Chapter 7

Integrative Analysis of Omics Big Data

Xiang-Tian Yu and Tao Zeng

Abstract

The diversity and huge omics data take biology and biomedicine research and application into a big data era, just like that popular in human society a decade ago. They are opening a new challenge from horizontal data ensemble (e.g., the similar types of data collected from different labs or companies) to vertical data ensemble (e.g., the different types of data collected for a group of person with match information), which requires the integrative analysis in biology and biomedicine and also asks for emergent development of data integration to address the great changes from previous population-guided to newly individual-guided investigations.

Data integration is an effective concept to solve the complex problem or understand the complicate system. Several benchmark studies have revealed the heterogeneity and trade-off that existed in the analysis of omics data. Integrative analysis can combine and investigate many datasets in a cost-effective reproducible way. Current integration approaches on biological data have two modes: one is "bottom-up integration" mode with follow-up manual integration, and the other one is "top-down integration" mode with follow-up in silico integration.

This paper will firstly summarize the combinatory analysis approaches to give candidate protocol on biological experiment design for effectively integrative study on genomics and then survey the data fusion approaches to give helpful instruction on computational model development for biological significance detection, which have also provided newly data resources and analysis tools to support the precision medicine dependent on the big biomedical data. Finally, the problems and future directions are highlighted for integrative analysis of omics big data.

Key words Integration, Omics, High throughput, Big data, Complex diseases, Bayesian, Matrix decomposition, Machine learning, Subtype, Precision medicine

1 Introduction

High-throughput screening is one of the primary technologies for exploring complex intracellular dynamics in modern biology, and the data produced by such approaches are usually called as omics data [1]. The intuitive omics on genome appeared from the Human Genome Project for obtaining the blueprint of complete human genetic information; after which, the transcriptome and proteome are also becoming available to measure the expression abundance of mRNA and protein, respectively [2]. Lately, the epigenomics was

Tao Huang (ed.), *Computational Systems Biology: Methods and Protocols*, Methods in Molecular Biology, vol. 1754, https://doi.org/10.1007/978-1-4939-7717-8_7,

developed to investigate the previously thought "dark matter" on genome (e.g., the potential regulatory elements located at noncoding sequences) [3, 4]. Along with the deep understanding of genotype-phenotype association, the metabolites have been widely applied to bridge the genome and phenome due to their outcome role of regulation [5], so that the metabonomics is increased to available for more accurate phenotype indication [6]. Meanwhile, the interactions or associations among different molecules are also confirmed and gathered in databases, which provide the metadata on molecule networks, so called as interactome [7, 8]. These diversity and huge omics data take biology and biomedicine research and application into a big data era (*see* **Note 1**), just like that popular in human society a decade ago [9]. They are opening a new challenge from horizontal data ensemble (e.g., the similar types of data collected from different labs or companies) to vertical data ensemble (e.g., the different types of data collected for a group of person with match information), which provide distinct but often complementary information [10] and are also helpful to address the great changes from previous population-guided to newly individual-guided investigations [11].

Integration is an effective concept to solve the complex problem or understand the complicate system [12]. In computational viewpoints, the data integration can make full use of complementary information [13], carry on necessary noise deduction [14], supply abstract of hidden factor [15], realize bias correction in analysis [16], and introduce common and diversity of data pattern [17]. Meanwhile in biological fields, the data integration is a multi-view investigation on the completeness and complexity of the biological system. Especially in the high-throughput cancer genomic studies, results from the analysis of single datasets often suffer from a lack of reproducibility because of small sample sizes, and the benchmark studies have revealed the heterogeneity and trade-off existed in the analysis of omics data [18, 19]. To address these problems, integrative analysis can effectively combine and investigate many datasets in a cost-effective way to improve reproducibility.

Briefly, current integrative analysis methods on biological data (e.g., omics data discussed in this paper) have two modes: one is "bottom-up integration" (i.e., data combination with follow-up manual integration), and the other one is "top-down integration" (i.e., data fusion with follow-up in silico integration). In the "bottom-up integration," the combination of large amounts of public data may allow us to examine general dynamical relationships during gene regulations [20] [21], e.g., combining different types of data provides a more comprehensive model of the cancer cell than that offered by any single type [22]. These combinatory analyses are expected to integrate the diverse data to reconstruct biologically meaningful networks and potentially provide a more reliable insight

into the underlying biological mechanisms [23]. By contrast, in the "top-down integration," the general integration idea is based on information fusion, where different data types can offer complementary perspectives on the same biological phenomenon. The integrative approaches would be more powerful when they can incorporate all data types simultaneously and generate a single integrated sample-cluster assignment, such as the statistic-based methods [10, 11, 24–26], the machine-learning-based methods [22, 27, 28], and the matrix-based methods [21, 29, 30]. Especially, the tensor structure is a basic feature of the multi-view data [31] to uncover shared signals across different high-dimensional data, and it is valuable to develop a model that applies a matrix decomposition to the gene expression matrix for each data type but with a linked individual (e.g., a set of latent components) [32].

Some review on integration study has shown the application potential of integrative analysis on high-dimensional genomic data [13, 33–37]. By contrast, this paper will firstly summarize the combinatory approaches to give candidate protocol on biological experiment design for effectively integrative study on genomics and next survey the data fusion approaches to give helpful instruction on computational model development for meaningful biological significance detection, which also provide new data resources and analysis tools to support the precision medicine dependent on big biomedical data. Below, we will introduce the data resources for integrative analysis, the batch effect removal in integration, the two integration modes, and the tool and visualization of integration analysis, respectively. Finally, we supply a few highlight notes on the problems and future directions for integrative analysis of omics big data.

2 Materials

Being the solid foundation of integrative biological analysis, the data sources, especially the online public data depositions, have provided enormous wealth of data and resources. According to the biological background of these data in databases, the widely accessible data can be summarized as several categories as shown in Table 1.

The genome sequencing technologies open the door to the high-throughput data in biology; thus, the human genome and other species' genomes have been sequenced and published with each passing year. The 1000 Genomes Project [38] has contributed great data on human genomics, and it is designed to supply the largest public human variation and genotype data. The ENCODE (Encyclopedia of DNA Elements) Consortium [39] is built to offer a comprehensive understanding on the functional elements in the human genome, which act/regulate at the DNA, RNA, or protein

Table 1
The category of data sources

Category	Database	URL
Genomics-focused	1000 Genomes [38]	http://www.1000genomes.org/
	Encode [39]	https://www.encodeproject.org/
	3CDB [42]	http://3cdb.big.ac.cn/
	4DGenome [43]	https://4dgenome.research.chop.edu/
Transcriptome-focused	NCBI GEO [44]	https://www.ncbi.nlm.nih.gov/geo/
	TCGA [45]	https://cancergenome.nih.gov/
	ICGC [46]	http://icgc.org/
Epigenomics-focused	miRBase [47]	http://www.mirbase.org/
	lncRNAdb [48]	http://www.lncrnadb.org/
	NGSmethDB [49]	http://bioinfo2.ugr.es:8080/NGSmethDB/
	MethylomeDB [50]	http://www.neuroepigenomics.org/methylomedb/
Metagenomics-focused	HMDB [51]	http://www.hmdb.ca/
	EBI metagenomics [52]	https://www.ebi.ac.uk/metagenomics/
Interactome-focused	BioGRID [58]	https://thebiogrid.org/
	STRING [59]	http://www.string-db.org/
	KEGG [60]	http://www.kegg.jp/
	Reactome [61, 62]	http://www.reactome.org/

levels when and where a gene is active. Beyond such coding information of biological sequences, the high-order structure of those sequences have also been resolved recently based on the development of Hi-C or similar technologies [40, 41]. A database of manually curated 3C data (3CDB) [42] is implemented to extract and store the contact frequencies between selected genomic sites in a cell population by literature review and manually extraction. Similarly, the 4DGenome [43] database stores chromatin interaction data compiled by literature curation or computational prediction, which would be efficient on investigating the spatial structure-and-function relationship of genomes.

After the sequence clarity, the detection and estimation of transcriptome have been widely studied based on the microarray or deep-sequencing technologies. NCBI GEO [44] is a well-known database to access the transcriptome data from many different biological experiments, focusing on different spices, different tissues, different cells, or different stresses. Particularly on the study of human cancer, the TCGA [45] and ICGC [46] have generated comprehensive, multidimensional maps of the key genomic changes in more than 30 types of cancer, which are public for assisting the cancer research community to improve the prevention, diagnosis, and treatment of cancer.

Recently, the conventional noncoding information or "dark matter" on genome has also been attractive and inspiring to recover many unknown regulatory factors. One is the miRNA, and the miRBase [47] database publishes predicted hairpin portion of a miRNA transcript, with information on the location and sequence of the mature miRNA sequence. Second is the lncRNA, and the lncRNAdb [48] is a manually curated reference database dependent on capturing a great proportion of the literature describing functions for individual eukaryotic lncRNAs. Third is the methylation, and the NGSmethDB [49] is a repository with single-base whole-genome methylome maps on the best-assembled eukaryotic genomes and the reliable and high-quality methylomes; meanwhile, the MethylomeDB [50] is an expert database containing genome-wide brain DNA methylation profiles of human and mouse brain specimens generated from in-house and collected from third-party publication.

Lately, along with the development of central dogma, the metabolism as the outcomes of regulation can reflect more phenotype-associated genetic information. For example, the Human Metabolome Database (HMDB) [51] is a free database gathering human-source small molecule metabolites, which contains or links chemical data, clinical data, and molecular biology/biochemistry data, and can be applied in biomarker discovery. Similarly, EBI metagenomics [52] is a freely available center for the storage and analysis of WGS sequenced meta-genomic/meta-transcriptomic data and also provides a standardized analysis workflow to produce rich taxonomic diversity and functional annotations with great consistence on different types of data.

In addition, from the systematical viewpoint on all biological elements, their associations or interactions can be summarized and abstracted as a network form, which inspire the network biology [53–57], and the integrative resources of such biological network knowledge can be obtained from several public databases, such as:

The Biological General Repository for Interaction Datasets (BioGRID) [58] which is an open access database dedicated to the annotation and archival of protein, genetic, and chemical interactions for all major model organism species and humans, by reviewing the biomedical literature for major model organism species.

The STRING database [59] which tries to provide a critical assessment and integration of protein-protein interactions, including direct (physical) as well as indirect (functional) associations, especially the inferred protein-protein associations from co-expression data.

The KEGG [60] which is an encyclopedia of genes and genomes, designed to assign functional meanings to genes and genomes both at the molecular and network level in the form of molecular interactions, reactions, and relations.

Table 2
The category of data structure

Data structure	Experimental protocol	Cases with visualization
Vector	Nucleic acid or amino acid Modification site	The UCSC Genome Browser database [63] MEXPRESS visualizing TCGA [64]
Matrix	Gene-sample Gene-time	Co-expression of gene profiles [65] AIE for cell cycle pattern [66]
Tensor	Gene-sample-source Gene-sample-time	Pan-cancer analysis on TCGA [68] Edge network modeling virus infection [113]
High-order cube	Gene-sample-source-time	Cross-tissue and cross-species transcriptome analysis [70]

The Reactome [61, 62] which plays both as an archive of biological processes by modeling signal transduction, transport, DNA replication, metabolism, and other cellular processes in an ordered network of molecular transformations and as a bioinformatics tool to discover unexpected functional relationships in biological data.

On the other hand, the data from above resources can have different data structures as listed in Table 2, which will determine the direction of follow-up integrative analysis. In mathematical terms, the data structures of such high-throughput data can usually take as a vector, a matrix, a tensor, and their combinations (Fig. 1). Simply, any sequence data (e.g., DNA sequencing) can transform to a (sequenced) vector; each element in a vector represents a nucleic acid or an amino acid or a modification site on the particular location of one sequence, e.g., the string consisted of (A,C,G,T) from 5′ to 3′ on DNA sequence [63], or the barcode-like signal of methylation level on CpG islands along the DNA sequence [64]. Meantime, the expression data of genes from a large cohort study can be organized as a matrix, where a row indicates a gene and a column indicates a sample, so that each element in a matrix represents one gene's expression level in one sample, e.g., the expression of genes in a group of individuals with the same disease [65] or the gene expression of cell cycle at consecutive time points [66]. Next, the triple-way biological experiment can produce data viewed in a cubic form and always be formalized as tensor, and there are two general types of such data [67]: one is "gene-sample-source", which collects the expression data from multiple samples under several biological conditions, e.g., an element in such tensor can point the expression level of one gene from one tissue of the same sample [68]; and the other one is "gene-sample-time", which gathers the expression data from a sample at a particular time point,

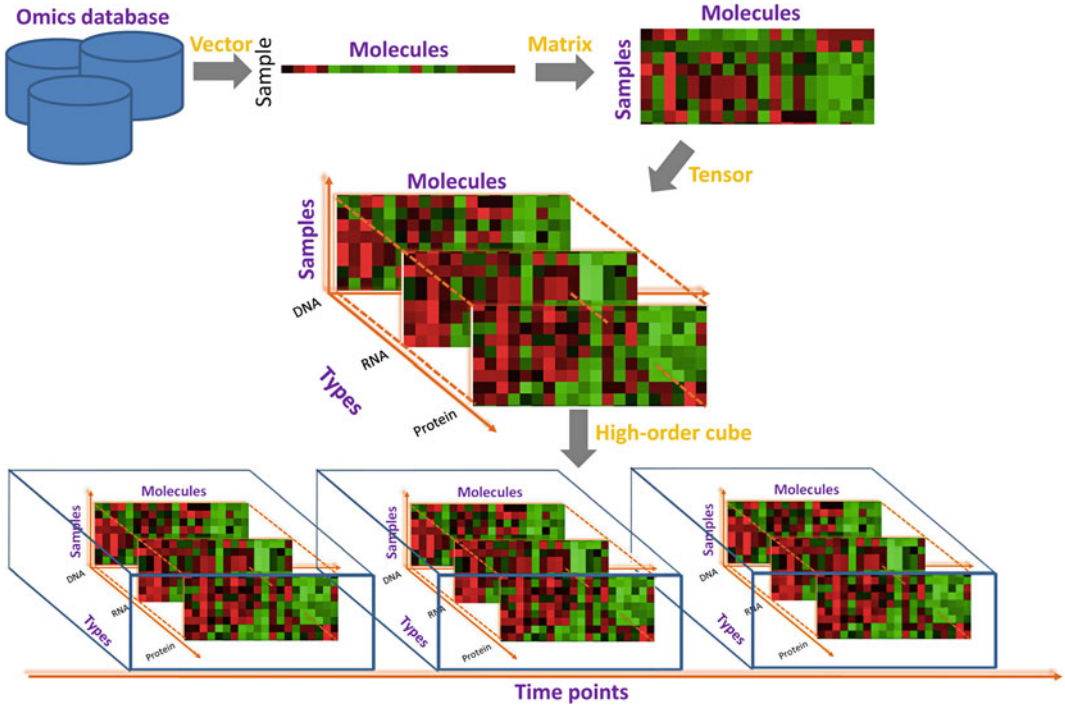

Fig. 1 The organization of data structures in omics big data

e.g., an element in such tensor can reflect the expression value of one gene of the same individual at an early or latter time point during virus infection [69]. Furthermore, nowadays, known as the era of big data, more delicate biological experiments can be carried on, and more complex data structure would be faced, e.g., the combination of tensor as "gene-sample-source-time", whose representative case is the cross-tissue gene expression analysis on evolutionary [70].

3 Methods

3.1 Batch Effect Removal Before Integration

The removal of batch effect should be an important step ahead of many integrative analyses on biological big data. Many variables will play in any given research, such as the influence of age or sex on the diseases. Especially there are many sources of variation when the expressions of thousands of genes are measured at once, so that the batch effects become more critical due to the complexity of genomes inside and environments outside [71]. In practice, the sequencing and microarray samples are usually collected or processed in multiple batches (e.g., at different times), which are easy to produce technical biases and incorrect results in the downstream analysis [72]. For example, to estimate expression correlation over

thousands of samples is possible nowadays because large amounts of expression data can be publicly available; however, extracting information from the correlation data is not straightforward due to the expression data generated by different laboratories from different cell types under different biological conditions [73]. To address those issues from batch effect, many computational approaches have been proposed. The "surrogate variable analysis" (SVA) is introduced to recover the effects of the important missed variables and essentially produce an analysis as if all relevant variables were included, which has shown the improved biological accuracy and reproducibility [71]. Meanwhile, the ComBat removes batch effects based on an empirical Bayes framework, which centers data to the overall grand mean of all samples and obtains an adjusted data without coinciding with the location of any original batches [74]. And a modified version of ComBat (M-ComBat) adopts to shift samples to the mean and variance of a "gold standard" as reference batch rather than the grand mean and pooled variance [75]. Next, an extension of PCA known as guided PCA (gPCA) has been proposed to quantify the existence of batch effects, and a new statistic is also designed to apply gPCA to test whether a batch effect exists in high-throughput data [76]. Further, a software pipeline, BatchQC, is implemented to use interactive visualizations and statistics to evaluate the impact of batch effects in a genomic dataset, which can also apply existing adjustment tools and allow to evaluate researchers' benefits interactively [72].

As an initiative integrative application related to batch effect removal, the conventional horizontal data ensemble needs to integrate the same type of data from different studies. For example, an integrative pre-screening is provided to reduce the dimensionality in cancer genomic studies for the analysis of multiple cancer genomic datasets, which can be coupled with existing analysis methods to identify cancer markers [77]. And by analyzing the accrued gene expression data in TCGA pan-cancer (PANCAN) data, the paired normal samples seem to be in general more informative on patient survival than tumors, whose analysis supports the importance of collecting and profiling matched normal tissues to gain more insights on disease etiology and patient progression [78].

3.2 Bottom-Up Integration

According to the combination of different types of high-throughput data, the "bottom-up integration" approaches have many particular analysis frameworks as summarized in Table 3. Generally, the mutation and transcriptome information are both considered, especially the mRNA expression is used in almost any analysis (*see* **Note 2**). Below, considering the usage of mutation or not, the integrative methods are introduced and discussed, respectively.

On one hand, the mutation-centered integration mainly tries to identify the genetic determinants of phenotype and its change,

Table 3
The representative approaches of bottom-up integration

Method	Level of omics								Biological purpose
	Mutation	mRNA	miRNA	Modification	Protein	Metabolite	Network	Annotation	
"eQTL-based" [80]	✓	✓							eQTL-based analysis
TieDIE [92]	✓	✓			✓				Robust synthetize of signaling network
"Network-based" [79]	✓						✓	✓	Identification of disrupted pathways
"Integrative network analysis" [114]	✓	✓		✓					Identifying important genetic and epigenetic features
"TCGA-based" [82–85]	✓	✓	✓	✓					Characterizing somatic alterations
"TCGA-based" [86–88]	✓	✓		✓					Characterizing the genomic/epigenomic landscape
"TCGA-based" [90, 91]	✓	✓	✓	✓	✓				Cancer subtypes caused by different subsets of genetic and epigenetic abnormalities
"Generalizable framework" [81]	✓	✓							Identifying pathogenetically relevant mutated genes
"Integrative framework" [89]	✓	✓			✓				Determining the prognostic, predictive, and therapeutic relevance of the functional proteome
"Pan-cancer initiative" [93]	✓	✓	✓	✓	✓				The Cancer Genome Atlas pan-cancer analysis project
dChip-GemiNI [98]		✓	✓				✓		Detecting feed-forward loops (FFLs) on TF-miRNA-mRNA network

(continued)

Table 3
(continued)

Method	Level of omics								Biological purpose
	Mutation	mRNA	miRNA	Modification	Protein	Metabolite	Network	Annotation	
"Semi-supervised normalization pipelines" [100]		✓			✓	✓	✓	✓	Training predictive cell models based on integrated data sources
"Integrative computational pipeline" [97]		✓	✓	✓					Dissecting the transcription factors (TFs) responsible for altered miRNA expression
"Data-driven discovery" [95]		✓						✓	Data-driven discovery of pain gene candidates
"Layers of regulation" [99]		✓			✓				Adaptive mechanisms include posttranscriptional and posttranslational events
pRSEM [96]		✓		✓					Estimating relative isoform abundances
SPIA [94]		✓						✓	Pan-cancer analysis on pathways

and other omics data can be assisted to recognize the sensitive mutation by removing the passenger mutations.

1. *The direct mapping of mutation information on pathway/network knowledge*. A network-based method has been used to integrate copy number alteration data with human protein-protein interaction networks and pathway databases to identify pathways that are commonly disrupted in many different types of cancer, which are likely essential for tumor formation in the majority of the cancers [79].
2. *The combination of mutation and transcriptome*. As a typical quantitative association approach, the eQTL (expression quantitative trait locus)-based analyses have been proposed to investigate the germ line determinants of gene expression in tumors by using the multilevel information from The Cancer Genome Atlas (TCGA) [80]. And in an investigation of the aggressive lung tumor subtype with poor prognosis, an integrated analyses have been conducted to identify pathogenetically relevant mutated genes by a generalizable framework for the identification of biologically relevant genes in the context of high mutational background [81].
3. *The further consideration of epigenetic influence*. A genome-scale analysis of 276 samples has been analyzed to characterize the somatic alterations in colorectal carcinoma, including exome sequence, DNA copy number, promoter methylation, and messenger RNA and microRNA expression [82–85]. Similarly, 178 lung squamous cell carcinomas have been deeply profiled to provide a comprehensive landscape of genomic and epigenomic alterations in squamous cell lung cancers and develop molecularly targeted agents for target treatment [86–88].
4. *The additional integration with protein expression*. The direct study of the functional proteome has the potential to provide a wealth of information that complements and extends genomic, epigenomic, and transcriptomic analysis. The resultant proteomic data in TCGA can be integrated with genomic and transcriptomic analyses of the same samples to identify commonalities, differences, emergent pathways, and network biology within and across tumor lineages [89]. By integrating information across platforms including reverse phase protein arrays, a hypothesis is held that much of the clinically observable plasticity and heterogeneity occurs within, and not across, the major biological subtypes of breast cancer [90, 91]. Besides, the Tied Diffusion through Interacting Events (TieDIE) is developed to integrate differentially expressed master transcriptional regulators, functionally mutated genes, and differentially activated kinases to synthesize a robust signaling

network consisting of druggable kinase pathways, which will be helpful for drug prioritization in individual patients [92].

5. *To provide basic data support.* The Cancer Genome Atlas (TCGA) Research Network has profiled and analyzed large numbers of human tumors to discover molecular aberrations at the DNA, RNA, protein, and epigenetic levels, whose resulting rich data provide a major opportunity to develop an integrated picture of commonalities, differences, and emergent themes across tumor lineages [45]. Dependent on TCGA, the pan-cancer initiative compares multiple tumor types, and the molecular aberrations and their functional roles across tumor types will enlighten how to extend therapies effective in one cancer type to others with a similar genomic profile [93].

   On the other hand, the transcriptome-centered integration mainly tries to identify the phenotype-associated genes by the complementary information from other omics data.

6. *The functional enrichment based on the expression abundance and its differential changes.* By a software package signaling pathway impact analysis (SPIA), all signaling pathways in the KEGG PATHWAY database have been widely investigated and obtained several notable findings concerning many pathways to be new discoveries, which imply many opportunities for laboratory and clinical follow-up studies [94]. Specially, a novel integrative paradigm has been applied for data-driven discovery of pain gene candidates, taking advantage of the vast amount of existing disease-related clinical literature and gene expression microarray data, which enables efficient biological studies validating additional candidates [95].

7. *The functional complementation between transcriptome and epigenome.* To improve the quantification accuracy of isoforms, a computational method as prior-enhanced RSEM (pRSEM) is proposed to use a complementary data type in addition to RNA-seq data, which shown to be superior than competing methods in estimating relative isoform abundances within or across conditions in qRT-PCR validations [96]. Another case is that an integrative computational pipeline has identified TFs with binding sites significantly overrepresented among miRNA genes overexpressed in ovarian carcinoma, and it can be applied to discover transcriptional regulatory mechanisms in other biological settings where analogous genomic data are available [97]. Besides, the dChip-GemiNI (Gene and miRNA Network-based Integration) method can statistically rank computationally predicted FFLs by accounting for differential gene and miRNA expression between two biological conditions such as normal and cancer and also derive potential TF-target genes and miRNA-mRNA interactions [98].

8. *The functional influence of protein on gene regulatory.* Stress responses were believed to be predominantly regulated at the transcriptional level; however, the adaptive mechanisms should include post-transcriptional and post-translational events. To address this issue, three layers of regulation have been integrated as transcriptome, translatome, and proteome, which is useful to gain a deeper understanding of how sophisticated regulation networks operate [99]. And semi-supervised normalization pipelines have been designed and performed experimental characterization to create a quality-controlled multi-omics compendium for *E. coli*, and a multi-scale model has further been trained by integrating four omics layers to predict genome-wide concentrations and growth dynamics [100].

3.3 Top-Down Integration

The standard "bottom-up integration" approach as above integrative clustering is usually to separate clustering followed by manual integration. By contrast, a more computational powerful approach would incorporate all data types simultaneously and generate a single integrated cluster assignment (*see* **Note 3**), which are thought as "top-down integration" as shown in Table 4.

1. *Statistic-based integration model.* One key integrative idea is unifying hidden factor from different types of data. A joint latent variable model as iCluster is developed for integrative clustering by incorporating flexible modeling of the associations between different data types and the variance-covariance structure within data types while simultaneously reducing the dimensionality of the datasets [24]. To extend the scope of integrative analysis for the inclusion of somatic mutation data, an expanded framework iCluster+ is further proposed to ensemble discrete and continuous variables that arise from integrated genomic, epigenomic, and transcriptomic profiling [11]. Similarly, a novel algorithm termed moCluster employs a multiblock multivariate analysis to define a set of latent variables representing joint patterns across input datasets, which is passed to an ordinary clustering algorithm in order to discover joint clusters [101]. The other important integrative idea is unifying data distribution under the theoretical framework around Bayesian principles. An integrative Bayesian analysis of genomics data (iBAG) framework is proposed to identify important genes/biomarkers by using hierarchical modeling to combine the data obtained from multiple platforms into one model [25]. And a Bayesian method referred as MDI (Multiple Dataset Integration) has been presented for the unsupervised integrative modeling, where each dataset is modeled using a Dirichlet-multinomial allocation (DMA) mixture model, with dependencies between these models captured

Table 4
The representative approaches of top-down integration

Categories	Methods	Computational instructions
Statistic (factor-centered)	"Residuals" [115]	A two-stage approach based on regularized singular value decomposition, and regularized estimation of prediction model
	iCluster [24]	A joint latent variable model incorporating the variance-covariance structures
	iCluster+ [11]	Joint modeling is proposed to ensemble discrete and continuous variables
	moCluster [101]	Multiblock multivariate analysis and an ordinary clustering algorithm
	iBAG [25]	Hierarchical modeling within Bayesian analysis
	MDI [10]	Dirichlet-multinomial allocation (DMA) mixture model within Bayesian analysis
	"Nonparametric Bayesian model" [26]	A hierarchy of Dirichlet processes within a nonparametric Bayesian model
	"Factor analysis" [116]	Factor analysis
Optimization (matrix-centered)	"Joint matrix factorization" [21]	Joint nonnegative matrix factorization
	"Multi-view bi-clustering" [30]	Rank matrix factorization
	GSVD [104] [105]	Higher-order generalized singular value decomposition
	"Ping-pong" [29]	Ping-pong algorithm
Machine learning (pattern-centered)	"Linear discriminant analysis" [22]	Factor analysis, combined with linear discriminant analysis
	"Kernel-based" [27]	Multiple kernel learning
	JointCluster [28]	Simultaneous clustering of multiple networks
	SNF [102]	Similarity network fusion based on theoretical multi-view learning framework
	PFA [103]	Pattern fusion analysis based on local tangent space alignment (LTSA) theory

through parameters that describe the agreement among the datasets [10]. Meanwhile, a nonparametric Bayesian model has been introduced to discover prognostic cancer subtypes by constructing a hierarchy of Dirichlet processes and has shown a good ability to distinguish concordant and discordant signals within each patient sample [26].

2. *Machine-learning-based integration model.* The main idea under such methods is to extract significant data pattern along with integrative analysis. An extended multiple kernel learning has been applied for dimensionality reduction approaches, and several kernels per data type are applicable to avoid the unnecessary choice of the best kernel functions and

kernel parameters for each data type beforehand [27]. And in a biological application, the high-throughput screens for mRNA, miRNA, and proteins have been jointly analyzed using factor analysis, combined with linear discriminant analysis (LDA), to identify the molecular characteristics of cancer [22]. Especially when focused on characterizing biological network, an algorithm JointCluster is implemented to find sets of genes that cluster well in multiple networks of interest, such as co-expression networks summarizing correlations among the expression profiles of genes and physical networks describing protein-protein and protein-DNA interactions among genes or gene products [28]. To produce a comprehensive view of a given disease by diverse types of genome-wide data, similarity network fusion (SNF) has been inspired from the theoretical multi-view learning framework to construct the networks of samples (e.g., patients) for each data type and fuse them into one network, which can represent the sample patterns underlying data [102]. Recently, a new framework called "pattern fusion analysis" (PFA) has been proposed to perform automated information alignment and bias correction and to fuse local sample patterns (e.g., from each data type) into a global sample pattern corresponding to phenotypes (e.g., across most data types). Particular, PFA can identify common and complementary sample patterns from different omics profiles by optimally adjusting the effects of each data type based on the local tangent space alignment (LTSA) theory [103].

3. *Matrix-based integration model.* Previously, the integrative scheme of ping-pong algorithm was proposed to integrate more than one type of data from the same biological samples, which is dependent on the usage of co-modules describing coherent patterns across paired datasets [29]. Actually, these methods can be included into several classes according to the type of applied matrix decomposition: one is a joint (nonnegative) matrix factorization technique that projects multiple types of genomic data onto a common coordinate system, in which heterogeneous variables weighted highly in the same projected direction form a multidimensional module (md-module) [21]; two is higher-order generalized singular value decomposition (GSVD), which is designed for efficient, parameter-free and reproducible identification of network modules simultaneously across multiple conditions [104, 105]; and three is rank matrix factorization as multi-view bi-clustering to model subtyping and recognize subtype-specific features simultaneously, e.g., integrate mutational and expression data while taking into account the clonal properties of carcinogenesis [30].

3.4 Tool and Visualization of Integration

Currently, the academic studies not only develop the biological or computational techniques for integrative analysis but also provide many software tools and visualization resources for iteratively review by biologist or clinician as listed in Table 5, to easily understand the complicate structure and information in multi-view data and their meta-outcome (*see* **Note 4**).

As the general applications of integrative analysis and visualization tool public accessible, Ensembl Genomes is an integrative resource for genome-scale data from non-vertebrate species [106], which exploits and extends technology developed in the context of the Ensembl project and provides a complementary set of resources for non-vertebrate species through a consistent set of programmatic and interactive interfaces. Similarly, the cBioPortal for Cancer Genomics provides a Web resource for exploring, visualizing, and analyzing multidimensional cancer genomics data [107], whose portal reduces molecular profiling data from cancer tissues and cell lines into readily understandable genetic, epigenetic, gene expression, and proteomic events.

Meanwhile, as expert approaches of integrative analysis and visualization tool online, a Web tool, named Integrated Clustering of Multidimensional biomedical data (ICM), can provide an interface from which to fuse, cluster, and visualize multidimensional biomedical data and knowledge or can explore the heterogeneity of a disease or a biological process by identifying subgroups of patients [108]. Next, an integrative meta-analysis of expression data (INMEX) is designed to support meta-analysis of multiple gene expression datasets, as well as datasets from gene expression and metabolomics experiments, whose statistical analysis module allows researchers to combine multiple datasets based on *P* values, effect sizes, rank orders, and other features [109]. Then, a Web server, SteinerNet, establishes a framework for integrating transcriptional, proteomic, and interactome data by searching for the solution to the prize-collecting Steiner tree problem [110]. Besides, a new data integration framework, Anduril, is introduced for translating fragmented large-scale data into testable predictions, and it allows rapid integration of heterogeneous data with state-of-the-art computational methods and existing knowledge in bio-databases [111].

Similarly, when taking particular focus on integrative analysis and visualization on TCGA data, Web-TCGA, a Web-based, freely accessible online tool, can also be run in a private instance, for integrated analysis of molecular cancer datasets provided by TCGA [68]. And MEXPRESS is developed as a straightforward and easy-to-use Web tool for the integration and visualization of the expression, DNA methylation, and clinical TCGA data on a single-gene level, which offers clinical researchers a simple way to evaluate the TCGA data for their genes or candidate biomarkers of interest [64]. And CrossHub software is developed to enable

Table 5
The representative approaches of integrative visualization

Methods	Description	URL
ICM [108]	Integrated clustering of multiple types of omics data is essential for developing individual-based treatments and precision medicine	http://biotech.bmi.ac.cn/icm/
MEXPRESS [64]	Offers clinical researchers a simple way to evaluate the TCGA data for their genes or candidate biomarkers of interest	http://mexpress.be
SteinerNet [110]	For researchers who would like to integrate their high-throughput data for a specific condition or cellular response and to find biologically meaningful pathway	http://fraenkel-nsf.csbi.mit.edu/steinernet/
CrossHub [112]	The contribution of different mechanisms to the regulation of gene expression varies for different tissues and tumors	https://sourceforge.net/projects/crosshub/
Anduril [111]	To translate the fragmented and heterogeneous datasets into knowledge	http://csbi.ltdk.helsinki.fi/anduril/
Web-TCGA [68]	Integrated analysis of molecular cancer datasets provided by TCGA	https://sourceforge.net/projects/webtcga/
Ensembl Genomes [106]	Participants in a growing range of collaborations involved in the annotation and analysis of genomes	http://www.ensemblgenomes.org
INMEX [109]	Properly combining or integrating the datasets with similar basic hypotheses can help reduce study bias, increase statistical power, and improve overall biological understanding	http://www.inmex.ca
cBioPortal [107]	To provide a practical guide to the analysis and visualization features of the cBioPortal for cancer genomics	http://cbioportal.org

two-way identification of most possible TF-gene interactions: on the basis of ENCODE ChIP-Seq binding evidence or Jaspar prediction and co-expression according to the data of the largest cancer omics resource [112].

4 Notes

This paper has given a comprehensive summary of data resources, data analysis, and data visualization supporting the integration of big biological data. Finally, we would like to list several notes on this review:

1. Conventional big data from society would have a large number of samples, and each sample has a few features/attributes. By contrast, the big biological data would supply not large but enough samples and test tens of thousands of features for each sample simultaneously. *This small-sample high-dimensional data requires new analytic approaches, including the data integration.*
2. "Bottom-up integration" mode with follow-up manual integration is always the hypothesis-driven approaches to extract the significant enriched or observed biological knowledge in data. The key of these methods is there should be clear and suitable biological hints on the experiments and outcome data, and then the data combination can extract the biological signals in each type of data and explain the same preset biological hypothesis in a single analysis framework. Although for different combinations on data types, there is already corresponding integrative analysis framework, it is still short of more general and flexible scheme to deal with the existing data types and potential new data types. *It is urgently required to design quantitative evaluation on the confidence of driver hypothesis ahead of data analysis and also on the contribution of different data types to the biological hypothesis.*
3. Meanwhile, "top-down integration" mode with follow-up in silico integration is usually the data-driven approaches to extract the most probable feature signals or sample patterns in data. The key of these methods is there must be efficient correction to reduce the noise and bias in different types of data, and then the data fusion can identify the coordinate data distribution or data correlation in multiple types of data in a unified mathematical model. Many techniques are available; however, they are used solid constraint on the union of data coordination, which limit their application on the diverse biological systems. *Thus, the more relaxations, e.g., soft-constraint-based approaches, will expand the power of data fusion in biological study and detect unseen biological patterns.*

4. Besides, a few tools are available for general study or special application, but the platform of benchmarking the integration methods requires further development, both on the "gold-standard" data and criteria. And the databases on storage and reanalysis of metadata of integration outcome also ask for attention, design, and advance.

References

1. Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, Millard S, Mugabushaka AM, Perrin N, Remacle JE, Remington K, Rocca-Serra P, Taylor CF, Thorley M, Tiwari B, Wilbanks J (2009) Megascience. 'Omics data sharing'. Science 326 (5950):234–236. https://doi.org/10.1126/science.1180598
2. Vo TV, Das J, Meyer MJ, Cordero NA, Akturk N, Wei X, Fair BJ, Degatano AG, Fragoza R, Liu LG, Matsuyama A, Trickey M, Horibata S, Grimson A, Yamano H, Yoshida M, Roth FP, Pleiss JA, Xia Y, Yu H (2016) A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. Cell 164 (1–2):310–323. https://doi.org/10.1016/j.cell.2015.11.037
3. Madhani HD, Francis NJ, Kingston RE, Kornberg RD, Moazed D, Narlikar GJ, Panning B, Struhl K (2008) Epigenomics: a roadmap, but to where? Science 322 (5898):43–44. https://doi.org/10.1126/science.322.5898.43b
4. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G (2015) Epigenomics: roadmap for regulation. Nature 518 (7539):314–316. https://doi.org/10.1038/518314a
5. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25(3):309–316. https://doi.org/10.1038/nbt1295
6. Nicholson JK, Lindon JC (2008) Systems biology: metabonomics. Nature 455 (7216):1054–1056. https://doi.org/10.1038/4551054a
7. Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis AR, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruyssinck E, Sahalie JM, Scholz A, Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda AO, Trigg SA, Twizere JC, Vega K, Walsh J, Cusick ME, Xia Y, Barabasi AL, Iakoucheva LM, Aloy P, De Las Rivas J, Tavernier J, Calderwood MA, Hill DE, Hao T, Roth FP, Vidal M (2014) A proteome-scale map of the human interactome network. Cell 159(5):1212–1226. https://doi.org/10.1016/j.cell.2014.10.050
8. Friedel CC, Zimmer R (2006) Toward the complete interactome. Nat Biotechnol 24 (6):614–615.; Author reply 615. https://doi.org/10.1038/nbt0606-614
9. Buxton B, Hayward V, Pearson I, Karkkainen L, Greiner H, Dyson E, Ito J, Chung A, Kelly K, Schillace S (2008) Big data: the next Google. Interview by Duncan Graham-Rowe. Nature 455(7209):8–9. https://doi.org/10.1038/455008a
10. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL (2012) Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 28(24):3290–3297. https://doi.org/10.1093/bioinformatics/bts595
11. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A 110(11):4245–4250. https://doi.org/10.1073/pnas.1208949110
12. Rapport DJ, Maffi L (2013) A call for integrative thinking. Science 339(6123):1032. https://doi.org/10.1126/science.339.6123.1032-a
13. Wen Y, Wei Y, Zhang S, Li S, Liu H, Wang F, Zhao Y, Zhang D, Zhang Y (2016) Cell subpopulation deconvolution reveals breast

cancer heterogeneity based on DNA methylation signature. Brief Bioinform. https://doi.org/10.1093/bib/bbw028

14. Voillet V, Besse P, Liaubet L, San Cristobal M, Gonzalez I (2016) Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. BMC Bioinformatics 17(1):402. https://doi.org/10.1186/s12859-016-1273-5
15. Weischenfeldt J, Simon R, Feuerbach L, Schlangen K, Weichenhan D, Minner S, Wuttig D, Warnatz HJ, Stehr H, Rausch T, Jager N, Gu L, Bogatyrova O, Stutz AM, Claus R, Eils J, Eils R, Gerhauser C, Huang PH, Hutter B, Kabbe R, Lawerenz C, Radomski S, Bartholomae CC, Falth M, Gade S, Schmidt M, Amschler N, Hass T, Galal R, Gjoni J, Kuner R, Baer C, Masser S, von Kalle C, Zichner T, Benes V, Raeder B, Mader M, Amstislavskiy V, Avci M, Lehrach H, Parkhomchuk D, Sultan M, Burkhardt L, Graefen M, Huland H, Kluth M, Krohn A, Sirma H, Stumm L, Steurer S, Grupp K, Sultmann H, Sauter G, Plass C, Brors B, Yaspo ML, Korbel JO, Schlomm T (2013) Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. Cancer Cell 23(2):159–170. https://doi.org/10.1016/j.ccr.2013.01.002
16. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, Ladanyi M, Sander C (2012) Integrative subtype discovery in glioblastoma using iCluster. PLoS One 7(4):e35236. https://doi.org/10.1371/journal.pone.0035236
17. Zeng T, Wang DC, Wang X, Xu F, Chen L (2014) Prediction of dynamical drug sensitivity and resistance by module network rewiring-analysis based on transcriptional profiling. Drug Resist Updates 17(3):64–76. https://doi.org/10.1016/j.drup.2014.08.002
18. Shi X, Shen S, Liu J, Huang J, Zhou Y, Ma S (2014) Similarity of markers identified from cancer gene expression studies: observations from GEO. Brief Bioinform 15(5):671–684. https://doi.org/10.1093/bib/bbt044
19. Shi X, Yi H, Ma S (2015) Measures for the degree of overlap of gene signatures and applications to TCGA. Brief Bioinform 16(5):735–744. https://doi.org/10.1093/bib/bbu049
20. Bebek G, Koyuturk M, Price ND, Chance MR (2012) Network biology methods integrating biological data for translational science. Brief Bioinform 13(4):446–459. https://doi.org/10.1093/bib/bbr075
21. Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res 40(19):9379–9391. https://doi.org/10.1093/nar/gks725
22. Liu Y, Devescovi V, Chen S, Nardini C (2013) Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. BMC Syst Biol 7:14. https://doi.org/10.1186/1752-0509-7-14
23. Hieke S, Benner A, Schlenl RF, Schumacher M, Bullinger L, Binder H (2016) Integrating multiple molecular sources into a clinical risk prediction signature by extracting complementary information. BMC Bioinformatics 17(1):327. https://doi.org/10.1186/s12859-016-1183-6
24. Shen R, Olshen AB, Ladanyi M (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics 25(22):2906–2912. https://doi.org/10.1093/bioinformatics/btp543
25. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do KA (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics 29(2):149–159. https://doi.org/10.1093/bioinformatics/bts655
26. Yuan Y, Savage RS, Markowetz F (2011) Patient-specific data fusion defines prognostic cancer subtypes. PLoS Comput Biol 7(10):e1002227. https://doi.org/10.1371/journal.pcbi.1002227
27. Speicher NK, Pfeifer N (2015) Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics 31(12):i268–i275. https://doi.org/10.1093/bioinformatics/btv244
28. Narayanan M, Vetta A, Schadt EE, Zhu J (2010) Simultaneous clustering of multiple gene expression and physical interaction datasets. PLoS Comput Biol 6(4):e1000742. https://doi.org/10.1371/journal.pcbi.1000742
29. Kutalik Z, Beckmann JS, Bergmann S (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data. Nat Biotechnol 26(5):531–539. https://doi.org/10.1038/nbt1397
30. Le Van T, van Leeuwen M, Carolina Fierro A, De Maeyer D, Van den Eynden J, Verbeke L, De Raedt L, Marchal K, Nijssen S (2016) Simultaneous discovery of cancer subtypes

and subtype features by molecular data integration. Bioinformatics 32(17):i445–i454. https://doi.org/10.1093/bioinformatics/btw434

31. Seely JS, Kaufman MT, Ryu SI, Shenoy KV, Cunningham JP, Churchland MM (2016) Tensor analysis reveals distinct population structure that parallels the different computational roles of areas M1 and V1. PLoS Comput Biol 12(11):e1005164. https://doi.org/10.1371/journal.pcbi.1005164
32. Hore V, Vinuela A, Buil A, Knight J, McCarthy MI, Small K, Marchini J (2016) Tensor decomposition for multiple-tissue gene expression experiments. Nat Genet 48(9):1094–1100. https://doi.org/10.1038/ng.3624
33. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L (2016) Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics 17(Suppl 2):15. https://doi.org/10.1186/s12859-015-0857-9
34. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform 17(4):628–641. https://doi.org/10.1093/bib/bbv108
35. Luo Y, Wang F, Szolovits P (2016) Tensor factorization toward precision medicine. Brief Bioinform. https://doi.org/10.1093/bib/bbw026
36. Vargas AJ, Harris CC (2016) Biomarker development in the precision medicine era: lung cancer as a case study. Nat Rev Cancer 16(8):525–537. https://doi.org/10.1038/nrc.2016.56
37. Lahti L, Schafer M, Klein HU, Bicciato S, Dugas M (2013) Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review. Brief Bioinform 14(1):27–35. https://doi.org/10.1093/bib/bbs005
38. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. https://doi.org/10.1038/nature11632
39. Gerstein M (2012) Genomics: ENCODE leads the way on big data. Nature 489(7415):208. https://doi.org/10.1038/489208b
40. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502(7469):59–64. https://doi.org/10.1038/nature12593
41. Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14(6):390–403. https://doi.org/10.1038/nrg3454
42. Yun X, Xia L, Tang B, Zhang H, Li F, Zhang Z (2016) 3CDB: a manually curated database of chromosome conformation capture data. Database (Oxford). https://doi.org/10.1093/database/baw044
43. Teng L, He B, Wang J, Tan K (2016) 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics 32(17):2727. https://doi.org/10.1093/bioinformatics/btw375
44. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 41(Database issue):D991–D995. https://doi.org/10.1093/nar/gks1193
45. Kim HS, Minna JD, White MA (2013) GWAS meets TCGA to illuminate mechanisms of cancer predisposition. Cell 152(3):387–389. https://doi.org/10.1016/j.cell.2013.01.027
46. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MM, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SO, Joly Y, Kato K, Kennedy KL, Nicolas P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clement B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayes M, Botwell DD, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, Lopez-Otin C, Majumder P, Marra M, McPherson

JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, Boutros PC, Campbell PJ, Flicek P, Getz G, Guigo R, Guo G, Haussler D, Heath S, Hubbard TJ, Jiang T, Jones SM, Li Q, Lopez-Bigas N, Luo R, Muthuswamy L, Ouellette BF, Pearson JV, Puente XS, Quesada V, Raphael BJ, Sander C, Shibata T, Speed TP, Stein LD, Stuart JM, Teague JW, Totoki Y, Tsunoda T, Valencia A, Wheeler DA, Wu H, Zhao S, Zhou G, Stein LD, Guigo R, Hubbard TJ, Joly Y, Jones SM, Kasprzyk A, Lathrop M, Lopez-Bigas N, Ouellette BF, Spellman PT, Teague JW, Thomas G, Valencia A, Yoshida T, Kennedy KL, Axton M, Dyke SO, Futreal PA, Gerhard DS, Gunter C, Guyer M, Hudson TJ, McPherson JD, Miller LJ, Ozenberger B, Shaw KM, Kasprzyk A, Stein LD, Zhang J, Haider SA, Wang J, Yung CK, Cros A, Liang Y, Gnaneshan S, Guberman J, Hsu J, Bobrow M, Chalmers DR, Hasel KW, Joly Y, Kaan TS, Kennedy KL, Knoppers BM, Lowrance WW, Masui T, Nicolas P, Rial-Sebbag E, Rodriguez LL, Vergely C, Yoshida T, Grimmond SM, Biankin AV, Bowtell DD, Cloonan N, deFazio A, Eshleman JR, Etemadmoghadam D, Gardiner BB, Kench JG, Scarpa A, Sutherland RL, Tempero MA, Waddell NJ, Wilson PJ, McPherson JD, Gallinger S, Tsao MS, Shaw PA, Petersen GM, Mukhopadhyay D, Chin L, DePinho RA, Thayer S, Muthuswamy L, Shazand K, Beck T, Sam M, Timms L, Ballin V, Lu Y, Ji J, Zhang X, Chen F, Hu X, Zhou G, Yang Q, Tian G, Zhang L, Xing X, Li X, Zhu Z, Yu Y, Yu J, Yang H, Lathrop M, Tost J, Brennan P, Holcatova I, Zaridze D, Brazma A, Egevard L, Prokhortchouk E, Banks RE, Uhlen M, Cambon-Thomsen A, Viksna J, Ponten F, Skryabin K, Stratton MR, Futreal PA, Birney E, Borg A, Borresen-Dale AL, Caldas C, Foekens JA, Martin S, Reis-Filho JS, Richardson AL, Sotiriou C, Stunnenberg HG, Thoms G, van de Vijver M, van't Veer L, Calvo F, Birnbaum D, Blanche H, Boucher P, Boyault S, Chabannon C, Gut I, Masson-Jacquemier JD, Lathrop M, Pauporte I, Pivot X, Vincent-Salomon A, Tabone E, Theillet C, Thomas G, Tost J, Treilleux I, Calvo F, Bioulac-Sage P, Clement B, Decaens T, Degos F, Franco D, Gut I, Gut M, Heath S, Lathrop M, Samuel D, Thomas G, Zucman-Rossi J, Lichter P, Eils R, Brors B, Korbel JO, Korshunov A, Landgraf P, Lehrach H, Pfister S, Radlwimmer B, Reifenberger G, Taylor MD, von Kalle C, Majumder PP, Sarin R, Rao TS, Bhan MK, Scarpa A, Pederzoli P, Lawlor RA, Delledonne M, Bardelli A, Biankin AV, Grimmond SM, Gress T, Klimstra D, Zamboni G, Shibata T, Nakamura Y, Nakagawa H, Kusada J, Tsunoda T, Miyano S, Aburatani H, Kato K, Fujimoto A, Yoshida T, Campo E, Lopez-Otin C, Estivill X, Guigo R, de Sanjose S, Piris MA, Montserrat E, Gonzalez-Diaz M, Puente XS, Jares P, Valencia A, Himmelbauer H, Quesada V, Bea S, Stratton MR, Futreal PA, Campbell PJ, Vincent-Salomon A, Richardson AL, Reis-Filho JS, van de Vijver M, Thomas G, Masson-Jacquemier JD, Aparicio S, Borg A, Borresen-Dale AL, Caldas C, Foekens JA, Stunnenberg HG, van't Veer L, Easton DF, Spellman PT, Martin S, Barker AD, Chin L, Collins FS, Compton CC, Ferguson ML, Gerhard DS, Getz G, Gunter C, Guttmacher A, Guyer M, Hayes DN, Lander ES, Ozenberger B, Penny R, Peterson J, Sander C, Shaw KM, Speed TP, Spellman PT, Vockley JG, Wheeler DA, Wilson RK, Hudson TJ, Chin L, Knoppers BM, Lander ES, Lichter P, Stein LD, Stratton MR, Anderson W, Barker AD, Bell C, Bobrow M, Burke W, Collins FS, Compton CC, DePinho RA, Easton DF, Futreal PA, Gerhard DS, Green AR, Guyer M, Hamilton SR, Hubbard TJ, Kallioniemi OP, Kennedy KL, Ley TJ, Liu ET, Lu Y, Majumder P, Marra M, Ozenberger B, Peterson J, Schafer AJ, Spellman PT, Stunnenberg HG, Wainwright BJ, Wilson RK, Yang H (2010) International network of cancer genome projects. Nature 464 (7291):993–998. https://doi.org/10.1038/nature08987

47. Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42(Database issue):D68–D73. https://doi.org/10.1093/nar/gkt1181
48. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res 43 (Database issue):D168–D173. https://doi.org/10.1093/nar/gku988
49. Lebron R, Gomez-Martin C, Carpena P, Bernaola-Galvan P, Barturen G, Hackenberg M, Oliver JL (2017) NGSmethDB 2017: enhanced methylomes and differential methylation. Nucleic Acids Res 45(D1):D97–D103. https://doi.org/10.1093/nar/gkw996

50. Xin Y, Chanrion B, O'Donnell AH, Milekic M, Costa R, Ge Y, Haghighi FG (2012) MethylomeDB: a database of DNA methylation profiles of the brain. Nucleic Acids Res 40(Database issue): D1245–D1249. https://doi.org/10.1093/nar/gkr1193
51. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorndahl T, Perez-Pineiro R, Eisner R, Allen F, Neveu V, Greiner R, Scalbert A (2013) HMDB 3.0—the human metabolome database in 2013. Nucleic Acids Res 41(Database issue): D801–D807. https://doi.org/10.1093/nar/gks1065
52. Mitchell A, Bucchini F, Cochrane G, Denise H, ten Hoopen P, Fraser M, Pesseat S, Potter S, Scheremetjew M, Sterk P, Finn RD (2016) EBI metagenomics in 2016—an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res 44(D1):D595–D603. https://doi.org/10.1093/nar/gkv1195
53. Friedman A, Perrimon N (2007) Genetic screening for signal transduction in the era of network biology. Cell 128(2):225–231. https://doi.org/10.1016/j.cell.2007.01.007
54. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113. https://doi.org/10.1038/nrg1272
55. Goymer P (2008) Network biology: why do we need hubs? Nat Rev Genet 9(9):650
56. Hu JX, Thomas CE, Brunak S (2016) Network biology concepts in complex disease comorbidities. Nat Rev Genet 17(10):615–629. https://doi.org/10.1038/nrg.2016.87
57. New AM, Lehner B (2015) Systems biology: network evolution hinges on history. Nature 523(7560):297–298. https://doi.org/10.1038/nature14537
58. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S, Theesfeld C, Sellam A, Stark C, Breitkreutz BJ, Dolinski K, Tyers M (2017) The BioGRID interaction database: 2017 update. Nucleic Acids Res 45(D1):D369–D379. https://doi.org/10.1093/nar/gkw1102
59. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43(Database issue):D447–D452. https://doi.org/10.1093/nar/gku1003
60. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1):D353–D361. https://doi.org/10.1093/nar/gkw1092
61. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P (2016) The reactome pathway knowledgebase. Nucleic Acids Res 44(D1):D481–D487. https://doi.org/10.1093/nar/gkv1351
62. Bohler A, Wu G, Kutmon M, Pradhana LA, Coort SL, Hanspers K, Haw R, Pico AR, Evelo CT (2016) Reactome from a WikiPathways perspective. PLoS Comput Biol 12(5):e1004941. https://doi.org/10.1371/journal.pcbi.1004941
63. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, Haeussler M, Heitner S, Hinrichs AS, Karolchik D, Lee BT, Lee CM, Nejad P, Raney BJ, Rosenbloom KR, Speir ML, Villarreal C, Vivian J, Zweig AS, Haussler D, Kuhn RM, Kent WJ (2017) The UCSC Genome Browser database: 2017 update. Nucleic Acids Res 45(D1):D626–D634. https://doi.org/10.1093/nar/gkw1134
64. Koch A, De Meyer T, Jeschke J, Van Criekinge W (2015) MEXPRESS: visualizing expression, DNA methylation and clinical TCGA data. BMC Genomics 16:636. https://doi.org/10.1186/s12864-015-1847-z
65. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536. https://doi.org/10.1038/415530a
66. Zeng T, Li J (2010) Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. Nucleic Acids Res 38(1):e1. https://doi.org/10.1093/nar/gkp822
67. Zeng T, Guo X, Liu J (2014) Negative correlation based gene markers identification in

integrative gene expression data. Int J Data Min Bioinform 10(1):1–17

68. Deng M, Bragelmann J, Schultze JL, Perner S (2016) Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. BMC Bioinformatics 17:72. https://doi.org/10.1186/s12859-016-0917-9
69. Huang Y, Zaas AK, Rao A, Dobigeon N, Woolf PJ, Veldman T, Oien NC, McClain MT, Varkey JB, Nicholson B, Carin L, Kingsmore S, Woods CW, Ginsburg GS, Hero AO III (2011) Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. PLoS Genet 7(8):e1002234. https://doi.org/10.1371/journal.pgen.1002234
70. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grutzner F, Bergmann S, Nielsen R, Paabo S, Kaessmann H (2011) The evolution of gene expression levels in mammalian organs. Nature 478 (7369):343–348. https://doi.org/10.1038/nature10532
71. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3 (9):1724–1735. https://doi.org/10.1371/journal.pgen.0030161
72. Manimaran S, Selby HM, Okrah K, Ruberman C, Leek JT, Quackenbush J, Haibe-Kains B, Bravo HC, Johnson WE (2016) BatchQC: interactive software for evaluating sample and batch effects in genomic data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btw538
73. Vandenbon A, Dinh VH, Mikami N, Kitagawa Y, Teraguchi S, Ohkura N, Sakaguchi S (2016) Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. Proc Natl Acad Sci U S A 113(17): E2393–E2402. https://doi.org/10.1073/pnas.1604351113
74. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1):118–127. https://doi.org/10.1093/biostatistics/kxj037
75. Stein CK, Qu P, Epstein J, Buros A, Rosenthal A, Crowley J, Morgan G, Barlogie B (2015) Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat. BMC Bioinformatics 16:63. https://doi.org/10.1186/s12859-015-0478-3
76. Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, de Andrade M, Kocher JP, Eckel-Passow JE (2013) A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. Bioinformatics 29 (22):2877–2883. https://doi.org/10.1093/bioinformatics/btt480
77. Song R, Huang J, Ma S (2012) Integrative prescreening in analysis of multiple cancer genomic studies. BMC Bioinformatics 13:168. https://doi.org/10.1186/1471-2105-13-168
78. Huang X, Stern DF, Zhao H (2016) Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival—evidence from TCGA pan-cancer data. Sci Rep 6:20567. https://doi.org/10.1038/srep20567
79. Hwang TH, Atluri G, Kuang R, Kumar V, Starr T, Silverstein KA, Haverty PM, Zhang Z, Liu J (2013) Large-scale integrative network-based analysis identifies common pathways disrupted by copy number alterations across cancers. BMC Genomics 14:440. https://doi.org/10.1186/1471-2164-14-440
80. Li Q, Seo JH, Stranger B, McKenna A, Pe'er I, Laframboise T, Brown M, Tyekucheva S, Freedman ML (2013) Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 152(3):633–641. https://doi.org/10.1016/j.cell.2012.12.034
81. Peifer M, Fernandez-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, Plenker D, Leenders F, Sun R, Zander T, Menon R, Koker M, Dahmen I, Muller C, Di Cerbo V, Schildhaus HU, Altmuller J, Baessmann I, Becker C, de Wilde B, Vandesompele J, Bohm D, Ansen S, Gabler F, Wilkening I, Heynck S, Heuckmann JM, Lu X, Carter SL, Cibulskis K, Banerji S, Getz G, Park KS, Rauh D, Grutter C, Fischer M, Pasqualucci L, Wright G, Wainer Z, Russell P, Petersen I, Chen Y, Stoelben E, Ludwig C, Schnabel P, Hoffmann H, Muley T, Brockmann M, Engel-Riedel W, Muscarella LA, Fazio VM, Groen H, Timens W, Sietsma H, Thunnissen E, Smit E, Heideman DA, Snijders PJ, Cappuzzo F, Ligorio C, Damiani S, Field J, Solberg S, Brustugun OT, Lund-Iversen M, Sanger J, Clement JH, Soltermann A, Moch H, Weder W, Solomon B, Soria JC, Validire P, Besse B, Brambilla E, Brambilla C, Lantuejoul S, Lorimier P, Schneider PM, Hallek M, Pao W,

Meyerson M, Sage J, Shendure J, Schneider R, Buttner R, Wolf J, Nurnberg P, Perner S, Heukamp LC, Brindle PK, Haas S, Thomas RK (2012) Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet 44 (10):1104–1110. https://doi.org/10.1038/ng.2396

82. Cancer Genome Atlas N (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487 (7407):330–337. https://doi.org/10.1038/nature11252
83. Cancer Genome Atlas Research N (2013) Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499 (7456):43–49. https://doi.org/10.1038/nature12222
84. Cancer Genome Atlas Research N (2014) Comprehensive molecular characterization of gastric adenocarcinoma. Nature 513 (7517):202–209. https://doi.org/10.1038/nature13480
85. Cancer Genome Atlas Research N (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 507 (7492):315–322. https://doi.org/10.1038/nature12965
86. Cancer Genome Atlas N (2015) Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517 (7536):576–582. https://doi.org/10.1038/nature14129
87. Cancer Genome Atlas Research N (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455(7216):1061–1068. https://doi.org/10.1038/nature07385
88. Cancer Genome Atlas Research N (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature 489 (7417):519–525. https://doi.org/10.1038/nature11404
89. Akbani R, Ng PK, Werner HM, Shahmoradgoli M, Zhang F, Ju Z, Liu W, Yang JY, Yoshihara K, Li J, Ling S, Seviour EG, Ram PT, Minna JD, Diao L, Tong P, Heymach JV, Hill SM, Dondelinger F, Stadler N, Byers LA, Meric-Bernstam F, Weinstein JN, Broom BM, Verhaak RG, Liang H, Mukherjee S, Lu Y, Mills GB (2014) A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat Commun 5:3887. https://doi.org/10.1038/ncomms4887
90. Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, Zhang H, McLellan M, Yau C, Kandoth C, Bowlby R, Shen H, Hayat S, Fieldhouse R, Lester SC, Tse GM, Factor RE, Collins LC, Allison KH, Chen YY, Jensen K, Johnson NB, Oesterreich S, Mills GB, Cherniack AD, Robertson G, Benz C, Sander C, Laird PW, Hoadley KA, King TA, Network TR, Perou CM (2015) Comprehensive molecular portraits of invasive lobular breast cancer. Cell 163(2):506–519. https://doi.org/10.1016/j.cell.2015.09.033
91. Cancer Genome Atlas N (2012) Comprehensive molecular portraits of human breast tumours. Nature 490(7418):61–70. https://doi.org/10.1038/nature11412
92. Drake JM, Paull EO, Graham NA, Lee JK, Smith BA, Titz B, Stoyanova T, Faltermeier CM, Uzunangelov V, Carlin DE, Fleming DT, Wong CK, Newton Y, Sudha S, Vashisht AA, Huang J, Wohlschlegel JA, Graeber TG, Witte ON, Stuart JM (2016) Phosphoproteome integration reveals patient-specific networks in prostate cancer. Cell 166 (4):1041–1054. https://doi.org/10.1016/j.cell.2016.07.007
93. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45 (10):1113–1120. https://doi.org/10.1038/ng.2764
94. Neapolitan R, Horvath CM, Jiang X (2015) Pan-cancer analysis of TCGA data reveals notable signaling pathways. BMC Cancer 15:516. https://doi.org/10.1186/s12885-015-1484-6
95. Ruau D, Dudley JT, Chen R, Phillips NG, Swan GE, Lazzeroni LC, Clark JD, Butte AJ, Angst MS (2012) Integrative approach to pain genetics identifies pain sensitivity loci across diseases. PLoS Comput Biol 8(6): e1002538. https://doi.org/10.1371/journal.pcbi.1002538
96. Liu P, Sanalkumar R, Bresnick EH, Keles S, Dewey CN (2016) Integrative analysis with ChIP-seq advances the limits of transcript quantification from RNA-seq. Genome Res 26(8):1124–1133. https://doi.org/10.1101/gr.199174.115
97. Knouf EC, Garg K, Arroyo JD, Correa Y, Sarkar D, Parkin RK, Wurz K, O'Briant KC, Godwin AK, Urban ND, Ruzzo WL, Gentleman R, Drescher CW, Swisher EM, Tewari M (2012) An integrative genomic approach identifies p73 and p63 as activators of miR-200 microRNA family transcription. Nucleic Acids Res 40(2):499–510. https://doi.org/10.1093/nar/gkr731

98. Yan Z, Shah PK, Amin SB, Samur MK, Huang N, Wang X, Misra V, Ji H, Gabuzda D, Li C (2012) Integrative analysis of gene and miRNA expression profiles with transcription factor-miRNA feed-forward loops identifies regulators in human cancers. Nucleic Acids Res 40(17):e135. https://doi.org/10.1093/nar/gks395
99. Berghoff BA, Konzer A, Mank NN, Looso M, Rische T, Forstner KU, Kruger M, Klug G (2013) Integrative "omics"–approach discovers dynamic and regulatory features of bacterial stress responses. PLoS Genet 9(6): e1003576. https://doi.org/10.1371/journal.pgen.1003576
100. Kim M, Rai N, Zorraquino V, Tagkopoulos I (2016) Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nat Commun 7:13090. https://doi.org/10.1038/ncomms13090
101. Meng C, Helm D, Frejno M, Kuster B (2016) moCluster: identifying joint patterns across multiple omics data sets. J Proteome Res 15 (3):755–765. https://doi.org/10.1021/acs.jproteome.5b00824
102. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A (2014) Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11(3):333–337. https://doi.org/10.1038/nmeth.2810
103. Shi Q, Zhang C, Peng M, Yu X, Zeng T, Liu J, Chen L (2017) Pattern fusion analysis by adaptive alignment of multiple heterogeneous omics data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx176
104. Lee CH, Alpert BO, Sankaranarayanan P, Alter O (2012) GSVD comparison of patient-matched normal and tumor aCGH profiles reveals global copy-number alterations predicting glioblastoma multiforme survival. PLoS One 7(1):e30098. https://doi.org/10.1371/journal.pone.0030098
105. Xiao X, Moreno-Moral A, Rotival M, Bottolo L, Petretto E (2014) Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. PLoS Genet 10(1): e1004006. https://doi.org/10.1371/journal.pgen.1004006
106. Kersey PJ, Staines DM, Lawson D, Kulesha E, Derwent P, Humphrey JC, Hughes DS, Keenan S, Kerhornou A, Koscielny G, Langridge N, McDowall MD, Megy K, Maheswari U, Nuhn M, Paulini M, Pedro H, Toneva I, Wilson D, Yates A, Birney E (2012) Ensembl genomes: an integrative resource for genome-scale data from non-vertebrate species. Nucleic Acids Res 40 (Database issue):D91–D97. https://doi.org/10.1093/nar/gkr895
107. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6(269):pl1. https://doi.org/10.1126/scisignal.2004088
108. He S, He H, Xu W, Huang X, Jiang S, Li F, He F, Bo X (2016) ICM: a web server for integrated clustering of multi-dimensional biomedical data. Nucleic Acids Res 44(W1): W154–W159. https://doi.org/10.1093/nar/gkw378
109. Xia J, Fjell CD, Mayer ML, Pena OM, Wishart DS, Hancock RE (2013) INMEX—a web-based tool for integrative meta-analysis of expression data. Nucleic Acids Res 41(Web Server issue):W63–W70. https://doi.org/10.1093/nar/gkt338
110. Tuncbag N, McCallum S, Huang SS, Fraenkel E (2012) SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways. Nucleic Acids Res 40(Web Server issue): W505–W509. https://doi.org/10.1093/nar/gks445
111. Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomaki V, Valo E, Nunez-Fontarnau J, Rantanen V, Karinen S, Nousiainen K, Lahesmaa-Korpinen AM, Miettinen M, Saarinen L, Kohonen P, Wu J, Westermarck J, Hautaniemi S (2010) Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Med 2(9):65. https://doi.org/10.1186/gm186
112. Krasnov GS, Dmitriev AA, Melnikova NV, Zaretsky AR, Nasedkina TV, Zasedatelev AS, Senchenko VN, Kudryavtseva AV (2016) CrossHub: a tool for multi-way analysis of The Cancer Genome Atlas (TCGA) in the context of gene expression regulation mechanisms. Nucleic Acids Res 44(7):e62. https://doi.org/10.1093/nar/gkv1478
113. Yu X, Li G, Chen L (2014) Prediction and early diagnosis of complex diseases by edge-network. Bioinformatics 30(6):852–859. https://doi.org/10.1093/bioinformatics/btt620
114. Zhang Q, Burdette JE, Wang JP (2014) Integrative network analysis of TCGA data for ovarian cancer. BMC Syst Biol 8:1338. https://doi.org/10.1186/s12918-014-0136-9

115. Zhu R, Zhao Q, Zhao H, Ma S (2016) Integrating multidimensional omics data for cancer outcome. Biostatistics 17(4):605–618. https://doi.org/10.1093/biostatistics/kxw010

116. Wang XV, Verhaak RG, Purdom E, Spellman PT, Speed TP (2011) Unifying gene expression measures from multiple platforms using factor analysis. PLoS One 6(3):e17691. https://doi.org/10.1371/journal.pone.0017691