OXFORD

Briefings in Bioinformatics, 2018, 1–16

doi: 10.1093/bib/bby014
Paper

It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data

Juan Xie, Anjun Ma, Anne Fennell, Qin Ma and Jing Zhao

Corresponding authors: Qin Ma, Department of Agronomy, Horticulture, and Plant Science, and Department of Mathematics and Statistics, BioSNTR, South Dakota State University, Brookings, SD 57006, USA. Tel: 1-605-688-6315; E-mail: qin.ma@sdstate.edu; Jing Zhao, Population Health Group, Sanford Research, Sioux Falls, SD, 57104, USA; Department of Internal Medicine, Sanford School of Medicine, University of South Dakota, Sioux Falls, SD, 57105, USA. Tel: 1-605-312-6468; E-mail: jing.zhao@sanfordhealth.org

Abstract

Biclustering is a powerful data mining technique that allows clustering of rows and columns, simultaneously, in a matrix-format data set. It was first applied to gene expression data in 2000, aiming to identify co-expressed genes under a subset of all the conditions/samples. During the past 17 years, tens of biclustering algorithms and tools have been developed to enhance the ability to make sense out of large data sets generated in the wake of high-throughput omics technologies. These algorithms and tools have been applied to a wide variety of data types, including but not limited to, genomes, transcriptomes, exomes, epigenomes, phenomes and pharmacogenomes. However, there is still a considerable gap between biclustering methodology development and comprehensive data interpretation, mainly because of the lack of knowledge for the selection of appropriate biclustering tools and further supporting computational techniques in specific studies. Here, we first deliver a brief introduction to the existing biclustering algorithms and tools in public domain, and then systematically summarize the basic applications of biclustering for biological data and more advanced applications of biclustering for biomedical data. This review will assist researchers to effectively analyze their big data and generate valuable biological knowledge and novel insights with higher efficiency.

Key words: biclustering; functional annotation; modularity analysis; network elucidation; disease subtype identification; biomarker and gene signatures detection; gene–drug association

Introduction

The advent of much-improved biotechnology and the decreased associated costs have generated a massive amount of biological and biomedical data. The next-generation sequencing (NGS) technology [1, 2] has higher resolution, improved accuracy, lower technical variation and other advantages in comparison with array-based counterparts [3–5]. NGS allows for rapid generation of larger volumes of biological information than ever before. Also, large amounts of patient clinical data are generated through NGS and electronic health record (EHR), which presents significant opportunities for knowledge discoveries in biomedical research

Juan Xie is a graduate student in the in the Department of Mathematics and Statistics at South Dakota State University, Brookings, SD, USA.
Anjun Ma is a graduate student in the in the Department of Mathematics and Statistics at South Dakota State University, Brookings, SD, USA.
Anne Fennell is a professor at the Department of Agronomy, Horticulture and Plant Science, BioSNTR, South Dakota State University, with expertise in Developmental Biology, Physiology and Molecular Biology
Qin Ma is the corresponding author. He is the director of the Bioinformatics and Mathematical Biosciences Lab and an assistant professor at the Department of Agronomy, Horticulture and Plant Science, South Dakota State University. He is also an adjunct faculty member of the Department of Mathematics and Statistics, BioSNTR and Sanford Research, USA.
Jing Zhao is the corresponding author. She is an assistant research scientist at Sanford Research, and an assistant professor at the Department of Internal Medicine, University of South Dakota Sanford School of Medicine.
Submitted: 15 November 2017; **Received (in revised form):** 16 January 2018

[6]. These complex and large volumes of data, collected from different sources, have changed the way biological and biomedical research is conducted [7, 8]. Effective utilization and interpretation of such data require advances in interdisciplinary sciences. The concept of big-data-to-knowledge relies extensively on biological, mathematical, statistical and computer sciences to extract usable information and generate new knowledge.

For example, the abundance of gene expression data sets provides an opportunity to identify genes with similar expression patterns across multiple conditions, i.e. co-expression gene modules (CEMs). These modules are crucial for inferring high-level functional machinery, e.g., regulatory and metabolic pathways. Microarray platforms have been the most widely used in generating gene expression data because of its easy accessibility and low cost. The high-throughput RNA sequencing (RNA-seq) is a revolutionary technology for gene expression profiling [9, 10], which promises a comprehensive picture of the transcriptome for a biological process, as it enables the complete quantification of all genes in a cell [9, 11]. Genome-scale identification of CEMs can be modeled by biclustering [12], which was introduced by Hartigan in 1972 [13] and applied to gene expression data analysis by Cheng and Church in 2000 [14]. Biclustering is a two-dimensional data mining technique that allows clustering of rows (representing genes) and columns (representing samples/conditions) in a gene expression matrix, simultaneously. The biclustering method can capture biologically meaningful and computationally significant CEMs, by identifying (possibly overlapped) homogeneous submatrices, subsets of rows with a coherent pattern across subsets of columns that satisfy specific quality metrics (e.g. mean squared residue used in [14] and MSE used in [15]). This unique feature makes it useful when applied to big gene expression data, as genes that participate in a cellular process are only active in specific conditions, thus are usually co-expressed under a subset of all conditions.

Furthermore, with the advancement of informatics technology, EHR contains sufficient information that can be transformed into disease phenotypes [16]. In this phenotyping process, a heuristic and the iterative searching algorithm is applied to search the large-scale EHR database with queries created by clinical experts and knowledgeable computational engineers [16], during which thousands of phenotypes generated for all the included individuals. These phenotype data can be organized into a matrix, with phenotype features as rows and individuals as columns, providing essential materials to identify a family of phenotype biclusters. The biclusters define a subgroup of patients from a subset of phenotypes, which are subject to detailed validation analysis to establish their relations with (i) prognostic or therapeutic characteristics of diseases [17–20], and (ii) genotype biclusters [16].

A substantial number of biclustering methods were developed during the past 17 years [14, 15, 21–38]. SAMBA [30], ISA [31], BIMAX [32], QUBIC [33] and FABIA [34] are some popular algorithms for general purpose. CCC-biclustering [39–41] and LateBiclustering [42] are designed for temporal data analysis, and BicPAM [43], BicNET [37, 44] and MCbiclust [45] are three recent tools. In addition, several tools (R packages, web servers, etc.) have been developed to facilitate users with a limited computational background [25, 46–52]. GEMS [49] is a web server for gene expression mining based on a Gibbs sampling paradigm, and biclust [50] and QUBICR [51] are two R packages integrating multiple existing algorithms, data preprocessing functions and interpretation and visualization of the results.

Several biclustering algorithm review studies have been conducted emphasizing different mechanistic perspectives [32, 53–57]. For example, Pontes *et al.* [58] presented a taxonomy of 47 biclustering algorithms according to their search strategies, and Busygin *et al.* [59] emphasized the mathematical models and concepts in biclustering techniques. Padilha *et al.* [56] claimed that an algorithm only achieved satisfactory results in a certain context, and the best algorithm choice depends on specific objectives. Eren *et al.* [60] compared 12 popular algorithms and concluded that QUBIC achieves the highest performance in synthetic data sets and captures a high proportion of enriched biclusters on real data sets. Adetayo *et al.* [61] presented an overview of data analysis using biclustering methods from a practical point of view, accompanied by R examples.

As far as we know, application of biclustering has not progressed in parallel with algorithm design. Considering all the biclustering-related publications, the portion of application studies has been much lower than that of algorithm development studies from the year 2000–17 (Figure 1). This situation is affected by multiple factors. First, there is a gap between tool development and the understanding of new biotechnologies and corresponding data properties. For example, microarray data are reflecting absolute gene expression with continuous fluorescence intensity values [62], while RNA-seq data measures the relative expression level using discrete, positive and highly skewed read counts [63–66]. Furthermore, there are abundant zeros in RNA-seq-based gene expression data, as not all the genes are expressed under a specific experimental condition, which is particularly true in single-cell RNA-seq (scRNA-seq) data [67, 68]. Hence, algorithms designed and evaluated using microarray data may not be suitable to be directly applied to RNA-seq data. RNA-seq and scRNA-seq data need the unique design of algorithm and tool development. However, contrary to the fact that RNA-seq is becoming more and more popular, few biclustering algorithms are explicitly designed for RNA-seq data [39, 40, 43, 44]. Second, there is a knowledge gap for applying biclustering tools and choosing the appropriate accompanying analytical tools for specific data analyses. Usually, biclustering is not a solo data analysis tool. Instead, it connects with other results annotation processes (e.g. DAVID and KOBAS), visualization programs (e.g. Cytoscape) and statistical methods (e.g. principal component analysis and regression analysis), to derive a more comprehensive interpretation. It is worth noting that organically integrating a biclustering algorithm and appropriate accompanying tools into a pipeline is not trivial. Construction of a unified pipeline requires a deeper understanding of underlying algorithm designs, data inputs and expected outputs.

The yearly proportion of biclustering references related to algorithm development and improvement and application studies is presented in Figure 1. The numbers of biclustering studies on algorithm design and application were similar at earliest stage when few tools were available. The proportion of application-related studies decreased relative to algorithm design until 2010. In the 1650 articles published in 2011, the number of studies related to algorithm design was almost nine times that of the application studies. Recently, more researchers have realized the biclustering application shortage and made significant efforts in this area. Between 2012 and 2016, the application publication proportion increased to 40%. There is still a considerable potential for more application-related studies; therefore, this review systematically summarizes the basic applications of biclustering in biological data and the advanced applications of biclustering in biomedical data. This information will enable biological researchers to select appropriate algorithms and computational tools for their various studies, effectively bridging the gap between big data and valuable biological knowledge

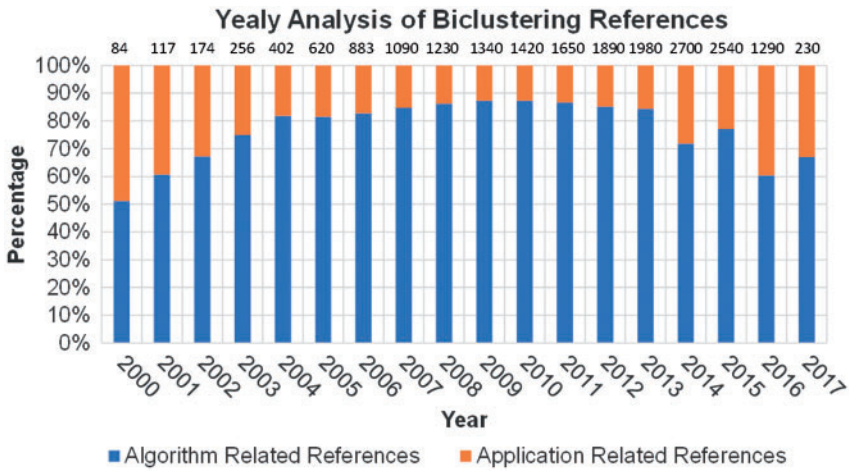

Figure 1. Yearly comparison of biclustering algorithm development and algorithm application related studies. The references in 2017 were collected as of 26 March 2017.

and efficiently providing novel data-driven insights. In the following, we will review how biclustering aids biological and biomedical data interpretation at the gene, module and network level, respectively.

Basic application of biclustering on biological data

It is well known that biological function can rarely be attributed to an individual molecule. Instead, most functions arise from complex interactions (as a whole system or module) among the cell's numerous components, such as protein, DNA, RNA and small molecules [69, 70]. Biotechnology has developed fast in the past two decades, from traditional arrays (e.g. microarray and tilling array) to NGS (e.g. DNA-seq, RNA-seq and chromatin immunoprecipitation sequencing (ChIP-seq)) to the third-generation long-read sequencing (e.g. PACBIO and Oxford Nanopore). The generated data provide unprecedented opportunity to understand the complex biological system at different levels, from basic mutation, gene and protein structure level, to pathway/module level, and even global networks. Biclustering analyses play a significant role in making sense out of various omics data toward the goal of generating a system-level understanding.

Functional annotation of unclassified genes

Functional annotation categorizes genes into one or multiple functional classes, which is an essential step for understanding the physiological purpose of target/interesting genes. However, a reliable functional assessment of a given gene can be carried out only if all its interacting genes are known in advance, as a gene can be involved in different pathways/networks to achieve specific biological functions [71]. These are typically not known for all genes or conditions. Biologists often deal with this challenge, in part, by taking advantage of the 'guilt-by-association' (GBA) principle. GBA assumes that functions can be transferred from one gene to another through biological association. Two kinds of information are required for a GBA-based functional annotation: known functional annotation in public domain and the associations between annotated and unannotated genes. NCBI, Gene Ontology (GO) [72] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [73] are three dominant representatives of such comprehensive databases; RegulonDB is one of the most widely used resources for *Escherichia coli* K-12 gene regulation [74]; The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) offers genomics, epigenomic and proteomic data for thousands of tumor samples across >20 types of cancer; and PlantTFDB provides comprehensive genomic transcriptional factor (TF) repertoires of green plants [75]. For unannotated genes, co-expression is one of the most widely used association indices, as gene expression profile collection is accessible and can be used to derive other associations, e.g. co-regulation [76, 77] and co-evolution [78, 79]. Biclustering can be used to identify co-expressed genes based on the similarity of their expression profiles across a wide range of conditions (e.g. treatments, tissues and samples), giving rise to a set of significant CEMs, i.e. biclusters [80]. Based on existing annotation databases and these CEMs, functional enrichment analysis is carried out to identify significantly overrepresented functions, using the hypergeometric distribution as a statistical test [81]. To be specific, the probability of an enriched function can be calculated as:

$$P(X = x|N, p, n) = \frac{\binom{pN}{x}\binom{(1-p)N}{n-x}}{\binom{N}{n}},$$

where x is the number of genes in a bicluster that belong to the certain pathway with size n, N is the total number of genes in the whole genome, p is the percentage of that pathway among all pathways in the whole genome and the P-value of getting such enriched or even more enriched module is calculated as:

$$P - \text{value} = P(X \geq x) = 1 - P(X < x) = 1 - \sum_{i=0}^{x-1} \frac{\binom{pN}{i}\binom{(1-P)N}{n-i}}{\binom{N}{n}}.$$

If the P-value is smaller than a specific cutoff (e.g. 0.01), then it concludes that the bicluster is enriched with that function. Highly enriched functions are assumed to be shared by all members in the obtained biclusters, and unannotated genes in those biclusters will be assigned to the most abundant functional class [82, 83]. It is noteworthy that biclustering is usually combined with the comparative genomics strategy in the case of gene annotation for new-sequenced organisms, which builds links between well-annotated model organisms and the new organisms [84].

Table 1. Case studies of functional annotation of unclassified genes

Data	Methods	Tools/databases	Outcomes	References
	Functional annotation of yeast			
Microarray (6200 ORFs under 515 conditions)	• Biclustering for gene classification • Functionally assign the unannotated genes in biclusters to the most abundant class • Cross-validation for annotation assessment	SAMBA SGD [88] –	2406 biclusters; 196 annotations of unknown genes	[30]
	Functional annotation of plant genomes			
Microarray (21 031 genes of *Arabidopsis* under 351 conditions)	• Biclustering on known PCW genes • Expand biclusters to include additional genes • Construct co-expression network • Predict and annotate motifs in promoter regions of co-expressed genes in each module	QUBIC QUBIC Cytoscape WeederTFBS MotifSampler CompariMotif PLACE AGRIS	417 seed biclusters; 2438 candidate PCW genes co-expressed with 349 PCW genes	[87]
Microarray (122 973 probes of Switchgrass, 94 conditions)	• Homologous mapping of identified PCW genes • Assign mapped genes to PCW-associated functions • Biclustering of mapped genes and expand for new candidates • Identify motifs for each bicluster • Validate prediction by annotated *Arabidopsis* PCW genes	Tblastn DAVID QUBIC – PCWGD[a]	991 homologs PCW genes; 104 clusters of co-expressed genes; 823 new PCW genes; 112 new genes	[84]
	Functional annotation of human and mouse			
A correlation matrix with associations among mouse long intergenic noncoding RNAs (lincRNA), protein-coding genes and lincRNAs	• Identify lincRNA • Create association matrix of lincRNA and protein-coding genes • Biclustering to identify functional modules consisting of lincRNAs and protein-coding genes • Assign putative functions to each lincRNA • Validate inferred biological functions for lincRNAs	ChIP-Seq GSEA SAMBA – –	Sets of lincRNAs associated with a diverse range of functions, including cell proliferation, immune surveillance, muscle development, etc.	[82]
65 human microarray data sets and GO function categories	• Discover network patterns based on frequent itemsets and biclustering • Design network topology statistic based on graph random walk • Assess functional annotation by a random forest method	– – –	1126 functions assigned to 895 genes (779 knowns and 116 unknowns)	[83]

Note: – denotes for no specific existed tools, and this also applies to all the following tables.
aPurdue Cell-Wall-Genomics Database (https://cellwall.genomics. purdue.edu). PCW: plant cell-wall; ORF: open reading frames.

Despite the high potential of this approach, it is essential to keep in mind that correlation does not guarantee causal relationships, i.e. genes with similar expression profiles may not have the same function. The results should be interpreted as preliminary computational predictions which provide useful hypothesis/candidates for future testing [85]. Thus, experimental validation of the predictions is needed. However, the percentage of unannotated genes is high even in well-studied model organisms [86] (e.g. the proportion of unannotated genes is around 40–50% in *E. coli*), and it is unrealistic to go through all the to-be-validated candidates exhaustively using experimental methods. Therefore, researchers usually just verify functions of a few genes of considerable interest [82], and in most cases, they rely on computational validation (e.g. cross-validation [30] and random forest [83]) and published literature support. This logic applies to all tables in this review, and will not be mentioned again.

The basic idea of computational validation is to mask the functions of some annotated genes in a CEM and check to see if the functions can be correctly assigned back to the masked genes. The validation could be conducted by assessing whether the genes share conserved sequence motifs, as it is believed that co-expressed genes tend to, although not necessarily, be transcriptionally co-regulated [87]. Recently, researchers proposed using genome-scale ChIP-seq data for the validation of the prediction of CEMs [84]. Table 1 summarizes five representative studies, which inferred the functions of unannotated genes from the well-annotated genes that they are co-expressed with. For each of five studies, we introduce the input data for the study (Data), biclustering algorithm and accompanying analysis methods (Methods), specific tool and software (Tools/Databases) used to accomplish the research, the output and results (Outcomes) and related references (Refs). All other tables in this study follow the same structure.

Modularity analysis

Compared with individual cellular components, modularity analysis puts more emphasis on the component's relationship and the topology of a module, i.e. a group of physically or functionally linked molecules that work together to achieve distinct functions [70]. Increasing evidence indicates that biological systems are inherently modular [89–91]. Therefore, modularity analysis has been widely applied to investigate the organization

and dynamics of biological systems at different levels, i.e. module identification, dynamic module analysis and module network reconstruction. Up to now, substantial efforts are devoted to the first level of modularity analysis, module identification.

Biclustering has been applied to identify different types of modules, which could be groups of interacting molecules (e.g. microRNA, miRNA, sponge modules in [92] and miRNA-mRNA modules in [93]), functionally related genes/proteins or any other manually defined clusters [94]. Depending on the target modules, different inputs and strategies are needed. For example, (i) scRNA-seq gene expression data were used to identify molecularly distinct subtypes of cells that contribute different brain functions [95]; (ii) an integrated correlation matrix was derived from expression data with target site information to predict miRNA-mRNA functional modules [93]; and (iii) time series expression data are often used to identify temporal transcriptional modules that consist of activated genes at consecutive time points [39]. As various modules are investigated, additional supporting data are often involved. For example, promoter sequences and integrated *de novo* motif detection are integrated with co-expression biclustering to identify regulatory modules [96]. Similar strategies have been implemented with the integration of other supporting data types (e.g. operon prediction, ChIP-seq data and network connections) [97].

With modules identified, further research concentrates on investigating the characteristics of modules. Applying functional annotation or enrichment analysis to these modules can illustrate/deduce their roles in biological processes [92, 93, 98]. Where expression profiles are available in multiple evolutionarily correlated species, researchers can conduct interspecific comparisons and investigate the underlying evolutionary story. For example, Waltman *et al.* [99] performed biclustering of multiple species data and then used a conservation score to identify conserved modules among these species. Based on co-regulation modules, Yang *et al.* [100] derived an expression-based quantity to characterize the functional constraint acting on a gene, and then tested the correlation of those quantities with gene sequence divergence rate to estimate the evolutionary potential of genes. With temporal modules, the dynamic regulatory interaction can be explored. Gonçalves *et al.* [101] ranked TFs targeting the modules at each time point and graphically depicted the regulatory activity in a module at consecutive time points. Other researchers examined the external relationship among modules, e.g. grouped modules of host proteins based on a distance measure to form higher-level subsystems [102]. Table 2 summarized four kinds of modularity analysis applications, including functional module identification, regulatory modules, evolution characteristic and module subsystem. Module-based network inference, as a higher level of modularity analysis, will be introduced in next section.

Biological networks elucidation

Biological interactions can be conceptualized as networks, with nodes representing biological entries and edges denoting relationships between nodes. For example, in protein–protein interaction (PPI) networks, nodes are proteins and edges represent physical interactions; in transcriptional regulatory networks (TRNs), nodes stand for regulators [TFs, microRNAs and long noncoding RNAs (lncRNAs)] and targets and edges are regulatory interaction directing from regulators to targets. Analyzing these networks provides systematic views and novel insights for understanding the underlying mechanisms controlling cellular processes. Table 3 shows examples in network analysis, which mainly focus on network inference and network decomposition.

Compared with random networks, one distinct characteristic of the biological networks is modularity, forming dense subgraphs [103, 104]. Several computational approaches have used the module-based method to infer networks. For example, in TRNs, one widely used approach is to group genes/regulators based on the similarity of their expression profile using biclustering, along with the modeling of the regulatory interactions between those modules to get a higher-level understanding of regulatory mechanisms [69]. This approach has been successfully applied in several other studies [105–107]. On the other hand, Tanay *et al.* [90] used the hierarchical topology of the biological networks. They first used biclustering to identify modules based on integrated heterogeneous experimental data, and then built a module graph, with nodes being modules and edge connected two modules whenever their genes intersect sufficiently. These small modules were clustered into supermodules based on their functional association. In this way, a hierarchical transcriptional network was built. It is noteworthy that researchers often integrate multiple sources of data, in the hope of getting a more comprehensive and accurate view of biological networks. For example, TRNs were constructed using expression data as well as sequence information and interaction data [105–107], and Tanay *et al.* [90] combined expression data, various interactions and phenotypes.

Network decomposition breaks a network down into simpler units or components, e.g. network motifs and modules, and is another hotspot in network analysis. Compared with the previous modularity analysis section where biclustering method is mainly applied to expression data, biclustering takes networks as input in decomposition. Decomposition reduces network complexity and facilitates the exploration of the underlying molecular mechanisms [108–110]. Henriques and Madeira [37] developed and applied a pattern-based biclustering algorithm to discover coherent modules from PPI and showed that most modules were significantly enriched with particular biological functions. Lakizadeh *et al.* integrated time series expression data and static PPI networks to extract dynamic PPI subnetwork and then detected protein complex based on these subnetworks. They concluded that this method could model the dynamicity inherent in static PPI networks [111].

Advanced application of biclustering in biomedical science

A genetic variation that contributes to a specific disease is usually detected through single-nucleotide polymorphisms (SNPs), insertion/deletions, variable number tandem repeats and copy number variants [112]. Besides, understanding the association between above genomic information and specific diseases has led to the discovery of new drugs [113]. However, the association studies are considered as complicated processes because disease risks are attributed to the combined effect of both multiple genetic variants and environmental factors. With the increasing application and decreasing cost of big data generation techniques in biomedical and health-care informatics, large volumes of biological and clinical data sets have become available in the public domain. On one hand, this advance provides materials to identify new therapeutic targets, drug indications and drug-response biomarkers; on the other hand, it also introduces more challenges to the data mining approaches

Table 2. Case studies of modularity analysis

Data	Methods	Tools/databases	Outcomes	References
	Functional module			
miRNA-mRNA regulatory score matrix derived from gene expression data	• Create miRNA-mRNA regulatory score matrix based on expression matrix and miRNA–target binding information	–	Four miRNA sponge modules	[92]
	• Biclustering on the score matrix to infer miRNA-mRNA biclusters	BCPlaid		
	• Filter biclusters using statistical methods and interaction information	–		
	• Functional annotation	GeneCodis		
	• Validation of predicted modules	–		
mRNA-miRNA association matrix derived from gene expression data	• Construct mRNA-miRNA association matrix based on expression data and miRNA target information	–	100 putative miRNA functional module	[93]
	• Biclustering to identify functional modules	BUBBLE		
	• Visualize and evaluate modules	miRMAP		
SC-RNA-seq (3005 mouse cortical cells)	• Biclustering	BackSPIN	47 distinct cell subclasses	[95]
	Regulatory modules			
Microarray data (*Saccharomyces cerevisiae* under 2200 conditions); upstream and downstream sequences	• Biclustering	COALESCE	450 regulatory modules	[96]
Microarray (*Mycobacterium tuberculosis* under 2325 measurements); and 154 TFs ChIP-seq data	• Biclustering	cMonkey2	600 modules	[97]
Time series microarray data for 2884 genes of *S. cerevisiae* in response to heat stress under five time points	• Biclustering	CCC-Biclustering	167 biclusters; Regulatory snapshots of documented regulators at each time point	[39, 101]
	• Ranking the prominent prioritized regulators targeting each of the modules at each time point	Regulatory Snapshots		
	• Graphically depict the regulatory activity in a module	Baiacu; BiGGEsTs		
	Evolutionary study			
Three normalized expression matrixes (*Bacillus subtilis*, *Bacillus anthracis* and *Listeria monocytogenes*); upstream sequences; metabolic and signaling pathways, co-membership in an operon and phylogenetic profile networks	• Biclustering on expression data	FD-MSCM	150 biclusters	[99]
	• Evaluate the conservation between biclusters	–		
Microarray (4117 orthologs in 15, 14 and 17 tissue groups in rice, maize and *Arabidopsis*, respectively)	• Biclustering to predict co-regulated modules	ISA	1181 modules	[100]
	• Quantify the functional constraint acting on a gene based on the modules (eFC)	–		
	• Correlate eFC with gene sequence divergence rate	–		
	Subsystem			
HIV-1, Human Protein Interaction Database (HHPID)	• Biclustering on the binary interaction matrix	Bimax	279 significant sets of host proteins show the same interaction to HIV-1	[102]
	• Construct bicluster distance matrix	–		
	• Construct neighbor-joining tree and designate host subsystem	–		

[113]. As the applications of biclustering in basic biological science lead to many discoveries and novel methodologies, there is a rapidly growing interest in extrapolating it into the big biomedical data. Biclustering is deemed as a powerful tool that could identify novel target genes, indicated drugs or biomarkers of drug responses, in which the principles of biclustering being used in functional annotation and modularity analysis of biological data are also applicable. In this section, we provide comprehensive guidance and discuss the applications of biclustering, particularly the integration with other methods, for

Table 3. Case studies of biological networks elucidation

Inputs	Methods	Tools/databases	Outputs	References
	Yeast transcriptional network			
Nearly 1000 *S. cerevisiae* expression profiles; 110 TF binding location profiles; 30 growth profiles; 1031 protein interaction; 4177 complex interactions and 1175 known interactions from MIPS	• Modeling genomic information as weighted graph • Biclustering • Generate module graph and explore associations between modules	– SAMBA	665 significant modules; Global Yeast molecular network	[90]
	Methanogenesis regulatory network			
Microarray (1661 methanogen genes under 58 conditions); upstream regions of all genes; operon prediction from MicrobesOnline; protein interactions from String	• Biclustering to Identify co-regulated gene subsets • Construct GRN to infer transcriptional influences of each bicluster • Visualize GRN • Use TF knockout experiment and extra data and to validate the GRN model	cMonkey Inferelator Cytoscape Gaggle –	166 biclusters; GRN model including a set of 1227 EF and TF regulatory influences that interlink the regulation of 1661 genes	[105]
	Mycobacterium tuberculosis regulatory network			
Microarray data (*M. tuberculosis* genes under 2325 conditions); upstream regions of all genes; ~5000 operon prediction from MicrobesOnline; ~250 000 protein interactions from String	• Biclustering to identify co-regulated gene subsets • Construct GRN model to infer transcriptional influences of each bicluster • Validate the GRN model using new data sets; visualize network	cMonkey Inferelator BioTapestry	598 biclusters; A global regulatory network covering 98% of MTB genes	[106]
	Phaeodactylum tricornutum regulatory network			
RNA-seq (1214 phaeodactylum tricornutum genes from 179 samples); genome annotation, chloroplastic and mitochondrial genomic information, functional annotation, PPIs	• Biclustering to identify putatively co-regulated genes • Construct regulatory network to infer regulatory influences • GO enrichment analysis to identify potential biological processes carried out by the co-regulated genes	cMonkey2 Inferelator –	121 biclusters covering 1214 metabolic genes and TFs	[107]
	Biological network decomposition			
Two gene interaction networks for yeast; two PPIs from *E. coli* and human	• Biclustering • Assess biological significance of retrieved modules	BicNET GOrilla	Modules with heightened biological significance	[37]
Yeast metabolic cycle expression matrix for 3553 genes under 12 time points; one yeast PPI network with 21 592 interactions among 4850 proteins	• Biclustering • Extract dynamic subnetworks from PPI • Detect protein complex	BiCAMWI – –	Protein complex	[111]

GRN: Gene regulatory network; MTB: Mycobacterium tuberculosis.

detecting disease subtype, identifying biomarker and gene signatures of disease and gene–drug association.

Disease subtype identification

Disease subtype could provide a framework for the development of more accurate biomarkers by stratification of patient populations [114]. It can be defined by related molecular characteristics or clinical features [115]. Gene expression data, depicted as a matrix with genes as columns, and subjects as rows (with known or unknown disease types), were widely used in molecular subtyping studies. This formulation is reasonable because pathways responding to specific disease subtypes may be activated across most the patients of the subtype, and the gene expression can be considered candidate signatures for subtypes [51]. With benchmark gene expression data sets and well-annotated disease subtype information, biclustering can discriminate biclusters from the gene expression matrix, containing genes that share similar expression patterns only in one or some specific subtypes [33, 116]. Hence, *de novo* identification of biclusters can be used to group subjects (patients) into disease subtypes, and these identified patient groups can be further evaluated by linking known clinical characteristics [117]. The evaluation process assumes that patients from different subtypes tend to have distinctive clinical features. In cancer subtyping study, survival time, neoplasm disease stage, tumor size, tumor grade, tumor nuclei percentage and patient age have been commonly used to assess the subtyping results [33, 117, 118]. Table 4 summed up those application studies in certain diseases, including leukemia, gastric cancer, breast cancer, lung cancer, etc.

For each characteristic, a dependence test, e.g. chi-square test, is used to examine the difference among all subtypes [119, 120]. To be specific, given a clinical characteristic (e.g. the presence of an adverse drug reaction), the null hypothesis of the test is that subtypes of a disease and the characteristic are independent, i.e. there are no differences among the subtypes regarding that characteristic. After summarizing the frequencies or counts of cases under different subtypes into a $r \times c$ contingency table

Table 4. Case studies of disease subtype identification

Data	Methods	Tools/databases	Outcomes	References
	Leukemia			
Microarray data with 12 533 probes from 72 patients of different subtypes of leukemia	• Biclustering by qualitative biclustering algorithm	QUBIC	Biclusters with cancer subtyping information	[33]
	Gastric cancer			
Microarray data for 80 paired gastric cancer and reference tissues from nontreated patients	• Biclustering on gene expression data for bicluster identification • Pathway enrichment analysis	QUBIC [33]; DAVID [122] KOBAS [123] HPID [124]	Pathways associated with cancer development; identified gastric cancer subtypes	[121]
	Breast cancer			
Microarray data with 7756 genes and matched clinical data for 437 primary breast tumor patients	• Adjust for cohort-correlated batch effect across the nonadjuvant-treated tumor data set • Biclustering to identify molecular-based tumor subgroup • Determine molecular classifiers for each bicluster	ComBat [125] cMonkey [126] PAM [127]	Similar clinical features associated with tumor within the same cluster	[118]
Microarray data with 17 814 genes across 547 samples and gene network consisted of 11 648 genes and 211 794 interactions	• Assign weights to genes based on impact in the network and expression variation • Weighted biclustering algorithm based on a semi-nonnegative matrix tri-factorization	PageRank [128] NCIS [117]	Cancer subtypes	[117]
	Colon and lung cancers			
290 colon cancer samples, each has 384 methylation probes covering 151 cancer-specific differentially methylated region Expression levels of 12 625 genes in 56 patients having lung cancer	• Heterogeneous sparse singular value decomposition-based biclustering	–	Variance biclusters of methylation data in cancer versus normal patients using colon cancer data cancer subtype patterns using lung cancer data	[116]

(r = number of rows, c = number of columns), the chi-square test statistic is calculated by using the formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E},$$

where O represents the observed frequency, and E represents the expected frequency under the null hypothesis, which is computed by:

$$E = \frac{\text{row total} \times \text{column total}}{\text{sample size}}.$$

The test statistics will be compared with the critical value of ${\chi_\alpha}^2(df = (r-1) \times (c-1))$. If $\chi^2 > {\chi_\alpha}^2$, the null hypothesis will be rejected, meaning that there are differences among subtypes regarding that characteristic (see details in Supplementary Example S1). Meanwhile, interpretation of the identified biclusters in gene dimension can be carried out, and more details of biomarker and gene signatures detection can be found in the next section.

Biomarker and gene signatures detection

Biclustering proved to be influential for mining information from elaborate biomedical data sets, especially in cancer research. Cancer is complicated because of the heterogeneity of tumor cells and is recognized as a system-level disease [129, 130]. Biclustering has been used with human gene expression data to identify cancer subtype patterns [33, 116–118, 131], metabolic pathways highly related to cancer progression [121], marker genes of a specific cancer type/subtype [95, 132] and clinical risk factors of cancer [133]. Also, studies of common or rare diseases have used biclustering of human gene expression data to identify phenotype–genotype associations [134, 135], dysregulated transcription modules [136] and genetic risk variants [137]. Depending on the available information, various levels of analyses can be conducted as summarized below.

Basically, given gene expression matrix with rows representing genes and columns representing patients, biclustering can identify co-expressed gene clusters that are specific to characteristics of patients, e.g. certain subtypes or disease stages. If genes included in the identified biclusters have differential expression patterns between different subtypes, then they can serve as candidate gene signatures or biomarkers for cancer staging and subtyping [121]. If predefined gene sets are given, and clinical characteristics/phenotype labels are also available, researchers can carry out gene set enrichment analysis (GSEA) first to investigate the correlation between gene sets and clinical characteristics/covariates (e.g. tumor grade, stage, age or hormone status). Based on these correlations results, a binary association matrix can be derived, with rows representing gene sets and columns

Table 5. Case studies of biomarker and gene signatures detection

Data	Methods	Tools/databases	Outcomes	References
	Breast cancer			
Association matrix of 1008 gene expression microarray profiles of primary breast tumors	• Biclustering binary data matrix	iBBiG	Modules associated with clinical covariates in breast cancer	[133]
Matrix of normalized miRNA-seq expression profiles	• Biclustering to evaluate miRNA deregulation • Validate each bicluster by an external repository of different groups of miRNAs in human species • Compare results with a different biclustering algorithm	ISA [31] MetaMirClust [142] UCSC [141] SAMBA [30]	12 different miRNA clusters	[131]
	Osteoporosis			
Regression coefficients matrix of 1109 unique SNPs associated with 23 studied traits from the GWAS data of the Framingham Osteoporosis Study	• GWAS database mining • Biclustering on matrix of SNPs against phenotypes • Gene annotation and identification of enriched canonical pathway and gene network inference	Tagger [143] Bayesian biclustering [144] UCSC [145] IPA	SNP–phenotype connections; highly genetically correlated traits; candidate genes identified for multiple bone traits	[134]
	Williams–Beuren syndrome			
Normalized skin fibroblast microarray data set including 9329 probe sets and 96 samples	• Identify transcriptional modules • Test modules containing at least 10 genes for dysregulation using hypergeometric distribution	ISA [31] –	72 dysregulated modules were found	[136]
	Schizophrenia			
8023 subjects, 4196 patients and 3827 controls, with 2891 SNPs in each subject	• Perform biclustering for both phenotype and genotype data • Cross-correlate phenotype and genotype biclusters • Organize and encode relations into topologically organized networks • Estimate genotype-associated disease risk	bioNMF [146] – PGMRA [135] SKAT [147]	Causally cohesive genotype–phenotype relations	[135]
	Complex diseases			
P-value matrix of 466 423 SNPs in 32 independent diseases/traits	• Identify biclusters of diseases/traits and SNPs • Map detected SNPs to genes	SparseBC [148] LAS [149] SSVD [150] –	Genetic risk variants for complex diseases	[137]

GWAS: Genome-wide association studies; PCW: plant cell-wall; CW: cell wall; MTB: Mycobacterium tuberculosis; ORF: open reading frames.

representing pairwise tests for phenotypes, the element '1' denoting significant association between gene set and pairwise test, and '0' denoting no significant association. Biclusters identified from this association matrix can represent modules that associated with known clinical covariates [133].

A matrix of SNPs or phenotypes and the extended matrices from them, including a matrix of regression coefficients of SNPs associated with traits and matrix of *P*-values of SNPs in traits, were subjected to biclustering to recognize the phenotype–genotype connections [134, 135, 137]. With the developments of RNA-seq, whole transcriptomic data are becoming available to characterize and quantify gene expression [138]. The recent advent of scRNA-seq technology has enabled researchers to study heterogeneity between individual cells and define cell type a based solely on its transcriptome [132]. Using biclustering, researchers can not only group cells into subpopulations but also identify biologically important gene signatures for each class simultaneously [95, 139]. For example, Zeisel *et al.* [95] recently classified single cells from the brain through biclustering, which identified numerous marker genes and highly restricted expression patterns of transcription factors for cell types. Kiselev *et al.* [132] developed a stable and accurate consensus tool, based on such scRNA-seq data, which can quantify the inherent heterogeneity of single cells, define the subclonal composition and identify marker genes [132]. Meanwhile, new biclustering applications are emerging, such as detecting disease marker genera from gut biome [140]. The gut microbiome is typically tricky to profile, and use of biclustering enhances identification of specific taxonomic signatures that can support the elucidation of disease risk [140].

These identified biclusters were subjected to downstream analysis of functional gene annotation [131, 134], gene network inference [134] or phenomic analysis [134, 135, 137]. Most of the gene functional annotations were done through the UCSC Genome Browser [141]. Gene networks among clustered genes were commonly constructed by the Ingenuity Pathways Analysis software developed by QIAGEN. Phenomic analysis performs pairwise genetic correlation of traits/phenotype against gene sets identified by biclustering, which is usually done using hypergeometric statistics or paired *t*-test. Table 5 gives an overview of biomarker/gene signature identification studies, with the detailed procedures regarding biclustering and accompanied analyses specified in the column 'Methods'. It is noteworthy that the application of biclustering in these biomedical studies is much more complicated

Table 6. Case studies of gene–drug association

Data	Methods	Tools/databases	Outcomes	References
	Drug–gene associations			
NCI-60 cancer cell line in drug response; gene expression data	• Identify co-modules of drugs and genes • Test drug–gene association	PPA [155] DrugBank [159] Connectivity Map [160]	859 co-modules were identified, and drug–gene associations were predicted more accurately than other algorithms	[155]
	Drug-induced transcriptional modules			
6100 gene expression profiles of human cancer cell treated with 1309 small molecules from CMap [160]1743 expression profiles from liver tissues of drug-treated rats [161]	• Biclustering drug-induced gene expression profiles [31] • Hypergeometric test for significance assessment of overlaps among gene members • Predict novel gene functions by comparing modules of human cancer and rat liver cell lines • Test enriched gene functions and identified biological themes among transcriptional modules	ISA [52] – STRING [156] DAVID [122]	Drug-induced transcriptional modules	[154]
	TFs for drug-associated gene modules			
7056 genome-wide expression profiles of five different human cell lines treated with 1309 chemical agents at different dosages from CMap [160]	• Identify drug–gene modules by biclustering method • Indicate GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) information associated with genes in modules • Use cumulative hypergeometric test to evaluate drug target enrichment	FABIA [34] DAVID [162] –	Links between 28 modules with 12 TFs were detected	[157]
	Transcriptomics and decision in early-stage of pharmaceutical drug discovery			
Transcriptomic profiles in eight drug discovery projects of oncology, virology, neuroscience and metabolic diseases	• Normalize and filtrate mRNA expression data • Identify transcriptional modules • Identify transcriptional modules related to the desired effect using target-related bioassay measurements	– FABIA [34] PSVM [163]	Transcriptional effects of compounds	[158]

compared with those in basic biological applications, regarding the data sources, data preprocessing methods and downstream statistical analyses.

Gene–drug association

In drug development, it is vital to understand the complicated responses in the human body to various drug treatments [151, 152]. However, rigorous testing of safety and efficacy of novel drug makes drug development time-consuming, expensive and often unsuccessful. Alternatively, computational drug repositioning is termed as an efficient way to identify new applications for current medicines [153]. By the advancement of biotechnologies, a significant amount of gene expression data becomes a paramount component in characterizing the human responses to drugs. Here, we review the applications of biclustering in the context that is considered appropriate in revealing the co-expression patterns encompassed in the drug-perturbed responses [154]. The genome-scale drug-treated gene expression data were served as raw materials for identification of co-expression modules using biclustering methods, where different drug treatments were conditions. Table 6 gave an overview of four typical studies that were examining the drug-induced co-expression modules. In these studies, information for both gene and drug members was mined to characterize the detected drug-induced modules. Conservation of identified biclusters was first evaluated across data sets through overlapping genes and drugs [154]. Then, genes and drugs in the bicluster were examined, respectively. Functional enrichment of these genes was tested using the DAVID knowledge base to determine the biological relevance of these biclusters [154, 155]. Enrichment of drug annotation terms can be assessed by various databases, such as STRING [156] and DAVID [122], for identification of TFs linked to these biclusters [154, 157, 158].

Conclusion and discussion

In summary, GBA is the basis of expression profile-based biclustering; however, co-expression does not guarantee co-regulation. One popular strategy to further elucidate co-regulation is to integrate supporting data that provide evidence of co-regulation with expression data, e.g. motif prediction and network connection. In support of a more comprehensive clarification of complex biological systems in a cell, existing biological network inference tools should embed multiple regulatory

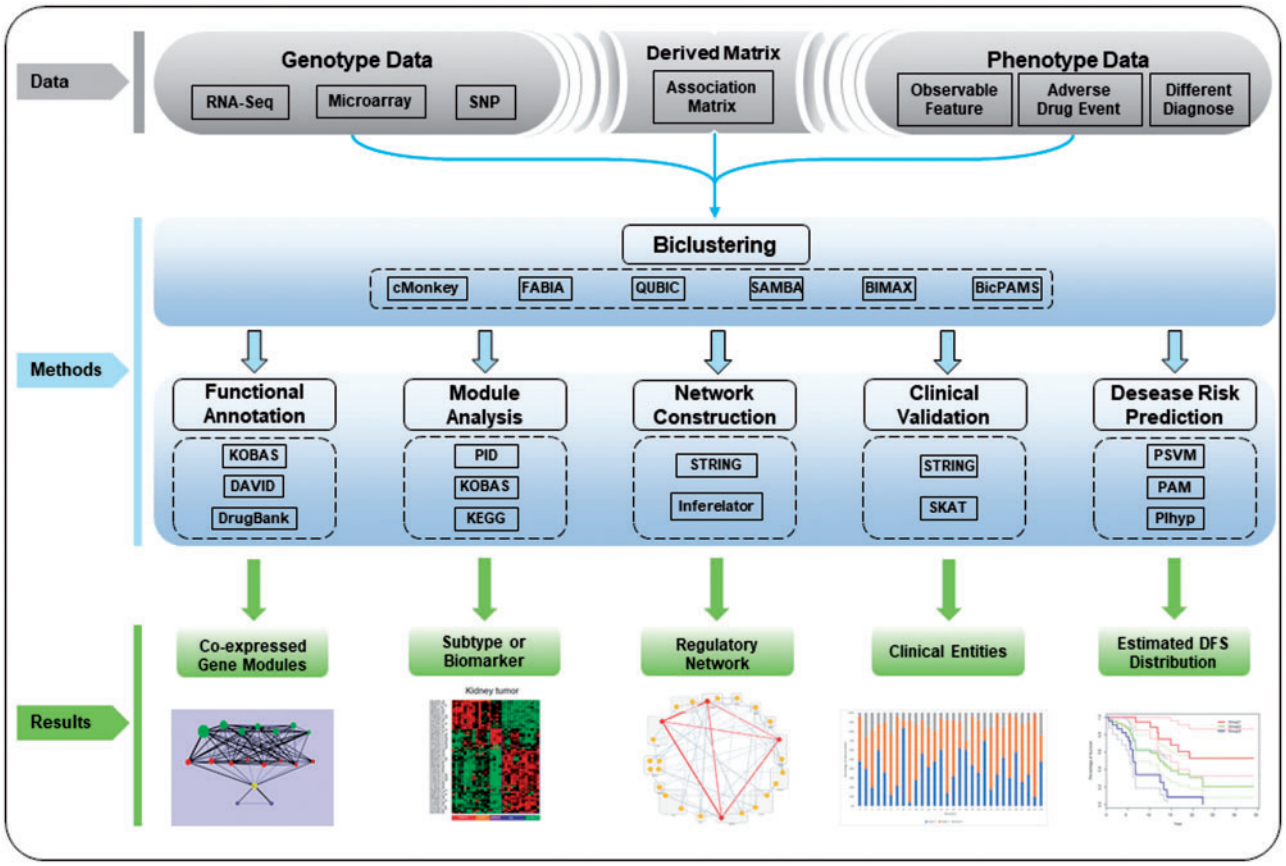

Figure 2. The overall workflow of biclustering application mechanism (related to upstream and downstream process). Three layers are shown to provide the path from raw data, appropriate analytical methods/tools to various cases of the result. The power of biclustering is illustrated by the ability to generate (from left to right in the figure) co-expressed gene modules, subtype or biomarker, regulatory networks, clinical entities and estimated disease free survival (DFS) distribution.

signals, e.g. TF, lncRNAs and miRNAs, and organically integrate biclustering within their network construction framework. Use of these methods and integration of well-annotated phenotypic data can enhance the identification of CEM and improve systems-level insights. Combination of biclustering of gene expression and clinical phenotype data with successive enrichment analyses has revealed disease subtype patterns and diseases biomarkers. Biclustering has contributed to drug development by exposing the co-expression patterns from the drug-treated gene expression data. Most uses of biclustering in biomedicine to date rely on a handful of conventional biclustering algorithms, as it remains unclear which are sufficiently accurate for any given data type.

A workflow of biclustering application is proposed here to integrate the methods and tools used in both biological and biomedical fields discussed above. As shown in Figure 2, there are three layers (Data, Methods and Results) in this workflow. The data sources in the first layer provide the information directly collected and derived from genotyping and phenotyping results. Different method combinations in layer two can be used for various analytical requirements. Biclustering can be used to analyze phenotype matrix, genotype matrix, as well as the derived association matrix of these two matrices. A few example tools were shown in the figure for biclustering methods, and a detailed table for the relevant tools can be found in Supplementary Table S1. These biclustering methods are often accompanied by downstream analysis, such as functional annotation, module analysis or network construction, to interpret the identified biclusters, together with statistical evaluation tools applied to demonstrate bicluster associations. Examples of results from a combination of the methods identified in layer two provide specific illustrations of corresponding outputs [33, 87, 118, 164, 165]. The connections between data and methods offer model analysis paths for researchers to use depending on the characteristics of their data.

The identified workflow guides many current studies; however, new biotechnologies are developing and emerging rapidly, while the corresponding biclustering tools are not evolving at a parallel pace. This situation is an important factor limiting the application of biclustering analysis to more complex data sets, e.g. multidimensional biological image data, requiring integration of multiple variables. Meanwhile, considering the variety and complexity of data from various platforms, the data integration and analyses are not trivial, and it is more challenge to combine multiple required computational techniques with biclustering analysis. Furthermore, different data types may need specifically designed biclustering algorithms. For example, scRNA-seq data exhibit higher heterogeneity than RNA-seq data and are increasing in popularity; however, few biclustering algorithms are explicitly designed for these new data. Hence, additional biclustering methods, which include specific design attributes taking into account the characteristics of biological and biomedical data, are still needed to facilitate larger-scale applications of biclustering.

Key Points

- This article provides a comprehensive review of the applications of biclustering in the biological and the biomedical fields.
- Biclustering has been widely used in GBA-based gene functional annotation. The documented functional information and the associations between annotated and unannotated genes are two kinds of essential information.
- Biclustering can be used for module identification. Depending on the to-be-identified modules, different information could be integrated with expression data. Once identified, further analysis of functional annotation, evolutional analysis and module network can be conducted.
- Biclustering analysis is often combined with network construction methods in module-based network inference, which facilitates the exploration of molecular mechanisms for biological process.
- With benchmark gene expression data sets and well-annotated disease subtype information, biclustering can group subjects/patients into disease subtypes, and dependence test can be applied to patient groups to investigate their clinical characteristics further.
- Biclustering of gene expression data in human yields biclusters of the subset of patients associated with a subset of genes, these genes are candidate biomarkers and the identified biclusters can provide other useful information like phenotype–genotype associations.
- Biclustering on drug-treated genome-wide expression data can recognize drug-induced modules. Following conservation analysis and enrichment analysis are often needed to verify gene–drug association.
- A workflow of biclustering application is generated, aiming to assist researchers to effectively derive biological knowledge and novel insights from their big data.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported by National Science Foundation/EPSCoR Award No. IIA-1355423, the State of South Dakota Research Innovation Center and the Agriculture Experiment Station of South Dakota State University (SDSU). Support for this project was also provided by the National Science Foundation of United States (grant number 1546869), the National Institutes of Health (U01 project, grant number 6U01HG007253-03) and Sanford Health–SDSU Collaborative Research Seed Grant Program. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (grant number ACI-1548562).

References

1. van Dijk EL, Auger H, Jaszczyszyn Y, *et al*. Ten years of next-generation sequencing technology. *Trends Genet* 2014;**30**(9):418–26.
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**(6):333–51.
3. Marioni JC, Mason CE, Mane SM, *et al*. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 2008;**18**(9):1509–17.
4. Miller JA, Menon V, Goldy J, *et al*. Improving reliability and absolute quantification of human brain microarray data by filtering and scaling probes using RNA-Seq. *BMC Genomics* 2014;**15**(1):154.
5. Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;**320**(5881):1344–9.
6. Luo J, Wu M, Gopukumar D, *et al*. Big data application in biomedical research and health care: a literature review. *Biomed Inform Insights* 2016;**8**:1–10.
7. Wu X, Zhu X, Wu G-Q, *et al*. Data mining with big data. *IEEE Trans Knowl Data Eng* 2014;**26**:97–107.
8. Swan M. The quantified self: fundamental disruption in big data science and biological discovery. *Big Data* 2013;**1**(2):85–99.
9. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**(1):57–63.
10. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011;**12**(2):87–98.
11. Garber M, Grabherr MG, Guttman M, *et al*. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* 2011;**8**(6):469–77.
12. Ulitsky I, Maron-Katz A, Shavit S, *et al*. Expander: from expression microarrays to networks and functions. *Nat Protoc* 2010;**5**(2):303–22.
13. Hartigan JA. Direct clustering of a data matrix. *J Am Stat Assoc* 1972;**67**(337):123–9.
14. Cheng Y, Church GM. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*. 2000;**8**:93–103.
15. Lazzeroni L, Owen A. Plaid models for gene expression data. *Stat Sin* 2002;**12**:61–86.
16. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013;**20**(1):117–21.
17. Burgel PR, Paillasseur JL, Roche N. Identification of clinical phenotypes using cluster analyses in COPD patients with multiple comorbidities. *Biomed Res Int* 2014;**2014**:1.
18. Han MK, Agusti A, Calverley PM, *et al*. Chronic obstructive pulmonary disease phenotypes: the future of COPD. *Am J Respir Crit Care Med* 2010;**182**:598–604.
19. Henriques R, Antunes C, Madeira SC. A structured view on pattern mining-based biclustering. *Pattern Recogn* 2015;**48**(12):3941–58.
20. Carreiro AV, Anunciacao O, Carrico JA, *et al*. Prognostic prediction through biclustering-based classification of clinical gene expression time series. *J Integr Bioinform* 2011;**8**:175.
21. Kluger Y, Basri R, Chang JT, *et al*. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res* 2003;**13**(4):703–16.
22. Murali T, Kasif S. Extracting conserved gene expression motifs from gene expression data. *Pac Symp Biocomput* 2003;**8**:77–88.
23. Gu J, Liu JS. Bayesian biclustering of gene expression data. *BMC Genomics* 2008;**9(Suppl 1)**:S4.
24. Chen Y, Mao F, Li G, *et al*. Genome-wide discovery of missing genes in biological pathways of prokaryotes. *BMC Bioinformatics* 2011;**12(Suppl 1)**:S1.
25. Zhou F, Ma Q, Li G, *et al*. QServer: a biclustering server for prediction and assessment of co-expressed gene clusters. *PLoS One* 2012;**7**(3):e32660.

26. Dhollander T, Sheng Q, Lemmens K, *et al*. Query-driven module discovery in microarray data. *Bioinformatics* 2007; **23**(19):2573–80.
27. De Smet R, Marchal K. An ensemble biclustering approach for querying gene expression compendia with experimental lists. *Bioinformatics* 2011;**27**(14):1948–56.
28. Zhao H, Cloots L, Van den Bulcke T, *et al*. Query-based biclustering of gene expression data using probabilistic relational models. *BMC Bioinformatics* 2011;**12(Suppl 1)**:S37.
29. Madeira SC, Oliveira AL. A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithms Mol Biol* 2009;**4**(1):8.
30. Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 2002;**18(Suppl 1)**:S136–44.
31. Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Stat Nonlin Soft Matter Phys* 2003;**67**(3 Pt 1):031902.
32. Prelić A, Bleuler S, Zimmermann P, *et al*. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 2006;**22**(9):1122–9.
33. Li G, Ma Q, Tang H, *et al*. QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res* 2009;**37**(15):e101.
34. Hochreiter S, Bodenhofer U, Heusel M, *et al*. FABIA: factor analysis for bicluster acquisition. *Bioinformatics* 2010;**26**(12): 1520–7.
35. Henriques R, Madeira SC. BiC2PAM: constraint-guided biclustering for biological data analysis with domain knowledge. *Algorithms Mol Biol* 2016;**11**(1):23.
36. Bunte K, Leppaaho E, Saarinen I, *et al*. Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics* 2016;**32**(16):2457–63.
37. Henriques R, Madeira SC. BicNET: flexible module discovery in large-scale biological networks using biclustering. *Algorithms Mol Biol* 2016;**11**(1):14.
38. Alzahrani M, Kuwahara H, Wang W, *et al*. Gracob: a novel graph-based constant-column biclustering method for mining growth phenotype data. *Bioinformatics* 2017;**33**(16): 2523–31.
39. Madeira SC, Teixeira MC, Sa-Correia I, *et al*. Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *IEEE/ACM Trans Comput Biol Bioinform* 2010;**7**(1):153–65.
40. Gonçalves JP, Madeira SC, Oliveira AL. BiGGEsTS: integrated environment for biclustering analysis of time series gene expression data. *BMC Res Notes* 2009;**2**(1):124.
41. Medina I, Carbonell J, Pulido L, *et al*. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res* 2010;**38(Suppl 2)**:W210–13.
42. Gonçalves JP, Madeira SC. Latebiclustering: efficient heuristic algorithm for time-lagged bicluster identification. *IEEE/ACM Trans Comput Biol Bioinform* 2014;**11**:801–13.
43. Henriques R, Madeira SC. BicPAM: pattern-based biclustering for biomedical data analysis. *Algorithms Mol Biol* 2014; **9**(1):27.
44. Henriques R, Ferreira FL, Madeira SC. BicPAMS: software for biological data analysis with pattern-based biclustering. *BMC Bioinformatics* 2017;**18**(1):82.
45. Bentham RB, Bryson K, Szabadkai G. MCbiclust: a novel algorithm to discover large-scale functionally related gene sets from massive transcriptomics data collections. *Nucleic Acids Res* 2017;**45**:8712–30.
46. Barkow S, Bleuler S, Prelic A, *et al*. BicAT: a biclustering analysis toolbox. *Bioinformatics* 2006;**22**(10):1282–3.
47. Cheng KO, Law NF, Siu WC, *et al*. BiVisu: software tool for bicluster detection and visualization. *Bioinformatics* 2007; **23**(17):2342–4.
48. Santamaria R, Theron R, Quintales L. BicOverlapper 2.0: visual analysis for gene expression. *Bioinformatics* 2014;**30**(12): 1785–6.
49. Wu CJ, Kasif S. GEMS: a web server for biclustering analysis of expression data. *Nucleic Acids Res* 2005;**33**:W596–9.
50. Kaiser S, Santamaria R, Theron R, *et al*. biclust: Bicluster algorithms. R package version 0.7 2009;2.
51. Zhang Y, Xie J, Yang J, *et al*. QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics* 2017;**33**:450–2.
52. Csardi G, Kutalik Z, Bergmann S. Modular analysis of gene expression data with R. *Bioinformatics* 2010;**26**(10): 1376–7.
53. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform* 2004;**1**(1):24–45.
54. Bozdağ D, Kumar AS, Catalyurek UV. Comparative analysis of biclustering algorithms. In: *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*. ACM, Niagara Falls, NY, USA, 2010, 265–274.
55. Chia BKH, Karuturi RKM. Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. *Algorithms Mol Biol* 2010;**5**(1):23.
56. Padilha VA, Campello RJ. A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics* 2017; **18**(1):55.
57. Li L, Guo Y, Wu W, *et al*. A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. *BioData Min* 2012;**5**(1):8.
58. Pontes B, Giraldez R, Aguilar-Ruiz JS. Biclustering on expression data: a review. *J Biomed Inform* 2015;**57**:163–80.
59. Busygin S, Prokopyev O, Pardalos PM. Biclustering in data mining. *Comput Oper Res* 2008;**35**(9):2964–87.
60. Eren K, Deveci M, Kucuktunc O, *et al*. A comparative analysis of biclustering algorithms for gene expression data. *Brief Bioinform* 2013;**14**(3):279–92.
61. Kasim A, Shkedy Z, Kaiser S, *et al*. *Applied Biclustering Methods for Big and High-Dimensional Data Using R*. Chapman & Hall/CRC, London, UK, 2016.
62. Yeung KY, Fraley C, Murua A, *et al*. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 2001;**17**(10):977–87.
63. Rau A, Maugis-Rabusseau C. Transformation and model choice for RNA-seq co-expression analysis. *Brief Bioinform* 2017, doi: 10.1093/bib/bbw128.
64. Bray NL, Pimentel H, Melsted P, *et al*. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;**34**(8):888–527.
65. Pachter L. Models for transcript quantification from RNA-Seq, *arXiv* preprint arXiv: 1104.3889 2011, in press.
66. Rau A, Maugis-Rabusseau C, Martin-Magniette ML, *et al*. Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics* 2015;**31**(9):1420–7.
67. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;**17**:75.
68. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* 2016;**17**:63.

69. Babu MM, Luscombe NM, Aravind L, *et al*. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 2004;**14**(3):283–91.
70. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
71. Gillis J, Pavlidis P. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Comput Biol* 2012;**8**(3):e1002444.
72. Consortium GO. Gene ontology consortium: going forward. *Nucleic Acids Res* 2015;**43**:D1049–56.
73. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**(D1):D353–61.
74. Gama-Castro S, Salgado H, Santos-Zavaleta A, *et al*. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res* 2016;**44**:D133–43.
75. Jin J, Tian F, Yang DC, *et al*. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 2017;**45**(D1):D1040–5.
76. Obayashi T, Kinoshita K, Nakai K, *et al*. ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Res* 2007;**35**:D863–9.
77. Yang MQ, Koehly LM, Elnitski LL. Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. *PLoS Comput Biol* 2007;**3**(4):e72.
78. Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci USA* 2006;**103**(47):17973–8.
79. Mezey JG, Nuzhdin SV, Ye F, *et al*. Coordinated evolution of co-expressed gene clusters in the *Drosophila* transcriptome. *BMC Evol Biol* 2008;**8**(1):2.
80. Ma Q, Yin Y, Schell MA, *et al*. Computational analyses of transcriptomic data reveal the dynamic organization of the *Escherichia coli* chromosome under different conditions. *Nucleic Acids Res* 2013;**41**(11):5594–603.
81. Castillo-Davis CI, Hartl DL. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* 2003;**19**(7):891–2.
82. Guttman M, Amit I, Garber M, *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009;**458**:223–7.
83. Huang Y, Li H, Hu H, *et al*. Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics* 2007;**23**(13):i222–9.
84. Chen X, Ma Q, Rao X, *et al*. Genome-scale identification of cell-wall-related genes in switchgrass through comparative genomics and computational analyses of transcriptomic data. *Bioenergy Res* 2016;**9**(1):172–80.
85. Horan K, Jang C, Bailey-Serres J, *et al*. Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol* 2008;**147**(1):41–57.
86. Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. *Nat Biotechnol* 2014;**32**(5):447–52.
87. Wang S, Yin Y, Ma Q, *et al*. Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis. *BMC Plant Biol* 2012;**12**(1):138.
88. Cherry JM, Adler C, Ball C, *et al*. SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 1998;**26**(1):73–9.
89. Wagner GP, Pavlicev M, Cheverud JM. The road to modularity. *Nat Rev Genet* 2007;**8**(12):921–31.
90. Tanay A, Sharan R, Kupiec M, *et al*. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc Natl Acad Sci USA* 2004;**101**(9):2981–6.
91. Purnick PE, Weiss R. The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol* 2009;**10**(6):410–22.
92. Zhang J, Le TD, Liu L, *et al*. Identifying miRNA sponge modules using biclustering and regulatory scores. *BMC Bioinformatics* 2017;**18**(S3):44.
93. Bryan K, Terrile M, Bray IM, *et al*. Discovery and visualization of miRNA-mRNA functional modules within integrated data using bicluster analysis. *Nucleic Acids Res* 2014;**42**(3):e17.
94. Wilson CM, Yang S, Rodriguez M, *et al*. Clostridium thermocellum transcriptomic profiles after exposure to furfural or heat stress. *Biotechnol Biofuels* 2013;**6**(1):131.
95. Zeisel A, Munoz-Manchado AB, Codeluppi S, *et al*. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 2015;**347**(6226):1138–42.
96. Huttenhower C, Mutungu KT, Indik N, *et al*. Detailing regulatory networks through large scale data integration. *Bioinformatics* 2009;**25**(24):3267–74.
97. Reiss DJ, Plaisier CL, Wu WJ, *et al*. cMonkey2: automated, systematic, integrated detection of co-regulated gene modules for any organism. *Nucleic Acids Res* 2015;**43**(13):e87.
98. Yang J, Worley E, Ma Q, *et al*. Nitrogen remobilization and conservation, and underlying senescence-associated gene expression in the perennial switchgrass Panicum virgatum. *New Pythol* 2016;**211**:75–89.
99. Waltman P, Kacmarczyk T, Bate AR, *et al*. Multi-species integrative biclustering. *Genome Biol* 2010;**11**(9):R96.
100. Yang R, Wang X. Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns. *Plant Cell* 2013;**25**(1):71–82.
101. Gonçalves JP, Aires RS, Francisco AP, *et al*. Regulatory snapshots: integrative mining of regulatory modules from expression time series and regulatory networks. *PLoS One* 2012;**7**(5):e35977.
102. MacPherson JI, Dickerson JE, Pinney JW, *et al*. Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Comput Biol* 2010;**6**(7):e1000863.
103. De Smet R, Marchal K. Advantages and limitations of current network inference methods. *Nat Rev Microbiol* 2010;**8**(10):717–29.
104. Wang X, Dalkic E, Wu M, *et al*. Gene-module level analysis: identification to networks and dynamics. *Curr Opin Biotechnol* 2008;**19**(5):482–91.
105. Yoon SH, Turkarslan S, Reiss DJ, *et al*. A systems level predictive model for global gene regulation of methanogenesis in a hydrogenotrophic methanogen. *Genome Res* 2013;**23**(11):1839–51.
106. Peterson EJ, Reiss DJ, Turkarslan S, *et al*. A high-resolution network model for global gene regulation in *Mycobacterium tuberculosis*. *Nucleic Acids Res* 2014;**42**(18):11291–303.
107. Levering J, Dupont CL, Allen AE, *et al*. Integrated regulatory and metabolic networks of the marine diatom phaeodactylum tricornutum predict the response to rising CO2 levels. *mSystems* 2017;**2**(1):e00142-16.
108. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol* 2007;**3**:88.

109. Liu G, Wang H, Chu H, *et al*. Functional diversity of topological modules in human protein-protein interaction networks. *Sci Rep* 2017;**7**(1):16199.
110. Zhang Y, Xuan J, de los Reyes BG, *et al*. Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data. *BMC Bioinformatics* 2008;**9**(1):203.
111. Lakizadeh A, Jalili S. BiCAMWI: a genetic-based biclustering algorithm for detecting dynamic protein complexes. *PLoS One* 2016;**11**(7):e0159923.
112. Lewis CM, Knight J. Introduction to genetic association studies. *Cold Spring Harb Protoc* 2012;**2012**(3):pdb.top068163.
113. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. *Clin Pharmacol Ther* 2016;**99**(3): 285–97.
114. Starmans MH, Boutros PC. Biomarkers and subtypes of cancer. *Aging* 2015;**7**(5):280–1.
115. Wang M, Spiegelman D, Kuchiba A, *et al*. Statistical methods for studying disease subtype heterogeneity. *Stat Med* 2016; **35**(5):782–800.
116. Chen G, Sullivan PF, Kosorok MR. Biclustering with heterogeneous variance. *Proc Natl Acad Sci USA* 2013;**110**(30): 12253–8.
117. Liu Y, Gu Q, Hou JP, *et al*. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics* 2014;**15**(1):37.
118. Wang YK, Print CG, Crampin EJ. Biclustering reveals breast cancer tumour subgroups with common clinical features and improves prediction of disease recurrence. *BMC Genomics* 2013;**14**(1):102.
119. Yeoh E-J, Ross ME, Shurtleff SA, *et al*. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002;**1**(2):133–43.
120. Parise CA, Bauer KR, Brown MM, *et al*. Breast cancer subtypes as defined by the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2) among women with invasive breast cancer in California, 1999-2004. *Breast J* 2009;**15**(6):593–602.
121. Cui J, Chen Y, Chou WC, *et al*. An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. *Nucleic Acids Res* 2011;**39**(4):1197–207.
122. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**(1):44–57.
123. Wu J, Mao X, Cai T, *et al*. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res* 2006;**34**:W720–4.
124. Schaefer CF, Anthony K, Krupa S, *et al*. PID: the pathway interaction database. *Nucleic Acids Res* 2009;**37**:D674–9.
125. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**(1):118–27.
126. Reiss DJ, Baliga NS, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics* 2006;**7**:280.
127. Tibshirani R, Hastie T, Narasimhan B, *et al*. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA* 2002;**99**(10):6567–72.
128. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. *Comput Netw ISDN Syst* 1998;**30**(1–7): 107–17.
129. Swanton C. Intratumor heterogeneity: evolution through space and time. *Cancer Res* 2012;**72**(19):4875–82.
130. Bedard PL, Hansen AR, Ratain MJ, *et al*. Tumour heterogeneity in the clinic. *Nature* 2013;**501**(7467):355–64.
131. Fiannaca A, La Rosa M, La Paglia L, *et al*. Analysis of miRNA expression profiles in breast cancer using biclustering. *BMC Bioinformatics* 2015;**16(Suppl 4)**:S7.
132. Kiselev VY, Kirschner K, Schaub MT, *et al*. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017; **14**(5):483–6.
133. Gusenleitner D, Howe EA, Bentink S, *et al*. iBBiG: iterative binary bi-clustering of gene sets. *Bioinformatics* 2012;**28**(19): 2484–92.
134. Gupta M, Cheung CL, Hsu YH, *et al*. Identification of homogeneous genetic architecture of multiple genetically correlated traits by block clustering of genome-wide associations. *J Bone Miner Res* 2011;**26**(6):1261–71.
135. Arnedo J, del Val C, de Erausquin GA, *et al*. PGMRA: a web server for (phenotype x genotype) many-to-many relation analysis in GWAS. *Nucleic Acids Res* 2013;**41**(W1):W142–9.
136. Henrichsen CN, Csárdi G, Zabot M-T, *et al*. Using transcription modules to identify expression clusters perturbed in Williams-Beuren syndrome. *PLoS Comput Biol* 2011;**7**(1): e1001054.
137. Teng B, Yang C, Liu J, *et al*. Exploring the genetic patterns of complex diseases via the integrative genome-wide approach. *IEEE/ACM Trans Comput Biol Bioinform* 2016;**13**(3): 557–64.
138. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**(7):621–8.
139. Shi F, Huang H. Identifying cell subpopulations and their genetic drivers from single-cell RNA-seq data using a biclustering approach. *J Comput Biol* 2017;**24**(7):663–74.
140. Falony G, Joossens M, Vieira-Silva S, *et al*. Population-level analysis of gut microbiome variation. *Science* 2016;**352**(6285): 560–4.
141. Fujita PA, Rhead B, Zweig AS, *et al*. The UCSC genome browser database: update 2011. *Nucleic Acids Res* 2011;**39**: D876–82.
142. Chan W-C, Ho M-R, Li S-C, *et al*. MetaMirClust: discovery of miRNA cluster patterns using a data-mining approach. *Genomics* 2012;**100**(3):141–8.
143. Barrett JC, Fry B, Maller J, *et al*. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; **21**(2):263–5.
144. Liang F, Wong WH. Evolutionary Monte Carlo: applications to C p model sampling and change point problem. *Stat Sin* 2000;**10**(2):317–42.
145. Kent WJ, Sugnet CW, Furey TS, *et al*. The human genome browser at UCSC. *Genome Res* 2002;**12**(6):996–1006.
146. Pascual-Montano A, Carmona-Saez P, Chagoyen M, *et al*. bioNMF: a versatile tool for non-negative matrix factorization in biology. *BMC Bioinformatics* 2006;**7**:366.
147. Wu MC, Kraft P, Epstein MP, *et al*. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 2010;**86**(6):929–42.
148. Tan KM, Witten DM. Sparse biclustering of transposable data. *J Comput Graph Stat* 2014;**23**(4):985–1008.
149. Shabalin AA, Weigman VJ, Perou CM, *et al*. Finding large average submatrices in high dimensional data. *Ann Appl Stat* 2009;**3**(3):985–1012.
150. Lee M, Shen H, Huang JZ, *et al*. Biclustering via sparse singular value decomposition. *Biometrics* 2010;**66**(4):1087–95.
151. Drews J. Drug discovery: a historical perspective. *Science* 2000;**287**(5460):1960–4.

152. Evans WE, McLeod HL. Pharmacogenomics–drug disposition, drug targets, and side effects. *N Engl J Med* 2003;**348**(6):538–49.
153. Rutherford KD, Mazandu GK, Mulder NJ. A systems-level analysis of drug-target-disease associations for drug repositioning. *Brief Funct Genomics* 2018;**17**(1):34.
154. Iskar M, Zeller G, Blattmann P, *et al*. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Mol Syst Biology* 2013;**9**:662.
155. Kutalik Z, Beckmann JS, Bergmann S. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat Biotechnol* 2008;**26**(5):531–9.
156. Szklarczyk D, Franceschini A, Kuhn M, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011;**39**:D561–8.
157. Xiong M, Li B, Zhu Q, *et al*. Identification of transcription factors for drug-associated gene modules and biomedical implications. *Bioinformatics* 2014;**30**(3):305–9.
158. Verbist B, Klambauer G, Vervoort L, *et al*. Using transcriptomics to guide lead optimization in drug discovery projects: lessons learned from the QSTAR project. *Drug Discov Today* 2015;**20**(5):505–13.
159. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006;**34**:D668–72.
160. Lamb J, Crawford ED, Peck D, *et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**(5795):1929–35.
161. Natsoulis G, Pearson CI, Gollub J, *et al*. The liver pharmacological and xenobiotic gene response repertoire. *Mol Syst Biol* 2008;**4**:175.
162. Dennis G, Jr., Sherman BT, Hosack DA, *et al*. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;**4**:P3.
163. Hochreiter S, Obermayer K. Support vector machines for dyadic data. *Neural Comput* 2006;**18**(6):1472–510.
164. Yang J, Chen X, McDermaid A, *et al*. DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses. *Bioinformatics* 2017;**33**:2586–8.
165. Liu B, Zhou C, Li G, *et al*. Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses. *Sci Rep* 2016;**6**(1):23030.