


# Analysis of cancer omics data: a selective review of statistical techniques

Chenjin Ma, Mengyun Wu and Shuangge Ma 

Corresponding author. Shuangge Ma, Department of Biostatistics, Yale University, New Haven CT, USA. Tel.: +1-203-785-3119; Fax: +1-203-785-6912.  
E-mail: [Shuangge.ma@yale.edu](mailto:Shuangge.ma@yale.edu)

## Abstract

Cancer is an omics disease. The development in high-throughput profiling has fundamentally changed cancer research and clinical practice. Compared with clinical, demographic and environmental data, the analysis of omics data—which has higher dimensionality, weaker signals and more complex distributional properties—is much more challenging. Developments in the literature are often ‘scattered’, with individual studies focused on one or a few closely related methods. The goal of this review is to assist cancer researchers with limited statistical expertise in establishing the ‘overall framework’ of cancer omics data analysis. To facilitate understanding, we mainly focus on intuition, concepts and key steps, and refer readers to the original publications for mathematical details. This review broadly covers unsupervised and supervised analysis, as well as individual-gene-based, gene-set-based and gene-network-based analysis. We also briefly discuss ‘special topics’ including interaction analysis, multi-datasets analysis and multi-omics analysis.

**Keywords:** cancer omics data, statistical analysis, selective review

## Introduction

Cancer is an omics disease. Changes and defects at multiple molecular levels (genetic, epigenetic, genomic, proteomic and metabolic) have been associated with cancer risk, prognosis, biomarkers and response to treatment [1, 2]. As well established in the literature, multiple types of omics measurements are not independent but rather interconnected—e.g. SNPs regulate gene expressions, and gene expressions encode proteins [3]. In the whole cancer formation and development process, omics measurements also interact with environmental risk factors [4, 5]. Research on cancer omics is vast, and for the biological, biomedical and bioinformatics aspects, we refer to published reviews [6, 7] and books [8, 9]. The analysis of cancer omics data can serve many purposes, including but not limited to understanding the mechanisms of regulation, identifying drivers for disease development and progression, modeling cancer heterogeneity and providing accurate descriptions of disease paths.

The analysis of cancer omics data is challenging with high dimensionality, low signal-to-noise ratios and profound interconnections among variables [10, 11]. To address these challenges, there have been significant statistical developments in the past two decades [12, 13]. However, such developments are often ‘scattered’, with

one publication usually focused on a single approach. There are a few review articles, however, they often focus on a single type of analysis (e.g. gene set enrichment analysis [14]). In addition, many published studies are not ‘friendly’ and inaccessible to cancer researchers with limited statistical training.

Our goal is to provide a high-end review of commonly adopted statistical techniques for cancer omics data. The target audience is cancer researchers who need to understand published data analysis, design analysis strategies for their own studies and understand data analysis conducted by statisticians. For professional statisticians, this article may serve as a refresher. We focus on intuition and concepts and refer to the original publications for mathematical details. This article may advance from the existing methodological publications (that focus on a single method) by providing a comprehensive review and from the existing review articles [15–17] in the following ways. First, there is a focus on cancer, and cancer data analysis has certain unique characteristics. For example, genetic interactions play an important role in cancer, but not some other diseases. Second, we strive to cover the whole spectrum of analysis, from the ‘basic’ yet still highly important marginal analysis to the recent/advanced deep learning. This level of comprehensiveness is

**Chenjin Ma** is an assistant professor in the College of Statistics and Data Science, Faculty of Science, Beijing University of Technology.

**Mengyun Wu** is an associate professor in the School of Statistics and Management, Shanghai University of Finance and Economics.

**Shuangge Ma** is a professor in the Department of Biostatistics at Yale School of Public Health.

Received: September 20, 2021. Revised: December 19, 2021. Accepted: December 20, 2021

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

**Table 1.** Numbers of omics measurements in four TCGA datasets

	BRCA n = 1098	COADREAD n = 633	KIRC n = 537	LUSC n = 504
Gene expression	17 268	17 518	17 243	17 268
Mutation	13 414	15 998	14 054	15 273
Copy number variation	20 871	20 871	21 526	20 871
Methylation	12 328	12 328	1678	12 328
MiRNA	398	299	353	366

not shared by most of the existing reviews. Third, there is an emphasis on statistical methods, whereas some existing review articles may emphasize other techniques such as machine learning and deep learning [18].

### Data sources

Based on our observation, a large amount of data still resides with individual researchers/groups, although there is an increasing effort in making data publicly available. The following is a list of publicly available multi-omics data sources.

- The Cancer Genome Atlas (TCGA, <https://gdc-portal.nci.nih.gov/>) has characterized over 20 000 primary tumors and matched normal samples, spanning over 37 cohorts and covering 33 different cancer types [19]. TCGA study has collected comprehensive data on whole exome (genome) sequencing, DNA copy number variation, DNA methylation, mRNA expression, microRNA, reverse-phase protein and clinical/demographic measurements [20]. Partial information on four representative cancers is provided in Table 1.
- The International Cancer Genome Consortium (ICGC, <https://dcc.icgc.org/>) is a comprehensive repository for cancer-specific multi-omics data. The portal currently contains data from 86 projects spanning over 35 tumor types, including 22 primary cancer sites, and from over 20 000 contributors [21].
- The Therapeutically Applicable Research to Generate Effective Treatments (TARGET, <https://ocg.cancer.gov/programs/target/>) has a goal of identifying therapeutic targets and prognostic markers to facilitate the development and application of novel and more effective treatment strategies. It currently hosts data on acute myeloid leukemia, osteosarcoma, kidney cancer, acute lymphoblastic leukemia and neuroblastoma [22].
- The UK Biobank (<https://www.ukbiobank.ac.uk/>) is a large-scale biomedical database, containing in-depth genetic and health information for about half a million UK participants. As of early 2020, it has processed and released data from >40 000 subjects.

There are also a few single-omics data sources.

- The Catalogue of Somatic Mutations in Cancer (COSMIC; <https://cancer.sanger.ac.uk/cosmic>) is the

world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

- The Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) is a public repository that archives and distributes high-throughput gene expression and other functional genomic data [23]. Beyond cancer, it also hosts data on other diseases.

More discussions on data resources (and some analysis tools) are available in the literature [7].

### Data preparation

In data preparation, challenges arise in quality control, accommodation of missing data, correction for batch effects, gene annotation and others [24, 25]. Data preparation is by no means trivial and is usually done by technicians and bioinformaticians. In some studies, to reduce computational cost and improve stability/interpretability, screening is conducted as the first step of the analysis. This can be done biologically, e.g. by retaining only genes with pathway information [26]. This can also be done statistically, which involves fitting marginal models and retaining genes with high marginal significance (*P*-values; [27]). Below we assume that data have been properly prepared and focus on high-end statistical analysis.

### Generic data exploration, visualization and analysis software

There are a few generically applicable software packages for data exploration, visualization and analysis. Some have been especially coupled with the TCGA, ICGC and TARGET. A few examples are as follows.

- The cBioPortal (<https://www.cbioportal.org/>) integrates genomic and clinical data and provides a suite of visualization and analysis options.
- The Genomic Data Commons (GDC, <https://portal.gdc.cancer.gov/>) is a cancer knowledge network that supports the hosting, standardization, and analysis of genomic, clinical and biospecimen data from cancer research programs.
- The FireBrowse (<http://firebrowse.org/>) is a companion portal that culls and analyzes data generated by online omics resources.
- The Driverdb3 (<http://driverdb.tms.cmu.edu.tw/>) integrates genomic and clinical data and publishes

bioinformatics algorithms to help researchers visualize the relationships between cancers and driver genes.

- The Cancer Proteome Atlas (TCPA, <https://tcpaportal.org/tcpa/>) is a comprehensive resource for accessing, visualizing and analyzing functional proteomics of patient tumor samples.
- The Regulome Explorer ([http://explorer.cancerregulome.org/all\\_pairs/](http://explorer.cancerregulome.org/all_pairs/)) is a web tool that integrates the associations between clinical and molecular features of the TCGA data.
- The Gene Expression Profiling Interactive Analysis (GEPIA, <http://gepia.cancer-pku.cn/>) is a web server for cancer and normal gene expression profiling and interactive analysis.
- The Circos (<http://circos.ca/>) is a software package for visualization, especially for genomic data.
- The Gitoools (<http://www.gitools.org/>) is a framework for the analysis and visualization of multidimensional genomic data using interactive heatmaps.

Other web-based tools include the Integrative Genomics Viewer, Cytoscape, Savant Genome Browser, StratomeX, IntOGen, UCSC Cancer Genomics Browser and Cancer Genome Workbench [7, 28, 29]. The aforementioned software packages/platforms can handle relatively simple data exploration and visualization but not complex analysis. For many of the analyses described below, more tailored software will be needed.

## Categorization of analysis

To facilitate presentation, we categorize the existing analyses as follows. Such a categorization is not absolute, and one study often contains multiple analyses. To simplify terminology, we generically use ‘gene’ when referring to an omics variable, and note that it can also represent, e.g. a SNP or a methylation locus.

- Depending on whether a cancer outcome/phenotype is present, analysis can be classified as unsupervised (which does not involve an outcome) and supervised (which involves an outcome). Recently, there has been increasing interest in semi-supervised analysis, which, however, remains limited in cancer literature.
- Depending on the ‘unit’, analysis can be classified as:
  - Individual-gene-based, under which genes behave exchangeably in analysis.
  - Gene-set-based, where functionally or statistically coordinated genes form sets, different sets may or may not have overlap, and analysis accounts for the gene set structure.
  - Gene-network-based, under which genes are represented using nodes in a network, and nodes are connected with edges if the corresponding genes are functionally or statistically connected.

- Depending on how many omics units are analyzed at a time, analysis can be classified as:
  - Marginal analysis, under which one unit or a small number of units are analyzed at a time.
  - Joint analysis, under which a large number of units are analyzed in a single step.

In addition, we single out the following analyses, which have unique biomedical significance, involve additional complexity, and have become standing-alone branches:

- Interaction analysis, including both G–E (gene–environment) and G–G (gene–gene) interaction analysis. Different from many other diseases, genetic interactions have played especially important roles in cancer.
- Multi-datasets analysis, under which multiple independent datasets are available, and the goal is to pool information, increase power and generate findings with higher quality. Such analysis has been made possible by the broad availability of cancer omics data.
- Multi-omics analysis, under which multiple types of omics data are available for the same subjects and collectively analyzed. Compared with other diseases, there is more multi-omics cancer data.

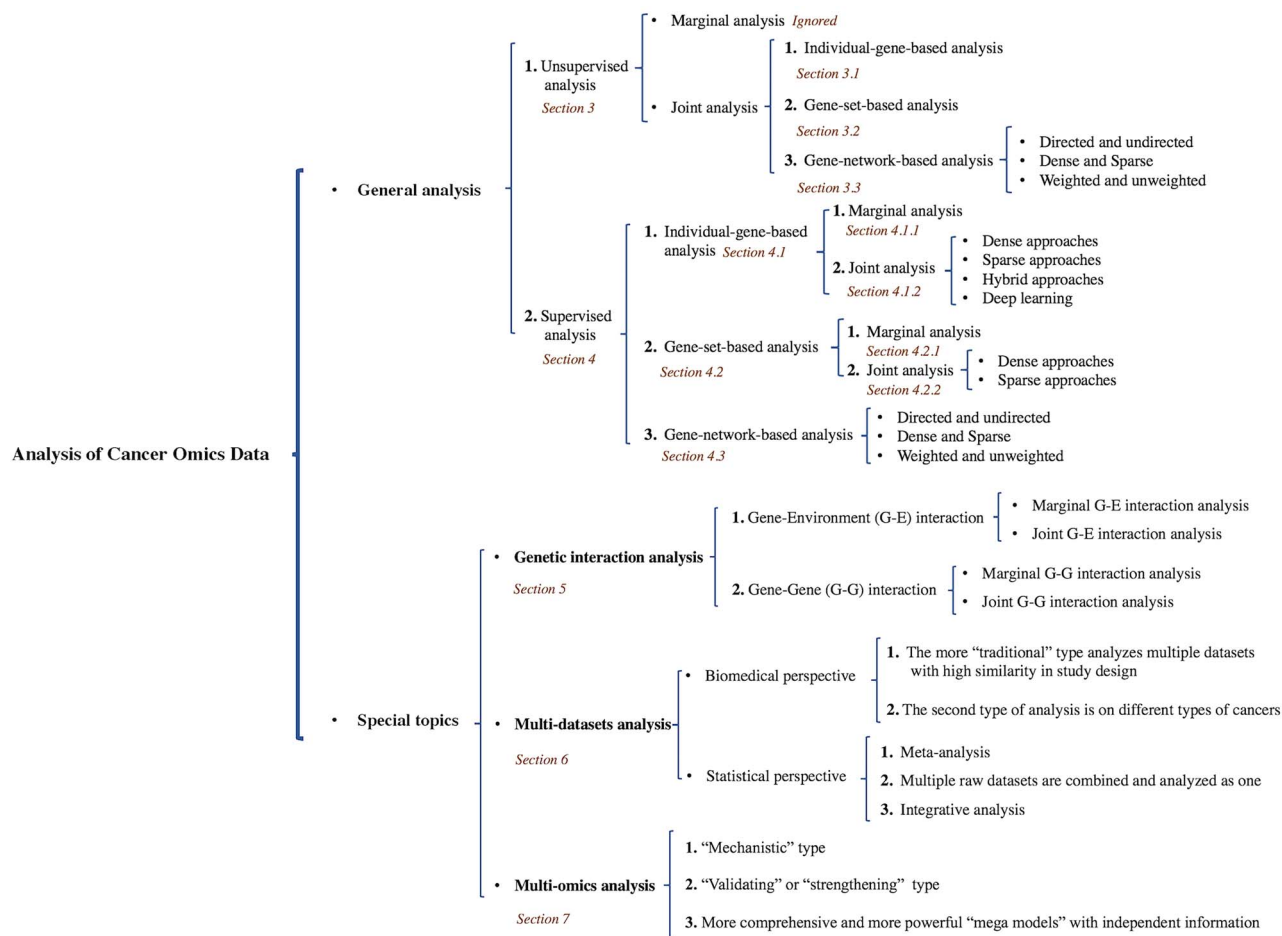
We note that these analyses are not independent of the above categorization system. For example, interaction analysis is a type of supervised analysis, is often individual-gene-based and can be both marginal and joint. Multi-omics analysis is usually joint but in principle can be marginal. To facilitate comprehension, the methods reviewed below are graphically organized in Figure 1.

## Unsupervised analysis

Without an outcome/phenotype, unsupervised marginal analyses are mostly exploratory and descriptive [30], e.g. calculation of the MAF (minor allele frequency) of a SNP or proportion of samples with sufficient expression for a gene. In the following subsections, we mostly focus on joint analysis.

### Individual-gene-based analysis

An important type of unsupervised analysis is clustering [31]. Cancer is a complex and heterogeneous disease. Sample clustering can assist, e.g. in identifying cancer subtypes, which can lead to more personalized treatment strategies. Gene clustering can assist, e.g. in identifying functional groups, which can lead to a better understanding of cancer biology. And bi-clustering [32] is a ‘marriage’ of these two. Most clustering methods are partitional (as represented by the K-means clustering), hierarchical (as represented by the hierarchical clustering) or density-based (which is less common with cancer omics data). As an example, we select 100 gene expressions from the TCGA SKCM data. The K-means and hierarchical clustering results are shown in Figure 2. Partitional methods



**Figure 1.** Organization of the reviewed methods.

lead to ‘cleaner’ results and can only answer Yes/No questions (whether two genes/samples belong to the same cluster), whereas hierarchical methods can lead to more informative structures but often highly imbalanced clusters. Clustering is a ‘classic’ statistical field [33] and can be realized using many existing packages in R [34], such as R functions *kmeans* and *hclust*. Utilization information and demonstrating examples are well available, e.g. in the R manuals (<https://cran.r-project.org>).

Clusters generated by most of the existing methods are non-overlapping. In the development and progression of cancer, a SNP, gene and protein can have multiple biological functions. To better accommodate multiple functionalities, overlapping or fuzzy clustering methods have been developed [35, 36]. Under such methods, each gene/sample is assigned a probability of belonging to each cluster. If such probabilities are dichotomized, fuzzy clustering reduces to partitional (here, we note that a partitional clustering cannot be easily ‘converted’ to a fuzzy clustering). On the negative side, fuzzy clustering results are not as easily interpretable. There are multiple fuzzy clustering packages in R, such as *fclust*, *cmeans* and function *fanny* in the *cluster* package. Among them, package *fclust* has been widely used. It implements the fuzzy k-means (FkM) approach and its variants as well

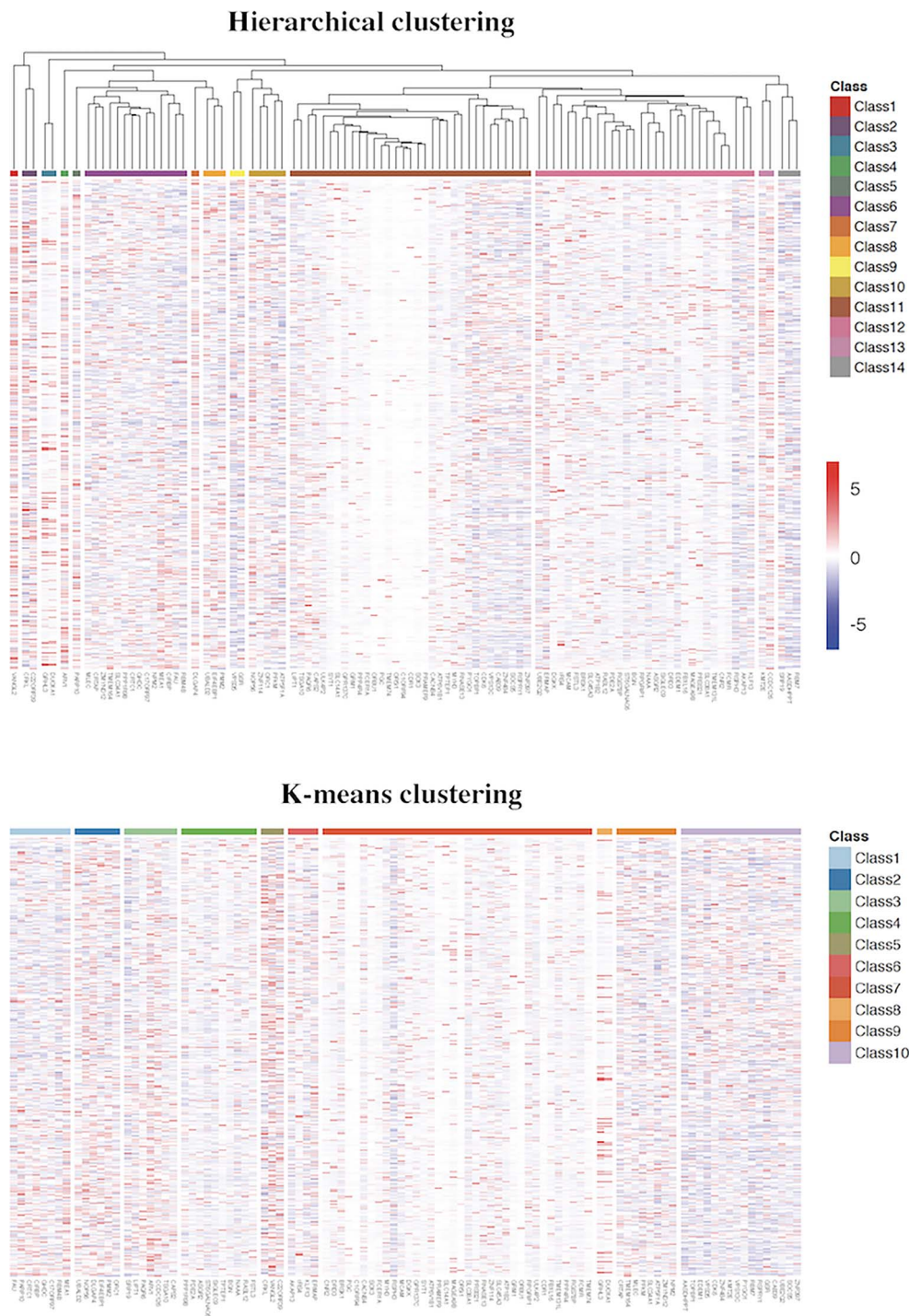
as the new fuzzy relational clustering algorithms and an improved version of the Gustafson–Kessel algorithm for avoiding singularity. Fuzzy cluster similarity measures, validity indices and visualization tools are also offered. This package also allows automatically selecting the number of clusters via the fuzzy cluster validity indices.

As a representative example, Curtis *et al.* analyzed copy number and gene expression data collected on a discovery set and a validation set of 997 and 995 primary breast tumors, respectively. The clustering analysis of paired DNA–RNA profiles revealed novel subgroups of breast cancer with distinct clinical outcomes beyond the classic expression subtypes [37].

## Gene-set-based analysis

Cancer development and progression are usually attributable to sets of coordinated genes. In this type of analysis, gene sets are first constructed, each set contains multiple genes and different sets can be non-overlapping or overlapping. In the literature, the definitions/constructions of gene sets are quite versatile. (i) In some studies, clustering (e.g. K-means or fuzzy clustering) is applied to construct gene sets. With this approach, all genes can be put into clusters, those in the same clusters are ‘statistically





**Figure 2.** Clustering analysis of 100 genes selected from the TCGA SKCM data: hierarchical (upper) and K-means (lower).

connected', and those in different clusters are statistically weakly/not connected. On the negative side, based on a single dataset, the gene set structure can be unstable. In addition, the sets may not have lucid biological interpretations. (ii) Gene sets can be constructed based on biological-information [38], e.g. pathway information extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG, <https://www.genome.jp/kegg/pathway.html>). The gene sets so constructed usually have sound

biological interpretations and do not change from data to data. On the negative side, not all genes have known biological functions, and different databases have different ways of defining biological functionalities. (iii) There are also 'customized' definitions. For example, SNPs of the same gene can be defined as a set, and genes mentioned in a specific publication can also be viewed as a set [39]. Such definitions are often narrower and less commonly used.

With predefined gene sets, we can examine distributional properties, e.g. properties of MAFs, correlations among gene expressions, etc. This type of analysis is simple and does not demand special techniques/software. Clustering analysis can be conducted based on gene sets. For example, sample clustering can be straightforwardly based on one or a few gene sets. Gene sets can be further clustered. For example, the distance between two gene sets can be computed as the average of individual-gene distances, and then methods such as K-means and hierarchical clustering can be applied. In the same spirit, bi-clustering can also be conducted based on gene sets.

### Gene-network-based analysis

Unlike some other diseases, cancer is genetically complex. For cancer, genes often function collaboratively, and one gene or a small number of genes may carry very limited information. Compared with the aforementioned analyses, gene-network-based analysis may excel by providing a system perspective. As sketched in Figure 3, in a gene network, a node corresponds to a gene, and two nodes are connected with an edge if the corresponding genes are interconnected. Based on key properties, gene networks can be classified as follows:

- (i) Directed and undirected: in a directed network, gene A may affect (be connected to) gene B, but not the other way around. In contrast, in an undirected network, the interconnection between genes A and B is the same as that between genes B and A. Comparatively, a directed network may better describe genetic regulations. Graphically, directions are often represented by arrows.
- (ii) Sparse and dense: in a dense network, all genes are interconnected, and the strengths of interconnection vary. In contrast, in a sparse network, only some are interconnected. A sparse network is simpler and has more lucid interpretations, whereas a dense network can also accommodate subtle gene interconnections.
- (iii) Weighted and unweighted: in an unweighted network, edges are binary (1/0), whereas, in a weighted network, each edge is associated with a continuous weight, describing the strength of interconnection. Graphically, weights can be represented by the thickness of edges. An unweighted network is simpler to construct and can be more robust. On the negative side, it contains less information.

As an example, the network in Figure 3 is undirected, sparse and weighted.

There are multiple ways for constructing gene networks:

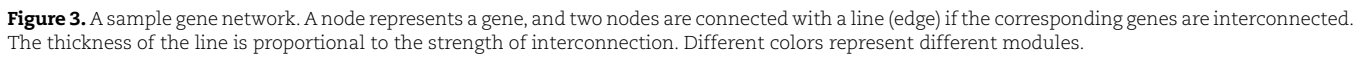
- (i) Biological construction, under which interconnections among genes are derived based on functional information [40]. For example, there are multiple protein-protein interaction databases [41]. The pros

and cons of this approach are similar to those of the biological-information-based gene set construction.

- (ii) Statistical construction, under which interconnections are derived statistically. This family further contains two types of approaches.

- (a) Unconditional construction, under which, when evaluating the interconnection between two genes, the other genes are 'ignored'. A representative approach is the WGCNA (weighted gene co-expression network analysis) [42], which was first developed for gene expression data and has been applied to other types of omics measurements. Its main steps are described in the Supplementary Materials. This approach can be realized using R package WGCNA, which provides functions on network construction, analysis and visualization [43]. We also refer to <https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/> for more information. A few important concepts arise in this construction. The first is adjacency, which is perhaps the most fundamental network measure and quantifies how strongly two genes are interconnected. The second is connectivity, which describes how strongly a gene is interconnected with the rest of the genes. The third is module, which is a set of tightly interconnected genes. For a specific gene, if the calculation of connectivity is limited to the other genes within the same module, the resulted measure is called intramodular connectivity, which is less affected by a large number of spurious 'connections'. Within a module, the gene with the highest intramodular connectivity is called the hub, which is the most important in a network sense.
- (b) Conditional construction, under which the adjacency between two genes quantifies whether they are interconnected *conditional* on the other genes. The most popular conditional construction is perhaps the Gaussian Graphical Model (GGM) approach [44], which can be realized using R packages GGMselect and ggm. Denote  $S$  as the sample Pearson correlation matrix for the  $d$  genes. The GGM approach estimates the precision matrix  $\Theta$ —the inverse of the correlation matrix—via minimizing  $L(\Theta) = -\log |\Theta| + \text{tr}(S\Theta)$ , where  $\text{tr}$  denotes trace. When the genes have a multivariate normal distribution, two genes are conditionally independent if and only if the corresponding element in  $\Theta$  is zero [45]. As such, determining the conditional adjacency structure amounts to a *sparse* estimation of  $\Theta$  (which has some components being exactly zero). For this purpose, the most popular approach is Graphical Lasso (the least absolute shrinkage and selection operator; [46]). Denote





to examine data properties (e.g. whether multivariate normality holds) to select proper approaches. Most of the existing gene-network-based unsupervised analyses are exploratory, identifying and examining network and module structures, hub genes, etc.

This type of analysis can, e.g. assist in identifying driver genes for specific cancer outcomes, so as to facilitate understanding disease biology and more importantly developing tailored treatment regimens.

Denote  $Y$  as the outcome variable, which can be continuous, categorical, count or censored survival. Denote  $X = (X_1, \dots, X_d)'$  as the  $d$  genes, and  $Z$  as a low-dimensional vector consisting of clinical/demographic/environmental risk factors.

The goal is to identify individual components of  $X$ , conditional on  $Z$ , that are significantly associated with  $Y$ . This analysis consists of the following steps:

- Comparatively, the unconditional construction is simpler and ‘cleaner’. It is often preferred by biomedical researchers. On the other hand, the conditional construction takes a more system perspective, which may be especially important for cancer. There are many other approaches for network construction, including Bayesian approaches [49], Boolean network construction [50], etc. We also refer to the aforementioned publications for software development and applications. There are a few studies that compare the performance of different network constructions [51, 52]. It is found that performance strongly depends on data characteristics, and it is not expected to have one approach dominating the others. With a practical cancer dataset, one needs

models such as the rank-based [53].  $\alpha_j$  is a vector of unknown coefficients,  $\alpha_j^T$  is its transpose and  $\beta_j$  is an unknown coefficient. For some models, an intercept is also needed. As each model is low-dimensional, estimation can be easily realized, e.g. using R functions *lm*, *glm* and *coxph*.

- (ii) Denote  $\{p_1, \dots, p_d\}$  as the collection of  $d$  P-values for the estimates of  $\beta_j$ 's.
- (iii) Conduct multiple comparison adjustment, and identify the significant ones. For this purpose, there are two popular approaches, namely Bonferroni [54] and FDR (false discovery rate) [55]. Comparatively, Bonferroni-type approaches are easier but more conservative. When  $d$  is large, FDR approaches can be preferred. More discussions are provided in the Supplementary Materials.

In some studies, there are two or more experimental conditions, and the goal is to identify differential genes [56]. This can also be cast in the regression framework: *gene*  $\sim$  *experimental conditions*. In practice, this is usually realized using hypothesis testing. Depending on distributional properties (e.g. continuous or categorical), commonly used techniques include t, Wilcoxon, Kolmogorov–Smirnov (KS), Chi-squared and Fisher's exact tests. Once P-values are obtained, multiple comparison adjustment can be conducted.

### Joint analysis

The goal is to construct a single model  $Y \sim f(\alpha'Z + \beta'X)$ , where notations have similar implications as in Section 'Marginal analysis'. In this analysis, the most significant challenge comes from the 'sample size  $\ll$  number of variables' problem. There are two main families of statistical approaches.

**Dense approaches** Here, all or a large number of genes are involved in the final model. A representative approach is based on ordinary PCA (principal component analysis; [57]) and consists of the following steps: (i) Conduct PCA with  $X$ , which can be realized using R functions *prcomp* and *svd*. Select the top PCs (principal components); and (ii) Construct an outcome model using  $Z$  and the selected PCs. Since this is a low-dimensional model, standard approaches and software can be used. Beyond PCA, other approaches include ICA (independent component analysis), PLS (partial least squares), SIR (sliced inverse regression) and others [58–60]. Such approaches have been referred to as *dimension-reduction*. Their essence is to identify a small number of 'latent genes'—linear combinations of the original genes—as new variables for modeling. The coefficients of genes in the linear combinations are referred to as loadings. Despite certain common ground, these dimension-reduction methods also have notable differences. For example, PCA removes correlations but not higher-order dependence, and different PCs have different levels of importance. In contrast, ICA removes correlations

and higher-order dependence, and all components are equally important. PLS has been observed to perform better than SIR when collinearity exists. Extensive discussions are available in the literature [61, 62]. The relative performance (e.g. prediction) of different approaches has been observed to be data-dependent, which is sensible as they are built on different data assumptions, and practical data do not fit in the same set of assumptions.

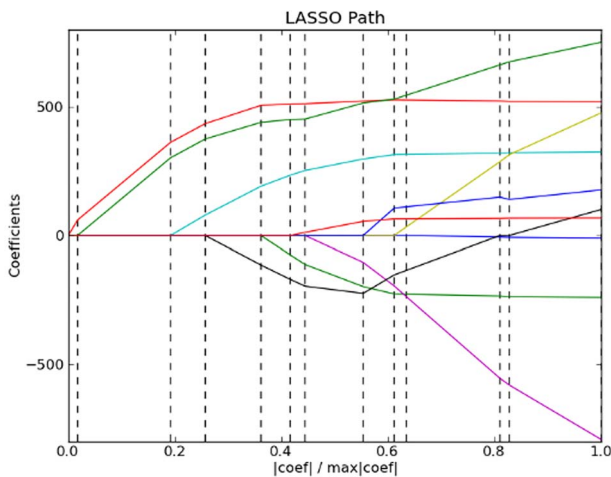
An even simpler but currently infrequently adopted dense approach is ridge penalized regression [63]. With model  $Y \sim f(\alpha'Z + \beta'X)$ , denote  $l(\alpha, \beta)$  as the log-likelihood function. The ridge estimate is defined as the maximizer of  $l(\alpha, \beta) - \lambda \sum_j \beta_j^2$ , where  $\lambda > 0$  is a tuning parameter and  $\beta_j$  is the  $j$ th component of  $\beta$ . Ridge regression can be realized using R functions such as *lm.ridge* and *glmnet*. Among them, *glmnet* can accommodate generalized linear and some other models, and is more broadly applicable. It can calculate estimates for a grid of tuning parameter values and select the optimal one.

The most significant advantage of dense approaches is simplicity: they can be relatively straightforwardly realized using well-known methods and software. In addition, they often have good prediction performance, as they can retain a large number of genes with small effects. On the other hand, it is generally believed that not all genes play a role in cancer. The number of 'cancer genes' identified to date is still small. As such, the fundamental assumption of dense models may be challenged. In addition, dense methods usually do not have lucid biological interpretations. Overall, dense models can be preferred when prediction but not interpretation is of main concern.

**Sparse approaches** With such approaches, only a subset of selected genes enter the outcome model. These approaches have been motivated by the 'belief' that only a small number of genes are cancer-associated—this is strongly supported by biological studies [64]. There is vast literature on variable/model selection. 'Classic' approaches such as step-up are still applicable. However, they are rarely used now because of unsatisfactory numerical results, in particular lack of stability.

A large family of methods, which were first developed in the 1990s, have attracted extensive attention in the statistics community, and are now gradually being accepted by the cancer community, are sparse penalization [10]. Use notations similar to in the previous subsection. A penalization approach maximizes the objective function:  $l(\alpha, \beta) - \lambda \sum_j \text{pen}(\beta_j)$ , where  $\lambda > 0$  is a tuning parameter, and  $\text{pen}(\cdot)$  is a penalty function. The most popular penalty is Lasso [65], where  $\text{pen}(\beta_j) = |\beta_j|$ . Alternatives include SCAD (smoothly clipped absolute deviation), bridge, MCP (minimax concave penalty) and others [66]. With proper tunings, some of the estimated  $\beta_j$ 's can be exactly zero. Genes with nonzero estimates are concluded as associated with the cancer outcome. Theoretically, Lasso can be estimation/variable selection biased, whereas the others can have consistency





**Figure 4.** A sample Lasso parameter path. Different colored lines correspond to different variables. The vertical lines correspond to different tuning parameter values.

properties. Computationally, Lasso can be preferred because of its computational and methodological simplicity. In applications to practical cancer data, the performance of different penalization approaches has been data-dependent. With penalization, an important concept is parameter path, which, as shown in Figure 4, is a plot of estimates as a function of tuning parameter (in Figure 4, the tuning parameter is equivalently defined using the magnitude of regression coefficients). From the parameter path, it is clear that, under certain tuning parameter values, some variables have exactly zero estimated coefficients, which automatically leads to variable selection. In addition, a parameter path can also show the ‘order’ of genes entering the cancer outcome model, and as such, it can be used to rank genes. Penalization achieves model estimation and variable selection simultaneously. This estimation-based selection is fundamentally different from the P-value-based. With penalization methods and high-dimensional variables, there are a few theoretically valid approaches for computing P-values [67, 68]. However, they have not been widely applied in practice. Built on the ‘base’ penalization methods mentioned above, ‘upgraded’ methods have been developed to tailor special data/models. For example, elastic net [69] can accommodate high correlations, and fused penalization [70] can accommodate certain adjacency structures. Such methods have been designed for special data settings. For example, for densely measured SNP data, fused penalization can outperform the other penalization methods. There are methodological articles, comprehensive reviews and book chapters on penalization [71, 72], which may include more technical details. Available packages include R *glmnet*, *penalized*, *ncvreg*, *ncpen* and others, some of which have comprehensive functions. For example, *ncpen* accommodates a variety of generalized linear (linear, logistic, logit, Poisson, etc.) and Cox models and various non-convex penalties (Lasso, SCAD, MCP, sparse bridge and others).

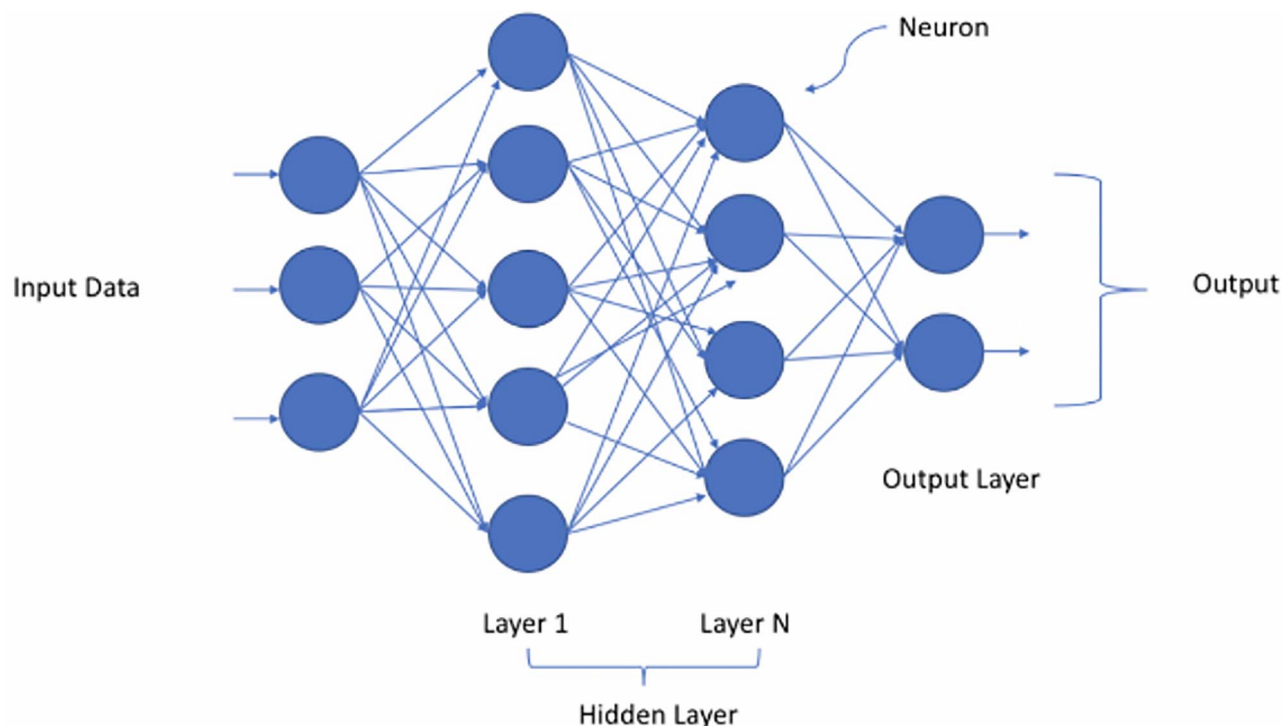
Another (smaller) family of sparse methods is built on thresholding [73]. A representative example is the TGDR (threshold gradient descent regularization; [74]). This approach is iterative and takes a strategy different from penalization. In each iteration, it first identifies the most important variables (with the largest derivatives), and then updates estimation only in those important directions. We refer to the Supplementary Materials for details. Pros of this approach (and the other thresholding ones) include simple computation, intuitive definition and often good numerical results. On the other hand, without a well-defined objective function, statistical properties have not been well established, making it less appealing to some statisticians.

Beyond penalization and thresholding, there are alternative families of approaches, including boosting [75], Bayesian [76] and others. Boosting methods have been popular with the machine learning community, and Bayesian methods can be especially useful when the size of the present data is small but there has been previous research (knowledge) on related problems. Published numerical studies, especially simulation, suggest that the performance of different approaches is data-dependent, and researchers are recommended to experiment with multiple approaches and select the optimal one based on, e.g. prediction.

Sparse approaches can identify a small number of important genes. This sparsity largely fits cancer biology and facilitates interpretation and downstream functional studies. To date, there may be more statistical research on this family than others. However, despite highly satisfactory theoretical results, in practice, it is often observed that the results are not sufficiently stable/replicable, and success in prediction is limited—this is sensible as sparse approaches can only pick up the strongest signals, whereas for cancer there may also be contributions from quite a few weak signals.

**Hybrid approaches** There are a few hybrid approaches, which ‘combine’ dense and sparse approaches. For example, in the PCA-based analysis, if the PCA step is replaced with the SPCA (sparse PCA [77]), then the resulted PCs are sparse in that some of the loadings are exactly zero. Here sparsity can be achieved, e.g. using penalization. There are also sparsified counterparts of PLS, ICA and other dense approaches. This family is relatively new and has not been extensively applied to cancer studies (and hence there is not enough evidence to assess performance). It can be ‘combined’ with either the sparse or dense family.

**Deep learning** This family of approaches is very new but has been growing fast in both methodological development and application [78, 79]. A deep neural network—possibly the most representative deep learning technique—is schematically shown in Figure 5. Loosely speaking, it contains multiple layers of linear combinations followed by parametric nonlinear transformations (known as activations). It has been adapted to continuous, categorical and censored survival outcomes. The



**Figure 5.** A deep neural network.

construction of a deep neural network may take a whole article or multiple articles to describe, and programming is highly nontrivial. We will not take on this daunting task and refer interested readers to recent publications [80, 81]. On one hand, it has been argued that ‘application of deep learning to genomic datasets is an exciting area ... and is primed to revolutionize genome analysis ...’ [78] and that ‘the magnitude and complexity of genomic data will ensure that deep learning will become an everyday tool for its analysis’ [79]. A recent application addressed the scarcity of data samples by exploiting transfer learning and fine-tuning and conducted survival analysis with TCGA data on twenty cancer types [82]. In another study [83], a deep neural network-based factorization model was developed to decipher the latent patterns in omics profiles, and data analysis on medulloblastoma cancer, leukemia, breast cancer and small-blue-round-cell cancer led to interesting findings. On the other hand, it has also been recognized that training a deep neural network may demand a large number (e.g. millions) of samples, even with a moderate number of input variables. With cancer omics data, there are concerns on ensuring stability/replicability and avoiding overfitting with deep learning. Deep learning mostly excels in prediction, and the estimation results are known to be black boxes. As such, it may be ideal for tasks like predicting response to treatment but not the identification of cancer driver genes.

### Gene-set-based analysis

In a few early studies [84], regression analysis and construction of gene sets are conducted simultaneously.

However, such methods have not been extensively applied. Here, we focus on the scenario where gene sets have been constructed *a priori*.

### Marginal analysis

Different from in Section ‘Marginal analysis’, the units for analysis are gene sets. Otherwise, many elements of this analysis are similar to those in Section ‘Marginal analysis’. In particular, a key goal is to identify which gene sets are significantly associated with the response variable, and there are two families of tests: self-contained and competitive [85]. In a self-contained test, the Null hypothesis is:  $H_0$ ={no genes in this set are significantly associated with the response}, and the conclusion of whether  $H_0$  holds is not impacted by the other sets. In comparison, in a competitive test, the Null hypothesis is:  $H_0$ ={the genes in this set are not more significantly associated with the response than the other sets}. Note that the two families to tests serve different purposes. For example, in a study with a candidate gene approach, a self-contained test may be more sensible, whereas, in a whole-genome study, a competitive test can be preferred. One of the most popular self-contained tests is the Global Test [86]. It can be realized using BioConductor package *globaltest*, which conducts the test, provides diagnostic plots, and contains functions to facilitate utilization along with the gene ontology (GO, <http://geneontology.org/>) and KEGG. This test can suggest whether a gene set is independent of a cancer outcome, and hence be used for initial screening. For gene-set-based analysis, the most extensively adopted is perhaps GSEA (gene set enrichment analysis) [87], which is a

competitive procedure. We refer to the Supplementary Materials for details.

In principle, it is possible to first apply dimension reduction techniques, e.g. PCA, to reduce the dimension of each set. Then determining whether a gene set is associated with an outcome can be realized using classic techniques such as ANOVA and likelihood ratio test. This approach is computationally simple. However, the drawbacks described in Section ‘Joint analysis (para, The goal is to construct...)’ may hold.

### Joint analysis

Assume that the  $d$  genes have been separated into  $S$  sets, and denote  $S(j)$  as the set membership of gene  $j$ . Here, we focus on the simple scenario with nonoverlapping gene sets and refer to the literature [88, 89] for developments with overlapping gene sets. For  $s = 1, \dots, S$ , use  $X_s$  to denote genes in set  $s$ . The outcome model can be written as  $Y \sim f(\alpha'Z + \sum_s \beta_s' X_s)$ , where notations have similar implications as above. Denote the log-likelihood function as  $l(\alpha, \beta_1, \dots, \beta_S)$ .

One family of dense approaches first applies dimension reduction, e.g. PCA, to each gene set. A small number of linear combinations are extracted for each set, and then joint model building is conducted with the linear combinations from all sets. If the number of sets and/or the selected number of linear combinations per set are large, techniques for individual-gene-based joint analysis, e.g. Lasso, may be needed. The pros and cons of this family are similar to those in Section ‘Joint analysis (para, The goal is to construct...)’.

In the statistical literature, there have been extensive developments on sparse techniques [90]. There are two main families of methods, and here we use penalization as an example for description.

- (i) The first family conducts *one-level* selection and estimation. The most famous is group Lasso [91], which has objective function  $l(\alpha, \beta_1, \dots, \beta_S) - \lambda \sum_s \|\beta_s\|$ , where  $\lambda > 0$  is the tuning parameter, and  $\|\beta_s\|$  is the  $l_2$  norm of  $\beta_s$ . In many ways, this method is similar to Lasso: some of the estimated  $\beta_s$ 's can be exactly zero and hence identified as not associated with cancer. Here ‘one-level’ means that if a gene set is identified, then all genes within this set have nonzero estimated coefficients (i.e. ‘all-in’); and if a gene set is not identified, then all its genes have exactly zero coefficients (i.e. ‘all-out’). This scheme is presented in the upper panel of Figure 6, where a square represents a gene, connected squares correspond to a set, and those in grey are identified as important. One-level penalization methods also include group SCAD, group MCP and others [92, 93]. The comparison of, e.g. group Lasso versus group SCAD is similar to that of Lasso versus SCAD. Software packages include R *grplasso*, *grpreg* and others, which conduct estimation for a grid of

tuning parameter values and can also select the optimal one. Some also include graphical functions. In cancer studies, it has been found that even a pathway with critical implications (e.g. DNA repair) contains ‘noises’. As such, this family of methods may lead to models with unnecessary noises and diminished interpretability.

- (ii) The second family conducts *two-level* selection and estimation. Such methods determine not only whether a gene set is important but also which genes in an important set are important (lower panel of Figure 6). Built on the group Lasso technique, two methods that can achieve two-level selection are:
  - (a) Sparse group Lasso [94], which has objective function  $l(\alpha, \beta_1, \dots, \beta_S) - \lambda \sum_s \|\beta_s\| - \tau \sum_{s,j} |\beta_{s,j}|$ . Here  $\tau$  is another tuning parameter, and  $\beta_{s,j}$  is the  $j$ th component of  $\beta_s$ .
  - (b) Composite penalization [95], which has objective function  $l(\alpha, \beta_1, \dots, \beta_S) - \lambda \sum_s (\sum_j |\beta_{s,j}|)^{1/2}$ .

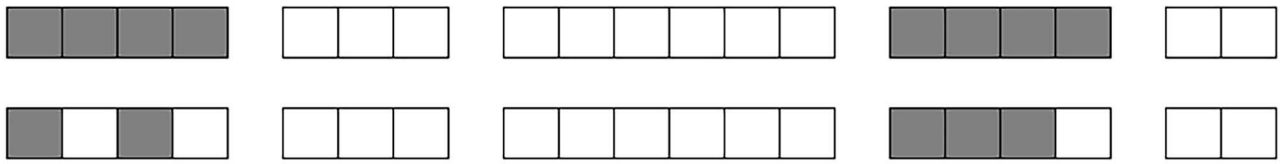
Built on base penalties including MCP, SCAD, bridge and others, quite a few two-level selection methods have been developed [96, 97]. Naturally, e.g. MCP-based methods have similar pros (e.g. consistency) and cons (e.g. computational complexity) as MCP. With two-level approaches, the resulted models are built on individual genes. A sensible question is: ‘If we simply look at the identification and estimation results, there seems no difference from the individual-gene-based analysis. Is there any benefit of accommodating the gene set structure?’ To address this, theoretical studies have been conducted, suggesting that, if gene sets are properly defined, improved identification and estimation results can be obtained.

Beyond penalization, other regularization techniques have also been extended from individual-gene-based analysis to accommodate gene set structures. Examples include the CTGDR (clustering TGDR; [98]), which can conduct one- and two-level selection. It has satisfactory numerical performance but less lucid statistical properties. As an extension of sparse boosting, the NSBoost method [99] first analyzes each gene set, selects important genes with each set, and constructs gene-set-specific models. Then the sets and set-level models are aggregated to generate the final selection and model building. It inherits some advantages and limitations of boosting. There are also approaches built on the Bayesian technique [100], which demand possibly subjective specification of priors.

### Gene-network-based analysis

As described in Section ‘Gene-network-based analysis’, in the construction of gene networks, the following important measures arise: adjacency, (intramodular) connectivity, hub and module. Methods have been developed to incorporate them in variable selection/model estimation.





**Figure 6.** Schemes of gene-set-based variable selection. One square represents one gene, connected squares correspond to a set, and those in grey are identified as important. One-level (upper) and two-level (lower) selection.

**Adjacency** Tightly interconnected genes often behave similarly in cancer models, and the ‘amount of similarity’ is positively related to the value of adjacency [101]. As a representative, we consider the Laplacian penalization method [102]. Denote  $a_{j,k} = a_{k,j} \geq 0$  as the adjacency between genes  $j$  and  $k$ , with a higher value indicating a tighter interconnection. The penalized objective function is  $l(\alpha, \beta) - \lambda \sum_j \text{pen}(\beta_j) - \tau \sum_{j,k} a_{j,k} (\beta_j - \beta_k)^2$ , where  $\tau \geq 0$  is a tuning parameter. The newly added penalty encourages those tightly connected to have similar regression coefficients. In practical data analysis, this method performs the best when there are a moderate number of tightly interconnected genes. On the other hand, it may fail when most interconnections are weak. This strategy has been extended to interaction analysis [28] and integrative analysis [103].

**Connectivity** Cancer outcomes are usually attributable to system-wide, as opposed to individual, molecular changes. Compared with ‘isolated’ genes, those well-connected in a network are more likely to be able to cause system-wide changes. As such, it has been suggested that genes with higher connectivity should have a higher priority in identification [104]. More discussions on incorporating connectivity are provided in the Supplementary Materials.

**Hub** Our literature search does not lead to approaches that explicitly accommodate individual hub genes in analysis. However, approaches such as the eigengene approach [105] are strongly related to the concept of hub. In particular, for each module, if PCA is conducted, then the first PC is referred to as eigengene. Under certain conditions, the eigengene has the highest connectivity. In a few studies [106], the collection of eigengenes is used in model building. Further advancing from this idea, it has been proposed that the first few, as opposed to the very first, PCs are used in model building [107].

**Module** The module structure provides an alternative way of generating gene sets. As such, once the modules are identified, gene-set-based analysis techniques as reviewed above can be applied.

**Remarks** A network contains rich information. Many measures (e.g. the shortest distance between two nodes) have not been accounted for in analysis. In addition, the aforementioned methods are all for undirected networks. A few studies have been conducted to accommodate directed networks [108]. For example,

it has been proposed that if gene  $A \rightarrow$  gene  $B$ , where ‘ $\rightarrow$ ’ indicates regulation, and if gene  $A$  has a zero coefficient, then gene  $B$  should automatically have a zero coefficient. Conceptually, this can be realized by imposing constraints in estimation. However, a large number of constraints will make computation intractable. Overall, accommodating more complex network structures and measures is still an ongoing effort.

### An application example

The methods reviewed above have been well applied. In most if not all of the methodological development studies, strong data analysis has been conducted. As a representative example, Tang and others assessed multiple supervised joint analysis methods and applied them to the TCGA data on three cancers [109]. A total of six methods, including a Bayesian hierarchical generalized linear model and five penalization methods (sparse group lasso, group lasso, group MCP, group SCAD and group cMCP), were applied. The analysis revealed that, for sarcoma, the identified genes are mainly associated with ATP. For ovarian cancer, the identified genes spread over a wide range of pathways. For breast cancer, five genes are identified, representing two pathways.

### Genetic interaction analysis

Cancer development and progression are associated with the combined effects of multiple clinical, environmental and genetic risk factors, as well as their interactions. Understanding genetic interactions can further advance our understanding of cancer biology and more importantly facilitate the development of interventions and treatments targeting *modifiable* risk factors. There are two types of interactions: gene–environment (G–E) and gene–gene (G–G) [110, 111]. With similar notations as above, the following interaction models have been considered:

- (i) Marginal G–E interaction analysis: for  $j = 1, \dots, d$ , consider the model  $\mathbf{Y} \sim f(\alpha'_j \mathbf{Z} + \beta_j \mathbf{X}_j + \gamma'_j \mathbf{Z} * \mathbf{X}_j)$ . Here,  $\alpha_j$  and  $\gamma_j$  are vectors, and  $\beta_j$  is a scalar. With the low dimensionality of  $\mathbf{Z}$ , each model is low-dimensional.
- (ii) Joint G–E interaction analysis: consider the model  $\mathbf{Y} \sim f(\alpha' \mathbf{Z} + \beta' \mathbf{X} + \gamma' \mathbf{Z} * \mathbf{X})$ , which is high-dimensional.
- (iii) Marginal G–G interaction analysis: for  $j, k = 1, \dots, d$ , consider the model  $Y \sim f(\alpha'_{j,k} \mathbf{Z} + \beta_j \mathbf{X}_j + \beta_k \mathbf{X}_k + \beta_{j,k} \mathbf{X}_j \mathbf{X}_k)$ . Usually, ‘self-interaction’ is excluded (i.e.  $j \neq k$ ). Each model is low-dimensional.

- (iv) Joint G–G interaction analysis:  $Y \sim f(\alpha'Z + \beta'X + \gamma'X * X)$ . Here  $\gamma'X * X = \gamma_{j,k}X_jX_k$  with  $j \neq k$  to exclude self-interaction. This is a high-dimensional model.

Many approaches have been developed, with different techniques but similar central strategies:

- (i) In marginal analysis, since each model is low-dimensional, estimation can be carried out in a standard manner, and  $P$ -values of interactions (and main effects) can be obtained. Then the Bonferroni and FDR techniques can be applied to identify significant effects. Some studies only focus on interactions in the identification procedure, whereas others consider interactions and main G effects simultaneously. Very recently, some studies have advocated that the ‘main effects, interactions’ identification hierarchy described below should be respected in marginal analysis [112].
- (ii) In joint analysis, the biggest challenge comes from high dimensionality, especially for G–G interaction analysis. A ‘simple’ solution is to first apply e.g. PCA, reduce the dimensionality of  $X$ , and then proceed with the low-dimensional reduced variables. This has been applied to both G–E and G–G analysis [113, 114]. The pros and cons of this approach are similar to those discussed above. Additional challenges arise when selection is needed along with estimation. In particular, in the recent literature [115, 116], there is a strong advocacy on the variable selection hierarchy:
  - (a) Weak hierarchy: if an interaction is selected, then at least one of its corresponding main effects must be selected.
  - (b) Strong hierarchy: if an interaction is selected, then both of its corresponding main effects must be selected.

Statistically speaking, the hierarchy makes modeling more appropriate [117]. However, biologically, it has been argued that it is possible to have interactions without corresponding main effects [111]—this debate is beyond our scope. The hierarchy brings additional constraints. With penalization, researchers have developed constraint [117], composite penalization [118], sparse group penalization [119] and other approaches to respect the hierarchy. In simulation, these methods have improved selection and estimation performance, and in data analysis, they have led to different findings.

To make joint analysis computationally more feasible, researchers have developed progressive approaches [120, 121]. They are usually based on an iteration strategy, where, in each iteration, only the interactions with the corresponding main effects being previously selected are considered. Thus, it guarantees the hierarchy. The biggest advantage of this approach is its simplicity. We refer to the Supplementary Materials for specifics.

Most of the existing genetic interaction analysis methods are individual-gene-based. There are also approaches that take into the gene set and network structures in estimation. For example, structured interaction analysis [28] accommodates the gene co-expression network structure via imposing adjacency-based penalization. The analysis of the TCGA data shows that it can better identify genes that are connected in the network. In addition, there are lucid biological interpretations.

There is a myriad of genetic interaction analyses in the literature. In a representative study [122], Yang and others analyzed a Taiwan dataset with 103 oral cancer cases and 98 controls. For oral cancer risk, DNA repair genes X-ray repair cross-complementing group (XRCCs) 1–4, environmental risk factors (including smoking, alcohol drinking and betel quid chewing), and their interactions—both G–G and G–E—were studied. Techniques used included logistic regression, multifactor dimensionality reduction and hierarchical interaction graph. It was suggested that the positive associations between the genotypes of XRCC1 rs1799782, XRCC2 rs2040639 DNA repair genes and oral cancer were enhanced by exposure to drinking and betel quid chewing.

### Multi-datasets analysis

For common cancers and outcomes, there are often multiple studies with somewhat similar designs, enabling the collective analysis of multiple independent datasets. Public databases described above, such as GEO, dbGaP and TCGA, have been used for such a purpose. Multi-datasets analysis inevitably faces additional challenges, e.g. selecting compatible datasets and matching variables across datasets. Such considerations have been extensively examined in cancer epidemiologic and bioinformatics studies [123] and will not be reiterated here.

Biomedically speaking, there are mainly two types of analysis. The first type is more ‘traditional’ and analyzes multiple datasets with high similarity in study design. For example, a study [124] analyzed four datasets on lung cancer overall survival with gene expression measurements. The goal of this type of study is ‘simple’: increase sample size to improve power/reliability. The second type of analysis is on, e.g. different types of cancers [125]. This is represented by the Pan-Cancer Atlas study (<https://gdc.cancer.gov/about-data/publications/pancanatlas>). For example, a study [126] analyzed data on nine cancer types, and genes with more fundamental roles in cancer development were identified.

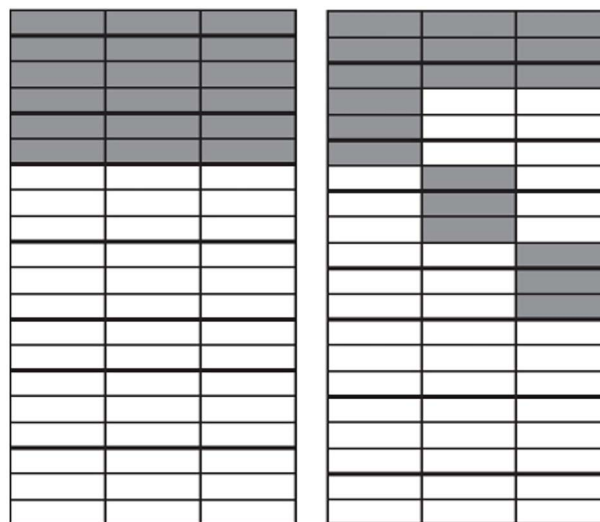
Statistically speaking, the existing analyses can be mostly classified into three categories:

- (i) Meta-analysis [127], under which each dataset is analyzed separately, and summary statistics are pooled to generate final estimation. Most meta-analysis methods designed for low-dimensional data can be directly applied here. For example, individual estimates can be averaged to generate final estimates. A key difference is that in some

analyses (e.g. penalized joint analysis),  $P$ -values are not available, limiting the application of some meta-analysis techniques.

- (ii) Multiple raw datasets are combined and analyzed as one [128]. Obviously, the individual-data analysis methods described above can be directly applied to the combined data. The biggest challenge is that multiple datasets need to be sufficiently comparable, which is often not true. This ‘sufficient comparability’ can be partly assessed with meta-data and distributional properties. In some (early) gene expression studies, effort was made to transform gene expressions of different studies into a common distribution, e.g. multivariate normal [129]. However, this can be computationally expensive and is not always feasible (e.g. with SNP data).
- (iii) Integrative analysis [130], under which multiple datasets are jointly analyzed in a single analysis, making it significantly different from meta-analysis. Also different from the second category of analysis, it is recognized that different datasets/studies may have unsolvable differences and hence are allowed to have different models. In the literature, two types of models have been proposed. In Figure 7, one row represents one gene, one column corresponds to one dataset and a gray block corresponds to one identified gene. The three datasets are allowed to have different models with different regression coefficients. Under the *homogeneity model* (left panel), multiple datasets/models have the same *sparsity structure*. That is, the same set of genes is identified as important in all datasets. Under the *heterogeneity model* (right panel), different sets of genes may be identified. For the homogeneity model, one row of regression coefficients (corresponding to the effects of the same gene in multiple datasets) can be viewed as a ‘group’, and the one-level gene-set variable selection methods (e.g. group Lasso) can be applied. For the heterogeneity model, we not only need to determine whether a gene is important at all but also in which dataset(s) it is important. This amounts to a two-level variable selection, and the methods reviewed in Section ‘Joint analysis (para. Assume that the genes...)’ (e.g. sparse group Lasso) can be applied.

Multi-datasets analysis can be unsupervised and supervised, marginal and joint, as well as individual-gene-, gene-set- and gene-network-based. Meta-analysis, partly because of its simplicity and long history, has been extensively conducted by biomedical researchers. However, it is recognized that, driven by small sample sizes, individual datasets may generate unsatisfactory results, and pooling multiple bad results may not lead to good ones. Integrative analysis is more sensible but computationally more complicated. Although some researchers have published research codes, there is still a



**Figure 7.** Variable selection scheme of integrative analysis. One row represents one gene, one column corresponds to one dataset, and a gray block corresponds to one identified gene. Homogeneity (left) and heterogeneity (right) model.

lack of user-friendly software. The existing developments have been mostly limited to the statistics community. However, with its promising performance, more integrative analysis in biomedical studies is expected.

## Multi-omics analysis

Multiple types of omics measurements have been implicated in cancer. In the past decade, as represented by TCGA, cancer studies are going multi-omics, collecting multiple types of omics measurements on the same subjects. Table 1 contains partial TCGA data information for four cancers. With the interconnections among different types of omics measurements, collectively analyzing multiple types of omics data can take a system perspective and provide insights not shared by the analysis of a single type of data. Multi-omics data have been analyzed in multiple ways.

The first type of analysis is ‘mechanistic’. For example, in [131], the regulations of gene expressions by genetic and epigenetic changes were analyzed. This amounts to an ultrahigh dimensional regression analysis, with both response (e.g. gene expressions) and covariates (e.g. SNPs) being high dimensional. Both ‘marginal regression analysis + multiple comparison adjustment’ and ‘joint analysis + regularization’ have been conducted. This type of analysis is usually unsupervised.

The second type of analysis targets validating or strengthening results from analysis with one type of omics data (e.g. gene expressions) by those from another type of data (e.g. SNPs). There are two main strategies. (i) The first is a post-analysis strategy [132]. Two or more types of data are analyzed separately, and then findings are compared. If, e.g. significant findings based on gene expressions match those based on the corresponding



SNPs, then they can be more trustworthy. This type of analysis can be ineffective if each individual analysis is too unsatisfactory. (ii) With the second strategy, multiple types of data are analyzed simultaneously [133, 134]. For example, in consensus clustering, clustering of samples is conducted using e.g. gene expressions, and another clustering is simultaneously conducted using e.g. SNPs. It is reinforced that the two clustering structures are identical. Overall, this type of analysis can be limited by the fact that different types of omics data have not only overlapping but also independent information [135], leading to the fundamental question of whether they should lead to identical results.

The third type of analysis puts more emphasis on the independent information contained in multiple types of omics data and seeks to build more powerful ‘mega’ models. Multiple strategies have been developed. (i) The first strategy stacks multiple types of data together and then applies the methods reviewed above [136]. It is simple but ignores the overlapping information contained in different types of data, which can be manifested as collinearity. (ii) The second strategy [137] first analyzes one type of data and applies the methods reviewed above. This step can reduce dimensionality to one (a linear combination of variables). Then *conditional on* this effect, another type of omics data is integrated. Since we now only need to deal with one type of high-dimensional data, the same technique as adopted in the previous step can be applied. This procedure is iterated until no data is left or no improvement in model estimation/prediction is observed. In this analysis, the order of different types of data (entering the model) can be determined statistically (as in a step-up analysis) or biologically (e.g. the type of omics data ‘closest to’ cancer outcome enters first). (iii) The third strategy [138] first decomposes multiple types of omics data. With a slight abuse of notation, use  $X$  and  $W$  to denote the *linear spaces* spanned by, say, gene expressions and SNPs, respectively.  $U = X \cap W$ , the overlap of the two spaces, is first computed. Then  $X \setminus U$  and  $W \setminus U$ , which are  $X$  and  $W$  ‘removing’  $U$ , are computed. Variable selection and model building are then based on  $\{U, X \setminus U, W \setminus U\}$  and the methods reviewed above. With this strategy, the three components contain independent information, solving the collinearity problem. On the other hand, e.g. one gene expression measurement may be decomposed into two components, making interpretation less clear.

Recently, reviews have been developed specifically for multi-omics data analysis centered on gene expression [132] and proteomic data [139]. The aforementioned techniques have also been extended, e.g. to accommodate genetic interactions [140]. A common concern shared by most of the existing multi-omics analysis is that introducing more data brings signals as well as extensive noises—the reliability of findings and whether the observed improvement in estimation/prediction is real are of concern. Overall, as multi-omics data are relatively new, methodological development is still a moving field.

Many multi-omics studies have been reported in the literature. As a representative example, Zhao and others analyzed TCGA data and associated cancer survival with clinical/demographic measurements and four types of omics measurements including gene expression, methylation, microRNA, and copy number alterations for invasive breast carcinoma, glioblastoma multiforme, acute myeloid leukemia and lung squamous cell carcinoma [141]. PCA, PLS and Lasso were applied along with Cox regression. It was found that, for most of the cancers, there was no substantial improvement in prediction performance when adding other omics measurements to the models with clinical/demographic variables and gene expressions.

## Final remarks

Omics data analysis is a big and evolving field. To differentiate this review from the other ones, we have focused on methods applicable to cancer data (which has certain unique properties), tried to cover the whole spectrum of analysis (as opposed to a certain type, e.g. gene-set analysis), and mostly focused on statistical methods. With cancer researchers as a major target, we have mostly referred to statistical publications, with which they may have less experience/knowledge. With our limited expertise, it is inevitable that some important analyses/methods are missed. In addition, with a focus on methodology, discussions on computational algorithms, software implementation, numerical comparison and data analysis examples are somewhat limited. We fully acknowledge the significance of these aspects; however, separate introductions/reviews would be needed.

For cancer omics researchers with possibly limited expertise in statistical analysis, we recommend that it is critically important to first understand the overall analysis framework and where each type of analysis is positioned. This review can be especially useful for this purpose. Next, understanding the objective and scheme of each analysis can be more important than the mathematical details. It should be recognized that many of the reviewed and other cancer omics data analyses are quite complicated, especially when there is no existing user-friendly software. The family of statistical methods for cancer omics data is still expanding fast. It is necessary to adopt cutting-edge new methods, which have the potential of more informatively describing cancer biology. Consultation with professional statisticians is recommended for properly selecting, conducting and interpreting analysis.

### Key points

- Cancer is an omics disease. A myriad of statistical methods has been developed, with diverse strategies, for analyzing cancer omics data.

- This review establishes the overall framework of analysis and comprehensively covers unsupervised and supervised analysis, as well as individual-gene-based, gene-set-based and gene-network-based analysis.
- It also covers special topics including genetic interaction analysis, multi-datasets analysis and multi-omics analysis.
- It can serve as a 'dictionary' for biomedical cancer researchers and a refresher for biostatisticians.

## Data availability

The TCGA data is publicly available at <https://portal.gdc.cancer.gov/>.

## Acknowledgements

We thank the editor and reviewers for their careful review and insightful comments, which have led to a significant improvement of this article.

## Conflict of interest

We declare that there is no conflict of interest.

## References

1. Yoo BC, Kim K-H, Woo SM, et al. Clinical multi-omics strategies for the effective cancer management. *J Proteomics* 2018;**188**: 97–106.
2. Chakraborty S, Hosen M, Ahmed M, et al. Onco-multi-OMICS approach: a new frontier in cancer research. *Biomed Res Int* 2018;**2018**:9836256.
3. Chen Y-X, Chen H, Rong Y, et al. An integrative multi-omics network-based approach identifies key regulators for breast cancer. *Comput Struct Biotechnol J* 2020;**18**:2826–35.
4. Koh EJ, Hwang SY. Multi-omics approaches for understanding environmental exposure and human health. *Mol Cell Toxicol* 2019;**15**(1):1–7.
5. Knox SS. From 'omics' to complex disease: a systems biology approach to gene-environment interactions in cancer. *Cancer Cell Int* 2010;**10**:11.
6. Yu KH, Snyder M. Omics profiling in precision oncology. *Mol Cell Proteomics* 2016;**15**(8):2525–36.
7. Das T, Andrieux G, Ahmed M, et al. Integration of online omics-data resources for cancer research. *Front Genet* 2020;**11**:578345.
8. Cho WC. *An Omics Perspective on Cancer Research*. Dordrecht: Springer, 2010.
9. Crowley J, Hoering A. *Handbook of Statistics in Clinical Oncology*. New York: Chapman and Hall/CRC, 2012.
10. Vinga S. Structured sparsity regularization for analyzing high-dimensional omics data. *Brief Bioinform* 2021;**22**(1):77–87.
11. Fan J, Han F, Liu H. Challenges of big data analysis. *Natl Sci Rev* 2014;**1**(2):293–314.
12. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Stat Sin* 2010;**20**(1):101–48.
13. Bühlmann P, Kalisch M, Meier L. High-dimensional statistics with a view toward applications in biology. *Annu Rev Stat Appl* 2014;**1**:255–78.
14. Hung J, Yang T, Hu Z, et al. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief Bioinform* 2012;**13**(3):281–91.
15. Chauvel C, Novoloaca A, Veyre P, et al. Evaluation of integrative clustering methods for the analysis of multi-omics data. *Brief Bioinform* 2020;**21**(2):541–52.
16. Altenbuchinger M, Weihs A, Quackenbush J, et al. Gaussian and mixed graphical models as (multi-) omics data analysis tools. *Biochim Biophys Acta Gene Regul Mech* 2020;**1863**(6): 194418.
17. Zhang Z, Zhao Y, Liao X, et al. Deep learning in omics: a survey and guideline. *Brief Funct Genomics* 2019;**18**(1):41–57.
18. Kaur P, Singh A, Chana I. Computational techniques and tools for omics data analysis: state-of-the-art, challenges, and future directions. *Arch Computat Methods Eng* 2021;**28**:4595–631.
19. Liu J, Lichtenberg T, Hoadley KA, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 2018;**173**(2):400–16.
20. Sun Q, Li M, Wang X. The cancer omics atlas: an integrative resource for cancer omics annotations. *BMC Med Genomics* 2018;**11**(1):63.
21. Zhang J, Bajari R, Andric D, et al. The international cancer genome consortium data portal. *Nat Biotechnol* 2019;**37**(4): 367–9.
22. Mao R, Hu S, Zhang Y, et al. Prognostic nomogram for childhood acute lymphoblastic leukemia: a comprehensive analysis of 673 patients. *Front Oncol* 2020;**10**:1673.
23. Clough E, Barrett T. The gene expression omnibus database. *Methods Mol Biol* 2016;**1418**:93–110.
24. Conesa A, Beck S. Making multi-omics data accessible to researchers. *Sci Data* 2019;**6**(1):1–4.
25. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 2017;**35**(6): 498–507.
26. Jiang CH, Yuan X, Li JF, et al. Bioinformatics-based screening of key genes for transformation of liver cirrhosis to hepatocellular carcinoma. *J Transl Med* 2020;**18**(1):40.
27. Carbone M, Arron ST, Beutler B, et al. Tumour predisposition and cancer syndromes as models to study gene–environment interactions. *Nat Rev Cancer* 2020;**20**(9):533–49.
28. Zhang Z, Li H, Jiang S, et al. A survey and evaluation of web-based tools/databases for variant analysis of TCGA data. *Brief Bioinform* 2018;**20**(4):1524–41.
29. Schroeder MP, Gonzalez-Perez A, Lopez-Bigas N. Visualizing multidimensional cancer genomics data. *Genome Med* 2013;**5**(1): 1–13.
30. González JR, Armengol L, Solé X, et al. SNPAssoc: an R package to perform whole genome association studies. *Bioinformatics* 2007;**23**(5):654–5.
31. Xu R, Wunsch DC. Clustering algorithms in biomedical research: a review. *IEEE Rev Biomed Eng* 2010;**3**:120–54.
32. Yu G, Yu X, Wang J. Network-aided bi-clustering for discovering cancer subtypes. *Sci Rep* 2017;**7**(1):1–5.
33. Kogan J. *Introduction to Clustering Large and High-Dimensional Data*. New York: Cambridge University Press, 2006.
34. Giordani P, Ferraro MB, Martella F. *An Introduction to Clustering with R*. Singapore: Springer, 2020.
35. Jiang Z, Li T, Min W, et al. Fuzzy c-means clustering based on weights and gene expression programming. *Pattern Recogn Lett* 2017;**90**:1–7.
36. Teran Hidalgo SJ, Zhu T, Wu M, et al. Overlapping clustering of gene expression data using penalized weighted normalized cut. *Genet Epidemiol* 2018;**42**(8):796–811.
37. Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012;**486**(7403):346–52.

38. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* 2005;**33**(suppl\_2):W741–8.
39. Schifano ED, Epstein MP, Bielak LF, et al. SNP set association analysis for familial data. *Genet Epidemiol* 2012;**36**(8):797–810.
40. Franke L, Van Bakel H, Fokkens L, et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* 2006;**78**(6):1011–25.
41. Martin A, Ochagavia ME, Rabasa LC, et al. BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics* 2010;**11**(1):91.
42. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;**4**(1): PMID: 16646834.
43. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**(1):559.
44. Chiquet J, Rigaiil G, Sundqvist M. A multiattribute Gaussian graphical model for inferring multiscale regulatory networks: an application in breast cancer. *Methods Mol Biol* 2019;**1883**:143–60.
45. Drton M, Maathuis MH. Structure learning in graphical modeling. *Annu Rev Stat Appl* 2017;**4**:365–93.
46. Zuo Y, Cui Y, Yu G, et al. Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. *BMC Bioinformatics* 2017;**18**(1):1–4.
47. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2008;**9**(3):432–41.
48. Xue L, Zou H. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann Stat* 2012;**40**(5):2541–71.
49. Wang J, Zuo Y, Man YG, et al. Pathway and network approaches for identification of cancer signature markers from omics data. *J Cancer* 2015;**6**(1):54.
50. Schwab JD, Ikononi N, Werle SD, et al. Reconstructing Boolean network ensembles from single-cell data for unraveling dynamics in the aging of human hematopoietic stem cells. *Comput Struct Biotechnol J* 2021;**19**:5321–32.
51. Kumari S, Nie J, Chen HS, et al. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS One* 2012;**7**(11):e50411.
52. Li P, Zhang C, Perkins EJ, et al. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC Bioinformatics* 2007;**8**:S13.
53. Song X, Ma S, Huang J, et al. A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics* 2007;**8**(2):197–211.
54. Kwon MS, Kim Y, Lee S, et al. Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC Genomics* 2015;**16**(9):S4.
55. Tan YD, Xu H. A general method for accurate estimation of false discovery rates in identification of differentially expressed genes. *Bioinformatics* 2014;**30**(14):2018–25.
56. Xi L, Feber A, Gupta V, et al. Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer. *Nucleic Acids Res* 2008;**36**(20):6535–47.
57. Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. *Brief Bioinform* 2011;**12**(6):714–22.
58. N, Nazarov PV, Czerwinski U, Cantini L, et al. Independent component analysis for unraveling the complexity of cancer omics datasets. *Int J Mol Sci* 2019;**20**(18):4414.
59. Sun S, Sun X, Zheng Y. Higher-order partial least squares for predicting gene expression levels from chromatin states. *BMC Bioinformatics* 2018;**19**(5):47–54.
60. Cook DR. Principal components, sufficient dimension reduction, and envelopes. *Annu Rev Stat Appl* 2018;**5**:533–59.
61. Ma Y, Zhu L. A review on dimension reduction. *Int Stat Rev* 2013;**81**(1):134–50.
62. Burges CJ. Dimension reduction: a guided tour. *Found Trends®. Mach Learn* 2010;**2**(4):275–365.
63. Vlaming R, Groenen PJ. The current and further use of ridge regression for prediction in quantitative genetics. *Biomed Res Int* 2015;**2015**:143712.
64. West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A* 2001;**98**(20):11462–7.
65. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B-Methodol* 1996;**58**(1):267–88.
66. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Statist* 2010;**38**(2):894–942.
67. Taylor J, Tibshirani R. Post-selection inference for penalized likelihood models. *Can J Stat* 2018;**46**(1):41–61.
68. Chai H, Zhang Q, Huang J, et al. Inference for low-dimensional covariates in a high-dimensional accelerated failure time model. *Stat Sin* 2019;**29**(2):877.
69. Ding MQ, Chen L, Cooper GF, et al. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Mol Cancer Res* 2018;**16**(2):269–78.
70. Tibshirani R, Saunders M, Rosset S, et al. Sparsity and smoothness via the fused lasso. *J R Stat Soc Ser B-Stat Methodol* 2005;**67**(1):91–108.
71. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinform* 2008;**9**(5):392–403.
72. Bühlmann P, Van De Geer S. *Statistics for High-Dimensional Data*. Berlin: Springer, 2011.
73. Shao J, Wang Y, Deng X, et al. Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann Statist* 2011;**39**(2):1241–65.
74. Ma S, Huang J. Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* 2005;**21**(24):4356–62.
75. Bühlmann P, Yu B. Sparse boosting. *J Mach Learn Res* 2006;**7**:1001–24.
76. O'Hara RB, Sillanpää MJ. A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal* 2009;**4**(1):85–117.
77. Zou H, Xue L. A selective overview of sparse principal component analysis. *Proc IEEE* 2018;**106**(8):1311–20.
78. Deep learning for genomics. *Nat Genet* 2019;**51**:1. <https://doi.org/10.1038/s41588-018-0328-0>.
79. Eraslan G, Avsec Ž, Gagneur J, et al. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**(7):389–403.
80. Lin Y, Zhang W, Cao H, et al. Classifying breast cancer subtypes using deep neural networks based on multi-omics data. *Genes* 2020;**11**(8):888.
81. Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**16**(3):841–50.



82. Kim S, Kim K, Choe J, et al. Improved survival analysis by learning shared genomic information from pan-cancer data. *Bioinformatics* 2020;**36**(Supplement\_1):i389–98.
83. Bau D, Zhu J-Y, Strobel H, et al. Understanding the role of individual units in a deep neural network. *Proc Natl Acad Sci U S A* 2020;**117**(48):30071–8. <https://doi.org/10.1073/pnas.1907375117>.
84. Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinform* 2007;**8**:60.
85. Leeuw CA, Neale BM, Heskes T, et al. The statistical properties of gene-set analysis. *Nat Rev Genet* 2016;**17**:353–64.
86. Goeman JJ, Van De Geer SA, De Kort F, et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;**20**(1):93–9.
87. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Nati Acad Sci* 2015;**102**(43):15545–50.
88. Yuan L, Liu J, Ye J. Efficient methods for overlapping group Lasso. *IEEE T Pattern Anal* 2013;**35**(9):2104–16.
89. Bauer S, Gagneur J, Robinson PN. GOing bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res* 2010;**38**(11):3523–32.
90. Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. *Stat Sci* 2012;**27**(4). <https://doi.org/10.1214/12-STS392>.
91. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Ser B Stat Methodol* 2006;**68**:49–67.
92. Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* 2007;**23**(12):1486–94.
93. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput* 2015;**25**:173–87.
94. Simon N, Friedman J, Hastie T, et al. A sparse-group Lasso. *J Comput Graph Stat* 2013;**22**:231–45.
95. Breheny P, Huang J. Penalized methods for bi-level variable selection. *Stat Interface* 2009;**2**(3):369–80.
96. Huang J, Ma S, Xie H, et al. A group bridge approach for variable selection. *Biometrika* 2009;**96**(2):339–55.
97. Xie G, Dong C, Kong Y, et al. Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features. *Genes* 2019;**10**(3):240.
98. Ma S, Huang J. Clustering threshold gradient descent regularization: with applications to microarray studies. *Bioinformatics* 2007;**23**(4):466–72.
99. Ma S, Huang Y, Huang J, et al. Gene network-based cancer prognosis analysis with sparse boosting. *Genet Res* 2012;**94**(4):205–21.
100. Cai M, Dai M, Ming J, et al. BIVAS: a scalable Bayesian method for bi-level variable selection with applications. *J Comput Graph Stat* 2020;**29**:40–52.
101. Cowen L, Ideker T, Raphael BJ, et al. Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* 2017;**18**(9):551–62.
102. Huang J, Ma S, Li H, et al. The sparse Laplacian shrinkage estimator for high-dimensional regression. *Ann Stat* 2011;**39**(4):2021–46.
103. Liu J, Huang J, Ma S. Incorporating network structure in integrative analysis of cancer prognosis data. *Genet Epidemiol* 2013;**37**(2):173–83.
104. Yang Y, Han L, Yuan Y, et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat Commun* 2014;**5**:3231.
105. Shen R, Ghosh D, Chinnaiyan A, et al. Eigengene-based linear discriminant model for tumor classification using gene expression microarray data. *Bioinformatics* 2006;**22**(21):2635–42.
106. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 2007;**1**:54.
107. Ma S, Kosorok MR, Huang J, et al. Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. *BMC Med Genomics* 2011;**4**:5.
108. Chekouo T, Stingo FC, Doecke JD, et al. miRNA–target gene regulatory networks: a Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics* 2015;**71**:428–38.
109. Tang Z, Shen Y, Li Y, et al. Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. *Bioinformatics* 2018;**34**(6):901–10.
110. Hunter J. Gene-environment interactions in human diseases. *Nat Rev* 2005;**6**:287–98.
111. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**:392–403.
112. Zhang S, Xue Y, Zhang Q, et al. Identification of gene-environment interactions with marginal penalization. *Genet Epidemiol* 2020;**44**(2):159–96.
113. D'Angelo GM, Rao D, Gu CC. Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC Proc* 2009;**3**:S62.
114. He Z, Zhang M, Lee S, et al. Set-based tests for the gene-environment interaction in longitudinal studies. *J Am Stat Assoc* 2017;**112**(519):966–78.
115. Wu M, Huang J, Ma S. Identifying gene-gene interactions using penalized tensor regression. *Stat Med* 2018;**37**:598–610.
116. Shan L, Chen Z. Sequential interaction group selection by the principle of correlation search for high-dimensional interaction models. *Stat Sinica* 2021;**31**:197–221.
117. Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions. *Ann Stat* 2013;**41**:1111–41.
118. Liu J, Huang J, Zhang Y, et al. Identification of gene-environment interactions in cancer studies using penalization. *Genomics* 2013;**102**:189–94.
119. Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization. *J Comput Graph Stat* 2015;**24**:627–54.
120. Hao N, Zhang HH. Interaction screening for ultrahigh-dimensional data. *J Am Stat Assoc* 2014;**109**(507):1285–301.
121. Zhu Q, Zhao H, Ma S. Identifying gene-environment and gene-gene interactions using a progressive penalization approach. *Genet Epidemiol* 2014;**38**:353–68.
122. Yang CH, Lin YD, Yen CY, et al. A systematic gene-gene and gene-environment interaction analysis of DNA repair genes XRCC1, XRCC2, XRCC3, XRCC4, and oral cancer risk. *OMICS* 2015;**19**(4):238–47.
123. Sun J, Zhao H, Lin S, et al. Integrative analysis from multi-centre studies identifies a function-derived personalized multi-gene signature of outcome in colorectal cancer. *J Cell Mol Med* 2019;**23**:5270–81.
124. Chen Y, Ke W, Chiu H. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput Biol Med* 2014;**48**:1–7.

125. Chen F, Zhang Y, Gibbons DL, et al. Pan-cancer molecular classes transcending tumor lineage across 32 cancer types, multiple data platforms, and over 10,000 cases. *Clin Cancer Res* 2018;**24**(9):2182–93.
126. Wang S, Wu M, Ma S. Integrative analysis of cancer omics data for prognosis modeling. *Gene* 2019;**10**:604.
127. Guerra R, Goldstein DR. *Meta-Analysis and Combining Information in Genetics and Genomics*. New York: Chapman and Hall/CRC, 2009.
128. Ye S, Cheng K, Hu J, et al. Development and validation of an individualized gene expression-based signature to predict overall survival in metastatic colorectal cancer. *Ann Transl Med* 2020;**8**(4):96.
129. Shabalín AA, Tjelmeland H, Fan C, et al. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* 2008;**24**(9):1154–60.
130. Zhao Q, Shi X, Huang J, et al. Integrative analysis of ‘-omics’ data using penalty functions. *WIREs Comput Stat* 2015;**7**(1):99–108.
131. Shi X, Zhao Q, Huang J, et al. Deciphering the associations between gene expression and copy number alteration using a sparse double Laplacian shrinkage approach. *Bioinformatics* 2015;**31**(24):77–3983.
132. Wu M, Yi H, Ma S. Vertical integration methods for gene expression data analysis. *Brief Bioinform* 2021;**22**(3):1–14.
133. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics* 2013;**29**(20):2610–6.
134. Li Y, Bie R, Hidalgo SJ, et al. Assisted gene expression-based clustering with AWNCut. *Stat Med* 2018;**37**(29):4386–403.
135. Rish A, Plass C. Lung cancer epigenetics and genetics. *Int J Cancer* 2008;**123**:1–7.
136. Kim S, Herazomaya JD, Kang DD, et al. Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genomics* 2015;**16**:924.
137. Wang S, Shi X, Wu M, et al. Horizontal and vertical integrative analysis methods for mental disorders omics data. *Sci Rep* 2019;**9**:13430.
138. Zhu R, Zhao Q, Zhao H, et al. Integrating multidimensional omics data for cancer outcome. *Biostatistics* 2016;**17**(4):605–18.
139. Wu M, Jiang Y, Ma S. Integration of proteomics and other omics data. *Methods Mol Biol* 2021;**2361**:307–24.
140. Xu Y, Wu M, Ma S. Multidimensional molecular measurements-environment interaction analysis for disease outcomes. *Biometrics* 2021. <https://doi.org/10.1111/biom.13526>.
141. Zhao Q, Shi X, Xie Y, et al. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Brief Bioinform* 2015;**16**(2):291–303.