied problems to determine whether they can reach the same conclusions as conventional statistical approximations. In general, ML approximations analyse the importance of each of the variables in the dataset without making any a priori assumptions, so the generalisation of the model does not have to be based on inherent biological knowledge of the data. Although there are ML approximations that base the selection of genes from each data platform to certain pathways of interest (*Seoane et al., 2013*), this field is still open field for new approximations.

One study observed that the ML algorithms reached similar conclusions and also provided a certain degree of diversity in the results (*Liñares Blanco et al., 2019*). This outcome aids the examination of new omics variables that might be of interest to study the development of cancer. Cancer is a multifactorial and complex disease, so it makes sense that the analysis should consider the differences that characterise the patients as a whole and not individually.

A deeper examination of the BRCA cohort

The TCGA consortium jointly analysed genomic DNA copy number arrays, DNA methylation, exome sequencing, mRNA arrays, miRNA sequencing and reverse-phase protein arrays (*The Cancer Genome Atlas Network, 2012b*). In this study, they demonstrated the existence of four main classes of breast cancer by combining data from five platforms; there was great heterogeneity. Mutations in only three genes (TP53, PIK3CA and GATA3) occurred in more than 10% of all the samples. In addition, they identified two new subgroups defined by protein expression—produced primarily by the tumour microenvironment. Besides, the comparison of basal-type breast tumours with high-grade serous ovarian tumours showed a myriad of molecular similarities, a finding that indicates a related aetiology and similar therapeutic opportunities.

In one study (*Ciriello et al., 2015*), the authors discovered that invasive lobular carcinoma (ILC) is a clinically and molecularly distinct disease. In this case, patients with ILC show CDH1 and PTEN loss, AKT activation and mutations in TBX3 and FOXA1. The proliferation and expression of genes related to the immune system defined three ILC subtypes.

The findings made by TCGA are leading the way in the search for new treatment and diagnostic opportunities for patients, in this case, with breast cancer. Although the work of the TCGA has been exhaustive, the possibilities offered by giving free access to its

data are enormous. For this reason, many researchers have taken these data as a reference and have reported results of great interest to the community.

We identified several publications that utilised ML to analyse TCGA BRCA data. There are published works using miRNA data (*Sherafatian, 2018*), methylation data (*Hao et al., 2017*), expression data (*Wen, Li & Chang, 2018*), integrative analysis of expression and methylation data (*Cappelli, Felici & Weitschek, 2018*) and even expression data from isomiRs (*Liao et al., 2018*). These works achieved prominent outcomes, notably the ability to infer that the problems of classification for diagnosis (healthy or disease patients) are problems that the ML algorithms solve quite easily, even with different types of data.

Several papers have been published to address this patient stratification. For example, to classify the subtypes of PR, ER and HER2 with miRNA data (*Sherafatian, 2018*; *Liao et al., 2018*), the status of the basal subtype through the analysis of images with deep learning algorithms (*Chidester, Do & Ma, 2018*) and the different subtypes of BRCA by the expression of molecular pathways (*Graudenzi et al., 2017*), mutation data (*Vural, Wang & Guda, 2016*) or even the integration of expression and methylation data (*List et al., 2014*). Cancer subtypes can be studied by unsupervised learning techniques and the integration of different data (expression, methylation, miRNA and CNV) (*Nguyen et al., 2017*).

Finally, other works have studied the interaction between miRNA and mRNA (*Koo, Zhang & Chaterji, 2018*; *Ghoshal et al., 2018*), the identification of altered pathways by mRNA expression data (*Klein, Stern & Zhao, 2017*) or by integrating expression and mutation data (*Rykunov et al., 2016*), the response to drugs in different cell lines (*Daemen et al., 2013*; *Geeleher et al., 2017*) and the identification of variants by means of genomic data (*Dong et al., 2016*) and by means of images with artificial vision techniques (*Sutton et al., 2017*).

CONCLUSIONS

Many studies on cancer have been performed in recent years with ML that uses molecular data. These data have mainly included diagnostic studies, prognosis or patient stratification. More recently, there have been promising results in response to drugs or genetic interactions. In this review, we investigated and identified those relevant works that have used TCGA data through algorithms or pipelines of analysis based on ML.

ML techniques can extract the underlying knowledge from a set of data, so it is relevant to understand the appropriateness of the data. In other words, these techniques must be used with certain precautions. Indeed, researchers should be aware that the conclusions they obtain may be biased due to poor data selection or analytical methodology. Among the different learning techniques, supervised learning has analysed the most problems using TCGA data. This endeavour has emphasised the use of genetic expression data through different variants of the SVM algorithm. There are still infinite opportunities and possibilities for the exploitation of TCGA data with ML. ML techniques can reach conclusions that are similar to conventional approaches and also to obtain a degree of variability that is extremely useful when searching for novel predictors.

It is clear that we are still at an early stage in the analysis of this pathology and it is necessary to develop and use more complex algorithms. For example, the use of

kernel-based models can integrate different datasets in the same process. The integration of data in the analysis of complex and multifactorial diseases continues to be a challenge for which it is necessary to invest even more time and money in finding better algorithms. As discussed above, the quantity of existing data will not stop growing and all derive from the same biological sample. Thus, it is expected that the connection between omics platforms can improve the performance of the models. It is still necessary to take a step forward in the development of multidimensional ML models for cancer research.

Complex problems, such as the prediction of different cell statuses (methylation, apoptosis or mutation), are already being tackled with promising results. We and others hope that the links between biological information extracted from the same patient will be further explored in order to elucidate the origin of the disease by ML techniques. Currently, the focus is on certain types and subtypes of cancer (e.g., BRCA, LUAD or OV), usually due to the number of people afflicted with it and the importance attributed by society. It is also necessary to increase investment in the generation of data that is related to relatively minor or especially aggressive cancer types in order to provide the algorithms with sufficient information in their learning phase and to avoid biases in their learning.

In this work, we exhaustively reviewed studies that have used ML techniques for the analysis of different types of cancer using TCGA data. In our opinion, the era of individual analysis has passed and we are entering the era of data integration studies—at the clinical-genomic level as well as medical imaging or evolution analysis by means of time series. We are working on the development of complex data integration algorithms in different fields, one of which is artificial intelligence. There are currently ML models that are demonstrating great effectiveness and are gaining followers. These methods include the aforementioned deep learning techniques, but research is required to render the results understandable and explain why a certain prediction is made, especially from a clinical point of view. The great challenge of the integration techniques is the incessant increase in the number of dimensions and the heterogeneity of the data sets generated from the same patient/biological process (*Kristensen et al., 2014*).

Finally, we hope that this review will serve as a starting point for researchers in bioinformatics and computer science who are interested in studying cancer, as well as those researchers who are more focused on the use of ML techniques to know the potential of their algorithms with TCGA data. More research and the development of new algorithms are required to overcome the disease.

ABBREVIATIONS

ML	Machine Learning
TCGA	The Cancer Genome Atlas
MSI	Microsatellite Instability
BRCA	Breast Cancer Adenocarcinoma
GBM	Glioblastoma Multiforme
SNP	single nucleotide polymorphisms
LUSC	Lung squamous cell carcinoma
OV	Ovarian Cancer

SVM	Support Vector Machines
RF	Random Forest
WGS	Whole Genome Sequencing
CNN	Convolutional Neural Networks
TIL	Tumour-infiltrating lymphocytes
MRI	Magnetic resonance imaging

ACKNOWLEDGEMENTS

We thank Dr. Jose A. Seoane for comments during the preparation of this review.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the "Collaborative Project in Genomic Data Integration (CICLOGEN)" PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER)—"A way to build Europe." and the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431D 2017/16), the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23) and Competitive Reference Groups (Ref. ED431C 2018/49). CITIC, as Research Center accredited by Galician University System, is funded by "Consellería de Cultura, Educación e Universidades from Xunta de Galicia", supported in an 80% through ERDF Funds, ERDF Operational Programme Galicia 2014–2020, and the remaining 20% by "Secretaría Xeral de Universidades" (Grant ED431G 2019/01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Carlos III Health Institute: PI17/01826.
European Regional Development Funds (FEDER).
General Directorate of Culture, Education and University Management of Xunta de Galicia: ED431D 2017/16.
Galician Network for Colorectal Cancer Research: ED431D 2017/23.
Competitive Reference Groups: ED431C 2018/49.
Consellería de Cultura, Educación e Universidades from Xunta de Galicia.
Secretaría Xeral de Universidades: ED431G 2019/01.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Jose Liñares-Blanco conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Alejandro Pazos conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Carlos Fernandez-Lozano conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This research is a Literature Review; there are no raw data or code files.

REFERENCES

Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, Annala M, Aprikian A, Armenia J, Arora A, Auman JT, Balasundaram M, Balu S, Barbieri CE, Bauer T, Benz CC, Bergeron A, Beroukhim R, Berrios M, Bivol A, Bodenheimer T, Boice L, Bootwalla MS, Borges dos Reis R, Boutros PC, Bowen J, Bowlby R, Boyd J, Bradley RK, Breggia A, Brimo F, Bristow CA, Brooks D, Broom BM, Bryce AH, Bubley G, Burks E, Butterfield YSN, Button M, Canes D, Carlotti CG, Carlsen R, Carmel M, Carroll PR, Carter SL, Cartun R, Carver BS, Chan JM, Chang MT, Chen Y, Cherniack AD, Chevalier S, Chin L, Cho J, Chu A, Chuah E, Chudamani S, Cibulskis K, Ciriello G, Clarke A, Cooperberg MR, Corcoran NM, Costello A J, Cowan J, Crain D, Curley E, David K, Demchok J A, Demichelis F, Dhalla N, Dhir R, Doueik A, Drake B, Dvinge H, Dyakova N, Felau I, Ferguson M L, Frazer S, Freedland S, Fu Y, Gabriel S B, Gao J, Gardner J, Gastier-Foster JM, Gehlenborg N, Gerken M, Gerstein MB, Getz G, Godwin AK, Gopalan A, Graefen M, Graim K, Gribbin T, Guin R, Gupta M, Hadjipanayis A, Haider S, Hamel L, Hayes DN, Heiman DI, Hess J, Hoadley KA, Holbrook AH, Holt RA, Holway A, Hovens CM, Hoyle AP, Huang M, Hutter CM, Ittmann M, Iype L, Jefferys SR, Jones CD, Jones SJM, Juhl H, Kahles A, Kane CJ, Kasaian K, Kerger M, Khurana E, Kim J, Klein RJ, Kucherlapati R, Lacombe L, Ladanyi M, Lai PH, Laird PW, Lander ES, Latour M, Lawrence MS, Lau K, LeBien T, Lee D, Lee S, Lehmann K-V, Leraas KM, Leshchiner I, Leung R, Libertino JA, Lichtenberg TM, Lin P, Linehan WM, Ling S, Lippman S M, Liu J, Liu W, Lochovsky L, Loda M, Logothetis C, Lolla L, Longacre T, Lu Y, Luo J, Ma Y, Mahadeshwar H S, Mallery D, Mariamidze A, Marra MA, Mayo M, McCall S, McKercher G, Meng S, Mes-Masson A-M, Merino MJ, Meyerson M, Mieczkowski PA, Mills GB, Shaw KRM, Minner S, Moinzadeh A, Moore R A, Morris S, Morrison C, Mose LE, Mungall AJ, Murray BA, Myers J B, Naresh R, Nelson J, Nelson M A, Nelson P S, Newton Y, Noble M S, Noushmehr H, Nykter M, Pantazi A, Parfenov M, Park PJ, Parker J S, Paulauskis J, Penny R, Perou C M, Piché A, Pihl T, Pinto P A, Prandi D, Protopopov A, Ramirez NC, Rao A, Rathmell WK, et al. 2015. The molecular taxonomy of primary prostate cancer. *Cell* **163(4)**:1011–1025 DOI 10.1016/j.cell.2015.10.025.

Agrawal N, Akbani R, Aksoy BA, Ally A, Arachchi H, Asa SL, Auman JT, Balasundaram M, Balu S, Baylin SB, Behera M, Bernard B, Beroukhim R, Bishop JA, Black AD, Bodenheimer T, Boice L, Bootwalla MS, Bowen J, Bowlby R, Bristow CA, Brookens R, Brooks D, Bryant R, Buda E, Butterfield YSN, Carling T, Carlsen R, Carter SL, Carty SE, Chan TA, Chen AY, Cherniack AD, Cheung D, Chin L, Cho J, Chu A, Chuah E, Cibulskis K, Ciriello G, Clarke A, Clayman GL, Cope L, Copland J A, Covington K, Danilova L, Davidsen T, Demchok JA,

DiCara D, Dhalla N, Dhir R, Dookran SS, Dresdner G, Eldridge J, Eley G, El-Naggar AK, Eng S, Fagin JA, Fennell T, Ferris RL, Fisher S, Frazer S, Frick J, Gabriel SB, Ganly I, Gao J, Garraway LA, Gastier-Foster JM, Getz G, Gehlenborg N, Ghossein R, Gibbs RA, Giordano TJ, Gomez-Hernandez K, Grimsby J, Gross B, Guin R, Hadjipanayis A, Harper HA, Hayes DN, Heiman DI, Herman JG, Hoadley KA, Hofree M, Holt RA, Hoyle AP, Huang FW, Huang M, Hutter CM, Ideker T, Iype L, Jacobsen A, Jefferys SR, Jones CD, Jones SJM, Kasaian K, Kebebew E, Khuri FR, Kim J, Kramer R, Kreisberg R, Kucherlapati R, Kwiatkowski DJ, Ladanyi M, Lai PH, Laird PW, Lander E, Lawrence MS, Lee D, Lee E, Lee S, Lee W, Leraas KM, Lichtenberg TM, Lichtenstein L, Lin P, Ling S, Liu J, Liu W, Liu Y, LiVolsi VA, Lu Y, Ma Y, Mahadeshwar HS, Marra MA, Mayo M, McFadden DG, Meng S, Meyerson M, Mieczkowski PA, Miller M, Mills G, Moore RA, Mose LE, Mungall AJ, Murray BA, Nikiforov YE, Noble MS, Ojesina AI, Owonikoko TK, Ozenberger BA, Pantazi A, Parfenov M, Park PJ, Parker JS, Paull EO, Pedamallu CS, Perou CM, Prins JF, Protopopov A, Ramalingam SS, Ramirez NC, Ramirez R, Raphael B J, Rathmell WK, Ren X, Reynolds SM, Rheinbay E, Ringel MD, Rivera M, Roach J, Robertson AG, Rosenberg MW, Rosenthal M, Sadeghi S, Saksena G, Sander C, Santoso N, Schein JE, Schultz N, Schumacher SE, Seethala RR, Seidman J, Senbabaoglu Y, Seth S, Sharpe S, Shaw KRM, Shen JP, Shen R, Sherman S, Sheth M, Shi Y, Shmulevich I, Sica GL, Simons JV, Sinha R, Sipahimalani P, Smallridge RC, Sofia HJ, Soloway MG, Song X, Sougnez C, Stewart C, Stojanov P, Stuart JM, Sumer SO, Sun Y, Tabak B, Tam A, Tan D, et al. 2014. Integrated genomic characterization of papillary thyroid carcinoma. *Cell* **159(3)**:676–690 DOI 10.1016/j.cell.2014.09.050.

Akbani R, Akdemir K C, Aksoy BA, Albert M, Ally A, Amin S B, Arachchi H, Arora A, Auman JT, Ayala B, Baboud J, Balasundaram M, Balu S, Barnabas N, Bartlett J, Bartlett P, Bastian BC, Baylin SB, Behera M, Belyaev D, Benz C, Bernard B, Beroukhim R, Bir N, Black AD, Bodenheimer T, Boice L, Boland GM, Bono R, Bootwalla MS, Bosenberg M, Bowen J, Bowlby R, Bristow CA, Brockway-Lunardi L, Brooks D, Brzezinski J, Bshara W, Buda E, Burns WR, Butterfield YSN, Button M, Calderone T, Cappellini GA, Carter C, Carter SL, Cherney L, Cherniack AD, Chevalier A, Chin L, Cho J, Cho RJ, Choi Y-L, Chu A, Chudamani S, Cibulskis K, Ciriello G, Clarke A, Coons S, Cope L, Crain D, Curley E, Danilova L, D'Atri S, Davidsen T, Davies MA, Delman KA, Demchok JA, Deng QA, Deribe YL, Dhalla N, Dhir R, DiCara D, Dinikin M, Dubina M, Ebrom JS, Egea S, Eley G, Engel J, Eschbacher JM, Fedosenko KV, Felau I, Fennell T, Ferguson ML, Fisher S, Flaherty KT, Frazer S, Frick J, Fulidou V, Gabriel SB, Gao J, Gardner J, Garraway LA, Gastier-Foster JM, Gaudioso C, Gehlenborg N, Genovese G, Gerken M, Gershenwald JE, Getz G, Gomez-Fernandez C, Gribbin T, Grimsby J, Gross B, Guin R, Gutschner T, Hadjipanayis A, Halaban R, Hanf B, Haussler D, Haydu LE, Hayes DN, Hayward NK, Heiman DI, Herbert L, Herman JG, Hersey P, Hoadley KA, Hodis E, Holt RA, Hoon DSB, Hoppough S, Hoyle AP, Huang FW, Huang M, Huang S, Hutter CM, Ibbs M, Iype L, Jacobsen A, Jakrot V, Janning A, Jeck WR, Jefferys SR, Jensen MA, Jones CD, Jones SJM, Ju Z, Kakavand H, Kang H, Kefford RF, Khuri FR, Kim J, Kirkwood J M, Klode J, Korkut A, Korski K, Krauthammer M, Kucherlapati R, Kwong L N, Kycler W, Ladanyi M, Lai PH, Laird PW, Lander E, Lawrence MS, Lazar AJ, Łaźniak R, Lee D, Lee J E, Lee J, Lee K, Lee S, Lee W, Leporowska E, Leraas K M, Li HI, Lichtenberg T M, Lichtenstein L, Lin P, Ling S, Liu J, Liu O, Liu W, Long G V, Lu Y, Ma S, Ma Y, Mackiewicz A, Mahadeshwar H S, Malke J, Mallery D, Manikhas GM, Mann GJ, Marra M A, Matejka B, Mayo M, Mehrabi S, Meng S, Meyerson M, Mieczkowski PA, Miller JP, Miller ML, Mills G B, Moiseenko F, Moore RA, Morris S, Morrison C, Morton D, Moschos S, et al. 2015. Genomic classification of cutaneous melanoma. *Cell* **161(7)**:1681–1696 DOI 10.1016/j.cell.2015.05.044.

Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, Holt RA, Jones SJM, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Schein JE, Sipahimalani P, Tam A, Thiessen N, Cheung D, Wong T, Brooks D, Robertson AG, Bowlby R, Mungall K, Sadeghi S, Xi L, Covington K, Shinbrot E, Wheeler DA, Gibbs RA, Donehower LA, Wang L, Bowen J, Gastier-Foster JM, Gerken M, Helsel C, Leraas KM, Lichtenberg TM, Ramirez NC, Wise L, Zmuda E, Gabriel SB, Meyerson M, Cibulskis C, Murray BA, Shih J, Beroukhim R, Cherniack AD, Schumacher SE, Saksena G, Pedamallu CS, Chin L, Getz G, Noble M, Zhang H, Heiman D, Cho J, Gehlenborg N, Saksena G, Voet D, Lin P, Frazer S, Defreitas T, Meier S, Lawrence M, Kim J, Creighton CJ, Muzny D, Doddapaneni HV, Hu J, Wang M, Morton D, Korchina V, Han Y, Dinh H, Lewis L, Bellair M, Liu X, Santibanez J, Glenn R, Lee S, Hale W, Parker JS, Wilkerson MD, Hayes DN, Reynolds SM, Shmulevich I, Zhang W, Liu Y, Iype L, Makhlouf H, Torbenson MS, Kakar S, Yeh MM, Jain D, Kleiner DE, Jain D, Dhanasekaran R, El-Serag HB, Yim SY, Weinstein JN, Mishra L, Zhang J, Akbani R, Ling S, Ju Z, Su X, Hegde AM, Mills GB, Lu Y, Chen J, Lee J-S, Sohn BH, Shim JJ, Tong P, Aburatani H, Yamamoto S, Tatsuno K, Li W, Xia Z, Stransky N, Seiser E, Innocenti F, Gao J, Kundra R, Zhang H, Heins Z, Ochoa A, Sander C, Ladanyi M, Shen R, Arora A, Sanchez-Vega F, Schultz N, Kasaian K, Radenbaugh A, Bissig K-D, Moore DD, Totoki Y, Nakamura H, Shibata T, Yau C, Graim K, Stuart J, Haussler D, Slagle BL, Ojesina AI, Katsonis P, Koire A, Lichtarge O, Hsu T-K, Ferguson ML, Demchok JA, Felau I, Sheth M, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Hutter CM, Sofia HJ, Verhaak RGW, Zheng S, Lang F, Chudamani S, Liu J, Lolla L, Wu Y, Naresh R, Pihl T, Sun C, Wan Y, Benz C, Perou AH, Thorne LB, Boice L, Huang M, Rathmell WK, Noushmehr H, Saggioro FP, Tirapelli DPDC, Junior CGC, Mente ED, Silva ODC Jr, Trevisan FA, Kang KJ, Ahn KS, Giama NH, Moser CD, Giordano TJ, Vinco M, Welling TH, Crain D, Curley E, Gardner J, Mallery D, Morris S, Paulauskis J, Penny R, et al. 2017. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169(7)**:1327–1341 DOI 10.1016/j.cell.2017.05.046.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, Ng PK-S, Jeong KJ, Cao S, Wang Z, Gao J, Gao Q, Wang F, Liu EM, Mularoni L, Rubio-Perez C, Nagarajan N, Cortés-Ciriano I, Zhou DC, Liang W-W, Hess JM, Yellapantula VD, Tamborero D, Gonzalez-Perez A, Suphavilai C, Ko JY, Khurana E, Park PJ, Van Allen EM, Liang H, Lawrence MS, Godzik A, Lopez-Bigas N, Stuart J, Wheeler D, Getz G, Chen K, Lazar AJ, Mills GB, Karchin R, Ding L, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L,

Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle KP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, et al. 2018. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173(2)**:371–385.

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* **58**:82–115.

Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, Rao A, Schultz A, Li X, Sumazin P, Williams C, Mestdagh P, Gunaratne PH, Yau C, Bowlby R, Robertson AG, Tiezzi DG, Wang C, Cherniack AD, Godwin AK, Kuderer NM, Rader JS, Zuna RE, Sood AK, Lazar AJ, Ojesina AI, Adebamowo C, Adebamowo SN, Baggerly KA, Chen T-W, Chiu H-S, Lefever S, Liu L, MacKenzie K, Orsulic S, Roszik J, Shelley CS, Song Q, Vellano CP, Wentzensen N, Weinstein JN, Mills GB, Levine DA, Akbani R, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, et al. 2018. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33(4)**:690–705.

Bismeijer T, Canisius S, Wessels LF. 2018. Molecular characterization of breast and lung tumors by integration of multiple data types with functional sparse-factor analysis. *PLOS Computational Biology* **14(10)**:e1006520 DOI 10.1371/journal.pcbi.1006520.

Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu C-J, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla S A, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L, Benz C, Barnholtz-Sloan J, Barrett W, Ostrom Q, Wolinsky Y, Black KL, Bose B, Boulos PT, Boulos M, Brown J, Czerinski C, Eppley M, Iacocca M, Kempista T, Kitko T, Koyfman Y, Rabeno B, Rastogi P, Sugarman M, Swanson P, Yalamanchii K, Otey IP, Liu YS, Xiao Y, Auman JT, Chen P-C,

Hadjipanayis A, Lee E, Lee S, Park PJ, Seidman J, Yang L, Kucherlapati R, Kalkanis S, Mikkelsen T, Poisson LM, Raghunathan A, Scarpace L, Bernard B, Bressler R, Eakin A, Iype L, Kreisberg RB, Leinonen K, Reynolds S, Rovira H, Thorsson V, Shmulevich I, Annala MJ, Penny R, Paulauskis J, Curley E, Hatfield M, Mallery D, Morris S, Shelton T, Shelton C, Sherman M, Yena P, Cuppini L, DiMeco F, Eoli M, Finocchiaro G, Maderna E, Pollo B, Saini M, Balu S, Hoadley KA, Li L, Miller CR, Shi Y, Topal MD, Wu J, Dunn G, Giannini C, O'Neill BP, Aksoy BA, Antipin Y, Borsu L, Berman SH, Brennan CW, Cerami E, Chakravarty D, Ciriello G, Gao J, Gross B, Jacobsen A, Ladanyi M, Lash A, Liang Y, Reva B, Sander C, Schultz N, Shen R, Socci ND, Viale A, Ferguson ML, Chen Q-R, Demchok JA, Dillon LAL, Shaw KRM, Sheth M, Tarnuzzer R, Wang Z, Yang L, Davidsen T, Guyer MS, Ozenberger BA, Sofia HJ, Bergsten J, Eckman J, Harr J, Myers J, Smith C, Tucker K, Winemiller C, Zach LA, Ljubimova JY, Eley G, Ayala B, Jensen MA, Kahn A, Pihl TD, Pot DA, Wan Y, Eschbacher J, Foltz G, Hansen N, Hothi P, Lin B, Shah N, Yoon J-G, Lau C, Berens M, Ardlie K, Beroukhim R, Carter SL, Cherniack AD, Noble M, Cho J, Cibulskis K, DiCara D, et al. 2013. The somatic genomic landscape of glioblastoma. *Cell* **155(2)**:462–477 DOI 10.1016/j.cell.2013.09.034.

Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS, Shukla SA, Guo G, Brooks AN, Murray BA, Imielinski M, Hu X, Ling S, Akbani R, Rosenberg M, Cibulskis C, Ramachandran A, Collisson EA, Kwiatkowski DJ, Lawrence MS, Weinstein JN, Verhaak RGW, Wu CJ, Hammerman PS, Cherniack AD, Getz G, Artyomov MN, Schreiber R, Govindan R, Meyerson M, Cancer Genome Atlas Research Network. 2016. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics* **48(6)**:607–616 DOI 10.1038/ng.3564.

Campbell JD, Yau C, Bowlby R, Liu Y, Brennan K, Fan H, Taylor AM, Wang C, Walter V, Akbani R, Byers LA, Creighton CJ, Coarfa C, Shih J, Cherniack AD, Gevaert O, Prunello M, Shen H, Anur P, Chen J, Cheng H, Hayes DN, Bullman S, Pedamallu CS, Ojesina AI, Sadeghi S, Mungall KL, Robertson AG, Benz C, Schultz A, Kanchi RS, Gay CM, Hegde A, Diao L, Wang J, Ma W, Sumazin P, Chiu H-S, Chen T-W, Gunaratne P, Donehower L, Rader JS, Zuna R, Al-Ahmadie H, Lazar AJ, Flores ER, Tsai KY, Zhou JH, Rustgi AK, Drill E, Shen R, Wong CK, Stuart JM, Laird PW, Hoadley KA, Weinstein JN, Peto M, Pickering CR, Chen Z, Van Waes C, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, et al. 2018. Genomic, pathway

network, and immunologic features distinguishing squamous carcinomas. *Cell Reports* **23(1)**:194–212.

Cappelli E, Felici G, Weitschek E. 2018. Combining dna methylation and rna sequencing data of cancer for supervised knowledge extraction. *BioData Mining* **11(1)**:22 DOI 10.1186/s13040-018-0184-6.

Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, Anjum S, Wang J, Manyam G, Zoppoli P, Ling S, Rao AA, Grifford M, Cherniack AD, Zhang H, Poisson L, Carlotti CG, Tirapelli DPDC, Rao A, Mikkelsen T, Lau CC, Yung WKA, Rabadan R, Huse J, Brat DJ, Lehman NL, Barnholtz-Sloan JS, Zheng S, Hess K, Rao G, Meyerson M, Beroukhim R, Cooper L, Akbani R, Wrensch M, Haussler D, Aldape KD, Laird PW, Gutmann DH, Noushmehr H, Iavarone A, Verhaak RGW, Anjum S, Arachchi H, Auman JT, Balasundaram M, Balu S, Barnett G, Baylin S, Bell S, Benz C, Bir N, Black KL, Bodenheimer T, Boice L, Bootwalla MS, Bowen J, Bristow CA, Butterfield YSN, Chen Q-R, Chin L, Cho J, Chuah E, Chudamani S, Coetzee SG, Cohen ML, Colman H, Couce M, D'Angelo F, Davidsen T, Davis A, Demchok JA, Devine K, Ding L, Duell R, Elder JB, Eschbacher JM, Fehrenbach A, Ferguson M, Frazer S, Fuller G, Fulop J, Gabriel SB, Garofano L, Gastier-Foster JM, Gehlenborg N, Gerken M, Getz G, Giannini C, Gibson WJ, Hadjipanayis A, Hayes DN, Heiman DI, Hermes B, Hilty J, Hoadley KA, Hoyle AP, Huang M, Jefferys SR, Jones CD, Jones SJM, Ju Z, Kastl A, Kendler A, Kim J, Kucherlapati R, Lai PH, Lawrence MS, Lee S, Leraas KM, Lichtenberg TM, Lin P, Liu Y, Liu J, Ljubimova JY, Lu Y, Ma Y, Maglinte DT, Mahadeshwar HS, Marra M A, McGraw M, McPherson C, Meng S, Mieczkowski PA, Miller CR, Mills GB, Moore RA, Mose LE, Mungall AJ, Naresh R, Naska T, Neder L, Noble MS, Noss A, O'Neill BP, Ostrom QT, Palmer C, Pantazi A, Parfenov M, Park PJ, Parker JS, Perou CM, Pierson CR, Pihl T, Protopopov A, Radenbaugh A, Ramirez NC, Rathmell WK, Ren X, Roach J, Robertson AG, Saksena G, Schein JE, Schumacher SE, Seidman J, Senecal K, Seth S, Shen H, Shi Y, Shih J, Shimmel K, Sicotte H, Sifri S, Silva T, Simons JV, Singh R, Skelly T, Sloan AE, Sofia HJ, Soloway MG, Song X, Sougnez C, Souza C, Staugaitis SM, Sun H, Sun C, Tan D, Tang J, Tang Y, Thorne L, Trevisan FA, Triche T, Van Den Berg DJ, Veluvolu U, Voet D, Wan Y, Wang Z, Warnick R, Weinstein JN, Weisenberger DJ, Wilkerson MD, Williams F, Wise L, Wolinsky Y, Wu J, Xu AW, et al. 2016. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164(3)**:550–563 DOI 10.1016/j.cell.2015.12.028.

Cheerla A, Gevaert O. 2019. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35(14)**:i446–i454 DOI 10.1093/bioinformatics/btz342.

Cheerla N, Gevaert O. 2017. Microrna based pan-cancer diagnosis and treatment recommendation. *BMC Bioinformatics* **18(1)**:32 DOI 10.1186/s12859-016-1421-y.

Chen H, Li C, Peng X, Zhou Z, Weinstein JN, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J,

Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, Liu X, Morgan M, Morton D, Fulton LA, Fulton RS, Kandoth C, Mardis ER, McLellan MD, Miller CA, et al. 2018a. A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell* **173(2)**:386–399.

Chen H-IH, Chiu Y-C, Zhang T, Zhang S, Huang Y, Chen Y. 2018b. Gsae: an autoencoder with embedded gene-set nodes for genomics functional characterization. *BMC Systems Biology* **12(8)**:142.

Chen L, Pan X, Hu X, Zhang Y-H, Wang SP, Huang T, Cai Y-D. 2018c. G ene expression differences among different msi statuses in colorectal cancer. *International Journal of Cancer* **143(7)**:1731–1740.

Chen L, Xuan J, Gu J, Wang Y, Zhang Z, Wang TL, Shih IM. 2012. Integrative network analysis to identify aberrant pathway networks in ovarian cancer. In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 17th Pacific Symposium on Biocomputing, PSB 2012*. 31–42.

Chen X, Duan Q, Xuan Y, Sun Y, Wu R. 2017. Possible pathways used to predict different stages of lung adenocarcinoma. *Medicine* **96(17)**.

Cheng P. 2018. A prognostic 3-long noncoding rna signature for patients with gastric cancer. *Journal of Cellular Biochemistry* **119(11)**:9261–9269 DOI 10.1002/jcb.27195.

Cherniack AD, Shen H, Walter V, Stewart C, Murray BA, Bowlby R, Hu X, Ling S, Soslow RA, Broaddus RR, Zuna RE, Robertson G, Laird PW, Kucherlapati R, Mills GB, Weinstein JN, Zhang J, Akbani R, Levine DA, Akbani R, Ally A, Auman JT, Balasundaram M, Balu S, Baylin SB, Beroukhim R, Bodenheimer T, Bogomolniy F, Boice L, Bootwalla MS, Bowen J, Bowlby R, Broaddus R, Brooks D, Carlsen R, Cherniack AD, Cho J, Chuah E, Chudamani S, Cibulskis K, Cline M, Dao F, David M, Demchok JA, Dhalla N, Dowdy S, Felau I, Ferguson ML, Frazer S, Frick J, Gabriel S, Gastier-Foster JM, Gehlenborg N, Gerken M, Getz G, Gupta M, Haussler D, Hayes DN, Heiman DI, Hess J, Hoadley KA, Hoffmann R, Holt RA, Hoyle AP, Hu X, Huang M, Hutter CM, Jefferys SR, Jones SJM, Jones CD, Kanchi RS, Kandoth C, Kasaian K, Kerr S, Kim J, Lai PH, Laird PW, Lander E, Lawrence MS, Lee D, Leraas KM, Leshchiner I, Levine DA, Lichtenberg TM, Lin P, Ling S, Liu J, Liu W, Liu Y, Lolla L, Lu Y, Ma Y, Maglinte DT, Marra MA, Mayo M, Meng S, Meyerson M, Mieczkowski PA, Mills GB, Moore RA, Mose LE, Mungall AJ, Mungall K, Murray BA, Naresh R, Noble MS, Olvera N, Parker JS, Perou CM, Perou AH, Pihl T, Radenbaugh AJ, Ramirez NC, Rathmell WK, Roach J, Robertson AG, Sadeghi S, Saksena G, Salvesen HB, Schein JE, Schumacher SE, Shen H, Sheth M, Shi Y, Shih J, Simons JV, Sipahimalani P, Skelly T, Sofia HJ, Soloway MG, Soslow RA, Sougnez C, Stewart C, Sun C, Tam A, Tan D, Tarnuzzer R, Thiessen N, Thorne LB, Tse K, Tseng J, Van Den Berg DJ, Veluvolu U, Verhaak RGW, Voet D, von Bismarck A, Walter V, Wan Y, Wang Z, Wang C, Weinstein JN, Weisenberger DJ, Wilkerson MD, Winterhoff B,

Wise L, Wong T, Wu Y, Yang L, Zenklusen JC, Zhang J, Zhang H, Zhang W, Zhu J-C, Zmuda E, Zuna RE, et al. 2017. Integrated molecular characterization of uterine carcinosarcoma. *Cancer Cell* **31(3)**:411–423 DOI 10.1016/j.ccell.2017.02.010.

Chidester B, Do MN, Ma J. 2018. Discriminative bag-of-cells for imaging-genomics. In: *PSB*. World Scientific, 319–330.

Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, Bowlby R, Shen H, Hayat S, Fieldhouse R, Lester SC, Tse GMK, Factor RE, Collins LC, Allison KH, Chen Y-Y, Jensen K, Johnson NB, Oesterreich S, Mills GB, Cherniack AD, Robertson G, Benz C, Sander C, Laird PW, Hoadley KA, King TA, Perou CM, Akbani R, Auman JT, Balasundaram M, Balu S, Barr T, Beck A, Benz C, Benz S, Berrios M, Beroukhim R, Bodenheimer T, Boice L, Bootwalla MS, Bowen J, Bowlby R, Brooks D, Cherniack AD, Chin L, Cho J, Chudamani S, Ciriello G, Davidsen T, Demchok JA, Dennison JB, Ding L, Felau I, Ferguson ML, Frazer S, Gabriel S B, Gao JJ, Gastier-Foster JM, Gatza ML, Gehlenborg N, Gerken M, Getz G, Gibson WJ, Hayes DN, Heiman DI, Hoadley KA, Holbrook A, Holt RA, Hoyle AP, Hu H, Huang M, Hutter CM, Hwang ES, Jefferys SR, Jones SJM, Ju Z, Kim J, Lai PH, Laird PW, Lawrence MS, Leraas KM, Lichtenberg TM, Lin P, Ling S, Liu J, Liu W, Lolla L, Lu Y, Ma Y, Maglinte DT, Mardis E, Marks J, Marra MA, McAllister C, McLellan M, Meng S, Meyerson M, Mills GB, Moore RA, Mose LE, Mungall AJ, Murray BA, Naresh R, Noble MS, Oesterreich S, Olopade O, Parker JS, Perou CM, Pihl T, Saksena G, Schumacher SE, Shaw KRM, Ramirez NC, Rathmell WK, Rhie SK, Roach J, Robertson AG, Saksena G, Sander C, Schein JE, Schultz N, Shen H, Sheth M, Shi Y, Shih J, Shelley CS, Shriver C, Simons J V, Sofia HJ, Soloway MG, Sougnez C, Sun C, Tarnuzzer R, Tiezzi DG, Van Den Berg DJ, Voet D, Wan Y, Wang Z, Weinstein JN, Weisenberger DJ, Wilkerson MD, Wilson R, Wise L, Wiznerowicz M, Wu J, Wu Y, Yang L, Yau C, Zack TI, Zenklusen JC, Zhang H, Zhang J, Zmuda E, et al. 2015. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163(2)**:506–519 DOI 10.1016/j.cell.2015.09.033.

Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A. 2018. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine* **24(10)**:1559–1567 DOI 10.1038/s41591-018-0177-5.

Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale A-L, Brenton JD, Tavaré S, Caldas C, Aparicio S. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486(7403)**:346.

Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, Pepin F, Durinck S, Korkola JE, Griffith M, Hur JS, Huh N, Chung J, Cope L, Fackler M, Umbricht C, Sukumar S, Seth P, Sukhatme VP, Jakkula LR, Lu Y, Mills GB, Cho RJ, Collisson EA, van't Veer LJ, Spellman PT, Gray JW. 2013. Modeling precision treatment of breast cancer. *Genome Biology* **14(10)**:R110 DOI 10.1186/gb-2013-14-10-r110.

Dai Y, Sun C, Feng Y, Jia Q, Zhu B. 2018. Potent immunogenicity in brca 1-mutated patients with high-grade serous ovarian carcinoma. *Journal of Cellular and Molecular Medicine* **22(8)**:3979–3986.

Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, Buhay C, Kang H, Kim SC, Fahey CC, Hacker KE, Bhanot G, Gordenin DA, Chu A, Gunaratne PH, Biehl M, Seth S, Kaipparettu BA, Bristow CA, Donehower LA, Wallen EM, Smith AB, Tickoo SK, Tamboli P, Reuter V, Schmidt LS, Hsieh JJ, Choueiri TK, Hakimi AA, Chin L, Meyerson M,

Kucherlapati R, Park W-Y, Robertson AG, Laird PW, Henske EP, Kwiatkowski DJ, Park PJ, Morgan M, Shuch B, Muzny D, Wheeler DA, Linehan WM, Gibbs RA, Rathmell WK, Creighton CJ, Creighton CJ, Davis CF, Morgan M, Gunaratne PH, Donehower LA, Kaipparettu BA, Wheeler DA, Gibbs RA, Signoretti S, Cherniack AD, Robertson AG, Chu A, Choueiri TK, Henske EP, Kwiatkowski DJ, Reuter V, Hsieh JJ, Hakimi AA, Tickoo SK, Ricketts C, Linehan WM, Schmidt LS, Gordenin DA, Bhanot G, Seiler M, Tamboli P, Rathmell WK, Fahey CC, Hacker KE, Smith AB, Wallen EM, Shen H, Laird PW, Shuch B, Muzny D, Buhay C, Wang M, Chao H, Dahdouli M, Xi L, Kakkar N, Reid JG, Downs B, Drummond J, Morton D, Doddapaneni H, Lewis L, English A, Meng Q, Kovar C, Wang Q, Hale W, Hawes A, Kalra D, Walker K, Murray BA, Sougnez C, Saksena G, Carter SL, Schumacher SE, Tabak B, Zack TI, Getz G, Beroukhim R, Gabriel SB, Meyerson M, Ally A, Balasundaram M, Birol I, Brooks D, Butterfield YSN, Chuah E, Clarke A, Dhalla N, Guin R, Holt RA, Kasaian K, Lee D, Li HI, Lim E, Ma Y, Mayo M, Moore RA, Mungall AJ, Schein JE, Sipahimalani P, Tam A, Thiessen N, Wong T, Jones SJM, Marra M A, Auman JT, Tan D, Meng S, Jones CD, Hoadley KA, Mieczkowski PA, Mose LE, Jefferys SR, Roach J, Veluvolu U, Wilkerson MD, Waring S, Buda E, Wu J, Bodenheimer T, Hoyle AP, Simons JV, Soloway MG, Balu S, Parker JS, Hayes DN, Perou CM, Weisenberger DJ, Bootwalla MS, Triche T Jr, Lai PH, Van Den Berg DJ, Baylin SB, Chen F, Coarfa C, Noble MS, DiCara D, Zhang H, Cho J, Heiman DI, Gehlenborg N, Voet D, Lin P, Frazer S, Stojanov P, Liu Y, Zou L, Kim J, Lawrence MS, Chin L, Yang L, Seth S, Bristow CA, Protopopov A, Song X, Zhang J, Pantazi A, Hadjipanayis A, Lee E, Luquette LJ, Lee S, Parfenov M, Santoso N, Seidman J, Xu AW, Kucherlapati R, Park PJ, Kang H, et al. 2014. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26(3)**:319–330 DOI 10.1016/j.ccr.2014.07.014.

Dong C, Guo Y, Yang H, He Z, Liu X, Wang K. 2016. icages: integrated cancer genome score for comprehensively prioritizing driver genes in personal cancer genomes. *Genome Medicine* **8(1)**:135.

Ellrott K, Bailey MH, Saksena G, Covingt, andoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, Sofia HJ, Hutter C, Getz G, Wheeler D, Ding L, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker WMD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA,

Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, et al. 2018. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Systems* **6(3)**:271–281.

Ertosun MG, Rubin DL. 2015. Automated grading of gliomas using deep learning in digital pathology images: A modular approach with ensemble of convolutional neural networks. In: *AMIA Annual Symposium Proceedings*. **2015**:American Medical Informatics Association, 1899.

Fan Z, Xue W, Li L, Zhang C, Lu J, Zhai Y, Suo Z, Zhao J. 2018. Identification of an early diagnostic biomarker of lung adenocarcinoma based on co-expression similarity and construction of a diagnostic model. *Journal of Translational Medicine* **16(1)**:205.

Farshidfar F, Zheng S, Gingras M-C, Newton Y, Shih J, Robertson AG, Hinoue T, Hoadley KA, Gibb EA, Roszik J, Covington KR, Wu C-C, Shinbrot E, Stransky N, Hegde A, Yang JD, Reznik E, Sadeghi S, Pedamallu CS, Ojesina AI, Hess JM, Auman JT, Rhie SK, Bowlby R, Borad MJ, Zhu AX, Stuart JM, Sander C, Akbani R, Cherniack AD, Deshpande V, Mounajjed T, Foo WC, Torbenson MS, Kleiner DE, Laird PW, Wheeler DA, McRee AJ, Bathe OF, Andersen JB, Bardeesy N, Roberts LR, Kwong LN, Akbani R, Allotey LK, Ally A, Alvaro D, Andersen JB, Appelbaum EL, Arora A, Auman JT, Balasundaram M, Balu S, Bardeesy N, Bathe OF, Baylin SB, Beroukhim R, Berrios M, Bodenheimer T, Boice L, Bootwalla MS, Borad MJ, Bowen J, Bowlby R, Bragazzi MC, Brooks D, Cardinale V, Carlsen R, Carpino G, Carvalho AL, Chaiteerakij R, Chandan VC, Cherniack AD, Chin L, Cho J, Choe G, Chuah E, Chudamani S, Cibulskis C, Cordes MG, Covington KR, Crain D, Curley E, De Rose AM, Defreitas T, Demchok JA, Deshpande V, Dhalla N, Ding L, Evason K, Farshidfar F, Felau I, Ferguson ML, Foo WC, Franchitto A, Frazer S, Fronick CC, Fulton LA, Fulton RS, Gabriel SB, Gardner J, Gastier-Foster JM, Gaudio E, Gehlenborg N, Genovese G, Gerken M, Getz G, Giama NH, Gibbs RA, Gingras M-C, Giuliante F, Grazi GL, Hayes DN, Hegde AM, Heiman DI, Hess JM, Hinoue T, Hoadley KA, Holbrook A, Holt RA, Hoyle AP, Huang M, Hutter CM, Jefferys SR, Jones SJM, Jones CD, Kasaian K, Kelley RK, Kim J, Kleiner DE, Kocher J-PA, Kwong LN, Lai PH, Laird PW, Lawrence MS, Leraas KM, Lichtenberg TM, Lin P, Liu W, Liu J, Lolla L, Lu Y, Ma Y, Mallery D, Mardis ER, Marra MA, Matsushita MM, Mayo M, McLellan MD, McRee AJ, Meier S, Meng S, Meyerson M, Mieczkowski PA, Miller CA, Mills GB, Moore RA, Morris S, Mose LE, Moser CD, Mounajjed T, Mungall AJ, Mungall K, Murray BA, Naresh R, Newton Y, Noble MS, O'Brien DR, Ojesina AI, Parker JS, Patel TC, Paulauskis J, Pedamallu CS, Penny R, Perou CM, Perou AH, Pihl T, Radenbaugh AJ, Ramirez NC, Rathmell WK, Reznik E, Rhie SK, Roach J, Roberts LR, Robertson AG, Sadeghi S, Saksena G, Sander C, Schein JE, Schmidt HK, Schumacher SE, Shelton C, Shelton T, Shen R, Sheth M, Shi Y, Shih J, Shinbrot E, Shroff R, Simons JV, et al. 2017. Integrative genomic analysis of cholangiocarcinoma identifies distinct idh-mutant molecular profiles. *Cell Reports* **18(11)**:2780–2794 DOI 10.1016/j.celrep.2017.02.033.

Fatai AA, Gamieldien J. 2018. A 35-gene signature discriminates between rapidly-and slowly-progressing glioblastoma multiforme and predicts survival in known subtypes of the cancer. *BMC Cancer* **18(1)**:377 DOI 10.1186/s12885-018-4103-5.

Feng Y, Dai Y, Gong Z, Cheng J-N, Zhang L, Sun C, Zeng X, Jia Q, Zhu B. 2018. Association between angiogenesis and cytotoxic signatures in the tumor microenvironment of gastric cancer. *OncoTargets and Therapy* **11**:2725 DOI 10.2147/OTT.

Fernandez-Lozano C, Gestal M, Munteanu CR, Dorado J, Pazos A. 2016. A methodology for the design of experiments in computational intelligence with multiple regression models. *PeerJ* **4**:e2721.

Fischer W, Moudgalya SS, Cohn JD, Nguyen NTT, Kenyon GT. 2018. Sparse coding of pathology slides compared to transfer learning with deep neural networks. *BMC Bioinformatics* **19(18)**:489.

Fishbein L, Leshchiner I, Walter V, Danilova L, Robertson AG, Johnson AR, Lichtenberg TM, Murray BA, Ghayee HK, Else T, Ling S, Jefferys SR, de Cubas AA, Wenz B, Korpershoek E, Amelio AL, Makowski L, Rathmell WK, Gimenez-Roqueplo A-P, Giordano TJ, Asa SL, Tischler AS, Pacak K, Nathanson KL, Wilkerson MD, Akbani R, Ally A, Amar L, Amelio AL, Arachchi H, Asa SL, Auchus RJ, Auman JT, Baertsch R, Balasundaram M, Balu S, Bartsch DK, Baudin E, Bauer T, Beaver A, Benz C, Beroukhim R, Beuschlein F, Bodenheimer T, Boice L, Bowen J, Bowlby R, Brooks D, Carlsen R, Carter S, Cassol CA, Cherniack AD, Chin L, Cho J, Chuah E, Chudamani S, Cope L, Crain D, Curley E, Danilova L, de Cubas AA, de Krijger RR, Demchok JA, Deutschbein T, Dhalla N, Dimmock D, Dinjens WNM, Else T, Eng C, Eschbacher J, Fassnacht M, Felau I, Feldman M, Ferguson ML, Fiddes I, Fishbein L, Frazer S, Gabriel SB, Gardner J, Gastier-Foster JM, Gehlenborg N, Gerken M, Getz G, Geurts J, Ghayee HK, Gimenez-Roqueplo A-P, Giordano TJ, Goldman M, Graim K, Gupta M, Haan D, Hahner S, Hantel C, Haussler D, Hayes DN, Heiman DI, Hoadley KA, Holt RA, Hoyle AP, Huang M, Hunt B, Hutter CM, Jefferys SR, Johnson AR, Jones SJM, Jones CD, Kasaian K, Kebebew E, Kim J, Kimes P, Knijnenburg T, Korpershoek E, Lander E, Lawrence MS, Lechan R, Lee D, Leraas KM, Lerario A, Leshchiner I, Lichtenberg TM, Lin P, Ling S, Liu J, LiVolsi VA, Lolla L, Lotan Y, Lu Y, Ma Y, Maison N, Makowski L, Mallery D, Mannelli M, Marquard J, Marra MA, Matthew T, Mayo M, Méatchi T, Meng S, Merino MJ, Mete O, Meyerson M, Mieczkowski PA, Mills GB, Moore RA, Morozova O, Morris S, Mose LE, Mungall AJ, Murray BA, Naresh R, Nathanson KL, Newton Y, Ng S, Ni Y, Noble MS, Nwariaku F, Pacak K, Parker JS, Paul E, Penny R, Perou CM, Perou AH, Pihl T, Powers J, Rabaglia J, Radenbaugh A, Ramirez NC, Rao A, Rathmell WK, Riester A, Roach J, Robertson AG, Sadeghi S, Saksena G, Salama S, Saller C, Sandusky G, Sbiera S, Schein JE, Schumacher SE, Shelton C, Shelton T, Sheth M, Shi Y, Shih J, Shmulevich I, Simons JV, Sipahimalani P, Skelly T, Sofia HJ, Sokolov A, Soloway MG, Sougnez C, Stuart J, Sun C, Swatloski T, Tam A, Tan D, Tarnuzzer R, Tarvin K, et al. 2017. Comprehensive molecular characterization of pheochromocytoma and paraganglioma. *Cancer Cell* **31(2)**:181–193 DOI 10.1016/j.ccell.2017.01.001.

Gao S, Qiu Z, Song Y, Mo C, Tan W, Chen Q, Liu D, Chen M, Zhou H. 2017. Unsupervised clustering reveals new prostate cancer subtypes. *Translational Cancer Research* **6(3)**:561–572.

Ge Z, Leighton JS, Wang Y, Peng X, Chen Z, Chen H, Sun Y, Yao F, Li J, Zhang H, Liu J, Shriver CD, Hu H, Piwnica-Worms H, Ma L, Liang H, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ,

Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, et al. 2018. Integrated genomic analysis of the ubiquitin pathway across cancer types. *Cell Reports* **23(1)**:213–226.

Geeleher P, Zhang Z, Wang F, Gruener RF, Nath A, Morrison G, Bhutra S, Grossman RL, Huang RS. 2017. Discovering novel pharmacogenomic biomarkers by imputing drug response in cancer patients from large genomics studies. *Genome Research* **27(10)**:1743–1751.

Ghoshal A, Zhang J, Roth MA, Xia KM, Grama AY, Chaterji S. 2018. A distributed classifier for microrna target prediction with validation through tcga expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **15(4)**:1037–1051.

Graudenzi A, Cava C, Bertoli G, Fromm B, Flatmark K, Mauri G, Castiglioni I. 2017. Pathway-based classification of breast cancer subtypes. *Frontiers in Bioscience* **22**:1697–1712.

Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, Flagg K, Hou J, Zhang H, Yi S, Jafari M, Lin D, Chung C, Caughey BA, Li G, Dhar D, Shi W, Zheng L, Hou R, Zhu J, Zhao L, Fu X, Zhang E, Zhang C, Zhu J-K, Karin M, Xu R-H, Zhang K. 2017. Dna methylation markers for diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences of the United States of America* **114(28)**:7414–7419 DOI 10.1073/pnas.1703577114.

Hmeljak J, Sanchez-Vega F, Hoadley KA, Shih J, Stewart C, Heiman D, Tarpey P, Danilova L, Drill E, Gibb EA, Bowlby R, Kanchi R, Osmanbeyoglu HU, Sekido Y, Takeshita J, Newton Y, Graim K, Gupta M, Gay CM, Diao L, Gibbs DL, Thorsson V, Iype L, Kantheti H, Severson DT, Ravegnini G, Desmeules P, Jungbluth AA, Travis WD, Dacic S, Chirieac LR, Fçoise G-Sallé, Fujimoto J, Husain AN, Silveira HC, Rusch VW, Rintoul RC, Pass H, Kindler H, Zauderer MG, Kwiatkowski DJ, Bueno R, Tsao AS, Creaney J, Lichtenberg T, Leraas K, Bowen J, Zenklusen JC, Akbani R, Cherniack AD, Byers LA, Noble MS, Fletcher JA, Robertson AG, Shen R, Aburatani H, Robinson BW, Campbell P, Ladanyi M. 2018. Integrative molecular characterization of malignant pleural mesothelioma. *Cancer Discovery* **8(12)**:1548–1565.

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, Akbani R, Bowlby R, Wong CK, Wiznerowicz M, Sanchez-Vega F, Robertson AG, Schneider BG, Lawrence MS, Noushmehr H, Malta TM, Stuart JM, Benz CC, Laird PW, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DAkbani R, Bowlby R, Wong CK, Wiznerowicz M, Sanchez-Vega F, Robertson AG, Schneider BG, Lawrence MS, Noushmehr H, Malta TM, Stuart JM, Benz CC, Laird PW, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y,

Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, et al. 2018. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173(2)**:291–304.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, Zhang J, Kandoth C, Akbani R, Shen H, Omberg L, Chu A, Margolin AA, van't Veer LJ, Lopez-Bigas N, Laird PW, Raphael BJ, Ding L, Robertson AG, Byers LA, Mills GB, Weinstein JN, Van Waes C, Chen Z, Collisson EA, Benz CC, Perou CM, Stuart JM. 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158(4)**:929–944.

Holzinger A, Malle B, Saranti A, Pfeifer B. 2021. Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. *Information Fusion* **71(7639)**:28–37 DOI 10.1016/j.inffus.2021.01.008.

Huang K-l, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, Scott AD, Krassowski M, Cherniack AD, Houlahan KE, Jayasinghe R, Wang L-B, Zhou DC, Liu D, Cao S, Kim YW, Koire A, McMichael JF, Hucthagowder V, Kim T-B, Hahn A, Wang C, McLellan MD, Al-Mulla F, Johnson KJ, Lichtarge O, Boutros PC, Raphael B, Lazar AJ, Zhang W, Wendl MC, Govindan R, Jain S, Wheeler D, Kulkarni S, Dipersio JF, Jüri R, Meric-Bernstam F, Chen K, Shmulevich I, Plon SE, Chen F, Ding L, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, et al. 2018. Pathogenic germline variants in 10,389 adult cancers. *Cell* **173(2)**:355–370.

Ing N, Huang F, Conley A, You S, Ma Z, Klimov S, Ohe C, Yuan X, Amin MB, Figlin R, Gertych A, Knudsen BS. 2017. A novel machine learning approach reveals latent vascular phenotypes predictive of renal cancer outcome. *Scientific Reports* **7(1)**:13190 DOI 10.1038/s41598-017-13196-4.

Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. 2019. Histoqc: an open-source quality control tool for digital pathology slides. *JCO Clinical Cancer Informatics* **3(3)**:1–7 DOI 10.1200/CCI.18.00157.

Jean-Quartier C, Jeanquartier F, Ridvan A, Kargl M, Mirza T, Stangl T, Markaĉ R, Jurada M, Holzinger A. 2021. Mutation-based clustering and classification analysis reveals distinctive age groups and age-related biomarkers for glioma. *BMC Medical Informatics and Decision Making* **21(1)**:1–14.

Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O, Sander C, The Cancer Genome Atlas Research Network. 2018. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* **34(2)**:211–224.

Kanas VG, Zacharaki EI, Thomas GA, Zinn PO, Megalooikonomou V, Colen RR. 2017. Learning mri-based classification models for mgmt methylation status prediction in glioblastoma. *Computer Methods and Programs in Biomedicine* **140**:249–257.

Karczewski KJ, Snyder MP. 2018. Integrative omics for health and disease. *Nature Reviews Genetics* **19(5)**:299.

Kim S, Oesterreich S, Kim S, Park Y, Tseng GC. 2017. Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics* **18(1)**:165–179 DOI 10.1093/biostatistics/kxw039.

Klein MI, Stern DF, Zhao H. 2017. Grape: a pathway template method to characterize tissue-specific functionality from gene expression profiles. *BMC Bioinformatics* **18(1)**:317 DOI 10.1186/s12859-017-1711-z.

Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, Greene CS, Liu Y, Akbani R, Feng B, Donehower LA, Miller C, Shen Y, Karimi M, Chen H, Kim P, Jia P, Shinbrot E, Zhang S, Liu J, Hu H, Bailey MH, Yau C, Wolf D, Zhao Z, Weinstein JN, Li L, Ding L, Mills GB, Laird PW, Wheeler DA, Shmulevich I, Monnat RJ, Xiao Y, Wang C, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, et al. 2018. Genomic and molecular

landscape of dna damage repair deficiency across The Cancer Genome Atlas. *Cell Reports* **23(1)**:239–254.

Kocak B, Yardimci AH, Bektas CT, Turkcanoglu MH, Erdim C, Yucetas U, Koca SB, Kilickesmez O. 2018. Textural differences between renal cell carcinoma subtypes: machine learning-based quantitative computed tomography texture analysis with independent external validation. *European Journal of Radiology* **107**:149–157.

Koo J, Zhang J, Chaterji S. 2018. Tiresias: context-sensitive approach to decipher the presence and strength of microrna regulatory interactions. *Theranostics* **8(1)**:277–291 DOI 10.7150/thno.22065.

Kristensen V, Lingjærde O, Russnes H, Vollan H, Frigessi A, Børresen-Dale A. 2014. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer* **14**:299–313.

Leung MK, Delong A, Alipanahi B, Frey BJ. 2015. Machine learning in genomic medicine: a review of computational problems and data sets. *Proceedings of the IEEE* **104(1)**:176–197 DOI 10.1109/JPROC.2015.2494198.

Levine DA, The Cancer Genome Atlas Research Network. 2013. Integrated genomic characterization of endometrial carcinoma. *Nature* **497(7447)**:67–73 DOI 10.1038/nature12113.

Liao Z, Li D, Wang X, Li L, Zou Q. 2018. Cancer diagnosis through isomir expression with machine learning method. *Current Bioinformatics* **13(1)**:57–63.

Liñares Blanco J, Gestal M, Dorado J, Fernandez-Lozano C. 2019. *Differential gene expression analysis of RNA-seq data using machine learning for cancer research*. Cham: Springer International Publishing, 27–65.

List M, Hauschild A-C, Tan Q, Kruse TA, Baumbach J, Batra R. 2014. Classification of breast cancer subtypes by combining gene expression and dna methylation data. *Journal of Integrative Bioinformatics* **11(2)**:1–14.

Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, Omberg L, Wolf DM, Shriver CD, Thorsson V, Hu H, Cancer Genome Atlas Research Network. 2018a. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173(2)**:400–416.

Liu Y, Sethi NS, Hinoue T, Schneider BG, Cherniack AD, Sanchez-Vega F, Seoane JA, Farshidfar F, Bowlby R, Islam M, Kim J, Chatila W, Akbani R, Kanchi RS, Rabkin CS, Willis JE, Wang KK, McCall SJ, Mishra L, Ojesina AI, Bullman S, Pedamallu CA, Lazar AJ, Sakai R, Thorsson V, Bass AJ, Laird RW, Cancer Genome Atlas Research Network. 2018b. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell* **33(4)**:721–735.

Mallavarapu T, Hao J, Kim Y, Oh JH, Kanga M. 2019. Pathway-based deep clustering for molecular subtyping of cancer. *Methods* **173**:24–31 DOI 10.1016/j.ymeth.2019.06.017.

Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O, Colaprico A, Czerwińska P, Mazurek S, Mishra L, Heyn H, Krasnitz A, Godwin AK, Lazar AJ, The Cancer Genome Atlas Research Network. 2018. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* **173(2)**:338–354.

Mo Q, Shen R, Guo C, Vannucci M, Chan KS, Hilsenbeck SG. 2017. A fully bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics* **19(1)**:71–86.

Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R. 2013. Pattern discovery and cancer gene identification in integrated cancer genomic data.

Proceedings of the National Academy of Sciences of the United States of America **110(11)**:4245–4250 DOI 10.1073/pnas.1208949110.

Muhamed Ali A, Zhuang H, Ibrahim A, Rehman O, Huang M, Wu A. 2018. A machine learning approach for the classification of kidney cancer subtypes using mirna genome data. *Applied Sciences* **8(12)**:2422.

Nair J, Jain P, Chandola U, Palve V, Vardhan NRH, Reddy RB, Kekatpure VD, Suresh A, Kuriakose MA, Panda B. 2015. Gene and mirna expression changes in squamous cell carcinoma of larynx and hypopharynx. *Genes & Cancer* **6(7–8)**:328–340 DOI 10.18632/genesandcancer.69.

Nguyen T, Tagett R, Diaz D, Draghici S. 2017. A novel approach for data integration and disease subtyping. *Genome Research* **27(12)**:2025–2039 DOI 10.1101/gr.215129.116.

Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, Verhaak RGW, Hoadley KA, Hayes DN, Perou CM, Schmidt HK, Ding L, Wilson RK, Van Den Berg D, Shen H, Bengtsson H, Neuvial P, Cope LM, Buckley J, Herman JG, Baylin SB, Laird PW, Aldape K. 2010. Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17(5)**:510–522 DOI 10.1016/j.ccr.2010.03.017.

Ou-Yang L, Zhang X-F, Wu M, Li X-L. 2017. Node-based learning of differential networks from multi-platform gene expression data. *Methods* **129(1)**:41–49 DOI 10.1016/j.ymeth.2017.05.014.

Park YW, Choi YS, Ahn SS, Chang JH, Kim SH, Lee S-K. 2019. Radiomics mri phenotyping with machine learning to predict the grade of lower-grade gliomas: A study focused on nonenhancing tumors. *Korean Journal of Radiology* **20(9)**:1381–1389.

Peng X, Chen Z, Farshidfar F, Xu X, Lorenzi PL, Wang Y, Cheng F, Tan L, Mojumdar K, Du D, Ge Z, Li J, Thomas GV, Birsoy K, Liu L, Zhang H, Zhao Z, Marchand C, Weinstein JN, Bathe OF, Liang H, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, et al. 2018. Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. *Cell Reports* **23(1)**:255–269.

Powell RT, Olar A, Narang S, Rao G, Sulman E, Fuller GN, Rao A. 2017. Identification of histological correlates of overall survival in lower grade gliomas using a bag-of-words paradigm: A preliminary analysis based on hematoxylin & eosin stained slides from the lower grade glioma cohort of The Cancer Genome Atlas. *Journal of Pathology Informatics* **8**:9.

Radovich M, Pickering CR, Felau I, Ha G, Zhang H, Jo H, Hoadley KA, Anur P, Zhang J, McLellan M, Bowlby R, Matthew T, Danilova L, Hegde AM, Kim J, Leiserson MDM, Sethi G, Lu C, Ryan M, Su X, Cherniack AD, Robertson G, Akbani R, Spellman P, Weinstein JN, Hayes DN, Raphael B, Lichtenberg T, Leraas K, Zenklusen JC, Fujimoto J, Scapulatempo-Neto C, Moreira AL, Hwang D, Huang J, Marino M, Korst R, Giaccone G, Gokmen-Polar Y, Badve S, Rajan A, Ströbel P, Girard N, Tsao MS, Marx A, Tsao AS, Loehrer PJ, Ally A, Appelbaum EL, Auman JT, Balasundaram M, Balu S, Behera M, Beroukhim R, Berrios M, Blandino G, Bodenheimer T, Bootwalla MS, Bowen J, Brooks D, Carcano FM, Carlsen R, Carvalho AL, Castro P, Chalabreysse L, Chin L, Cho J, Choe G, Chuah E, Chudamani S, Cibulskis C, Cope L, Cordes MG, Crain D, Curley E, Defreitas T, Demchok JA, Detterbeck F, Dhalla N, Dienemann H, Edenfield WJ, Facciolo F, Ferguson ML, Frazer S, Fronick CC, Fulton LA, Fulton RS, Gabriel SB, Gardner J, Gastier-Foster JM, Gehlenborg N, Gerken M, Getz G, Heiman DI, Hobensack S, Holbrook A, Holt RA, Hoyle AP, Hutter CM, Ittmann M, Jefferys SR, Jones CD, Jones SJM, Kasaian K, Kim J, Kimes PK, Lai PH, Laird PW, Lawrence MS, Lin P, Liu J, Lolla L, Lu Y, Ma Y, Maglinte DT, Mallery D, Mardis ER, Marra MA, Martin J, Mayo M, Meier S, Meister M, Meng S, Meyerson M, Mieczkowski PA, Miller CA, Mills GB, Moore RA, Morris S, Mose LE, Muley T, Mungall AJ, Mungall K, Naresh R, Newton Y, Noble MS, Owonikoko T, Parker JS, Paulaskis J, Penny R, Perou CM, Perrin C, Pihl T, Radenbaugh A, Ramalingam S, Ramirez N, Rieker R, Roach J, Sadeghi S, Saksena G, Schein JE, Schmidt HK, Schumacher SE, Shelton C, Shelton T, Shi Y, Shih J, Sica G, Silveira HCS, Simons JV, Sipahimalani P, Skelly T, Sofia HJ, Soloway MG, Stuart J, Sun Q, Tam A, Tan D, Tarnuzzer R, Thiessen N, Van Den Berg DJ, Vasef MA, Veluvolu U, Voet D, Walter V, Wan Y, Wang Z, Warth A, Weis C-A, Weisenberger DJ, Wilkerson MD, Wise L, Wong T, Wu H-T, Wu Y, Yang L, Zhang J, Zmuda E, et al. 2018. The integrated genomic landscape of thymic epithelial tumors. *Cancer Cell* **33(2)**:244–258 DOI 10.1016/j.ccell.2018.01.003.

Raphael BJ, Hruban RH, Aguirre AJ, Moffitt RA, Yeh JJ, Stewart C, Robertson AG, Cherniack AD, Gupta M, Getz G, Gabriel SB, Meyerson M, Cibulskis C, Fei SS, Hinoue T, Shen H, Laird PW, Ling S, Lu Y, Mills GB, Akbani R, Loher P, Londin ER, Rigoutsos I, Telonis AG, Gibb EA, Goldenberg A, Mezlini AM, Hoadley KA, Collisson E, Lander E, Murray BA, Hess J, Rosenberg M, Bergelson L, Zhang H, Cho J, Tiao G, Kim J, Livitz D, Leshchiner I, Reardon B, Van Allen E, Kamburov A, Beroukhim R, Saksena G, Schumacher SE, Noble MS, Heiman DI, Gehlenborg N, Kim J, Lawrence MS, Adsay V, Petersen G, Klimstra D, Bardeesy N, Leiserson MDM, Bowlby R, Kasaian K, Birol I, Mungall KL, Sadeghi S, Weinstein JN, Spellman PT, Liu Y, Amundadottir LT, Tepper J, Singhi AD, Dhir R, Paul D, Smyrk T, Zhang L, Kim P, Bowen J, Frick J, Gastier-Foster JM, Gerken M, Lau K, Leraas KM, Lichtenberg TM, Ramirez NC, Renkel J, Sherman M, Wise L, Yena P, Zmuda E, Shih J, Ally A, Balasundaram M, Carlsen R, Chu A, Chuah E, Clarke A, Dhalla N, Holt RA, Jones SJM, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Brooks D, Auman JT, Balu S, Bodenheimer T, Hayes DN, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou CM, Perou AH, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Parker JS, Wilkerson MD, Korkut A, Senbabaoglu Y, Burch P, McWilliams R, Chaffee K, Oberg A, Zhang W, Gingras M-C, Wheeler DA, Xi L, Albert M, Bartlett J, Sekhon H, Stephen Y, Howard Z, Judy M, Breggia A, Shroff RT, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Jennifer S, Roggin K, Becker K-F, Behera M, Bennett J, Boice L, Burks E, Carlotti Junior

CG, Chabot J, Pretti da Cunha Tirapelli D, Sebastião dos Santos J, Dubina M, Eschbacher J, Huang M, Huelsenbeck-Dill L, Jenkins R, Karpov A, Kemp R, Lyadov V, Maithel S, Manikhas G, Montgomery E, Noushmehr H, Osunkoya A, Owonikoko T, Paklina O, Potapova O, Ramalingam S, Rathmell WK, Rieger-Christ K, Saller C, Setdikova G, Shabunin A, Sica G, Su T, Sullivan T, Swanson P, Tarvin K, Tavobilov M, Thorne LB, Urbanski S, Voronina O, Wang T, Crain D, et al. 2017. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32(2)**:185–203 DOI 10.1016/j.ccell.2017.07.007.

Rendleman MC, Buatti JM, Braun TA, Smith BJ, Nwakama C, Beichel RR, Brown B, Casavant TL. 2019. Machine learning with the tcga-hnsc dataset: improving usability by addressing inconsistency, sparsity, and high-dimensionality. *BMC Bioinformatics* **20(1)**:339 DOI 10.1186/s12859-019-2929-8.

Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, Hinoue T, Laird PW, Hoadley KA, Akbani R, Castro MAA, Gibb EA, Kanchi RS, Gordenin DA, Shukla SA, Sanchez-Vega F, Hansel DE, Czerniak BA, Reuter VE, Su X, de Sa Carvalho B, Chagas VS, Mungall KL, Sadeghi S, Pedamallu CS, Lu Y, Klimczak LJ, Zhang J, Choo C, Ojesina AI, Bullman S, Leraas KM, Lichtenberg TM, Wu CJ, Schultz N, Getz G, Meyerson M, Mills GB, McConkey DJ, Weinstein JN, Kwiatkowski DJ, Lerner SP, Akbani R, Al-Ahmadie H, Albert M, Alexopoulou I, Ally A, Antic T, Aron M, Balasundaram M, Bartlett J, Baylin SB, Beaver A, Bellmunt J, Birol I, Boice L, Bootwalla MS, Bowen J, Bowlby R, Brooks D, Broom BM, Bshara W, Bullman S, Burks E, Cárcano FM, Carlsen R, Carvalho BS, Carvalho AL, Castle EP, Castro MAA, Castro P, Catto JW, Chagas VS, Cherniack AD, Chesla DW, Choo C, Chuah E, Chudamani S, Cortessis VK, Cottingham SL, Crain D, Curley E, Czerniak BA, Daneshmand S, Demchok JA, Dhalla N, Djaladat H, Eckman J, Egea SC, Engel J, Felau I, Ferguson ML, Gardner J, Gastier-Foster JM, Gerken M, Getz G, Gibb EA, Gomez-Fernandez CR, Gordenin DA, Guo G, Hansel DE, Harr J, Hartmann A, Herbert LM, Hinoue T, Ho TH, Hoadley KA, Holt RA, Hutter CM, Jones SJM, Jorda M, Kahnoski RJ, Kanchi RS, Kasaian K, Kim J, Klimczak LJ, Kwiatkowski DJ, Lai PH, Laird PW, Lane BR, Leraas KM, Lerner SP, Lichtenberg TM, Liu J, Lolla L, Lotan Y, Lu Y, Lucchesi FR, Ma Y, Machado RD, Maglinte DT, Mallery D, Marra MA, Martin SE, Mayo M, McConkey DJ, Meraney A, Meyerson M, Mills GB, Moinzadeh A, Moore RA, Mora Pinero EM, Morris S, Morrison C, Mungall KL, Mungall AJ, Myers JB, Naresh R, O'Donnell PH, Ojesina AI, Parekh DJ, Parfitt J, Paulauskis JD, Sekhar Pedamallu C, Penny RJ, Pihl T, Porten S, Quintero-Aguilo ME, Ramirez NC, Rathmell WK, Reuter VE, Rieger-Christ K, Robertson AG, Sadeghi S, Saller C, Salner A, Sanchez-Vega F, Sandusky G, Scapulatempo-Neto C, Schein JE, Schuckman AK, Schultz N, Shelton C, Shelton T, Shukla SA, Simko J, Singh P, Sipahimalani P, Smith ND, Sofia HJ, Sorcini A, Stanton ML, Steinberg GD, Stoehr R, Su X, Sullivan T, Sun Q, Tam A, Tarnuzzer R, Tarvin K, Taubert H, Thiessen N, Thorne L, Tse K, Tucker K, Van Den Berg DJ, van Kessel KE, Wach S, Wan Y, Wang Z, et al. 2017a. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* **171(3)**:540–556 DOI 10.1016/j.cell.2017.09.007.

Robertson AG, Shih J, Yau C, Gibb EA, Oba J, Mungall KL, Hess JM, Uzunangelov V, Walter V, Danilova L, Lichtenberg TM, Kucherlapati M, Kimes PK, Tang M, Penson A, Babur O, Akbani R, Bristow CA, Hoadley KA, Iype L, Chang MT, Cherniack AD, Benz C, Mills GB, Verhaak RGW, Griewank KG, Felau I, Zenklusen JC, Gershenwald JE, Schoenfield L, Lazar AJ, Abdel-Rahman MH, Roman-Roman S, Stern M-H, Cebulla CM, Williams MD, Jager MJ, Coupland SE, Esmaeli B, Kandoth C, Woodman SE, Abdel-Rahman MH, Akbani R, Ally A, Auman JT, Babur O, Balasundaram M, Balu S, Benz C, Beroukhim R, Birol I, Bodenheimer T, Bowen J, Bowlby R, Bristow CA, Brooks D, Carlsen R, Cebulla CM, Chang MT, Cherniack AD, Chin L, Cho J, Chuah E, Chudamani S, Cibulskis C, Cibulskis K, Cope L, Coupland SE, Danilova L, Defreitas T, Demchok JA, Desjardins L, Dhalla N, Esmaeli B,

Felau I, Ferguson ML, Frazer S, Gabriel SB, Gastier-Foster JM, Gehlenborg N, Gerken M, Gershenwald JE, Getz G, Gibb EA, Griewank KG, Grimm EA, Hayes DN, Hegde AM, Heiman DI, Helsel C, Hess JM, Hoadley KA, Hobensack S, Holt RA, Hoyle AP, Hu X, Hutter CM, Jager MJ, Jefferys SR, Jones CD, Jones SJM, Kandoth C, Kasaian K, Kim J, Kimes PK, Kucherlapati M, Kucherlapati R, Lander E, Lawrence MS, Lazar AJ, Lee S, Leraas KM, Lichtenberg TM, Lin P, Liu J, Liu W, Lolla L, Lu Y, Iype L, Ma Y, Mahadeshwar HS, Mariani O, Marra MA, Mayo M, Meier S, Meng S, Meyerson M, Mieczkowski PA, Mills GB, Moore RA, Mose LE, Mungall AJ, Mungall KL, Murray BA, Naresh R, Noble MS, Oba J, Pantazi A, Parfenov M, Park PJ, Parker JS, Penson A, Perou CM, Pihl T, Pilarski R, Protopopov A, Radenbaugh A, Rai K, Ramirez NC, Ren X, Reynolds SM, Roach J, Robertson AG, Roman-Roman S, Roszik J, Sadeghi S, Saksena G, Sastre X, Schadendorf D, Schein JE, Schoenfield L, Schumacher SE, Seidman J, Seth S, Sethi G, Sheth M, Shi Y, Shields C, Shih J, Shmulevich I, Simons JV, Singh AD, Sipahimalani P, Skelly T, Sofia H, Soloway MG, Song X, Stern M-H, Stuart J, Sun Q, Sun H, Tam A, Tan D, Tang M, Tang J, Tarnuzzer R, Taylor BS, Thiessen N, Thorsson V, Tse K, Uzunangelov V, Veluvolu U, Verhaak RGW, Voet D, Walter V, Wan Y, Wang Z, Weinstein JN, Wilkerson MD, Williams MD, et al. 2017b. Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer cell* **32(2)**:204–220 DOI 10.1016/j.ccell.2017.07.003.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)* **115(3)**:211–252 DOI 10.1007/s11263-015-0816-y.

Rykunov D, Beckmann ND, Li H, Uzilov A, Schadt EE, Reva B. 2016. A new molecular signature method for prediction of driver cancer pathways from transcriptional data. *Nucleic Acids Research* **44(11)**:e110.

Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, Samaras D, Shroyer KR, Zhao T, Batiste R, Van Arnam J, Shmulevich I, Rao AUK, Lazar AJ, Sharma A, Vésteinn T, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J,

Korchina V, Lee S, et al. 2018. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Reports* **23(1)**:181–193.

Salvucci M, Würstle ML, Morgan C, Curry S, Cremona M, Lindner AU, Bacon O, Resler AJ, Murphy ÁC, O'Byrne R, Flanagan L, Dasgupta S, Rice N, Pilati C, Zink E, Schöller LM, Toomey S, Lawler M, Johnston PG, Wilson R, Camilleri-Broët S, Salto-Tellez M, McNamara DA, Kay EW, Laurent-Puig P, Van Schaeybroeck S, Hennessy BT, Longley DB, Rehm M, Prehn JHM. 2017. A stepwise integrated approach to personalized risk predictions in stage iii colorectal cancer. *Clinical Cancer Research* **23(5)**:1200–1212 DOI 10.1158/1078-0432.CCR-16-1084.

Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, Chakravarty D, Daian F, Gao Q, Bailey MH, W-W Liang W-W, Foltz SM, Shmulevich I, Ding L, Heins Z, Ochoa A, Gross B, Gao J, Zhang H, Kundra R, Kandoth C, Bahceci I, Dervishi L, Dogrusoz U, Zhou W, Shen H, Laird P, Way GP, Greene CS, Liang H, Xiao Y, Wang C, Iavarone A, Berger AH, Bivona TG, Lazar AJ, Hammer GD, Giordano T, Kwong LN, McArthur G, Huang C, Tward AD, Frederick MJ, McCormick F, Meyerson M, Van Allen EM, Cherniack AD, Ciriello G, Sander C, Schultz N, Cancer Genome Atlas Research Network. 2018. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173(2)**:321–337.

Schaub FX, Dhankani V, Berger AC, Trivedi M, Richardson AB, Shaw R, Zhao W, Zhang X, Ventura A, Liu Y, Ayer DE, Hurlin PJ, Cherniack AD, Eisenman RN, Bernard B, Grandori C, Network Cancer Genome Atlas. 2018. Pan-cancer alterations of the myc oncogene and its proximal network across The Cancer Genome Atlas. *Cell Systems* **6(3)**:282–300.

Seoane JA, Day INM, Gaunt TR, Campbell C. 2013. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics* **30(6)**:838–845.

Shen H, Shih J, Hollern DP, Wang L, Bowlby R, Tickoo SK, Thorsson V, Mungall AJ, Newton Y, Hegde AM, Armenia J, Sánchez-Vega F, Pluta J, Pyle LC, Mehra R, Reuter VE, Godoy G, Jones J, Shelley CS, Feldman DR, Vidal DO, Lessel D, Kulis T, Cárcano FM, Leraas KM, Lichtenberg TM, Brooks D, Cherniack AD, Cho J, Heiman DI, Kasaian K, Liu M, Noble MS, Xi L, Zhang H, Zhou W, ZenKlusen JC, Hutter CM, Felau I, Zhang J, Schultz N, Getz G, Meyerson M, Stuart JM, Akbani R, Wheeler DA, Laird PW, Nathanson KL, Cortessis VK, Hoadley KA, Cancer Genome Atlas Research Network. 2018. Integrated molecular characterization of testicular germ cell tumors. *Cell Reports* **23(11)**:3392–3406.

Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, Ladanyi M, Sander C. 2012. Integrative subtype discovery in glioblastoma using icluster. *PLOS ONE* **7(4)**:e35236.

Shen R, Olshen AB, Ladanyi M. 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25(22)**:2906–2912 DOI 10.1093/bioinformatics/btp543.

Sherafatian M. 2018. Tree-based machine learning algorithms identified minimal set of mirna biomarkers for breast cancer diagnosis and molecular subtyping. *Gene* **677(2)**:111–118 DOI 10.1016/j.gene.2018.07.057.

Srivastava S, Wang W, Manyam G. 2013. et al.Integrating multi-platform genomic data using hierarchical bayesian relevance vector machines. *EURASIP Journal on Bioinformatics and Systems Biology* **2013(1)**:9 DOI 10.1186/1687-4153-2013-9.

Stephen RP, Lewis JF. 2013. Clinical and molecular models of glioblastoma multiforme survival. *International Journal of Data Mining and Bioinformatics* **7(3)**:245–265 DOI 10.1504/IJDMB.2013.053310.

Sun D, Chen J, Liu L, Zhao G, Dong P, Wu B, Wang J, Dong L. 2018a. Establishment of a 12-gene expression signature to predict colon cancer prognosis. *PeerJ* **6**:e4942.

Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, Verlingue Lïc, Brandao D, Lancia A, Ammari S, Hollebecque A, Scoazec J-Y, Marabelle A, Massard C, Soria J-C, Robert C, Paragios N, Deutsch E, Ferté C. 2018b. A radiomics approach to assess tumour-infiltrating cd8 cells and response to anti-pd-1 or anti-pd-l1 immunotherapy: an imaging biomarker, retrospective multicohort study. *The Lancet Oncology* **19(9)**:1180–1191 DOI 10.1016/S1470-2045(18)30413-3.

Sutton EJ, Huang EP, Drukker K, Burnside ES, Li H, Net JM, Rao A, Whitman GJ, Zuley M, Ganott M, Bonaccio E, Giger ML, Morris EA, On behalf of the TCGA group. 2017. Breast mri radiomics: comparison of computer-and human-extracted imaging phenotypes. *European Radiology Experimental* **1(1)**:22 DOI 10.1186/s41747-017-0025-2.

Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, Lazar AJ, Cherniack AD, Beroukhim R, Meyerson M, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V, Lee S, Lewis L, Li W, et al. 2018. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer cell* **33(4)**:676–689.

The Cancer Genome Atlas Network. 2012a. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487(7407)**:330–337 DOI 10.1038/nature11252.

The Cancer Genome Atlas Network. 2012b. Comprehensive molecular portraits of human breast tumours. *Nature* **490(7418)**:61–70 DOI 10.1038/nature11412.

The Cancer Genome Atlas Network. 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517(7536)**:576–582 DOI 10.1038/nature14129.

The Cancer Genome Atlas Research Network. 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine* **368(22)**:2059–2074 DOI 10.1056/NEJMoa1301689.

The Cancer Genome Atlas Research Network. 2015. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine* **372(26)**:2481–2498 DOI 10.1056/NEJMoa1402121.

The Cancer Genome Atlas Research Network. 2016. Comprehensive molecular characterization of papillary renal-cell carcinoma. *New England Journal of Medicine* **374(2)**:135–145 DOI 10.1056/NEJMoa1505917.

The Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455(7216)**:1061–1068 DOI 10.1038/nature07385.

The Cancer Genome Atlas Research Network. 2011. Integrated genomic analyses of ovarian carcinoma. *Nature* **474(7353)**:609–615 DOI 10.1038/nature10166.

The Cancer Genome Atlas Research Network. 2012c. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489(7417)**:519–525 DOI 10.1038/nature11404.

The Cancer Genome Atlas Research Network. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499(7456)**:43–49 DOI 10.1038/nature12222.

The Cancer Genome Atlas Research Network. 2014a. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513(7517)**:202–209 DOI 10.1038/nature13480.

The Cancer Genome Atlas Research Network. 2014b. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507(7492)**:315–322 DOI 10.1038/nature12965.

The Cancer Genome Atlas Research Network. 2014c. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511(7511)**:543–550 DOI 10.1038/nature13385.

The Cancer Genome Atlas Research Network. 2017a. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* **171(4)**:950–965.e28 DOI 10.1016/j.cell.2017.10.014.

The Cancer Genome Atlas Research Network. 2017b. Integrated genomic and molecular characterization of cervical cancer. *Nature* **543(7645)**:378–384 DOI 10.1038/nature21386.

The Cancer Genome Atlas Research Network. 2017c. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541(7636)**:169–175 DOI 10.1038/nature20805.

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Yang T-HO, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Mokrab Y, Newman AM, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedamallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CS, Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich I, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M,

Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Laird PW, Shen H, et al. 2018. The immune landscape of cancer. *Immunity* **48(4)**:812–830.

Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN. 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer Cell* **17(1)**:98–110 DOI 10.1016/j.ccr.2009.12.020.

Vural S, Wang X, Guda C. 2016. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Systems Biology* **10(3)**:62 DOI 10.1186/s12918-016-0306-z.

Wang C, Liang C. 2018. Msipred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine. *Scientific Reports* **8(1)**:17546 DOI 10.1038/s41598-018-35682-z.

Wang X, Han L, Zhou L, Wang L, Zhang L-M. 2018. Prediction of candidate rna signatures for recurrent ovarian cancer prognosis by the construction of an integrated competing endogenous rna network. *Oncology Reports* **40(5)**:2659–2673 DOI 10.3892/or.2018.6707.

Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, Sander C, Cherniack AD, Mina M, Ciriello G, Schultz N, Sanchez Y, Greene CS, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, de Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ, Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, Laird PW, Shen H, Zhou W, Bellair M, Chang K, Covington K, Creighton CJ, Dinh H, Doddapaneni HV, Donehower LA, Drummond J, Gibbs RA, Glenn R, Hale W, Han Y, Hu J, Korchina V,

Lee S, Lewis L, Li W, Liu X, et al. 2018. Machine learning detects pan-cancer ras pathway activation in The Cancer Genome Atlas. *Cell Reports* **23(1)**:172–180.

Wei D. 2018. A multigene support vector machine predictor for metastasis of cutaneous melanoma. *Molecular Medicine Reports* **17(2)**:2907–2914.

Wen J-X, Li X-Q, Chang Y. 2018. Signature gene identification of cancer occurrence and pattern recognition. *Journal of Computational Biology* **25(8)**:907–916 DOI 10.1089/cmb.2017.0261.

Wilop S, Chou W-C, Jost E, Crysandt M, Panse J, Chuang M-K, Brümmendorf TH, Wagner W, Tien H-F, Kharabi Masouleh B. 2016. A three-gene expression-based risk score can refine the european leukemianet aml classification. *Journal of Hematology & Oncology* **9(1)**:78 DOI 10.1186/s13045-016-0308-8.

Wong KK, Rostomily R, Wong ST. 2019. Prognostic gene discovery in glioblastoma patients using deep learning. *Cancers* **11(1)**:53 DOI 10.3390/cancers11010053.

Xie H, Xu H, Hou Y, Cai Y, Rong Z, Song W, Wang W, Li K. 2019. Integrative prognostic subtype discovery in high-grade serous ovarian cancer. *Journal of Cellular Biochemistry* **120(11)**:18659–18666 DOI 10.1002/jcb.29049.

Xu G, Zhang M, Zhu H, Xu J. 2017. A 15-gene signature for prediction of colon cancer recurrence and prognosis based on svm. *Gene* **604**:33–40 DOI 10.1016/j.gene.2016.12.016.

Yang S, Xu J, Zeng X. 2018. A six-long non-coding rna signature predicts prognosis in melanoma patients. *International Journal of Oncology* **52(4)**:1178–1188 DOI 10.3892/ijo.2018.4268.

Yang W, Yoshigoe K, Qin X, Liu JS, Yang JY, Niemierko A, Deng Y, Liu Y, Dunker AK, Chen Z, Wang L, Xu D, Arabnia HR, Tong W, Yang MQ. 2014. Identification of genes and pathways involved in kidney renal clear cell carcinoma. *BMC Bioinformatics* **15(17)**:S2 DOI 10.1186/1471-2105-15-S17-S2.

Yasser E-M, Hsieh T-Y, Shivakumar M, Kim D, Honavar V. 2018. Min-redundancy and max-relevance multi-view feature selection for predicting ovarian cancer survival using multi-omics data. *BMC Medical Genomics* **11(3)**:71.

Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M. 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications* **7**:12474.

Zhang Y, Li A, Peng C, Wang M. 2016. Improve glioblastoma multiforme prognosis prediction by using feature selection and multiple kernel learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13(5)**:825–835 DOI 10.1109/TCBB.2016.2551745.

Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, Lerario AM, Else T, Knijnenburg TA, Ciriello G, Kim S, Assie G, Morozova O, Akbani R, Shih J, Hoadley KA, Choueiri TK, Waldmann J, Mete O, Robertson AG, Wu H-T, Raphael B J, Shao L, Meyerson M, Demeure MJ, Beuschlein F, Gill AJ, Sidhu S B, Almeida M Q, Fragoso MCBV, Cope LM, Kebebew E, Habra MA, Whitsett TG, Bussey KJ, Rainey WE, Asa SL, Bertherat J, Fassnacht M, Wheeler DA, Hammer GD, Giordano TJ, Verhaak RGW, Zheng S, Verhaak RGW, Giordano TJ, Hammer GD, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, Lerario AM, Else T, Knijnenburg TA, Ciriello G, Kim S, Assié G, Morozova O, Akbani R, Shih J, Hoadley KA, Choueiri TK, Waldmann J, Mete O, Robertson AG, Wu H-T, Raphael BJ, Meyerson M, Demeure MJ, Beuschlein F, Gill AJ, Sidhu SB, Almeida M, Barisson Fragoso MC, Cope LM, Kebebew E, Habra MA, Whitsett TG, Bussey KJ, Rainey WE, Asa SL, Bertherat J, Fassnacht M, Wheeler DA, Benz C, Ally A, Balasundaram M, Bowlby R, Brooks D, Butterfield YSN, Carlsen R, Dhalla N, Guin R, Holt RA, Jones SJM, Kasaian K, Lee D, Li HI, Lim L, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Park PJ, Kroiss M, Gao J, Sander C, Schultz N, Jones

CD, Kucherlapati R, Mieczkowski PA, Parker JS, Perou CM, Tan D, Veluvolu U, Wilkerson MD, Hayes DN, Ladanyi M, Quinkler M, Auman JT, Latronico AC, Mendonca BB, Sibony M, Sanborn Z, Bellair M, Buhay C, Covington K, Dahdouli M, Dinh H, Doddapaneni H, Downs B, Drummond J, Gibbs R, Hale W, Han Y, Hawes A, Hu J, Kakkar N, Kalra D, Khan Z, Kovar C, Lee S, Lewis L, Morgan M, Morton D, Muzny D, Santibanez J, Xi L, Dousset B, Groussin L, Libé R, Chin L, Reynolds S, Shmulevich I, Chudamani S, Liu J, Lolla L, Wu Y, Yeh JJ, Balu S, Bodenheimer T, Hoyle AP, Jefferys SR, Meng S, Mose LE, Shi Y, Simons JV, Soloway MG, Wu J, Zhang W, Mills Shaw KR, Demchok JA, Felau I, Sheth M, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J, Davidsen T, Crawford C, Hutter CM, Sofia HJ, Roach J, Bshara W, Gaudioso C, Morrison C, Soon P, Alonso S, Baboud J, Pihl T, et al. 2016. Comprehensive pan-genomic characterization of adrenocortical carcinoma. *Cancer Cell* **29(5)**:723–736 DOI 10.1016/j.ccell.2016.04.002.

Zhou J, Li L, Wang L, Li X, Xing H, Cheng L. 2018. Establishment of a svm classifier to predict recurrence of ovarian cancer. *Molecular Medicine Reports* **18(4)**:3589–3598.