Integrative Analyses of Cancer Data: A Review from a Statistical Perspective

Yingying Wei

Department of Statistics, The Chinese University of Hong Kong, Shatin, Hong Kong.

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

ABSTRACT: It has become increasingly common for large-scale public data repositories and clinical settings to have multiple types of data, including high-dimensional genomics, epigenomics, and proteomics data as well as survival data, measured simultaneously for the same group of biological samples, which provides unprecedented opportunities to understand cancer mechanisms from a more comprehensive scope and to develop new cancer therapies. Nevertheless, how to interpret a wealth of data into biologically and clinically meaningful information remains very challenging. In this paper, I review recent development in statistics for integrative analyses of cancer data. Topics will cover meta-analysis of homogeneous type of data across multiple studies, integrating multiple heterogeneous genomic data types, survival analysis with high- or ultrahigh-dimensional genomic profiles, and cross-data-type prediction where both predictors and responses are high- or ultrahigh-dimensional vectors. I compare existing statistical methods and comment on potential future research problems.

KEYWORDS: integrative analysis, cancer genomics, survival analysis, high-dimensional data, ultrahigh-dimensional data

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Wei. Integrative Analyses of Cancer Data: A Review from a Statistical Perspective. *Cancer Informatics* 2015:14(S2) 173–181 doi: 10.4137/CIN.S17303.

RECEIVED: November 18, 2014. **RESUBMITTED:** February 01, 2015. **ACCEPTED FOR PUBLICATION:** February 09, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Review

FUNDING: Author discloses no funding sources.

COMPETING INTERESTS: Author discloses no potential conflicts of interest.

CORRESPONDENCE: ywei@sta.cuhk.edu.hk

Introduction

With the rapid development and decreasing cost of high-throughput technologies, cancer biology has moved from a data-poor field to a data-abundant field. To illustrate, to date, more than 1,000,000 samples have been stored in Gene Expression Omnibus[1] and ArrayExpress[2]; meanwhile, over 2,500 sequencing samples are deposited in the ENCyclopedia of DNA elements (ENCODE) project[3] and the Sequence Read Archive[4]. Moreover, multiple types of genomics, epigenomics, and proteomics data, together with clinical data such as survival data, are simultaneously measured for cancer patients in the Cancer Genome Project (CGP),[5] the International Cancer Genome Consortium (ICGC),[6] and The Cancer Genome Atlas (TCGA). Consequently, the large volume of data provides unprecedented opportunities as well as challenges for using integrative analysis to reveal cancer mechanisms. Novel statistical methods and theories are in urgent demand for interpreting the wealth of data into biologically and clinically meaningful information while avoiding the "blind men and an elephant" scenario.

When tracing the history of the past two decades, it can be seen that high-throughput biology has triggered the active statistical research in high-dimensional data. Recently, to handle the real dimension of genomic data, which is usually beyond the scale of hundreds of variables, methods handling ultrahigh-dimensional data are emerging. The diversity of cancer data types together with the availability of related studies on similar types of cancers adds another two dimensionalities of complexity. It is of critical clinical and biological interests to understand what subtypes a cancer have, how genomic profiles and survival rates of patients vary among subtypes, whether a patient's survival can be predicted from his or her genomic profiles, and how one type of genomic profile is correlated with another type of genomic profile. No doubt, the abundance and sophisticated structures of cancer data will drive a whole class of exciting statistical problems in the coming years. In this paper, I review recent developments in statistical methods for integrative analyses of cancer data. This review complements an earlier review this year[7] from a statistical perspective, with a more detailed comparison of statistical methods, a broader range of topics such as integration of homogeneous type of genomic profiles across studies and integration of genomic profiles with survival data, as well as comments on potential extension of current methods (see Table 1). With increasing feasibility to access cancer data from those public repositories[8] so that anyone interested in

the data can easily work on them, we hope the current review will arouse interests in developing new statistical tools and theories for integrative genomic analyses in cancer.

The rest of the paper is organized as follows. In the next section, I review models for integrating a single type of genomic profile across multiple studies to improve signal detection. This is followed by a section that is devoted to the integration of multiple types of genomic profiles. The next two sections present, respectively, integrating genomic data with survival analysis and the cross-data-type prediction problem where both the responses and predictors are high dimensional. Finally, the last section concludes the paper.

Integration of a Single Genomic Data Type

High-throughput technologies often have low signal-to-noise ratio. Consequently, results obtained from analysis based on a single study often suffer from low reproducibility either because of the low sample size or the heterogeneity of the datasets. With the rapid accumulation of related studies in public data repositories as mentioned above, it is more cost effective to borrow information across studies to improve signal detection. Nevertheless, caution should be paid when pooling datasets together to account for systematic biases such as batch effects as well as study specificity.

Batch effects. Batch effects are widespread in high-throughput biology. They are artifacts not related to the biological variation of scientific interests. For instance, two microarray experiments on the same technical replicates processed on two different days might present different results due to factors such as room temperature or the two technicians who did the two experiments. Batch effects can substantially confound the downstream analysis, especially meta-analysis across studies. Moreover, even more recent technologies such as next-generation sequencing cannot eliminate batch effects.[9] Therefore, it is crucial to correct batch effects for valid integration across studies.

Table 1. Summary of the main reviewed methods.

NAME	INTEGRATION TYPE	CORE STATISTICAL METHOD	REFERENCE
Combat	Single data type, multiple studies	Empirical Bayes	10
SVA	Single data type, multiple studies	Surrogate variable analysis	11,12
svaseq	Single data type, multiple studies	Surrogate variable analysis	13
RUV	Single data type, multiple studies	Generalized linear model	14
Consistent DE	Single data type, multiple studies	Bayesian hierarchical model	15
EBarrays	Single data type, multiple studies	Bayesian hierarchical model	16–19
XDE	Single data type, multiple studies	Bayesian hierarchical model	20
Cormotif	Single data type, multiple studies	Bayesian hierarchical model	21
2-Norm group bridge	Single data type, multiple studies	Penalized method	22
iCluster	Multiple data types, single study	Matrix factorization	33
Joint Bayesian factor	Multiple data types, single study	Matrix factorization	38
JIVE	Multiple data types, single study	Matrix factorization	42
md-module	Multiple data types, single study	Matrix factorization	45
MDI	Multiple data types, single study	Bayesian hierarchical model	51
Prob_GBM	Multiple data types, single study	Bayesian hierarchical model	53
Consensus clustering	Multiple data types, single study	Bayesian hierarchical model	54
SNF	Multiple data types, single study	Network fusion	54
Multi-attribute graph	Multiple data types, single study	Network fusion	57,58
Penalized survival	Single data type with survival	Penalized method	69–71
Network penalized survival	Single data type with survival	Penalized method	73,74
SIS survival	Single data type with survival	Sure independence screening	76
PSIS	Single data type with survival	Sure independence screening	77
FAST	Single data type with survival	Sure independence screening	80
Bagging survival trees	Single data type with survival	Bootstrap	82
Survival ensembles	Single data type with survival	Inverse probability weighting	85
RIST	Single data type with survival	Imputation	88
T_SVD	Multiple data types multiple studies	Neural network	92

For microarray data, when the batches are known, a location and scale adjustment method, combat, was developed to adjust for batch effects.[10] The core idea of combat[10] was that the observed measurement Y_{ijg} for the expression value of gene g for sample j from batch i can be expressed as

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}, \quad (1)$$

where X consist of covariates of scientific interests while γ_{ig} and δ_{ig} characterize the additive and multiplicative batch effects of batch i for gene g. After obtaining the estimators from the above linear regression, the raw data Y_{ijg} can be adjusted to Y^*_{ijg} :

$$Y^*_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g + X\hat{\beta}_g. \quad (2)$$

For real application, an empirical Bayes method was applied for parameter estimation.

When batches were unknown, the surrogate variable analysis (SVA)[11,12] was developed. The main idea was to separate the effects caused by covariates of our primary interests from the artifacts not modeled. Parallel to Equation (1), now the raw expression value Y_{jg} of gene g in sample j can be formulated as

$$Y_{jg} = \alpha_g + X\beta_g + \sum_{k=1}^{K} \lambda_{kg} h_{kj} + \varepsilon_{jg}, \quad (3)$$

where h_{kj}s represent the unmodeled factors and are called as "surrogate variables". Once again, the basic idea was to estimate h_{kj}s and adjust them accordingly. An iterative algorithm based on singular value decomposition was derived to iterate between estimating the main effects $\hat{\alpha}_g + X\hat{\beta}_g$ given the estimation of surrogate variables and estimating surrogate variables from the residuals $r_{jg} = Y_{jg} - \hat{\alpha}_g - X\hat{\beta}_g$.

For sequencing data, svaseq, the generalized version of SVA, suggested applying a moderated log transformation to the count data or fragments per kilobase of exon per million fragments mapped (FPKM) first to account for the nature of discrete distributions,[13] thus updating Equation (3) to:

$$log(Y_{jg} + c) = \alpha_g + X\beta_g + \sum_{k=1}^{K} \lambda_{kg} h_{kj} + \varepsilon_{jg}, \quad (4)$$

where c is a small positive constant. Instead of a direct transformation on the raw counts or FPKM, remove unwanted variation (RUV) adopted a generalized linear model for Y_{jg} with the conditional mean specified as:

$$log(E(Y_{jg} \mid X, h_{1j}, \ldots, h_{Kj}, \varepsilon_{jg})) = \alpha_g + X\beta_g + \sum_{k=1}^{K} \lambda_{kg} h_{kj} + \varepsilon_{jg}. \quad (5)$$

RUV also allowed the use of negative control genes and control samples with details listed in its online methods.[14]

Hierarchical model. Differential expression detection between cancer patients and control samples is usually the first step to screen for risk genes and drug targets. However, as mentioned in the beginning of this section, gene expression microarrays suffer from noisy measurements, especially when only a small number of samples are available. Consequently, it is appealing to pool information across related studies or related cancer types to borrow strength. Specifically, within each study $d = 1,\ldots,D$, we have n_{0d} control samples and n_{1d} cancer patients. The gene expression for total G genes is measured for each sample. Our task is to determine whether each gene g is differentially expressed in a given study d. Hereafter, we assume that data have already been properly normalized and adjusted for batch effects.

The simplest method of pooling information is to assume that a gene is either differentially expressed in all studies or none of the studies.[15] However, it fails to allow genes to be differentially expressed in only a subset of studies, thus losing study specificity. A more flexible model EBarrays[16–18] included all the possible differential expression patterns into the mixture model and fitted the model with an empirical Bayes approach. The Markov-chain Monte Carlo (MCMC) algorithm was also developed for model fitting along the line.[19] EBarrays performs well when the total number of studies integrated D is small, but it encounters the barrier of exponential growth of parameters when D is large, as it has to enumerate all 2^D possible patterns. XDE[20] did not have the exponential growth parameter space problem, but its Bayesian hierarchical model assumed that each gene had the same prior probability to be differentially expressed within a given study. To tackle the exponential growth of the parameter space while still allowing heterogeneity among genes, Cormotif[21] adopted a small number of latent probability vectors to capture the correlation among studies while still able to regenerate all 2^D differential expression patterns.

Ma et al.[22] considered a more general case where a response variable Y_{jd} was available for each sample j within every study d. The task was to build a regression model $f((\beta^d)^T \boldsymbol{x}^{(jd)})$, where $\boldsymbol{x}^{(jd)}$ is the gene expression profile for sample j within study d. Differential expression detection is a subclass of this problem, with Y_{jd} being binary. The authors adopted a penalized approach to select genes whose coefficients were nonzeros. Although the penalty functions were designed to enforce the same set of genes to have nonzero coefficients across all studies, the magnitudes of coefficients were allowed to vary across studies. It would be of interest to investigate a more flexible model where the set of genes with nonzero coefficients is also allowed to vary from study to study.

Despite the refining methods for detecting differential expression from a single sequencing experiment, including DEseq[23] and edgeR,[24–26] the sequencing data version of hierarchical models for integration of multiple studies

still requires development to address the typical discrete distributions observed for count and FPKM data. Before more fine-tailored methods becoming available, one potential easy approach might be to conduct a moderate log transformation as svaseq[13] and then apply the aforementioned microarray-based methods.

Integration of Multiple Genomic Data Types

Due to the decreasing cost of high-throughput technologies, more and more studies now measure multiple heterogeneous genomic profiles simultaneously for the same set of samples (patients and controls) such as gene expression, gene mutations, copy number alterations, and DNA methylations, where each data type consists of tens of thousands of measurements. A key problem of heterogeneous data type integration is how to characterize the common structure shared by all the data types as well as the individual data-type-specific variation.

In this review, I will focus on the recent statistical methodology development for integrative analyses of cancer data. Meanwhile, many well-developed machine learning algorithms, such as boosting,[27,28] random forest,[29] and support vector machine,[30] have also been increasingly applied to cancer data and proven good prediction performance although with less interpretability. Readers may consult the corresponding review papers and the reference therein for details.[7,31,32] The recently developed statistical methods can in general be categorized into three classes: matrix factorization, Bayesian models, and network fusion. In many scenarios, sparsity assumptions are also incorporated for regularization purpose to select a more parsimonious set of features. Here, I let $\{\boldsymbol{X}^{(d)}\}_{d=1,\ldots,D}$ represent D different data types. $\boldsymbol{X}^{(d)}_{p_d \times N}$ are the measurements for p_d genomic features on N objects for data type d.

Matrix factorization. Matrix factorization aims at decomposing the variation in the datasets with lower rank matrix approximation. Assuming there are a set of "fundamental" common factors $\boldsymbol{F}$ determining the values of all the original genomic features, the iCluster model was developed as[33]

$$\boldsymbol{X}^{(d)} = \boldsymbol{L}^{(d)}\boldsymbol{F} + \boldsymbol{E}^{(d)}, \boldsymbol{F} \sim \boldsymbol{N}(0, \boldsymbol{I}). \quad (6)$$

Here, $\boldsymbol{F}$ are the K underlying factors; $\boldsymbol{L}^{(d)}$ is a $p_d \times K$ matrix containing the factor loadings specific to data type d; and $\boldsymbol{E}^{(d)} \sim N(0, \boldsymbol{\Psi}^{(d)})$ are the residual terms after accounting for the common factors. Sparsity was imposed on the loadings $\boldsymbol{L}^{(d)}$. To accommodate different characteristics of heterogeneous data types, different types of penalty functions (the lasso penalty,[34] the fused lasso penalty,[35] and the elastic net penalty[36]) were applied to different data types. For instance, the fused lasso penalty was specifically suitable for DNA copy number data, as it accounted for spatial dependence along the genome. Treating $\boldsymbol{F}$ as "missing data", an Expectation-Maximization algorithm[37] was applied to the penalized complete likelihood for model fitting. Cancer subtypes were determined according to a standard K-means clustering on $E(\boldsymbol{F}|\boldsymbol{X})$. A resampling-based criterion measuring cluster reproducibility was used to choose the tuning parameters for the penalty parameters and the number of latent factors K.

Along this line, Ray et al.[38] generalized the above model to the following factorization:

$$\boldsymbol{X}^{(d)} = \boldsymbol{L}^{(d)}(\boldsymbol{F}^{(c)} + \boldsymbol{F}^{(d)}) + \boldsymbol{E}^{(d)}. \quad (7)$$

$\boldsymbol{F}^{(c)}$ represent the factor scores shared by all data types, and $\boldsymbol{F}^{(d)}$ are the factor scores specific to data type d. The model further assumed sparsity on both the factors scores $\boldsymbol{F}^{(c)}$ and $\boldsymbol{F}^{(d)}$, as well as factor loadings $\boldsymbol{L}^{(d)}$. For selection of the number of factors K, a finite beta-Bernoulli process was employed as an approximation to the Indian buffet process[39,40] for the binary indicators of the nonzero components in $\boldsymbol{F}^{(c)}$ and $\boldsymbol{F}^{(d)}$. After specifying the priors for all the parameters in the model, a Gibbs sampler[41] was used for posterior inference.

Instead of the same sharing factor loadings $\boldsymbol{L}^{(d)}$ for both $\boldsymbol{F}^{(c)}$ and $\boldsymbol{F}^{(d)}$, Joint and Individual Variation Explained (JIVE) proposed a similar model where data-type-specific loadings were also allowed for the common factors $\boldsymbol{F}^{(c)}$.[42] In other words, the model can now be factored as

$$\boldsymbol{X}^{(d)} = \boldsymbol{W}^{(d)}\boldsymbol{F}^{(c)} + \boldsymbol{L}^{(d)}\boldsymbol{F}^{(d)} + \boldsymbol{E}^{(d)}. \quad (8)$$

Denoting $\boldsymbol{J}_d = \boldsymbol{W}^{(d)}\boldsymbol{F}^{(c)}$ and $\boldsymbol{A}_d = \boldsymbol{L}^{(d)}\boldsymbol{F}^{(d)}$, $\boldsymbol{J} = [\boldsymbol{J}_1, \ldots, \boldsymbol{J}_D]^T$ and $\boldsymbol{A}_d, d = 1\ldots, D$ were allowed to have different ranks. A permutation testing approach was used to select the number of factors. With the orthogonal constraint that $\boldsymbol{J}\boldsymbol{A}_d^T = 0, d = 1, \ldots, D$, the joint structure $\boldsymbol{J}$ and the individual structure $\boldsymbol{A}_d$, $d = 1\ldots, D$ were fitted iteratively by fixing one at a time and minimizing the square norm of the residual matrices. To induce sparsity, L_1 penalties were placed on the loading matrices $\boldsymbol{W}^{(d)}$ and $\boldsymbol{L}^{(d)}$ and incorporated into the iterative estimating algorithm.

Nonnegative matrix factorization (NMF) attempts to decompose a nonnegative matrix into nonnegative loadings and nonnegative factors, thus describing the non-subtractive patterns in the data.[43,44] Zhang et al.[45] generalized the single matrix NMF to integrative analysis of multidimensional genomic data. After transforming the raw data into input data fulfilling the constraints of nonnegativity as Kim et al.[43], the following squared Euclidean error loss function was optimized[45]:

$$min \sum_{d=1}^{D} ||\boldsymbol{X}^{(d)} - \boldsymbol{L}^{(d)}\boldsymbol{F}||^2, \boldsymbol{F} \geq 0, \boldsymbol{L}^{(d)} \geq 0, d = 1, \ldots, D. \quad (9)$$

One drawback of the NMF decomposition lies in the time complexity of the fitting algorithms, which is on the scale of $O(tK(N + \sum_{d=1}^{D} p_d))$, with t being the iteration number for the fitting algorithm. Consequently, for a large number of genomic features, data reduction techniques such

as principal component analysis[46] were required in the data preprocessing step, which might result in loss of information. Moreover, network information can be incorporated into NMF. Network-based stratification (NBS)[47] minimized the following objective function in order to cluster tumors into subtypes according to somatic mutation profiles with $\boldsymbol{K}$ being an adjacency matrix encoding network information:

$$min\,||\,\boldsymbol{X}-\boldsymbol{LF}\,||^2+trace(\boldsymbol{L}^T\boldsymbol{KL}),\boldsymbol{F}\geq 0,\boldsymbol{L}\geq 0. \quad (10)$$

As pointed out by the authors, NBS can be further generalized to integrate multiple layers of information[47]; thus I expect a loss function as a combination of Equations (9) and (10).

A major issue with all the factorization approaches mentioned above is that they require proper normalization across data types. Generally, different data types have different distributions, different variability, and different numbers of genomic features. For instance, without proper scaling, as pointed out by Lock et al.[42], it is very likely that "the largest data set wins". JIVE attempted to handle that issue with normalization first across each row and then scaling across data types. On the other hand, as mentioned above, iCluster[33] tried to use different penalty functions to take care of different data features. However, it still failed to distinguish between binary, categorical, and continuous data types. The method proposed by Mo et al.[48] can be viewed as a generalization of iCluster[33] by incorporating different distribution assumptions while still assuming the same common latent factors for all types of data. Specifically, with i indexing patient and j indexing genomic feature, for binary outcome, it rephrased Equation (6) as

$$log\frac{P(x_{ij}^{(d)}=1\,|\,\boldsymbol{F}_i)}{1-P(x_{ij}^{(d)}=1\,|\,\boldsymbol{F}_i)}=\alpha_j^{(d)}+\boldsymbol{L}_j^{(d)}\boldsymbol{F}_i. \quad (11)$$

Similarly, for multicategory outcomes, with $P(x_{ij}^{(d)}=c\,|\,F_i)$, $c=1,\ldots,C$ denoting the probability for each category, Equation (6) became

$$P(x_{ij}^{(d)}=c\,|\,\boldsymbol{F}_i)=\frac{exp(\alpha_{jc}^{(d)}+\boldsymbol{L}_{jc}^{(d)}\boldsymbol{F}_i)}{\sum_{l=1}^{C}exp(\alpha_{jl}^{(d)}+\boldsymbol{L}_{jl}^{(d)}\boldsymbol{F}_i)} \quad (12)$$

Likelihood for count outcome with Poisson distribution and continuous variables with normal distribution can be derived accordingly. Lasso penalty was also placed on $\boldsymbol{L}^{(d)}$ for regularization. The tuning parameters for regularization was chosen by Bayesian information criterion (BIC), and the model was fitted by the modified Monte Carlo Newton–Raphson algorithm.[49,50]

A potential future research problem would be how to adapt different distribution assumptions into a more flexible factorization framework such as the joint Bayesian factor model[38] and JIVE.[42] Moreover, other than Bayesian framework, how to conduct statistical inference including significance tests and confidence intervals for factor models, especially with penalization methods, would also be an important future research problem. Another problem worth investigation in real application, as pointed out by the referee, is the choice of the number of components or clusters K; the authors of the above models have tried resampling-based criterion measuring cluster reproducibility, permutation based testing approach, Indian buffet process, and BIC, whereas the Akaike information criterion (AIC) and Bayesian factor might also be of interest.

Bayesian models. Bayesian hierarchical models are another set of popular tools for integrative analysis of heterogeneous data types. They offer the flexibility to model different data-type-specific distributions as well as various types of correlation among data types.

In multiple dataset integration (MDI),[51] the authors considered the case where multiple genomic data types were measured under a single biological condition for a common set of genomic features. For instance, gene expression data, protein–DNA interaction data, and protein–protein interaction data were measured simultaneously for the same group of genes. The model assumed that each data type followed a K-component mixture model.

Let c_{id} indicate the class membership of feature i in dataset d. Then, MDI modeled the associations among datasets via the following conditional prior for data-type-specific class memberships:

$$p(c_{i1},\ldots,c_{iD}\,|\,\phi)\propto\prod_{d=1}^{D}\pi_{c_{id}d}\prod_{d=1}^{D-1}\prod_{l=d+1}^{D}(1+\phi_{dl}1(c_{id}=c_{il})),\phi_{dl}\geq 0. \quad (13)$$

Here, $1(\cdot)$ is the indicator function, and ϕ_{dl} characterizes the pairwise association among multiple datasets. MDI was further extended by incorporating a feature selection step in modeling its data-type-specific distributions and applied to gene expression, copy number variation, methylation, and microRNA data of 277 glioblastoma samples from TCGA.[52] In MDI, ϕ_{dl} describes the global association between two datasets for all the features. A more flexible model might allow the association to vary from a cluster of features (genes) to another cluster of features (genes).

Instead of modeling associations among different data types, Prob_GBM[53] modeled the associations among patients using a patient-similarity network. It first discretized all the genomic features and concatenated them into one vector for each patient. Next, for each patient, it assumed that each genomic feature was generated from a multinomial distribution whose parameters were determined by a K-dimensional Dirichlet distribution. Consequently, the likelihood can be written out in a similar fashion as that of MDI, where ϕ are now determined by the binary links in the patient-similarity network. One drawback of this approach is that it requires the

discretization of each data type, which may lose a substantial amount of information.

The Bayesian consensus clustering was proposed to model the overall clustering consensus among different data types rather than pairwise associations among data types. Therefore, an overall single clustering can be achieved at patient level, resulting in cancer subtype discoveries.[54] Denoting the overall clustering labels as $C = (C_1, \ldots, C_N)$, then compared to Equation (13), the data-type-specific conditional model can now be formulated as

$$P(c_{id} = k \mid C_i) = v(k, C_i, \alpha_d) = \begin{cases} \alpha_d & if \quad c_{id} = C_i \\ \dfrac{1-\alpha_d}{K-1} & otherwise \end{cases}, \quad (14)$$

where α_d regulates the consensus between the clustering for dataset d and the overall clustering. So far, software has been developed with the data-specific distribution specified as normal distribution.

All the above models were embedded into the Bayesian framework. Consequently, one main challenge lies in the computation of the MCMC algorithm for model fitting. Generally speaking, compared to matrix factorization methods, the Bayesian hierarchical model provides more flexibility to model data-type-specific distributions and various dependence structures. Nevertheless, it remains challenging to build models that comprehensively capture the association among different data types, among patients, and among different clusters of genomic features.

Network fusion. Another emerging approach for identifying cancer subtypes is to construct networks for patients and then conduct clustering according to the obtained network graph. Similarity network fusion (SNF)[55] first constructed a similarity network of patients for each data type, where each node represented a patient and the weight on each edge indicated the similarity between two patients. Then, SNF normalized each network $\boldsymbol{W}^{(d)}$ into a matrix $\boldsymbol{P}^{(d)}$ that captured the global similarities among patients with row sums being 1 and a matrix $\boldsymbol{S}^{(d)}$ that described only the local similarities among the K nearest neighbors of each patient. By iteratively updating $\boldsymbol{P}^{(d)} = \boldsymbol{S}^{(d)} \times \boldsymbol{P}^{(d')} \times (\boldsymbol{S}^{(d)})^T$, $d' \neq d$ until convergence, SNF fused multiple networks $\boldsymbol{P}^{(d)}$ into a single network and used spectral clustering[56] to obtain clusters of nodes (patients).

Instead of building a graph for each type of data, Katenka et al.[57] stacked $\boldsymbol{X}^{(d)}$ to $\boldsymbol{X} = ((\boldsymbol{X}^{(1)})^T, \ldots, (\boldsymbol{X}^{(d)})^T)^T$. A hypothesis testing approach was used to construct an association network according to canonical correlation between two groups of attributes. Kolar et al.[58] continued on the same line and assumed $\boldsymbol{X}$ followed a joint multivariate Gaussian distribution. Then, a penalized log likelihood was optimized to estimate the partial canonical correlations for constructing a Markov graph.[59] Finally, nodes (patients) were clustered using a heuristic based on the edge weights in the obtained graph, as in Katenka et al.[57]

SNF lacks a rigorous probabilistic model to fuse multiple graphs; the methods of both Katenka et al.[57] and Kolar et al.[58] required $\boldsymbol{X}^{(d)}$ to be continuous, which might not be suitable for some types of genomic data such as copy number variation. Given the burst of statistical literature on multiple graphs estimation,[60–66] though usually for single data type across multiple conditions, I expect estimation of multiple graphs constructed from multiple data types and construction of a single graph from heterogeneous data types with data-type-specific distributions will call for novel statistical models, methods, and theories for network research.

Integration of Genomic Data with Survival Data

One of the major goals of cancer research is to identify the survival curves for cancer patients. Therefore, statistical methods for studying the relationship between survival data and high-dimensional genomic data are of vital clinical importance. Here, I briefly review recent development in integrating genomic data with survival data.

Let T_i and C_i denote the true underlying failure time and censoring time. However, we only see observed failure time $Y = min(T,C)$, and I use $\delta = 1(T \leq C)$ to indicate whether the observation is censored or not. $\boldsymbol{X} = (X_1, \ldots, X_p)$ are the p-dimensional covariates. Conditional-independent censoring mechanism given the covariates is usually assumed. Our goal is to reveal the dependence of survival time T on covariates $\boldsymbol{X}$ with the censored data $(Y, \delta, \boldsymbol{X})$. Two main approaches to model survival data with high-dimensional genomic data are penalization-based variable selection methods and tree-based ensemble learning methods.

Variable selection methods. The Cox proportional hazard model[67] is one of the most widely used models for survival data. It assumes that the hazard at time t for x^i is

$$\begin{aligned} \lambda(t \mid \boldsymbol{x}^i) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t \mid T \geq t, \boldsymbol{X} = \boldsymbol{x}^i) \\ &= \lambda_0(t) exp(\sum_{j=1}^{p} x^i_j \beta_j), \end{aligned} \quad (15)$$

where $\lambda_0(t)$ is the unspecified baseline hazard function. Then, for model fitting, the partial likelihood[68] can be derived as

$$L(\beta) = \prod_{i \in D} \frac{exp(\beta^T \boldsymbol{x}^i)}{\sum_{l \in R_i} exp(\beta^T \boldsymbol{x}^l)}, \quad (16)$$

where D is the set of indices of observed failures, and R_i is the set of indices for subjects who are at risk at failure time Y_i of subject i. $\hat{\beta}$ can be achieved by maximizing the log partial likelihood. For high-dimensional covariates $\boldsymbol{X}$, penalty functions for β such as lasso penalty[69] and smoothly clipped absolute deviation[70,71] can be applied to the log partial likelihood. For penalized variable selection methods as well as other dimension reduction methods developed for survival

data before 2009, see Witten and Tibshirani[72] for a detailed review. Along the same road map, when biological pathway information is available, penalty functions were also designed to conduct both group-level selection and within-group-level variable selection[73] as well as enforcing smoothness of regression coefficients for genes connected in a network.[74]

Parallel to the development of methods for linear models moving from high dimension to ultrahigh dimension, defined as the dimensionality growing exponentially with the sample size in Fan and Lv,[75] sure independence screening (SIS) type of methods have also been developed for survival data. Given outcome Y and ultrahigh-dimensional covariates $\boldsymbol{X} = (X_1,\ldots,X_p)$, SIS first screened $\boldsymbol{X} = (X_1,\ldots,X_p)$ according to their marginal correlation with Y to a subset of $\boldsymbol{X}_s = \{X_i\}_{i\varepsilon S}$ and then built a regression model for Y with the selected set $\boldsymbol{X}_s$ using various penalized approaches. For survival data, Fan et al.[76], extended the marginal correlation screening to screening on the marginal utility, defined as the maximum of the partial likelihood achieved by each single covariate for the censored outcome. The principled Cox sure independence screening procedure (PSIS)[77] screened on the standardized coefficient $I_j(\hat{\beta}_j)^{1/2}\hat{\beta}_j$, where $I_j(\hat{\beta}_j)^{-1}$ is the variance estimate, and further incorporated a false discovery rate control[78,79] procedure to determine the cutoff threshold automatically with theoretical justification provided. In contrast, rather than using the proportional hazard model, the feature aberration at survival times (FAST) statistic was developed in Gorst-Rasmussen and Scheike[80] as a measure of the aberration of each covariate relative to its at-risk average. Specifically, let $N(t) = 1(T \wedge C \leq t,\ T \leq C)$ be the counting process for the number of failures up to time t, and $Y(t) = 1(T \wedge C \leq t)$. Then, abbreviating $\bar{X} = \frac{\sum_{i=1}^{n} X_i Y_i(t)}{\sum_{i=1}^{n} Y_i(t)}$, FAST is defined as

$$d = n^{-1}\int_0^{\tau}\sum_{i=1}^{n}\{\boldsymbol{X}_i - \bar{\boldsymbol{X}}(t)\}dN_i(t). \qquad (17)$$

Theoretical justification of the SIS property for the FAST statistics within a class of single-index hazard rate models was provided.

Ensemble learning. Ensemble learning methods such as random forest[29] and boosting[81] are well known for offering outstanding prediction accuracy. Several methods have attempted to handle the missingness caused by censoring and thus generalized ensemble learning methods to survival data. Hothorn et al.[82] first drew multiple bootstrap samples[83] with replacement and constructed a survival tree for each bootstrap sample. Given a new observation, its survival function is estimated by the Kaplan–Meier curve[84] for all data points in all the trees that belonged to the same leaf node as the new observation. In Hothorn et al.[85], the authors first log-transformed Y, then missingness was accounted by adding inverse probability of censoring (IPC) weights[86] to the loss function for either random forest or gradient boosting.[87] The recursively imputed survival trees (RIST),[88] on the other hand, attempted to impute the censoring data and ran extremely randomized trees (ERT), a tree-based ensemble method with a higher degree of randomization than random forests, on the imputed complete data. RIST iterated between imputing censored data using conditional survival distributions and refitting the conditional survival distributions by pooling all the trees with imputed data. Despite the wide use of random forest, theoretical analyses of its consistency and asymptotics[89–91] are just emerging. Therefore, at present theoretical properties of tree-based ensemble methods remain significant challenges. Moreover, generalization to scenarios where covariates $\boldsymbol{X}$ consist of multiple data types as discussed in section "Integration of multiple genomic data types" is also of great interests.

Cross-Data-Type Prediction

An ultimate goal of genomics is to demystify the regulation program of different functional genomic profiles. How is DNA methylation related to gene expression? How does transcription factor binding control gene expression? What is the relationship between chromatin status and methylation status? The core problem underlying all these questions is whether we can predict one type of genomic profile from another, where both the response and predictor variables are multivariate with at least tens of thousands of variables. In such scenarios, we surpass simple or multiple linear regressions, penalized approaches such as lasso, and sure independence screening for ultrahigh dimensions in that the response variable itself is also an ultrahigh-dimension vector rather than a scalar one. The small sample size adds another dimension of challenge for inferring the relations between tens of thousands of responses and predictors.

The thresholding singular value decomposition (T_SVD) regression[92] is among the very first to study this problem. T_SVD actually adopted a standard single-layer neural network model to link the high-dimensional predictors $\boldsymbol{X}$ with the high-dimensional responses $\boldsymbol{Y}$. To bring this to light, the regression model can be formulated as

$$E(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{XC} = \boldsymbol{X}\sum_{j=1,\ldots,r} d_j \boldsymbol{u}_j \boldsymbol{v}_j^t = \boldsymbol{XUDV}^T, \qquad (18)$$

where $\boldsymbol{u}_j$ is the input weights for the j^{th} hidden-intermediate node in the neural network while $\boldsymbol{v}_j$ is the output weights for the j^{th} node. Consequently, a sparse orthogonal decomposition algorithm preserving sparsity in $\boldsymbol{U}$ and $\boldsymbol{V}$ was developed to estimate $\boldsymbol{U}$ and $\boldsymbol{V}$ iteratively.

It can be seen that the cross-data-type prediction will open another new field for statistical methodology and theoretical research, given that both the predictors and the responses can not only be ultrahigh dimensional but also consist of multiple data types.

Conclusions

More and more efforts have been devoted to the development of statistical models and methods for integrative cancer data. Nevertheless, research on integrative analyses for cancer is still in its infancy with many open problems. How can systematic biases such as batch effects be detected and corrected in each new type of high-throughput technology so that meta-analysis across studies can be conducted? How can cancer subtypes be classified according to multiple genomic profiles jointly or determined by only a subset of genomic profiles? How can a single network be constructed with multidimensional genomic profiles? How can networks constructed from different types of data be modeled jointly? How can survival time of cancer patients be predicted by multiple types of ultrahigh-dimensional genomic profiles? How can one ultrahigh-dimensional vector be predicted from another ultrahigh-dimensional vector, one maybe continuous while the other discrete? All these questions are of vital clinical importance for identifying risk factors, drug targets, cancer diagnosis, survival prediction, and therapy selection toward a personalized approach. Naturally, they urge the demand for developing valid statistical methods with outstanding practical performance as well as solid theoretical foundations. I anticipate a wealth of new computationally efficient, interpretable, and robust statistical methods for integrative cancer analyses in the near future, which will thereby significantly promote cancer research and therapeutic development.

Acknowledgments

I wish to acknowledge the Guest Editor Dr. Hao Wu for help in preparing this manuscript.

Author Contributions

Conceived and designed the experiments: YW. Analyzed the data: YW. Wrote the first draft of the manuscript: YW. Contributed to the writing of the manuscript: YW. Agree with manuscript results and conclusions: YW. Jointly developed the structure and arguments for the paper: YW. Made critical revisions and approved final version: YW. The author reviewed and approved of the final manuscript.

REFERENCES

1. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30:207–10.
2. Parkinson H, Kapushesky M, Shojatalab M, et al. Arrayexpressa public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*. 2007; 35:D747–50.
3. Consortium EP. The encode (encyclopedia of dna elements) project. *Science*. 2004;306:636–40.
4. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res*. 2010;39:D19–21.
5. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. 2009;458: 719–24.
6. International Cancer Genome Consortium, Hudson TJ, Anderson W, et al. International network of cancer genome projects. *Nature*. 2010;464:993–8.
7. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer*. 2014;14:299–313.
8. Zhu Y, Qiu P, Ji Y. Tcga-assembler: open-source software for retrieving and processing tcga data. *Nat Methods*. 2014;11:599–600.
9. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11: 733–9.
10. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*. 2007;8:118–27.
11. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. 2007;3:e161.
12. Leek JT, Storey JD. A general framework for multiple testing dependence. *Proc Natl Acad Sci U S A*. 2008;105:18718–23.
13. Leek JT. Svaseq removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res*. 2014;42(21):1–9.
14. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of rna-seq data using factor analysis of control genes or samples. *Nat Biotechnol*. 2014;32:896–902.
15. Conlon EM, Song JJ, Liu JS. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*. 2006;7:247.
16. Kendziorski C, Newton M, Lan H, Gould M. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med*. 2003;22:3899–914.
17. Yuan M, Kendziorski C. A unified approach for simultaneous gene clustering and differential expression identification. *Biometrics*. 2006;62:1089–98.
18. Ruan L, Yuan M. An empirical bayes' approach to joint analysis of multiple microarray gene expression studies. *Biometrics*. 2011;67:1617–26.
19. Jensen ST, Erkan I, Arnardottir ES, Small DS. Bayesian testing of many hypotheses x many genes: a study of sleep apnea. *Ann Appl Stat*. 2009;3:1080–101.
20. Scharpf RB, Tjelmeland H, Parmigiani G, Nobel AB. A bayesian model for cross-study differential gene expression. *J Am Stat Assoc*. 2009;104:1295–310.
21. Wei Y, Tenzen T, Ji H. Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics*. 2015;16:31–46.
22. Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics*. 2011;12:763–75.
23. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11:R106.
24. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007;23:2881–7.
25. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012;40(10):4288–97.
26. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in rna sequencing data using observation weights. *Nucleic Acids Res*. 2014;42:e91.
27. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55:119–39.
28. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann Stat*. 2000;28:337–407.
29. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32.
30. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. 1992. In: Proceedings of the fifth annual workshop on Computational learning theory, ACM: 144–52.
31. De Bin R, Sauerbrei W, Boulesteix A-L. Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Stat Med*. 2014;33:5310–29.
32. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics*. 2012;99:323–9.
33. Shen R, Wang S, Mo Q. Sparse integrative clustering of multiple omics data sets. *Ann Appl Stat*. 2013;7:269.
34. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. 1996;58:267–88.
35. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *J R Stat Soc Series B Stat Methodol*. 2005;67:91–108.
36. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67:301–20.
37. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *J R Stat Soc Series B Stat Methodol*. 1977;39:1–38.
38. Ray P, Zheng L, Lucas J, Carin L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*. 2014;30:1370–6.
39. Ghahramani Z, Griffiths TL. Conference on NIPS. Infinite latent feature models and the indian buffet process. 2005. In: Advances in Neural Information Processing Systems: 475–82.
40. Thibaux R, Jordan MI. Conference on AI Statistics. Hierarchical beta processes and the indian buffet process. 2007. In: International Conference on Artificial Intelligence and Statistics: 564–71.
41. Geman S, Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell*. 1984;6(6):721–41.
42. Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann Appl Stat*. 2013;7:523–42.

43. Kim PM, Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res*. 2003;13:1706–18.
44. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*. 2004;101:4164–69.
45. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multidimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*. 2012;40:9379–91.
46. Jolliffe I. *Principal Component Analysis*. England: Wiley Online Library; 2005.
47. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods*. 2013;10:1108–15.
48. Mo Q, Wang S, Seshan VE, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci U S A*. 2013;110:4245–50.
49. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1.
50. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *J Am Stat Assoc*. 1997;92:162–70.
51. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*. 2012;28:3290–7.
52. Savage R, Ghahramani Z, Griffin J, Kirk P, Wild D. Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. 2013. In: *International Conference on Machine Learning (ICML)* 2012, *Workshop on Machine Learning in Genetics and Genomics*.
53. Cho D-Y, Przytycka TM. Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model. *Nucleic Acids Res*. 2013;41(17):8011–20.
54. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics*. 2013; 29(20):2610–6.
55. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11:333–7.
56. Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. *Adv Neural Inf Process Syst*. 2002;2:849–56.
57. Katenka N, Kolaczyk ED. Inference and characterization of multi-attribute networks with application to computational biology. *Ann Appl Stat*. 2012; 6:1068–94.
58. Kolar M, Liu H, Xing EP. Graph estimation from multi-attribute data. *J Mach Learn Res*. 2014;15:1713–50.
59. Frank O, Strauss D. Markov graphs. *J Am Stat Assoc*. 1986;81:832–42.
60. Kolar M, Song L, Ahmed A, Xing EP. Estimating time-varying networks. *Ann Appl Stat*. 2010;4:94–123.
61. Xing EP, Fu W, Song L. A state-space mixed membership blockmodel for dynamic network tomography. *Ann Appl Stat*. 2010;4:535–66.
62. Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. *Biometrika*. 2011;98(1):1–15.
63. Peterson C, Stingo F, Vannucci M. Bayesian inference of multiple gaussian graphical models. *J Am Stat Assoc*. 2014. [In press].
64. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. *J R Stat Soc Series B Stat Methodol*. 2014;76:373–97.
65. Zhu Y, Shen X, Pan W. Structural pursuit over multiple undirected graphs. *J Am Stat Assoc*. 2014;109(508):1683–96.
66. Mohan K, London P, Fazel M, Witten D, Lee S-L. Node-based learning of multiple gaussian graphical models. *J Mach Learn Res*. 2014;15:445–88.
67. David CR. Regression models and life tables (with discussion). *J R Stat Soc Series B Stat Methodol*. 1972;34:187–220.
68. Cox DR. Partial likelihood. *Biometrika*. 1975;62:269–76.
69. Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med*. 1997;16:385–95.
70. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96:1348–60.
71. Fan J, Li R. Variable selection for cox's proportional hazards model and frailty model. *Ann Stat*. 2002;30:74–99.
72. Witten DM, Tibshirani R. Survival analysis with high-dimensional covariates. *Stat Methods Med Res*. 2010;19(1):29–51.
73. Wang S, Nan B, Zhu N, Zhu J. Hierarchically penalized cox regression with grouped variables. *Biometrika*. 2009;96:307–22.
74. Zhang W, Ota T, Shridhar V, Chien J, Wu B, Kuang R. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput Biol*. 2013;9:e1002975.
75. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Series B Stat Methodol*. 2008;70:849–911.
76. Fan J, Feng Y, Wu Y. High-dimensional variable selection for coxs proportional hazards model. James O. Berger, Tony T. Cai and Iain M. Johnstone, editors. *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*. Hong Kong: Institute of Mathematical Statistics; 2010:70–86.
77. Zhao SD, Li Y. Principled sure independence screening for cox models with ultra-high-dimensional covariates. *J Multivar Anal*. 2012;105:397–411.
78. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995;57:289–300.
79. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29:1165–88.
80. Gorst-Rasmussen A, Scheike T. Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J R Stat Soc Series B Stat Methodol*. 2013;75:217–45.
81. Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. James O. Berger, Tony T. Cai and Iain M. Johnstone, editors. *Computational Learning Theory*. New York: Springer; 1995:23–37.
82. Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. Bagging survival trees. *Stat Med*. 2004;23:77–91.
83. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7: 1–26.
84. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457–81.
85. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. *Biostatistics*. 2006;7:355–73.
86. Laan MJ, Robins JM. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer; 2003.
87. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29:1189–232.
88. Zhu R, Kosorok MR. Recursively imputed survival trees. *J Am Stat Assoc*. 2012;107:331–40.
89. Biau G, Devroye L, Lugosi G. Consistency of random forests and other averaging classifiers. *J Mach Learn Res*. 2008;9:2015–33.
90. Scornet E, Biau G, Vert J-P. *Consistency of Random Forests*. Ithaca: arXiv preprint; 2014. [arXiv:1405.2881].
91. Scornet E. *On the Asymptotics of Random Forests*. Ithaca: arXiv preprint; 2014. [arXiv:1409.2090].
92. Ma X, Xiao L, Wong WH. Learning regulatory programs by threshold svd regression. *Proc Natl Acad Sci U S A*. 2014;111:15675–80.