South Dakota State University

Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange

Electronic Theses and Dissertations

2018

Development of Biclustering Techniques for Gene Expression Data Modeling and Mining

Juan Xie

South Dakota State University

Follow this and additional works at: https://openprairie.sdstate.edu/etd

Part of the Genetics and Genomics Commons, and the Statistics and Probability Commons

Recommended Citation

Xie, Juan, "Development of Biclustering Techniques for Gene Expression Data Modeling and Mining" (2018). *Electronic Theses and Dissertations*. 2960.
https://openprairie.sdstate.edu/etd/2960

This Thesis - Open Access is brought to you for free and open access by Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Open PRAIRIE: Open Public Research Access Institutional Repository and Information Exchange. For more information, please contact michael.biondo@sdstate.edu.

DEVELOPMENT OF BICLUSTERING TECHNIQUES FOR GENE EXPRESSION DATA MODELING AND MINING

BY

JUAN XIE

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Statistics

South Dakota State University

2018

DEVELOPMENT OF BICLUSTERING TECHNIQUES FOR GENE EXPRESSION DATA MODELING AND MINING

JUAN XIE

This thesis is approved as a creditable and independent investigation by a candidate for the Master of Science in Statistics degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this thesis does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Qin Ma, PhD.
Thesis Advisor Date

Kurt Cogswell, PhD.
Head, Mathematics & Statistics Date

Dean, Graduate School Date

ACKNOWLEDGEMENTS

First and foremost, endless thanks to my advisor, Dr. Qin Ma, who encouragingly guided me through the course of this dissertation. Thanks for providing such an excellent opportunity to work in your lab. I still remembered that at the beginning I knew nothing about programming and often felt frustrated, he gave me enough time to learn from the scratch and always acknowledged my every little progress, thanks for the patience and encouragement. In the past two years, he taught me how to learn a new skill, how to give presentations, how to read and write papers, and all the necessary and essential quantities that a graduate student should have.

Besides my advisor, I would like to thank the rest of my committees: Prof. Anne Fennell, Dr. Xijin Ge, and Dr. David Wiltse. Thank you for being my committees, thanks for your precious time and valuable comments. Thank you, Prof. Fennell, for sponsoring me to the PAG conference, which enabled me to know excellent research work and presentations. Thank you, Dr. Ge, for pointing out my stupid and fatal mistake on enrichment analysis. I might never realize that mistake and got the wrong results.

My sincere thanks also go the all the BMBL members. Adam gave me many insightful pieces of advice and comments regarding my research and presentations, and he showed me an example of a excellent presenter and project leader; Jingyu’s high-quality figures inspired me a lot; Anjun, Cankun, ShaoPeng, Minxuan, Yiran, Yirong, and Xiaozhu, thank you all for your support and comments.

Finally, infinite gratitude to my family. Mom, thank you for traveling tens of thousands of miles to take care of Helen and me during the first few months after my delivery. Without your support and encouragement, it might be impossible for me to return to school. My husband, Fangjun Li gave me unfailing support and continuous encouragement through the process of researching and writing this thesis. He taught me how to program and told me many good programming habits. He spent much time on the housework and accompanied our child a lot, even when he had his dissertation to finish. Special thanks to my little sweetheart, Helen, you bring me sunshine and laugh.

CONTENTS

LISTS OF FIGURES viii

LIST OF TABLES x

ABSTRACT xi

CHAPTER 1: Introduction 1

- **1.1 Gene Expression Data** 1
- **1.2 Biclustering Techniques** 2
- **1.3 QUBIC** 8
- **1.4 Qserver** 10

CHAPTER 2: QUBIC2—A Novel Biclustering Algorithm for Large-scale RNA-Seq Data Analysis 13

- **2.1 Overall Design of QUBIC2** 13
- **2.2 Detailed Methods in QUBIC2** 17
    - 2.2.1 Left Truncated Mixed Gaussian (LTMG) Model and Qualitative Representation 17
    - 2.2.2 KL Score 19
    - 2.2.3 QUBIC2 Algorithm 20
    - 2.2.4 Size-based *P*-value 21
- **2.3 Functional Gene Modules Detection from RNA-Seq Data** 22
    - 2.3.1 Data Acquisition 22

2.3.2 Simulation of Co-regulated Gene Expression Data 23

2.3.3 Evaluation of Functional Modules 25

2.3.4 Biclustering Parameters 26

2.3.5 Results 27

2.4 A Statistical Evaluation Framework for Identified Biclusters 31

2.4.1 Methods 31

2.4.2 Results 32

2.5 Cell Type Classification Based on scRNA-Seq Data 34

2.5.1 Cell Type Classification Pipeline 34

2.5.2 Data, Biclustering Parameters and Evaluation Criteria 35

2.5.3 Results 38

2.6 Application of QUBIC2 on Temporal and Spatial scRNA-Seq Data 39

2.6.1 Data 39

2.6.2 Results 40

2.7 Summary 44

CHAPTER 3: QUBICR- A Biconductor Package for Qualitative Biclustering Analysis of Gene Co-expression Data 47

3.1 Implementation 49

3.2 Functions 50

3.3 Summary 56

CHAPTER 4: Application of Biclustering on Biological and Biomedical Data 57

4.1 Basic Application of Biclustering on Biological Data 60

4.1.1 Functional Annotation of Unclassified Genes... 61

4.1.2 Modularity Analysis 64

4.1.3 Biological Networks Elucidation.. 68

4.2 Advanced Application of Biclustering in Biomedical Science........................... 71

4.2.1 Disease Subtype Identification 72

4.2.2 Biomarker and Gene Signatures Detection 74

4.2.3 Gene-drug Association 78

4.3 Summary.. 80

REFERENCES 84

APPENDIX 1: QUBIC2 tutorial.. 96

APPENDIX 2: Data Links 100

APPENDIX 3: Citation Map for QUBIC 101

APPENDIX 4: Plan of Study.. 102

APPENDIX 5: Curriculum Vitae... 103

LISTS OF FIGURES

Figure 1. Workflow of QUBIC.. 10

Figure 2. An example workflow of using Qserver. 12

Figure 3. QUBIC2 workflow. . 15

Figure 4. Overall performance comparison between QUBIC2 and five popular biclustering methods based on the agreement between identified biclusters and known modules. 29

Figure 5.A. The distribution of correlation coefficients(ρ) between P-value obtained from enrichment analysis and size-based P-value.; B. Scatter plot of ρ and p-value.. 32

Figure 6. The relationship between biclustering parameter f and correlation coefficient. 33

Figure 7. A. Computational pipeline for cell type classification.; B. Benchmark of QUBIC2 against five popular biclustering algorithms; C. Sankey diagram. 39

Figure 8. A. Visualization of three biclusters (BC002, BC003, and BC004) selected based on the specificity to time point; B. Time-dependent distribution of cells in six selected biclusters identified in the LPS data.. 42

Figure 9. A. The coordinates of cells correspond to five morphological layers; B. The coordinates of cells from three selected biclusters; C. The spatial coordinates of samples in the four biclusters identified in wild-type 1 mouse; D. In addition to the coordinates of bicluster samples, the yellow cubes represent significant outlier samples; E. The same information as in C except the samples are from wild-type 2 mouse; F. The same information as in D except the samples are from wild-type 2 mouse. 43

Figure 10. Performance of QUBIC2, QUBIC, two decomposition methods and two clustering methods in term of F score on a human scRNA-Seq data. 46

Figure 11. A. Comparison of QUBIC-R and 6 R packages in biclust; B. Comparison of running time among four recommended programs; C. Heatmap visualization of two biclusters identified in E. coli data; D. Co-expression networks of Figure 11C biclusters. 48

Figure 12. Data limit test of QUBICR on simulated datasets.. 50

Figure 13. Heatmap for the 4th bicluster identified in the E. coli data............................ 54

Figure 14. Network for the 4th bicluster identified in the E. coli data. 55

Figure 15. Yearly comparison of biclustering algorithm development and algorithm application related studies.. 60

Figure 16. The overall workflow of biclustering application mechanism related to upstream and downstream process.. 82

Figure 17. Citation map for QUBIC 101

LIST OF TABLES

Table 1. Summary of biclustering algorithms and tools, sorted in the decreasing order of their numbers of citation since published 5
Table 2. Main parameter adjusted for each algorithm 26
Table 3. Parameter ranges for each biclustering algorithm used in the cell type classification section 35
Table 4. Summary of GSE60402 40
Table 5. Case studies of Functional annotation of unclassified genes 63
Table 6. Case studies of Modularity analysis. 66
Table 7. Case studies of Biological networks elucidation. 70
Table 8. Case studies of disease subtype identification. 73
Table 9. Case studies of Biomarker and gene signatures detection. 77
Table 10. Case studies of gene-drug association. 79
Table 11. Data links 100

ABSTRACT

DEVELOPMENT OF BICLUSTERING TECHNIQUES FOR GENE EXPRESSION DATA MODELING AND MINING

JUAN XIE

2018

The next-generation sequencing technologies can generate large-scale biological data with higher resolution, better accuracy, and lower technical variation than the array-based counterparts. RNA sequencing (RNA-Seq) can generate genome-scale gene expression data in biological samples at a given moment, facilitating a better understanding of cell functions at genetic and cellular levels. The abundance of gene expression datasets provides an opportunity to identify genes with similar expression patterns across multiple conditions, i.e., co-expression gene modules (CEMs). Genome-scale identification of CEMs can be modeled and solved by biclustering, a two-dimensional data mining technique that allows clustering of rows and columns in a gene expression matrix, simultaneously. Compared with traditional clustering that targets global patterns, biclustering can predict local patterns. This unique feature makes biclustering very useful when applied to big gene expression data since genes that participate in a cellular process are only active in specific conditions, thus are usually co-expressed under a subset of all conditions.

The combination of biclustering and large-scale gene expression data holds promising potential for condition-specific functional pathway/network analysis. However, existing biclustering tools do not have satisfied performance on high-resolution

RNA-Seq data, majorly due to the lack of (i) a consideration of high sparsity of RNA-Seq data, especially for scRNA-Seq data, and (ii) an understanding of the underlying transcriptional regulation signals of the observed gene expression values. QUBIC2, a novel biclustering algorithm, is designed for large-scale bulk RNA-Seq and single-cell RNA-seq (scRNA-Seq) data analysis. Critical novelties of the algorithm include (i) used a truncated model to handle the unreliable quantification of genes with low or moderate expression; (ii) adopted the Gaussian mixture distribution and an information-divergency objective function to capture shared transcriptional regulation signals among a set of genes; (iii) utilized a Dual strategy to expand the core biclusters, aiming to save dropouts from the background; and (iv) developed a statistical framework to evaluate the significances of all the identified biclusters. Method validation on comprehensive data sets suggests that QUBIC2 had superior performance in functional modules detection and cell type classification. The applications of temporal and spatial data demonstrated that QUBIC2 could derive meaningful biological information from scRNA-Seq data.

Also presented in this dissertation is QUBICR. This R package is characterized by an 82% average improved efficiency compared to the source C code of QUBIC. It provides a set of comprehensive functions to facilitate biclustering-based biological studies, including the discretization of expression data, query-based biclustering, bicluster expanding, biclusters comparison, heatmap visualization of any identified biclusters, and co-expression networks elucidation.

In the end, a systematical summary is provided regarding the primary applications of biclustering for biological data and more advanced applications for biomedical data. It

will assist researchers to effectively analyze their big data and generate valuable biological knowledge and novel insights with higher efficiency.

CHAPTER 1: Introduction

1.1 Gene Expression Data

Gene expression is the process by which information from a gene is used in the synthesis of a functional product, that is, a molecule needed to perform a job in the cell (e.g., protein). The process mainly consists of two steps: transcription and translation. In transcription, the DNA sequence of a gene is copied to make an RNA molecule. In translation, the sequence of the mRNA is decoded to specify the amino acid sequence of a polypeptide. Since genes encode proteins and proteins dictate cell functions, the genes expressed in a cell determine what the cell can do.

Many biotechnologies are available to profile gene expression. Microarrays emerged in the late 1990s, which is the first high-throughput technology that enables the researchers to monitor the expression level of tens of thousands of genes simultaneously [1]. Microarrays are typically microscope slides that are printed with thousands of tiny spots in ordered positions, with each spot containing a known DNA sequence or gene. After steps of mRNA extraction, cDNA synthesis, cDNA fragmentation, and fluorescent labeling, the relative abundance of genes is quantified by detecting fluorescent intensity, which is continuous and positive. Due to its easy accessibility and low cost, microarrays have been the most widely used platforms in generating gene expression. However, microarrays need a reference genome and transcriptome to be available; thus, their application is confined to organisms whose genome have already been sequenced.

With the advent of massively parallel sequencing, next-generation sequencing (NGS) technologies have become more affordable. Compared to the array-based counterparts, NGS has higher resolution, better accuracy, lower technical variation and many other advantages [2, 3]. It allows for a much faster-paced accumulation of large-scale biological data. The high-throughput RNA-sequencing (RNA-Seq) is a revolutionary technology for gene expression profiling and promises a comprehensive picture of the transcriptome for a biological process [4, 5]. Unlike microarrays, RNA-Seq can be used to new organisms whose genome has not been sequenced yet. It extracts usable information from the mature mRNA within a biological source and generates a massive number of short segments (reads, 100-250 bps), which enable the discrete quantification of all genes expressed in a cell [5, 6]. Currently, researchers can either analyze a large sample of cells from a single organism in the form of bulk RNA-Seq data or isolate individual cells from complex organisms and measure their transactional activity through single-cell RNA-sequencing (scRNA-Seq). Such gene expression data from individual cells promises to provide a better understanding of cell functions at genetic and cellular levels[7] . In short, these biotechnologies have generated large genome-scale gene expression data in the public domain, and their tremendous values have been confirmed in many research areas such as elucidation of cell-type-specific regulatory networks [8, 9] and cancer & complex diseases studies [10-12].

1.2 Biclustering Techniques

The abundance of gene expression datasets provides an opportunity to identify genes with similar expression patterns across multiple conditions, i.e., co-expression gene modules (CEMs). The genes in these modules tend to be functionally related or co-regulated by the same transcriptional regulatory signals (TRSs). Thus, they enable the higher-level interpretation of gene expression data, improve functional annotation, facilitate inference of gene regulatory mechanisms, and are useful for a better understanding of disease/cancer mechanisms. Genome-scale identification of CEMs can be modeled and solved by biclustering [13], which was introduced by Hartigan in 1972 [14] and applied to gene expression data analysis by Cheng and Church in 2000 [15]. 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 [15] and MSE used in [16]). This unique feature makes it very useful when applied to big gene expression data since genes that participate in a cellular process are only active in specific conditions, thus are usually co-expressed under a subset of all conditions.

Besides the identification of CEMs, scRNA-Seq data enables studies of individual cells or cell types as well as their complex interactions under specific stimuli, e.g., cell types classification and clustering. In multicellular organisms, biological function emerges when various cell types form complex organs [17]. Investigations into organ development, cell function, and disease mechanisms highly depend upon accurate identification and categorization of cell types, sometimes along with their temporal and spatial features [18]. Traditionally, cell type was defined based on morphological properties or marker proteins, yet this method failed to characterize the full diversity of cells. scRNA-Seq data provides the possibility to group cells based on their genome-wide transcriptome profiles, and several studies have already been carried out using scRNA-Seq data to identify novel cell types, proving its power to unravel the full diversity of cells in human and mouse [8]. Mathematically, the problem of cell types classification can be treated as biclustering problems, as the essence is to find sub-populations of cells sharing common expression patterns among subsets of genes.

A substantial number of biclustering methods were developed during the past 18 years [15, 16, 19-36]. SAMBA [28], ISA [29], Bimax [30], QUBIC [31], and FABIA [37] are some popular algorithms for general purpose. CCC-biclustering [38-40] is designed for temporal data analysis, and BicPAM [41], BicNET [35, 42] and MCbiclust [43] are three recent studies. Besides, several tools (R packages, web servers, etc.) have been developed to facilitate users with a limited computational background [23, 44-50]. GEMS [47] is a web server for gene expression mining based on a Gibbs sampling paradigm; and biclust [48] and QUBICR [49] are two R packages integrating multiple existing algorithms, data preprocessing functions, and interpretation & visualization of the results. A list of some highly cited or recently published biclustering algorithms and tools is shown in Table1.

Table 1. Summary of biclustering algorithms and tools, sorted in the decreasing order of their numbers of citation since published. Application or usage was noted for some of the algorithms. Citations were collected via Google Scholar as of Sep 2018

Algorithms/ Tools	**Citations***	**Published Year**	**Review comments***	**Notes**
SAMBA [28]	939	2002	-	-
Bimax[30]	874	2006	[E] Choice for constant-upregulated biclusters	-
ISA [29]	414	2002	[E] Choice for constant-upregulated biclusters; [NC] Performs well on synthetic data	-
Plaid[16]	717	2002	[E] Choice for constant-upregulated biclusters; Has the highest enriched bicluster ratio in real datasets	
Spectral[51]	654	2003	[NC] Performs well on human and synthetic data	
cMonkey[52] / cMonkey2 [53]	257/ 21	2006/ 2015	-	Integrates various orthogonal pieces of information which support evidence of gene co-regulation, and optimizes biclusters to be supported simultaneously by one or more of these prior constraints
FABIA [32]	198	2010	[E] Choice for constant-upregulated biclusters; [NC] Performs well on synthetic data	-
SSVD [54]	192	2010	-	-

QUBIC [31]	167	2009	E Choice for constant-upregulated biclusters; Has the highest enriched bicluster ratio in real datasets NC Performs well on synthetic and human data	
BBC[55]	126	2008	E Best one for plaid biclusters	-
CPB[56]	40	2009	E Best one for constant, scale, shift and shift-scale datasets	
LAS [57]	113	2009	-	Discovery of biologically relevant structures in high dimensional data; Significant results highlighted with a large negative average image for easy observation.
BackSPIN [58]	830	2015	-	First biclustering algorithm for scRNA-Seq data
PPA [59]	93	2008	-	-
CCC-Biclustering [60]	95	2010	-	Coherent biclusters with maximal contiguous columns in linear time; Combining time-series expression with the regulatory network.
COALESCE [61]	80	2009	E Choice for constant-upregulated biclusters	Efficient enough to discover expression biclusters and putative regulatory motifs in metazoan genomes and very large microarray compendia (>10,000 conditions)
BioNMF [62]	79	2006	-	-
BiGGEsTs [39]	51	2009	-	Suitable for temporal biclustering

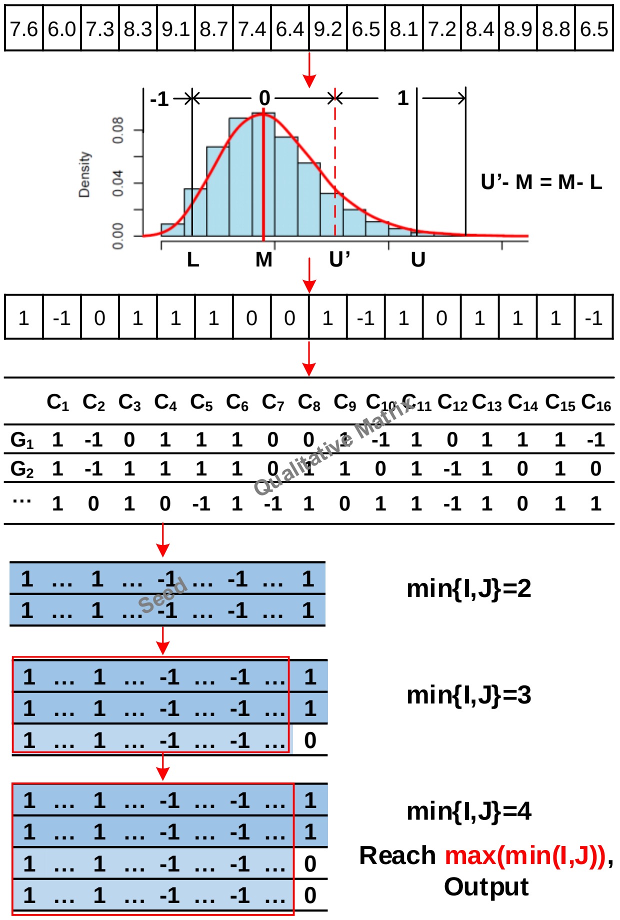
NCIS [63]	44	2014	-	Identification of cancer subtypes
FD-MSCM [64]	35	2010	-	-
BicPAM [41]	28	2014	-	Biclustering for biomedical data analysis; Suitable for non-constant biclusters
IBBiG [65]	24	2012	-	-
BUBBLE [66]	14	2006	-	Based on bottom-up search strategy; Using mean squared residue measurement.
SparseBC [67]	21	2014	-	-
BicNET [35]	13	2016	-	Discovery of non-trivial modules directly for biological network construction; Noisy and missing interaction fix; Analysis of protein interaction and gene interaction networks
MCbiclust [68]	3	2017	-	-

Note: In the Reviewer Comments column, algorithms/tools mentioned by [69] are denoted by ‘E’, mentioned by [70] are denoted by ‘NC’

Several review studies of biclustering have been carried out in different perspectives [30, 71-75]. For example, Pontes *et al.* presented a taxonomy of 47 biclustering algorithms according to their search strategies [76], and Busygin *et al.* emphasized the mathematical models and concepts in biclustering techniques [77]. Padilha *et al.* claimed that an algorithm only achieved satisfactory results in a specific context and the best choice depends on particular objectives [74]. Eren *et al.* compared 12 popular algorithms and concluded that QUBIC is one of the best as it achieves the highest performance in synthetic datasets and captures a high proportion of enriched biclusters on real datasets, and Plaid, FABIA, ISA and Bimax are the recommended tools for capturing upregulated biclusters [78]. Adetayo *et al.* presented an overview of data analysis using biclustering methods from a practical point of view, accompanied by R examples [79]. In 2018, Saelens *et al*. ranked Spectral, ISA, FABIA and QUBIC as the top biclustering methods regarding predicting gene modules from human and/or synthetic data [70].

1.3 QUBIC

QUBIC (Qualitative BIClustering algorithm) is a qualitative biclustering algorithm, which was first introduced in 2009. It assumes that a gene has three expression states under all the conditions, i.e., highly-expressed, lowly-expressed, and normally-expressed. The values in the first two expression states are so-called affected values. QUBIC employs a framework to identify dynamic cutoffs and corresponding affected values for different genes (**Figure 1**). A discretized qualitative matrix ($\boldsymbol{M_R}$) can be generated after applying the above process to each gene, with non-zero integers representing affected values and 0s being background. Then a weighted graph is constructed based on this matrix, where each node corresponds to a gene, and each edge has a weight indicating the similarity level between the two corresponding genes. The aim is to search biclusters corresponding to induced heavy subgraphs, which is an NP-hard problem. QUBIC heuristically iterates a seed list (***S***), where a seed represents a pair of genes, and its weight is the number of conditions under which they have the same values in M_R. In each iteration, it starts from a feasible seed with the highest weight, then expands vertically and horizontally to recruit more genes and conditions. Finally, QUBIC outputs a bicluster with max (min (I, J)), where I and J being the number of rows and columns of the bicluster (**Figure 1**).



	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
G_1	1	-1	0	1	1	1	0	0	1	-1	1	0	1	1	1	-1
G_2	1	-1	1	1	1	1	0	1	1	0	1	-1	1	0	1	0
…	1	0	1	0	-1	1	-1	1	0	1	1	-1	1	0	1	1

Figure 1. Workflow of QUBIC. QUBIC sorts the expression values of the gene i under all given conditions in an increasing order: $v_{i1} \cdots v_{i,s-1} v_{is} \cdots v_{i,c-1} v_{ic} v_{i,c+1} \cdots v_{i,m-s+1} v_{i,m-s+2} \cdots v_{im}$, where c=m/2 and s-1= m×q (q=6% by default). Then it selects initial bounds **L=** $v_{i,s-1}$ and **U**= $v_{i,m-s+1}$. QUBIC adjusts the bounds based on their distance from the median (M= v_{ic}), e.g., if (U-M) > (M-L), then use **U'** = (M-L) + M = 2M –L as the new upper bound. The values less or equal to L are labeled as -1, those greater or equal to **U'** are labeled as 1, and those fall between L and U' are labeled as 0. Repeat this process for each gene in the dataset, a representing matrix $\mathbf{M_R}$ can be generated.

1.4 Qserver

QUBIC has been proved to be able to solve more general biclustering problems than previous biclustering algorithms[31]. To fully utilize the analysis power of QUBIC, a web server named Qserver (Qualitative BIClustering server) was developed in 2011[23]. Qserver integrates capabilities of biclustering with *cis*-regulatory motifs prediction and functional enrichment analyses. Specifically, Qserver provides the following functionalities: (i) biclustering analysis using QUBIC; (ii) prediction and assessment of conserved *cis*-regulatory motifs in promoter sequences of the predicted co-expressed genes; (iii) functional enrichment analyses of the predicted co-expressed gene clusters using Gene Ontology (GO) terms, and (iv) visualization capabilities in support of interactive biclustering analyses.

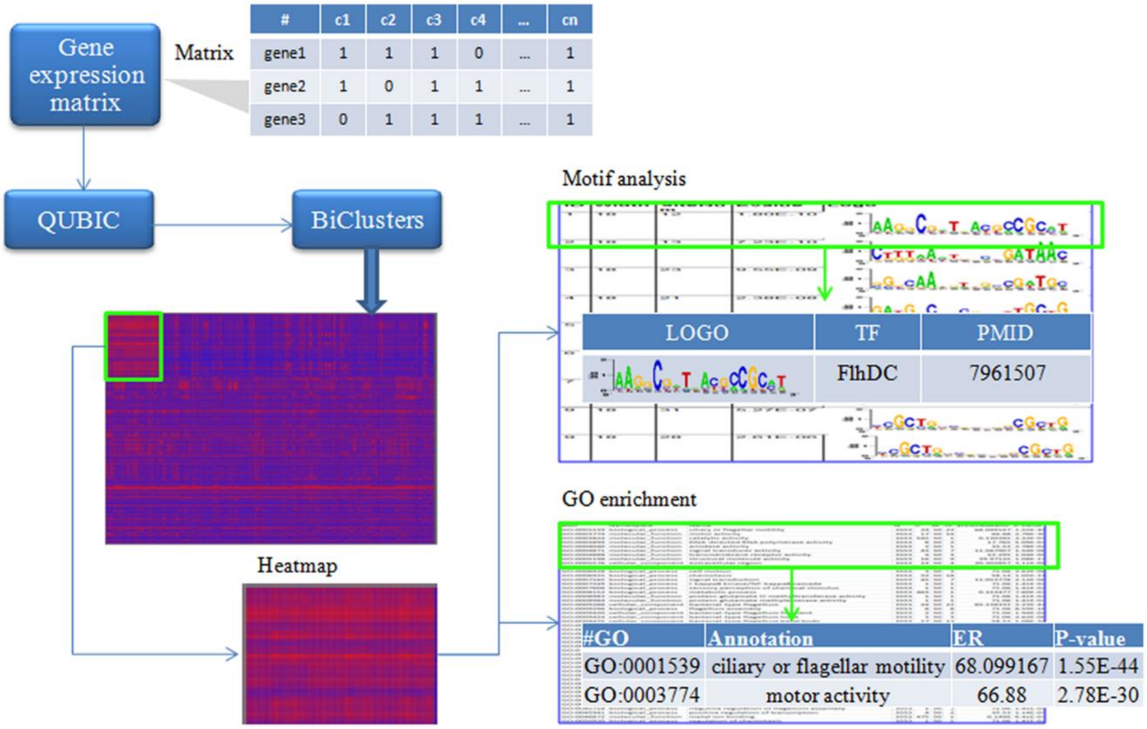
For biclustering analysis, QUBIC algorithm is implemented. Users can provide continuous or discretized gene expression matrix as input. If continuous data is provided, Qserver will automatically discretize it qualitatively. Qserver allows users to adjust the main parameters in QUBIC, and suggestion regarding how to change for different applications is provided in the Help page.

After obtaining sets of biclusters, QServer allows computationally validating of the biclusters by predicting conserved cis-regulatory motifs among the promoter sequences automatically extracted from the upstream sequences (the default value is 300 bps long) of the co-expressed genes. Two motif prediction programs, BOBRO [80] and MEME[81] are provided, both of which attempt to find conserved sequences among a set of given promoter sequences using different strategies, and both offer a statistical significance score for each predicted motif.

For the predicted biclusters, Qserver can also conduct functional enrichment analysis based on GO classification. Specifically, given a bicluster, Qserver will check if it is enriched with a GO term, compared against the background gene distribution, i.e., the whole genome. A *P*-value and enrichment ratio of that GO term will be provided.

It is common that different sets of gene expression data may use different naming conventions for genes. To deal with this issue, Qserver collected three gene/protein naming systems (i.e., GI, locus, and RefSeq) so that it can automatically detect the naming system used in an expression matrix. It also collected the genome sequences and the gene annotations from the NCBI Genome database in support of motif prediction and functional enrichment analysis, covering human, mouse, Arabidopsis, B subtilis, Synechocystis sp. PCC6803, Synechococcus sp. WH8102 and *E. coli* K12. For other organisms, Qserver will only do biclustering analysis and plot the heatmaps for biclusters.

In summary, Qserver provides three functional modules for the expression data. First, the input matrix is subject to biclustering analysis using QUBIC. For each bicluster, *cis*-regulatory motifs are then identified in the promoter regions of its component genes, using either MEME or BOBRO. Qserver will also provide the detailed information of each identified motif, including its *P*-value and the logo plot. The third module is to identify enriched GO categories among genes in each bicluster. The workflow of Qserver is shown in **Figure 2**.



#	c1	c2	c3	c4	...	cn
gene1	1	1	1	0	...	1
gene2	1	0	1	1	...	1
gene3	0	1	1	1	...	1

LOGO	TF	PMID
	FlhDC	7961507

#GO	Annotation	ER	P-value
GO:0001539	ciliary or flagellar motility	68.099167	1.55E-44
GO:0003774	motor activity	66.88	2.78E-30

Figure 2**.** An example workflow of using Qserver.

CHAPTER 2: QUBIC2—A Novel Biclustering Algorithm for Large-scale RNA-Seq Data Analysis

Although numerous algorithms and tools have been developed for gene expression data analysis, most existing biclustering algorithms are designed and evaluated using microarray rather than RNA-Seq data. One of the unique features of gene expression data derived from RNA-Seq, especially the scRNA-Seq data, is the massive zeros (up to 60% of all the genes in a cell have read counts being zeros) [82, 83]. The normalized read counts roughly follow lognormal distributions; however, the raw zero counts of specific genes will lead to negative infinity after logarithmic transformation [84-87], resulting in unquantifiable errors. Therefore, the biclustering methods that are successful for microarray cannot be directly applicable to RNA-Seq data [88], and novel methods taking full consideration of characteristics of RNA-Seq data are urgently needed in the public domain. In this chapter, I will present QUBIC2, a novel biclustering algorithm developed for large-scale RNA-Seq data analysis.

2.1 Overall Design of QUBIC2

Inheriting the qualitative representation and graph-theory based model from QUBIC [31], QUBIC2 has four unique features: (*i*) developed a rigorous truncated model to handle the unquantifiable errors caused by zeros, and used a reliable qualitative representation of gene expression to reflect expression states corresponding to various TRSs; (*ii*) integrated an information-divergence objective function in the biclustering framework in support of functional gene modules identification; (*iii*) employed a Dual strategy to expand the cores, aiming to save dropouts from the background.; and (*iv*) developed a robust P-value framework to support statistical evaluation of all the identified biclusters. Details of these four features are showcased as follows.

A mixture of left-truncated Gaussian distributions (LTMG) model was designed to fit the RNA-Seq data, rather than discarding zeros or adding a small constant to original counts [85, 89]. The basic idea is to treat the large number of observed zeros and low expressions as left censored data in the mixture Gaussian model of each gene [90, 91], assuming that the observed frequency of expressions on the left of the censoring point should be equal to the area of the cumulative distribution function of the mixture Gaussian distribution left of the censoring point. Furthermore, we assumed that a gene should receive K possible TRSs under all the conditions, and its expression profile would follow a mixture of K left truncated Gaussian distributions. The LTMG model was applied to fit the expression value of each gene, and the gene expression value under a specific condition was labeled to the most likely distribution. Accordingly, a row consisting of discrete values $(1,2,\cdots,K)$ for each gene was generated (**Figure 3A**). Then this qualitative row was split into K new rows, such that in the i^{th} row those previously labeled as i are labeled as 1, while the rest were labeled as 0. Finally, a binary representing matrix M_R was generated.

A weighted graph $G=(V,\ E)$ was constructed based on M_R, where nodes V correspond to genes, edges E connecting every pair of genes (**Figure 3B**). The edge weight indicates the similarity between the two corresponding genes, which is defined as the number of conditions in which the two genes have 1s in M_R. Intuitively, two genes from a bicluster should have a heavy edge in G innately while two random genes may have a heavy edge only accidentally. Hence, a bicluster should correspond to a maximal subgraph of G, with edges typically heavier than the edges of an arbitrary subgraph. Identifying all the biclusters equals to identifying all the heavy subgraphs in G, which is an NP-hard problem. Therefore, a heuristic strategy was designed as follows.

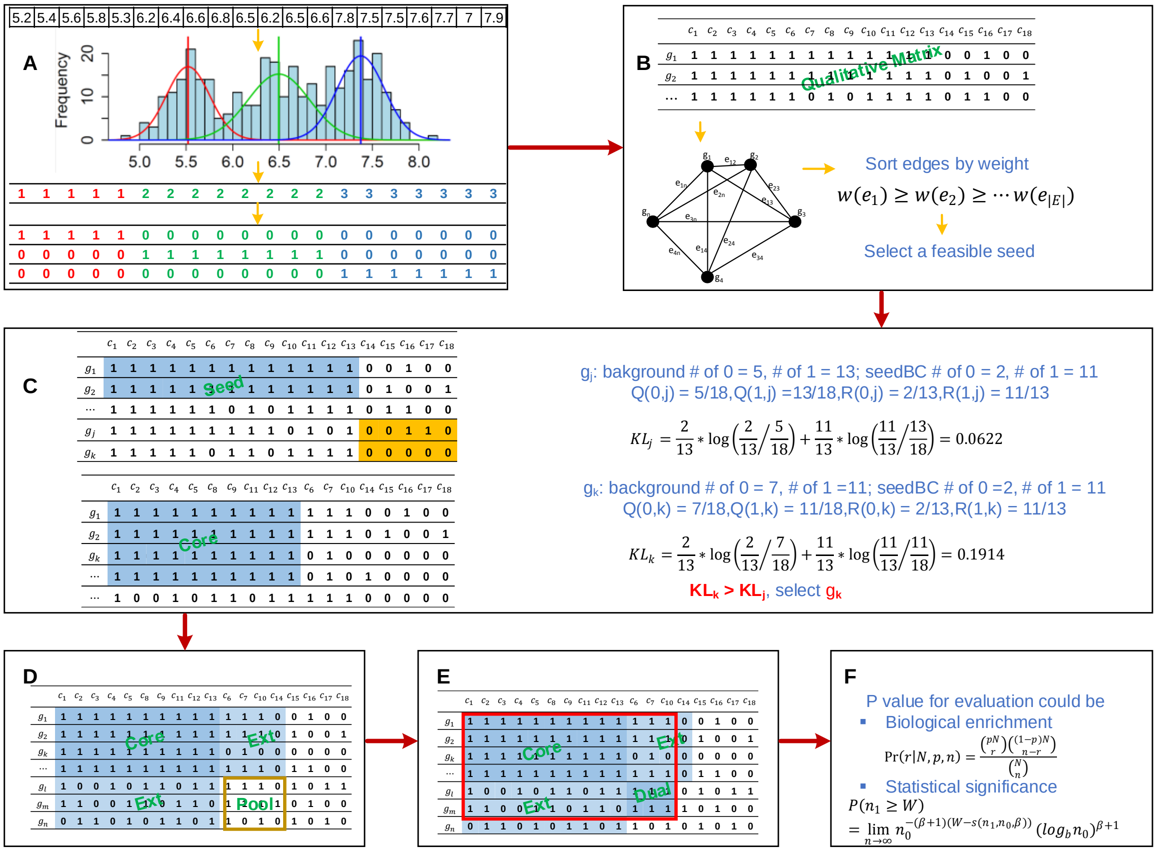

A
5.2 5.4 5.6 5.8 5.3 6.2 6.4 6.6 6.8 6.5 6.2 6.5 6.6 7.8 7.5 7.5 7.6 7.7 7 7.9
Frequency
0 10 20
5.0 5.5 6.0 6.5 7.0 7.5 8.0
1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3
B
Qualitative Matrix
Sort edges by weight
$w(e_1) \geq w(e_2) \geq \cdots w(e_{|E|})$
Select a feasible seed
C
Seed
Core
g_j: bakground # of 0 = 5, # of 1 = 13; seedBC # of 0 = 2, # of 1 = 11
Q(0,j) = 5/18,Q(1,j) =13/18,R(0,j) = 2/13,R(1,j) = 11/13
$KL_j = \frac{2}{13} * \log\left(\frac{2}{13} \Big/ \frac{5}{18}\right) + \frac{11}{13} * \log\left(\frac{11}{13} \Big/ \frac{13}{18}\right) = 0.0622$
g_k: background # of 0 = 7, # of 1 =11; seedBC # of 0 =2, # of 1 = 11
Q(0,k) = 7/18,Q(1,k) = 11/18,R(0,k) = 2/13,R(1,k) = 11/13
$KL_k = \frac{2}{13} * \log\left(\frac{2}{13} \Big/ \frac{7}{18}\right) + \frac{11}{13} * \log\left(\frac{11}{13} \Big/ \frac{11}{18}\right) = 0.1914$
$KL_k > KL_j$, select g_k
D
Core
Ext
Ext
Pool
E
Core
Ext
Ext
Dual
F
P value for evaluation could be
Biological enrichment
$\Pr(r|N,p,n) = \frac{\binom{pN}{r}\binom{(1-p)N}{n-r}}{\binom{N}{n}}$
Statistical significance
$P(n_1 \geq W) = \lim_{n\to\infty} n_0^{-(\beta+1)(W-s(n_1,n_0,\beta))} (\log_b n_0)^{\beta+1}$

Figure 3. QUBIC2 workflow. A. Discretization of gene expression data. Each gene's expression profile is fitted by the LTMG model and discretized qualitatively. Finally, a binary representing matrix is generated; B. Graph construction and seed selection. A weighted group is constructed based on the representing matrix. Then a feasible seed is selected from the seed list; C. Build an initial core based on the seed. QUBIC2 will recruit genes with higher weight with the seed. If two genes have the same weight, the one with higher KL score will be selected; D. Expand core and determine pool. QUBIC2 will expand the core vertically and horizontally to recruit more genes and conditions, respectively. The intersected zone created by extended genes and conditions as a Dual searching pool; E. Dual search in the pool and output the bicluster with genes and conditions that come from Core and Dual as final bicluster (red box); F. Statistical evaluation of identified biclusters based on either biological annotations or the size of the bicluster.

The algorithm would iterate a seed list (S), which is the sorted list of edges in G in the decreasing order of their weights (i.e., $w(e_1) \geq w(e_2) \geq \cdots, w(e_{|E|})$). An edge $e_{ij} = g_i g_j$ is selected as a seed if and only if at least one of g_i and g_j is not in any previously identified biclusters, or g_i and g_j are in two nonintersecting biclusters in terms of genes. QUBIC2 first built a core bicluster from a seed and then expanded to recruit more genes and conditions into a to-be-identified bicluster, until the Kullback-Leibler divergence score (KL score) was locally optimized. It was proposed based on the assumption that the difference between a bicluster and its background should be larger than the difference between an arbitrary same-size submatrix and its background. The KL score of a bicluster was designed to quantify this difference as the larger of the difference was, the larger of the score is (Figure 3C. See Section 2.2.2 for details).

The previous steps predict an all-1 core. We believe that some 0s outside the cores are dropouts and therefore we need to expand the cores. Since it is difficult to determine the cutoffs for expansion, we first expand the core both horizontally and vertically, and then heuristically search another core in the expanded region. Specifically, during expansion, the algorithm will control the consistency level for a bicluster, which is defined as the minimum ratio of the number of 1s in a column/row and the number of rows/columns in the bicluster. Then QUBIC2 will adopt the same strategy as it used for predicting Cores to search another core in the expanded region (**Figure 3D-E**), giving rise to a submatrix (I, J) of M_R (i.e., a bicluster) with optimized consistency level and maximal KL score can be identified. It is assumed that 0s induced in this way are more likely to be dropouts.

Furthermore, for the first time, a statistical framework based on the size of the biclusters was implemented to calculate a P-value for each of the identified biclusters. The problem of assessing the significance of identified biclusters was formulated as calculating the probability of finding at least one submatrix enriched by 1 from a binary matrix with given size, with a beta distribution employed during the process. This P-value framework enables users systematically evaluate the statistical significance of all the identified biclusters, especially for those from less-annotated organisms (**Figure 3F**).

2.2 Detailed Methods in QUBIC2

2.2.1 Left Truncated Mixed Gaussian (LTMG) Model and Qualitative Representation

To accurately model the gene expression profile of RNA-Seq and scRNA-Seq data, we explicitly developed a mixed Gaussian model with left truncation assumption. Denotes the log-transformed FPKM, RPKM or CPM expression values of gene X over N conditions as $\mathrm{X} = \{x_1, \ldots x_n\}$, we assumed that $x_j \in X$ follows a mixture of k Gaussian distributions, corresponding to k possible TRSs. The density function of x_j is:

$$p(x_j; \Theta) = \sum_{i=1}^{k} \alpha_i p(x_j; \theta_i) = \sum_{i=1}^{k} \alpha_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-(x_j-\mu_i)^2}{2\sigma_i{}^2}}$$

And the density function of X is:

$$p(\mathrm{X}; \Theta) = \prod_{j=1}^{n} p(x_j; \Theta) = \prod_{j=1}^{n} \sum_{i=1}^{k} \alpha_i p(x_j; \theta_i) = \prod_{j=1}^{n} \sum_{i=1}^{k} \alpha_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-(x_j-\mu_i)^2}{2\sigma_i{}^2}} = L(\Theta; X)$$

where α_i is the mixing weight, μ_i and σ_i are the mean and standard deviation of i^{th} Gaussian distribution, which can be estimated by the EM algorithm with given X:

$$\Theta^* = {}^{\arg\max_{\Theta} L(\Theta; X)}$$

To model the errors at zero and the low expression values, we introduce a parameter Z_{cut} for each gene expression profile and consider the expression values smaller than Z_{cut} as left censored data. With the left truncation assumption, the gene expression profile is split into M truly measured expression value (> Z_{cut}) and $N-M$ left censored gene expressions ($\leq Z_{cut}$) for the N conditions. Latent variables y_j and Z_j are introduced to estimate Θ by the following Q function:

$$Q(\Theta;\Theta^{t-1}) = \sum p\left(y_j \middle| x_j;\Theta^{t-1}\right)\sum_{j=1}^{m}\sum_{i=1}^{k}\log(\alpha_i p(x_j;\mu_i,\sigma_i))$$

$$+\sum p\left(y_j \middle| z_j;\Theta^{t-1}\right)\sum_{j=m+1}^{n}\sum_{i=1}^{k}\log(\alpha_i p(z_j;\mu_i,\sigma_i))$$

To estimate the parameters Θ that maximizes the likelihood function, we have Maximization step of the EM algorithm as [92]:

$$a_i^t = \frac{1}{N}(\sum_{j=1}^{M} P\left(i \middle| x_j,\Theta^{t-1}\right) + \sum_{j=M+1}^{N} P(i|Z_j,Z_{cut},\Theta^{t-1}))$$

$$u_i^t = \frac{\sum_{j=1}^{M} x_j P\left(i \middle| x_j,\Theta^{t-1}\right) + \sum_{j=M+1}^{N}(u_i^{t-1} - \sigma_i^{t-1} H(\frac{Z_{cut}-u_i^{t-1}}{\sigma_i}))P(i|Z_j,Z_{cut},\Theta^{t-1})}{\sum_{j=1}^{M} P(i|x_j,\Theta^{t-1}) + \sum_{j=M+1}^{N} P(i|Z_j,Z_{cut},\Theta^{t-1})}$$

$$\sigma_i^{t^2}$$

$$= \frac{\sum_{j=1}^{M} P\left(i \middle| x_j,\Theta^{t-1}\right)(x_j-u_i^{t-1})^2 + \sigma_i^{t-1^2}\sum_{j=M+1}^{N}(1-\frac{Z_{cut}-u_i^{t-1}}{\sigma_i} * H(\frac{Z_{cut}-u_i^{t-1}}{\sigma_i})) * P(i|Z_j,Z_{cut},\Theta^{t-1})}{\sum_{j=1}^{M} P(i|x_j,\Theta^{t-1}) + \sum_{j=M+1}^{N} P(i|Z_j,Z_{cut},\Theta^{t-1})}$$

where $P\left(i \middle| Z_j,Z_{cut},\Theta^{t-1}\right) = \frac{P(-\infty<Z_j<Z_{cut}|u_i^{t-1},\sigma_i^{t-1})}{\sum_{i=1}^{K} P(-\infty<Z_j<Z_{cut}|u_i^{t-1},\sigma_i^{t-1})}$, $H(x) = \frac{\phi(x)}{\Phi(x)}$, $\phi(x)$ and $\Phi(x)$ are the pdf and cdf of standard normal distribution.

Parameters Θ can be estimated by iteratively running the estimation (E) and maximization (M) steps. In this study, Z_{cut} is set for each gene as the logarithm of the minimal non-zero RPKM/FPKM/TPM value in the gene's expression profile. The EM algorithm is conducted for K = 1, …, 9 to fit the expression profile of each gene and the K that gives the best fit is selected according to the Bayesian Information Criterion (BIC):

$$BIC = -2\ln(\Theta^*) + 3K\ln(N)$$

where K is the number of TRS, K is the number of conditions. K that minimizes the BIC will be selected.

Then the original gene expression values will be labeled to the most likely distribution under each condition. In detail, the probability that x_j belongs to distribution i is formulated by:

$$p\left(x_j \in TRS\ i \middle| K, \Theta^*\right) \propto \frac{\alpha_i}{\sqrt{2\pi\sigma_j{}^2}} e^{\frac{-(x_j-\mu_i)^2}{2\sigma_i{}^2}}$$

And x_j is labeled by TRS i if $p\left(x_j \in TRS\ i \middle| K, \Theta^*\right) = \max\limits_{i=1,\cdots,K}\left(p\left(x_j \in TRS\ i \middle| K, \Theta^*\right)\right)$. In such a way, a row consisting of discrete values (1,2, …, K) for each gene will be generated.

2.2.2 KL Score

A Kullback-Leibler divergence score (**KL** score) is introduced in QUBIC 2 to guide candidate-selection and biclustering optimization. The KL score of a bicluster is defined as:

$$KL_B = \frac{1}{N}\sum_{j=1}^{N}\sum_{i\in\{0,1\}} R(i,j) \times log\frac{R(i,j)}{Q(i,j)} + \frac{1}{M}\sum_{k=1}^{M}\sum_{i\in\{0,1\}} C(i,k) \times log\frac{C(i,k)}{P(i,k)}$$

where N and M are the numbers of rows and columns of a submatrix B in M_R, respectively. $R(i,j)$ represents the proportion of element i in row j of B, $Q(i,j)$ is the proportion of i in the entire corresponding row, $C(i,k)$ is the proportion of i in column k of B, and $P(i,k)$ is the proportion of i in the entire corresponding column.

Meanwhile, the KL score for a gene quantify the similarity between a candidate gene j and a bicluster, which is defined as follows:

$$KL_j = \sum_{i \in \{0,1\}} R(i,j) \times log \frac{R(i,j)}{Q(i,j)}$$

where $R(i,j)$ represent the proportion of i under corresponding columns of the current bicluster.

2.2.3 QUBIC2 Algorithm

The QUBIC2 algorithm concludes as follows:

Step 1 (Data discretization and qualitative representation): Given an expression matrix with log-transformed FPKM, RPKM or CPM value for genes, use LTMG model to fit data. Label the values to the most likely distribution to get a representing row for each gene. Split these rows into multiple rows to get the representative matrix M_R (**Figure 3A**).

Step 2 (Graph construction and seed selection): Construct a weighted graph for M_R. Select a feasible seed from the seed list; Stop if the seed list is empty (**Figure 3B**).

Step 3 (Build core bicluster): Build an initial bicluster by finding all the conditions under which the two genes of the seed have 1s in M_R. Set these columns of the two genes as the current bicluster B = (*I*, *J*). Expand B by adding a new gene that has the most 1s in *J*, giving rise to a new bicluster B’ = (*I’*, *J’*), where *I’* is *I* after adding the new gene and *J’* is *J* by deleting those columns with 0s. If two genes have the same number of 1s in *J*, choose the one with larger KL similarity with B (**Figure 3C**). If $KL_{B'} > KL_B$, set B to B’ and repeat Step 2, otherwise stop and denote B as **Core**. Go to Step 4.

Step 4 (Core expansion): Expand the Core horizontally and vertically under preset consistency level as follows: for each gene(row) i not in B, if the ratio between the number of 1s in row i under J and |J| is ≥c, mark it as an extended gene; for each condition (column) j not in B, if the ratio between the number of 1s in the column j among I and |I| is ≥c, mark it as an extended condition. (**Figure 3D**). Mark the intersected zone created by extended genes and conditions as a Dual searching pool (brown box in **Figure 3D**). Go to Step 5.

Step 5 (Search Dual): Search **Dual** in the intersected expanded zone, using the same process in Step 3, output the bicluster with genes and conditions that come from Core and Dual (red box in **Figure 3E**). Delete current seed, go to step 1.

2.2.4 Size-based *P*-value

For well-annotated organisms, the *P*-value of an identified bicluster enriching with a specific regulatory pathway can be calculated based on a hypergeometric distribution. However, the known experimental annotation is currently limited, even for most well-studied model organisms (about half of the protein-coding genes of *E. coli* have solid experimental evidence for their function in KEGG and GO) [93]. This status still limits the capability of a systematic evaluation of all the identified biclusters. To fill this gap, we calculate an alternative size-based *P*-value as follows. For a binary representing matrix M_R, containing m_0 rows and n_0 columns, suppose we obtain an m_1-by-n_1 bicluster M_1 with all the elements be 1s. The probability of $n_1 \geq W$ can be assessed by the following formula [94], giving rise to a *P*-value of the bicluster M_1:

$$P(n_1 \geq W) = \lim_{n\to\infty} {n_0}^{-(\beta+1)(W-s(n_1,n_0,\beta))} (\log_b n_0)^{\beta+1}$$

where $\alpha = \frac{m_0}{n_0}$, $\beta = \frac{m_1}{n_1}$, $b = \frac{1}{p}$, $p = P\left(M_{i,j} = 1\right) = 1 - P\left(M_{i,j} = 0\right)$ for $\forall i, j$

$$s(n_1, n_0, \beta) = \frac{\beta+1}{\beta}\log_b n_0 - \frac{\beta+1}{\beta}\log_b\left(\frac{\beta+1}{\beta}\log_b n_0\right) + \log_b \alpha + \frac{(1+\beta)\log_b e - \beta\log_b \beta}{\beta}$$

2.3 Functional Gene Modules Detection from RNA-Seq Data

2.3.1 Data Acquisition

A total of four expression datasets were used in this section, that is, one synthetic RNA-Seq data (22,846 rows × 100 columns), one bulk RNA-Seq dataset from *Escherichia coli* (*E. coli,* 4,497 rows × 155 columns), a bulk RNA-Seq dataset from TCGA (3,084 rows × 8,555 columns), and a scRNA-Seq dataset from human embryos (3,798 genes × 90 cells). The synthetic dataset was simulated using our in-house simulation method (see **Section 2.3.2**). It contains 22,846 genes and 100 samples. A total of 10 co-regulated modules was embedded in this dataset, covering 2,240 up-regulated genes. The *E. coli* RNA-Seq data consists of 4,497 genes and 155 samples, which was integrated and aggregated by our group. In short, 155 fastq files were downloaded from ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ using the sratoolkit (v2.8.1, https://github.com/ncbi/sra-tools/wiki/Downloads), and they are processed following quality check (FastQC), reads trimming (Btrim), reads mapping (HISAT2) and transcript counting (HTSeq). Then, raw read counts were RPKM normalized. The human RNA-Seq data contains 3,084 genes and 8,555 samples, which was obtained from [70]. The scRNA-Seq data was downloaded from [95] as an RPKM expression matrix with 20,214 gene and 90 cells, and then 3,798 genes were kept for the analysis in this study by removing the genes without annotation.

Multiple sets of known modules/biological pathways were provided or collected to support the enrichment analysis of the above four datasets. For synthetic data, the ten groups of pre-defined up-regulated genes were used as co-regulated modules. For *E. coli* data, we used five kinds of biological pathways, which are complex regulons and regulons extracted from the RegulonDB database (version 9.4, accessed on 05/08/2017), KEGG pathways collected from the KEGG database (accessed on 08/08/2017), SEED subsystems from the SEED genomic database (accessed on 08/08/2017) [96], and EcoCyc pathways from the EcoCyc database (version 21.1, as of 08/08/2017) [97]. Complex regulons (ComTF) were defined as a group of genes that are regulated by the same transcription factor (TF) or the same set of TFs. In total, 457 complex regulons, 204 regulons, 123 KEGG pathways, 316 SEED subsystems, and 424 EcoCyc pathways were retrieved, respectively. For the human TCGA and scRNA-Seq data, we used three sets of modules provided by [70].

2.3.2 Simulation of Co-regulated Gene Expression Data

We utilized a single cell RNA-Seq dataset of human melanoma [98] (with 22,846 genes and 4,645 cells) to simulate bulk tissue RNA-Seq data with known co-regulated modules. Specifically, a single cell RNA-Seq pool consists counts data of 4,466 cells of six annotated cell types namely B-, T-, endothelial, fibroblast, macrophage, and cancer cells were constructed. The top 1,000 cell type specifically expressed genes of each cell type were identified by using Z score of the mean of each gene’s expression level in each cell type.

For each round of simulation, the number of to be simulated bulk tissue samples and co-regulation modules is first defined. Then the genes of each co-regulation module denoted as X_k will be specified by randomly selecting M_k genes from the top 1,000 cell type specifically expressed genes of one cell type. A co-regulation strength matrix P is then simulated from a bimodal distribution over (0,1), with $P[i,k]$ denotes the proportion of cells with the transcriptional regulatory signal of co-regulation module k in bulk sample i. A bulk tissue data is simulated by randomly drawing cells from the cell pool by following a multinomial distribution, with predefined parameters and the total number of cells. For co-regulation module k in bulk sample i, genes X_k in a proportion $P[i,k]$ of the selected cells of the cell type corresponds to k are perturbed by an X-fold increase of the gene expression. Then the bulk data i with simulated co-regulations are formed by summing the perturbed gene expression profile the selected cells and normalized to RPKM expression scale. The Pseudo code of the simulation approach is provided as follows:

```
For k in 1 to #co − regulation modules
    $X_k \triangleq$ Randomly select $M_k$ genes from the cell type specifically expressed genes of one cell type
For i in 1 to #Bulk tissue data
    For k in 1 to #co − regulation modules
        Randomly simulate $P[i,k] \triangleq$ proportion of cells with a perturbed expression
For i in 1 to #Bulk tissue data
    Randomly select N cells from the cell pool with a multinomial distribution with replacement
    For k in 1 to #co − regulation modules
        Choose $P[i,k]$ proportion of cells of the cell type coresponds to the coregulation module k
            Perturb expression of $X_k$ in the chosen cells by a X − fold increase
    Simulate the ith bulk tissue data by sum of the perturbed gene expression of the selected N cells
```

The rationales of this simulation approach include (1) gene expression level and noise in the bulk data are purely simulated by sum of real single-cell data, without using artificially assigned expressions scale and noise; (2) co-regulation genes are modeled as a specific fold increase of a number of cell-type-specific genes in a particular subset of the cells, which characterizes the heterogeneity of transcriptional regulation among cells in a tissue; (3) multiple co-regulation modules in specific to different cell types can be simultaneously simulated. Hence, we believe the gene expression data simulated by this way can satisfactorily reflect genes co-regulated by a perturbed transcriptional regulation signal in real bulk tissue data.

2.3.3 Evaluation of Functional Modules

The capability of algorithms to recapitulate known functional modules are assessed using precision and recall. First, for each identified bicluster, we use the P-value of its most enriched functional class (biological pathway) as the P-value of the bicluster. Specifically, the probability of having x genes of the same functional class in a bicluster of size n from a genome with a total of N genes can be computed using the following hypergeometric function[99]:

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

where p is the percentage of that pathway among all pathways in the whole genome. The P-value of getting such enriched or even more enriched bicluster is calculated as:

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

The bicluster is deemed enriched with that function if its P-value is smaller than a specific cutoff (e.g., 0.05).

Given a group of biclusters identified by a tool under a parameter combination, the precision is defined as the fraction of observed biclusters significantly enriched with the one biological pathway/known modules (Benjamini-Hochberg adjusted p<0.05),

$$Precision = \frac{\#\ of\ significant\ biclusters}{\#\ of\ biclusters}$$

For recall, we compute the fraction of known modules that were rediscovered by the algorithms,

$$Recall = \frac{\#\ of\ significant\ modules}{\#\ of\ modules}$$

Finally, the harmonic mean of precision and recall were calculated to represent the performance of an algorithm on a given dataset and parameter setting, denoted as *F* score:

$$F = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

Note that the number of biclusters used to calculate precision and recall may affect the results. To make sure the evaluation is as fair as possible, for each dataset, we select the first 30 biclusters.

2.3.4 Biclustering Parameters

To assess the robustness, each tool is run multiple times by varying parameters that affect the size and number of biclusters. In general, parameters are adjusted around their default or recommended (if available) value. The parameters varied as well as details about the range and increment are listed in Table2.

Table 2. Main parameter adjusted for each algorithm

Algorithm	**Implementation**	**Parameters**	**Note**
Bimax	R package 'biclust'	***minr*** ranges from 10~60(increment 5) ***minc*** ranges from 10~45 (increment 5) **number** set to 100	Need discretized data as input. For each dataset, take the discretized data from QUBIC as input. No recommendation provided by the author or biclust manual. Default: ***minr***=2, ***minc***=2
ISA	R package 'isa2'	***set.seed*** ranges from 10~600, increment 10	ISA is stochastic, by setting different seeds

			may obtain different biclusters

FABIA	R package fabia'	***alpha*** ranges from 0~0.05, increament:0.01;
		spl ranges from 0~2, increment 0.5;
		spz ranges from 0~2, increment 0.5;
		cyc=100, ***p***=100
			default: ***alpha***=0.1, ***spl*** =0, ***spz***=0.5, ***cyc***=500, ***p***=5

Plaid	R package 'biclust'	both ***row.release*** and ***col.release*** range from 0.5~0.7, increment be 0.05 ***max.layer*** 10~100	for ***row.release*** and ***col.release***, 0.5~0.7 is the recommended range
QUBIC	R package 'QUBIC'	***f*** 0.1~1.0, increment 0.05 ***c*** 0.8~1.0, increment 0.05 ***k*** 3~23, increment 5	default: ***f***=1.0, ***c***=0.95, ***k***=ncol/20
QUBIC2	C++	***f*** 0.25~1.0, increment 0.05 ***k*** 5~23, increment 5	

2.3.5 Results

Compared with five biclustering algorithms (Bimax [30], ISA [100], FABIA [37], Plaid [16], and QUBIC [31]), the performance of QUBIC2 in identifying FGMs was systematically evaluated using four gene expression datasets. For the identified biclusters from a specific tool, *precision* showcases the fraction of biclusters whose genes are significantly enriched with specific biological pathways (i.e., relevance), and *recall* reflects the fraction of captured known modules/pathways among all known modules in a functional annotation database, e.g., KEGG [101] and RegulonDB [102] (i.e., diversity). The harmonic mean value of precision and recall, referred to as the *F* score, was used as the integrated criteria in performance evaluation.

Evaluation studies usually used default parameters of the to-be-analyzed tools, which were optimized for specific benchmark datasets. However, when applied to datasets coming from a different organism (e.g., *E. coli* vs. human), or be acquired by other technologies (e.g., microarray vs. RNA-Seq), the default parameters often fail to achieve satisfying performance and need further optimization/adjustment. To minimize the biases in performance comparison among multiple tools, for each of the four datasets, we run the six tools under more than 50 parameter combinations by adjusting their critical parameters around default/recommended values. Then the *F* score of identified biclusters under each parameter combination was calculated. In this way, we can test a tool's robustness and infer how sensitive of its performance is to parameter adjustment, besides the basic performance comparison among different tools.

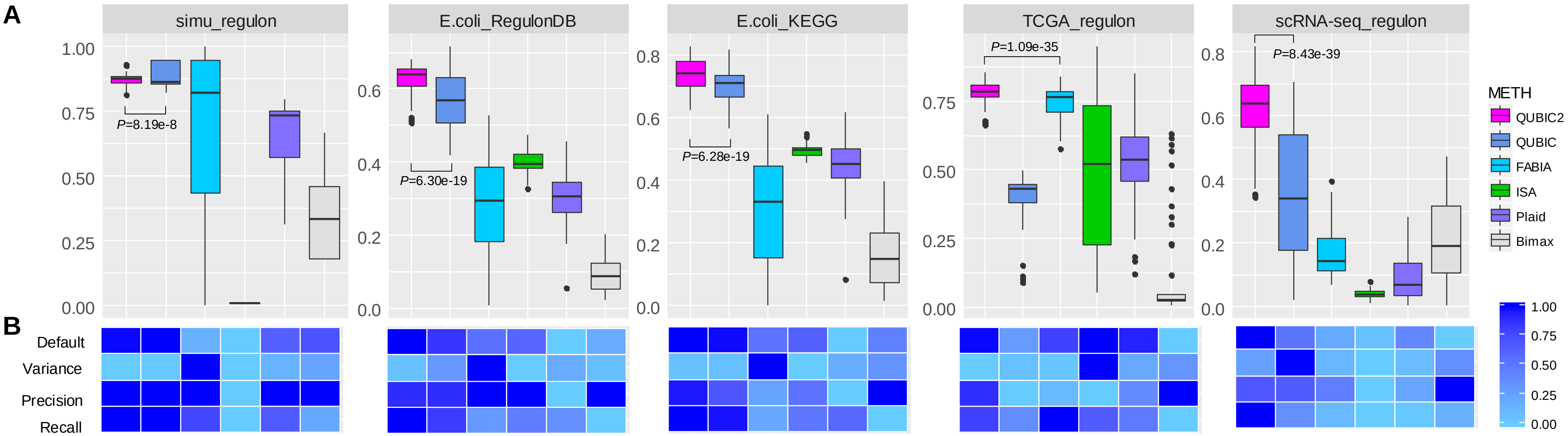


Figure 4. Overall performance comparison between QUBIC2 and five popular biclustering methods based on the agreement between identified biclusters and known modules. A. Distribution of F scores on each of the four datasets under multiple runs (n>40). Black line in the box denote median value, whiskers denote 10% and 90% percentiles, while the box denotes 25% and 75% percentiles; B. relative performance of six algorithms in terms of F score under default parameters, variance of F scores under multiple sets of parameters, median value for the precision and median value for the recall, respectively (normalized over six algorithms). Note that the variance of F scores depends on the increment of parameters, and therefore only indicative.

As showcased in Figure 4, QUBIC2 achieved the highest median *F* scores and the highest *F* scores with the default parameter on all the four datasets, and its *F* scores were significantly higher than the second-best algorithms in all the comparison circumstances (Wilcoxon test *P*-value <0.01). QUBIC2 performed well in both precision and recall, indicating that the identified FGMs are relevant and diverse; and it had relatively small variance, while the performance of some algorithms on specific dataset was susceptible to parameter change (e.g., FABIA on *E. coli*). Regarding median *F* scores, QUBIC was the second-best algorithm on simulated data, *E. coli* RNA-Seq data, and human scRNA-Seq data, while FABIA was the second-best one for TCGA data. As regards the default settings, QUBIC ranked as the top ones on simulated data and *E. coli* data, and ISA and Plaid had relative higher rank on TCGA data. ISA was generally very stable, and its variances were the smallest on three datasets. As for Bimax, although its recall was relatively low, it was characterized with high precision on the four datasets. It is noteworthy that QUBIC2 is the only program, among all the six biclustering algorithms, which did not encounter a dramatic performance drop on scRNA-Seq data compared to RNA-Seq data, suggesting the unique applicative power of QUBIC2 on FGMs detection from scRNA-Seq data.

Furthermore, the performance of all the biclustering algorithms on *E. coli* data was better than on human data, with the possible reason that *E. coli* data has more completed functional annotation and affects the evaluation of module significance. Therefore, for less annotated organisms, we need a statistical evaluation framework for all the identified biclusters.

2.4 A Statistical Evaluation Framework for Identified Biclusters

The significances of gene modules from the identified biclusters were usually evaluated by pathway enrichment analysis. However, many organisms (including human) have limited functional annotations supported by experimentally verifications, which makes a systematic evaluation of all identified biclusters non-trivial. To fill this gap, a statistical method was proposed in QUBIC2, which can calculate a *P*-value for a bicluster purely based on their size. To evaluate the validity of the proposed method, a Spearman correlation test was conducted.

2.4.1 Methods

QUBIC2 was run on the *E. coli* RNA-Seq data under 63 parameter settings. For each setting, around 100 biclusters were identified. Five sets of regulatory or metabolic pathways were extracted from four databases of *E. coli* (RegulonDB, KEGG, SEED [96] and EcoCyc [97]) to support this association study. In specific, for each set of ~100 biclusters obtained under the same settings, six groups of *P*-values for all these biclusters were calculated, with five groups of *P*-values derived via pathway enrichment analysis (named knowledge-based *P*-values) and one group of *P*-values computed using our size-based method. Spearman correlation test was conducted to investigate the rank-order correlation among the six groups of *P*-values. Five correlation coefficients (ρ), which demonstrated the extent of correlation between size-based *P*-values and five biological knowledge-based *P*-values, as well as five corresponding *p*-values, were recorded from the test. Note that the *p*-value of correlation test denotes the probability of observing such a correlation or even stronger correlation, under the null hypothesis that no correlation exists. For simplicity, the correlation coefficient between the size-based *P*-value and biological knowledge-based *P*-value was prefixed with the name of a pathway, e.g., TF_ρ and KEGG_ ρ . In the end, a total of 5 × 63 ρ (63 parameter settings, each with five ρs) and the same number of *p*-values were obtained.

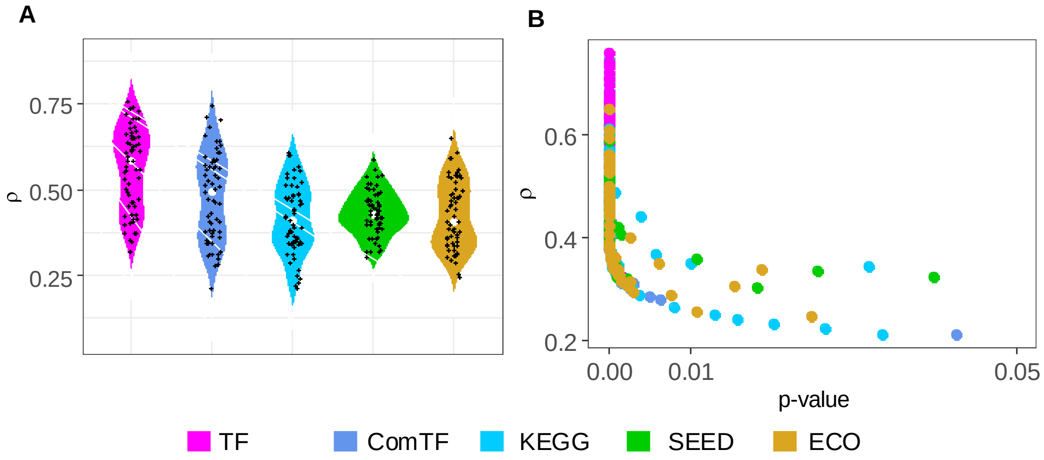


Figure 5.A. The distribution of correlation coefficients(ρ) between *P*-value obtained from enrichment analysis and size-based *P*-value. We run QUBIC2 under 63 different parameter settings, and ρ was calculated under each run; B. Scatter plot of ρ and p-value. The y-axis denotes ρ, the correlation coefficient for the Spearman association test, the x-axis denotes the p-value of the association test. Note that to distinguish, italic lowercase p was used to denote the p-value of the Spearman correlation test, while italic uppercase P was used to denote the significance of biclusters.

2.4.2 Results

Interestingly, we found that there is a strong association between the knowledge-based *P*-value and the corresponding size-based *P*-values. The average Spearman correlation coefficients (ρ) were higher than 0.40 (ComTF_ρ =0.48, TF_ρ=0.56, KEGG_ρ =0.42, SEED_ρ=0.43 and ECO_ρ =0.42), and the average *p*-values for the correlation test were smaller than 0.01. As showcased in Figure 5A, all the ρs in the five groups are positive. Besides, ρs related with regulatory pathways (i.e., TF_ ρ and ComTF_ ρ) were generally larger than ρ s those related to metabolic pathways (i.e., KEGG_ ρ and SEED_ ρ). This indicated that the size-based *P*-value seemed to be more suitable for the evaluation of biclusters' regulatory significance. Furthermore, all the corresponding *p*-values were less than 0.05 (**Figure 5B**), suggesting that the correlations between knowledge-based *P*-values and size-based *P*-values were statistically significant at the 0.05 level. Also, the parameter *f* which controls the level of overlaps between biclusters had a negative association with ρ (**Figure 6**), suggesting that the size-based *P*-values would have a stronger association with knowledge-based *P*-values when the overlaps between biclusters are relatively low.

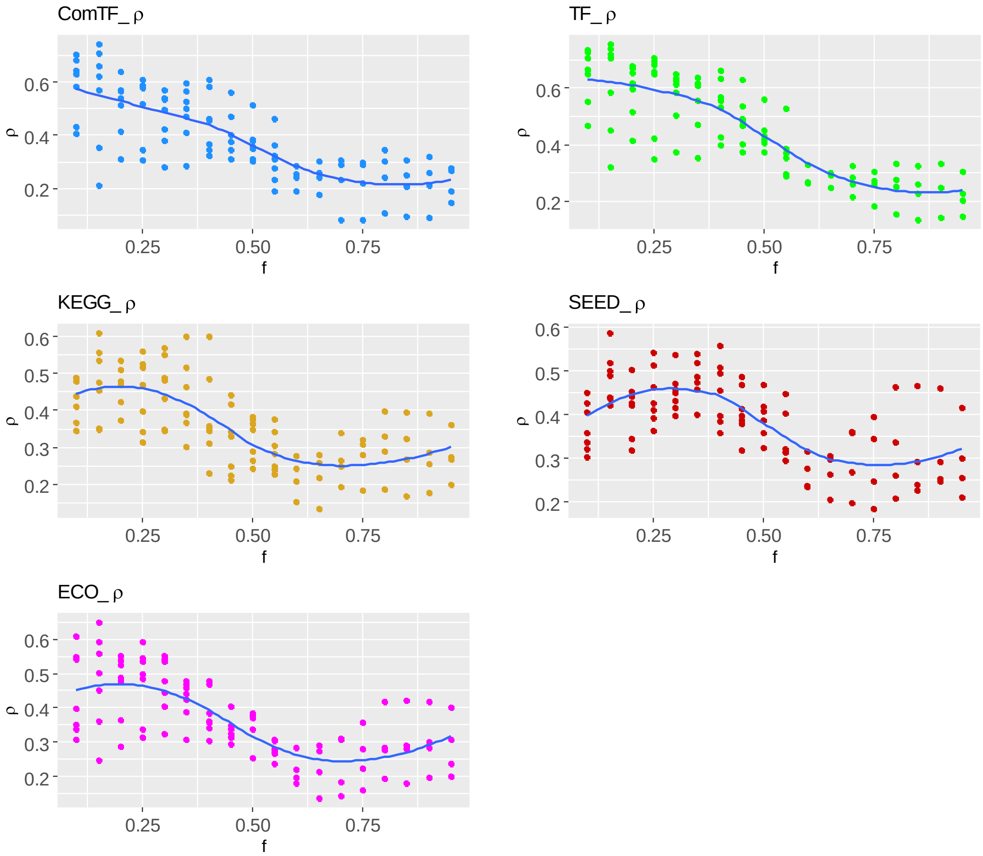


Figure 6. The relationship between biclustering parameter f and correlation coefficient that indicates the association between biological knowledge-based *P*-value and size-based *P*-value. The blue line in each plot corresponds to the Loess smooth line.

2.5 Cell Type Classification Based on scRNA-Seq Data

The above sections demonstrated the outstanding performance of QUBIC2 on FGMs identification and its unique feature of statistical evaluation for all the identified biclusters. In this section, we showed the predictive power of biclustering methods on cell types identification from scRNA-Seq data.

2.5.1 Cell Type Classification Pipeline

By using biclustering, we can group genes and cells simultaneously. However, since biclustering aims to find sets of genes that are co-expressed across a subset of conditions, it is possible that genes may co-expressed across multiple cell types. Therefore, one bicluster may consist of cells from different types, and cells from the same types may appear in different biclusters. In a word, it is not guaranteed that one bicluster corresponds to one cell type. However, it is assumed that two cells from a bicluster are more likely to be of the same subtypes than the two cells that are randomly selected. It is believed that biclusters can capture this feature to some extent. If there are multiple biclusters and when we condense them together, we can distinguish sets of cells belonging to the same type from sets of cells that are grouped by chance.

Based on the above idea, we developed a pipeline to obtain cell type classification based on biclustering results (**Figure 7A**). First, a biclustering tool was applied to the expression data (rows represent genes and columns represent cells) to identify a set of biclusters. Then a weighted graph $G = (C,\ E)$ was constructed to model the relationship between cell pairs among biclusters. A node c_i in G represented a cell, and $e_{i,j}$ represented the edge connecting c_i and c_j, where $i \neq j$. We assigned weight $w_{i,j}$ to $e_{i,j}$ to represent the number of biclusters that contain both c_i and c_j. Intuitively, a higher $w_{i,j}$ value indicates that c_i and c_j are simultaneously involved in more biclusters, hence, are more likely to be the same cell type than cell pairs with lower weight. A symmetrical cell-cell matrix with diagonal as 0 was then constructed to record $w_{i,j}$ and Markov Cluster Algorithm (**MCL**) was performed to cluster cells into cell types and produce cell labels. In specific, the MCL clustering was run 100 times by varying inflation factor from 1 to 100, resulting 100 cell labels. A binary similarity matrix was constructed for each cell label: if two cells belong to the same cluster, their similarity is 1; otherwise, the similarity is 0. Then a consensus matrix was built by averaging all similarity matrices. The resulting consensus matrix was clustered using hierarchical clustering with complete agglomeration, and the clusters were inferred at the k level of the hierarchy, where k is the chosen based on the average silhouette score of that 100 MCL clustering results.

2.5.2 Data, Biclustering Parameters and Evaluation Criteria

One golden-standard scRNA-Seq data [95] was used. It consists of 20,214 genes and 90 cells, where the cells were assigned into seven subgroups with the true cell subtypes information provided in [95].

For each of the six biclustering methods, we applied the classification pipeline to the above dataset. Each tool was run under multiple parameter settings. The details about the range of parameters are given in Table3.

Table 3. Parameter ranges for each biclustering algorithm used in the cell type classification section

Algorithm	Parameters	Note
Bimax	***minr*** 10~200, increment 10 ***minc*** 10~30, increment 10 ***number***=2000	

ISA	***set.seed*** ranges from 10~600, increment 10	
FABIA	***alpha*** 0.01~0.5, increment 0.01 ***spl*** 0~2, increment 0.05; ***spz*** 0~2, increment 0.05 ***p***=50	tried to set ***p***=100, 1000, 2000, but got error message ‘too many biclusters’ and aborted
Plaid	***row.release*** 0.5~0.7 ***col.release*** 0.5~0.7 ***max.layer*** 10~100	

QUBIC	***f*** 0.5~1.0, increment 0.05;
	c 0.8~0.95, increment 0.05;
	k =13;
	o =2000
		default ***o***=100

QUBIC2	***f*** 0.6~1.0, increment 0.05;
	c 0.8~0.95, increment 0.05;
	k = 4,13
	o = 2000	

The Adjusted Rand Index (ARI) was adopted as the evaluation criteria to access the agreement between predicted cell types and these ‘ground truth' [103]. Two more external validation criteria, namely Jaccard Index (JI) and Fowlkes Mallows Index (FMI), were also used here aiming to provide a comprehensive evaluation.

Specifically, external validation measures the extent to which cluster labels match externally supplied class labels. Generally, they are based on counting the pairs of points on which two classifiers agree/disagree. Denote two partitions of the same data set as R and Q. The reference partition, R, encode the class labels, i.e., it partitions the data into k known classes. Partition Q, in turn, partitions the data into v categories, which is the one to be evaluated.

Adjusted Rand Index (ARI) is defined as

$$ARI = \frac{a - \frac{(a+c)(a+b)}{d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{d}}$$

a: Number of pairs of data objects belonging to the same class in R and the same cluster in Q.

b: Number of pairs of data objects belonging to the same class in R and different clusters in Q.

c: Number of pairs of data objects belonging to different classes in R and the same cluster in Q.

d: Number of pairs of data objects belonging to different classes in R and different clusters in Q.

Terms a and d are measures of consistent classifications (agreements), whereas terms b and c are measures of inconsistent classifications (disagreements).

Jaccard Index is defined as:

$$JI = \frac{a}{a+b+c}$$

The Jaccard Index can be seen as a proportion of good pairs with respect to the sum of non-neutral (good plus bad) pairs.

Folkes-Mallow's index is defined as

$$FI = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Fowlkes–Mallow's index can be seen as a non-linear modification of the Jaccard coefficient that also keeps normality.

2.5.3 Results

The performance of QUBIC2 was compared with five biclustering methods (QUBIC, FABIA, ISA, Plaid and Bimax) and three cell type prediction methods (SC3[95], SINCERA[104], and SNN-Cliq[105]). It is found that the average ARI score, as a representative, of QUBIC2 was 37%, 220%, 632%, 151%, and 185% higher than the other five biclustering methods, respectively; and was 30%, 67% and 62% higher than the three cell type prediction methods, respectively. QUBIC2 and QUBIC were the top two biclustering tools, respectively, in terms of median values on the three criteria. Both surpassed the performance of SC3 (median value from 100 runs, denoted by the red dash line in each panel of Figure 7B). Besides, ISA always demonstrated the smallest variance across the three validation criteria. The FMI values of each tool were more stable than the other two values. Figure 7C showcased one cell type classification result obtained by QUBIC2. The result was in good agreement with the reference cell labels and QUBIC2 correctly grouped the three major cell types (8_cell_embryo, Morulae, and late_blastoCyst).

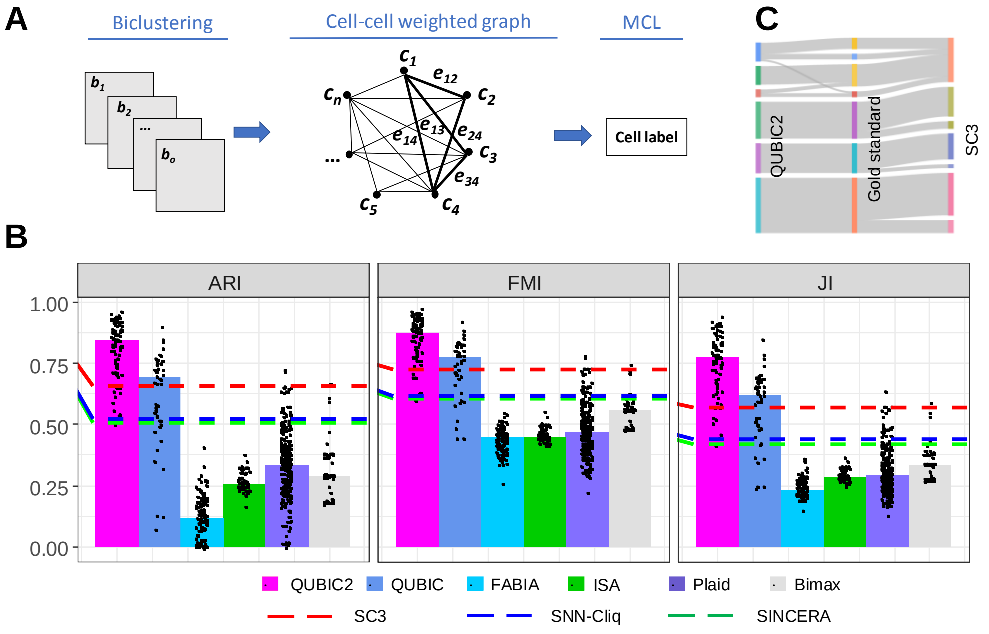


Figure 7. A. Computational pipeline for cell type classification. This pipeline consists of three steps: biclustering, generation of weighted cell-cell matrix and clustering using MCL; B. Benchmark of QUBIC2 against five popular biclustering algorithms. Each panel shows the similarity between the inferred labels and the reference labels quantified by ARI, FW and JI, respectively. Each algorithm was applied >40 times to the same dataset. The three indices were calculated for each run of the respective methods (black dots). Bars represent the median of the distribution of black dots. The red dash lines correspond to the benchmark performance of SC3 (ARI: 0.6549, FMI: 0.7243, JI: 0.5671); C. Sankey diagram comparing the 7 clusters obtained with SC3 (right layer) and 6 clusters obtained with QUBIC2 (left layer). The middle layer corresponds to the seven reference clusters. The widths of the lines linking nodes from two layers correspond to the number of cells they have in common.

2.6 Application of QUBIC2 on Temporal and Spatial scRNA-Seq Data

When spatial and temporal information is available, scRNA-Seq can reveal more biological insights beyond cell types. In this section, QUBIC2 was applied on two temporal (and) and two spatial scRNA-Seq datasets, respectively, to explore the temporal and spatial organization of cells.

2.6.1 Data

The time series lung scRNA-Seq dataset (GSE52583) with 152 cells and 15,174 genes from was downloaded from http://www.cs.cmu.edu/~jund/scdiff/download/data/. The cells were collected at three time points: E14, E16, and E18. Another time series scRNA-Seq data with 527 cells and 13991 genes (GSE48968) was downloaded from the GEO database, in which the RPKM values are available.

The Mouse olfactory bulb spatial transcriptomic data was downloaded from [106], which contains 280 cells and 15,981 genes. Ståhl *et al.* [106] classified the cells into five clusters that correspond to well-defined morphological layers. The cells use coordinates as IDs, and the cell layers information was manually extracted using the ST viewer (https://github.com/SpatialTranscriptomicsResearch/st_viewer), based on the coordinate information. The raw reads of mouse spatial scRNA-Seq data GSE60402 were retrieved from the SRA database [107], and the RPKM values for it were calculated using software packages TopHat [108] and Cufflink [109]. GSE60402 was split into three subsets according to sample information. The detailed information of the selected and divided datasets is listed in Table 4.

Table 4. Summary of GSE60402

GEO Accession ID	Data ID	Description	#Cells	#Genes
GSE60402	GSE60402-Mutant	From Gfra1 mutant sample	94	11094
GSE60402	GSE60402-Wildtype1	From wild type mouse 1	124	10037
GSE60402	GSE60402-Wildtype2	From wild type mouse 2	94	10714

2.6.2 Results

QUBIC2 identified five biclusters from GSE52583. Three of the five biclusters contain time-specific cells. In particular, bicluster BC002 consists of cells exclusively from E14; bicluster BC003 includes cells that only from E16; and bicluster BC004 has cells coming from E18 (**Figure 8A**). Functional enrichment analyses of the component genes from these three biclusters were carried out based on DAVID [110], and the results showed that genes in BC002 mainly related to cell cycle, cell division, and mitosis; BC003 genes were enriched with ribosome, translation, and structural constituent of ribosome; and spliceosome-related genes were grouped in BC004.

In addition to identifying biclusters corresponding to specific time point, QUBIC2 can also be used to find biclusters with time-dependent patterns. Here QUBIC2 was used to analyze a scRNA-Seq data with mouse dendritic cells (DCs) collected at 1h, 2h, 4h and 6h after treatment with pathogenic agent lipopolysaccharide (LPS) and untreated controls (GSE48968) [111]. In total, 51 biclusters were identified in the datasets treated with LPS. For each bicluster, the Fisher exact test was conducted on its constituting samples to assess if significant over-representation by any time points could be found within the bicluster. For those biclusters showing significant association with the time-course, a pathway enrichment analysis was conducted to infer the biological characteristics of the bicluster. In detail, pathway enrichment analysis is undertaken and the statistical significance of each enriched pathway is assessed by using a hypergeometric test (statistical significance cutoff = 0.005) against 4,725 curated gene sets in the MsigDB database, which includes 1,330 canonical KEGG, Biocarta and Reactome pathways, and 3,395 gene sets representing expression signatures derived from experiments with genetic and chemical perturbations, together with 6,215 Mouse GO terms each containing at least 5 genes [112, 113]. In the end, 30 biclusters that are significantly over-represented by one or several consecutive time points were identified in the LPS dataset ($\alpha=0.005$, $P<1\text{e-}22$), and six of them showed clear time dependence (**Figure 8B**). Specifically, bicluster BC013 consists of untreated samples and samples collected at 1h, which represents the earliest response to LPS and enriches multiple immune response pathways. Bicluster BC005 consists mainly of untreated samples and samples collected at 1h and 2h, which also is enriched with immune response pathways but with more responses to a virus, T cell chemotaxis and so on. BC009 and BC001 are enriched by samples collected at 1h and 2h, covering a wider range of stress-response pathways, suggesting that the activation of stress response pathways and altered metabolisms as secondary responses after the early immune response. BC025 and BC002 consist of samples collected at 4h and 6h, and their genes enrich pathways associated with alterations in cell morphogenesis, migration, cell-cell junction and so on. Overall these observations suggest that our analysis can identify all the major responses to the LPS treatment in a time-dependent manner.

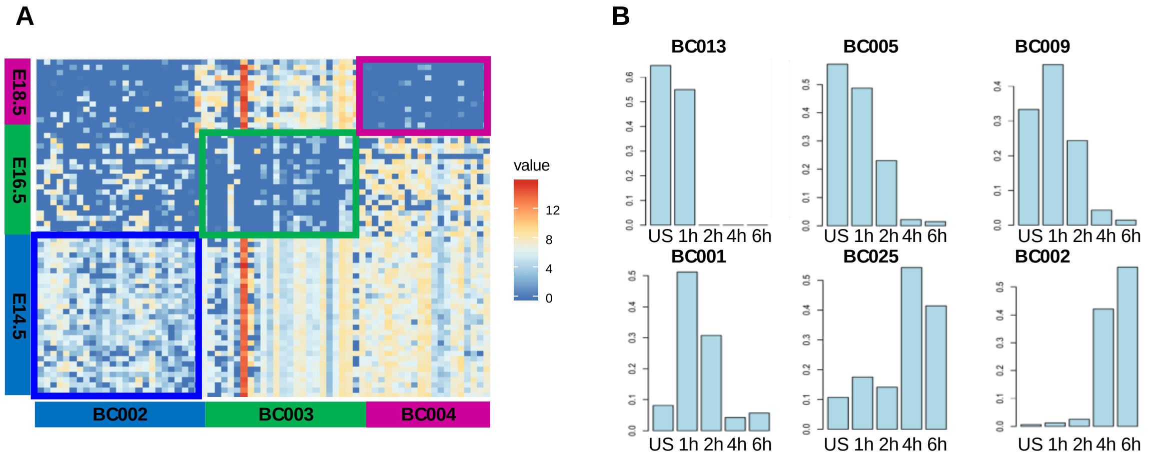


Figure 8. A. Visualization of three biclusters (BC002, BC003, and BC004) selected based on the specificity to time point; B. Time-dependent distribution of cells in six selected biclusters identified in the LPS data. In each histogram, the five bars from left to right show the proportion of the untreated samples and samples collected at 1h, 2h, 4h and 6h after the LPS treatment.

Then QUBIC2 was applied to a mouse spatial scRNA-Seq dataset with 280 cells. The cells were classified into five clusters that correspond to five distinct morphological layers in [106] (**Figure 9A**). Five biclusters were predicted. Among them, the bicluster BC000 consists of cells mainly from the granular layer; the bicluster BC001 contains cells from the mitral layer and glomerular layer; the bicluster BC002 includes cells mostly from the olfactory nerve layer (**Figure 9B**). Functional annotation showed that BC000 mainly enriches plasma membrane, cell membrane, and cell projection; BC001 enriches synapse, neuron projection, and cell projection; and BC002 enriches cell projection.

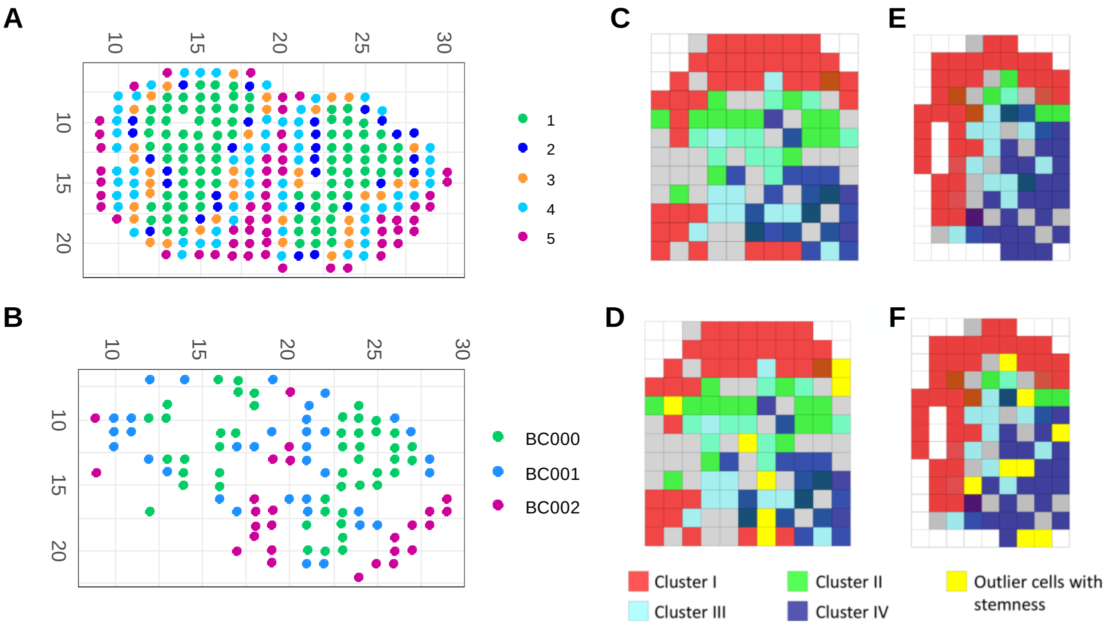


Figure 9. A. The coordinates of cells correspond to five morphological layers (1. Granular cell layer; 2. Mitral cell layer; 3. Outer plexiform layer; 4. Glomerular layer; 5. Olfactory nerve layer); B. The coordinates of cells from three selected biclusters; C. The spatial coordinates of samples in the four biclusters identified in wild-type 1 mouse; Colors red, green, cyan and dark blue represent samples in four different biclusters; D. In addition to the coordinates of bicluster samples, the yellow cubes represent significant outlier samples; E. The same information as in C except the samples are from wild-type 2 mouse; F. The same information as in D except the samples are from wild-type 2 mouse.

Finally, another spatial scRNA-Seq dataset (GSE60402) with samples dissected from three mouse medial ganglionic eminence tissues and known spatial coordinates was analyzed. QUBIC2 was applied, and 37, 40, and 120 biclusters were identified in the mutant, wild-type 1, and wild-type 2 datasets, respectively. Further investigation on the spatial distribution of cells in each bicluster showed that all the four spatial biclusters with distinct expression patterns by cell cycle, cell morphogenesis, and neuron development genes, as reported in the original study [114], were identified by QUBIC2. It is noteworthy that the outliers with highly expressed stem cell markers tend to be located at the intermediate region between two adjacent (or overlapping) biclusters in the three datasets as shown in Figure 9D and 9F. Our interpretation is that these location-dependent expression patterns may be caused by parallel and independent differentiations from common stem cells.

2.7 Summary

The combination of biclustering and large-scale gene expression data holds a promising potential in elucidating the functional pathways/networks encoded in a genome. However existing biclustering tools fail to generate satisfactory results from high-resolution RNA-Sequencing (RNA-Seq) data due to the lack of full consideration of (*i*) intrinsic characteristics of RNA-Seq data, e.g., the massive zeros in both bulk and scRNA-Seq data, and (*ii*) the underlying transcriptional regulation signals of gene expression. Here we presented a novel biclustering algorithm, QUBIC2, for the analysis of large-scale bulk RNA-Seq and scRNA-Seq data. QUBIC2 (*i*) used a truncated model to handle the unquantifiable errors caused by zeros, (*ii*) adopted an information-divergency objective function to optimize to-be-identified biclusters, (*iii*) utilized a Core-Dual strategy to recruit novel genes and optimize parameters in identifying a bicluster, and (*iv*) developed a size-based *P*-value calculation method to evaluate the statistical significances of all the identified biclusters.

Our method validation on comprehensive data sets showed that QUBIC2 had significant advantages in the functional module detection area, outperforming five widely-used biclustering methods. The proposed *P*-value calculation method based on bicluster size did make sense, which may facilitate the evaluation of all the identified biclusters, especially from less-annotated organisms. The cell type classification pipeline, based on QUBIC2, worked well and outperformed the state-of-the-art performance of SC3. By utilizing time-dependent data, QUBIC2 discovered biclusters specific to time point and identified a cascade of immune responses to the external pathogenic treatment. From the spatial transcriptomic data, QUBIC2 discovered that spatially adjacent single cells might have high co-expression patterns, and particularly, two distinct spatially clustered cells may be derived initially from the same stem cell. We believe that QUBIC2 can serve biologists as a useful tool to extract novel biological insights from large-scale RNA-Seq data.

Although the advantages mentioned above, to fully excavate the potential of scRNA-Seq data, there are several shortcomings needed to be overcome. First, as sequencing costs decrease, larger scRNA-Seq datasets will become increasingly common; thus, the scalability to large dataset and efficiency of tools will become more and more critical. Currently, the discretization and Dual searching functions of QUBIC2 are time-consuming on large-scale datasets. Based on our test, it takes 17 minutes to discretize a dataset with 4,297 rows and 466 columns (a desktop with 48.0GB memory, Intel Core i7-6700, and 3.40GHz). Given a dataset with 22,846 genes and 100 conditions, the running time while using Dual strategy are generally 2 minutes longer than that without Dual. The OpenMP method will be implemented in the EM steps for discretization, and more efficient heuristics algorithm will be designed to optimize the dual searching of biclustering.

Another challenge involves the interpretation of time-series and spatial data. For example, in the GSE52583 data, QUBIC2 could only separate cells collected at different time points, yet the further differentiation stage information was not captured. For the mouse olfactory bulb data, QUBIC2 did not separate cells from adjacent layers. To deal with this drawback, we need to combine biclustering with other statistical methods specifically designed for time series and spatial gene expression data.

It is noteworthy that many other kinds of methods can be used for gene expression data analysis. Forty-two module detection tools covering five main approaches were reviewed in [70], and the authors concluded that decomposition methods outperformed all other strategies, including biclustering methods. Meanwhile, they also observed that QUBIC and FABIA had higher performance on human and synthetic data. We compared two top-rated decomposition methods and two top clustering methods with QUBIC2 and QUBIC on a human scRNA-Seq data; the results showed that QUBIC2 surpassed both decomposition and clustering methods (**Figure 10**). In the future, we will carry out a more comprehensive comparison between QUBIC2 and other decomposition and network-based methods, aiming to give a systematical evaluation of the power of computational techniques on scRNA-Seq data.

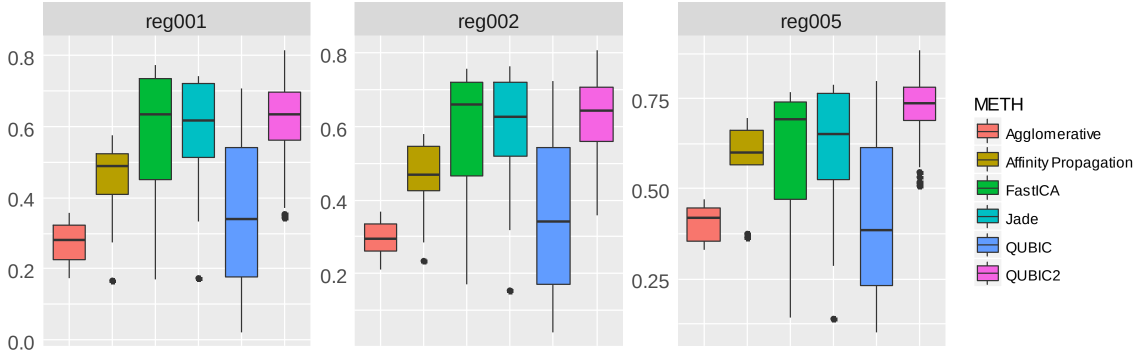
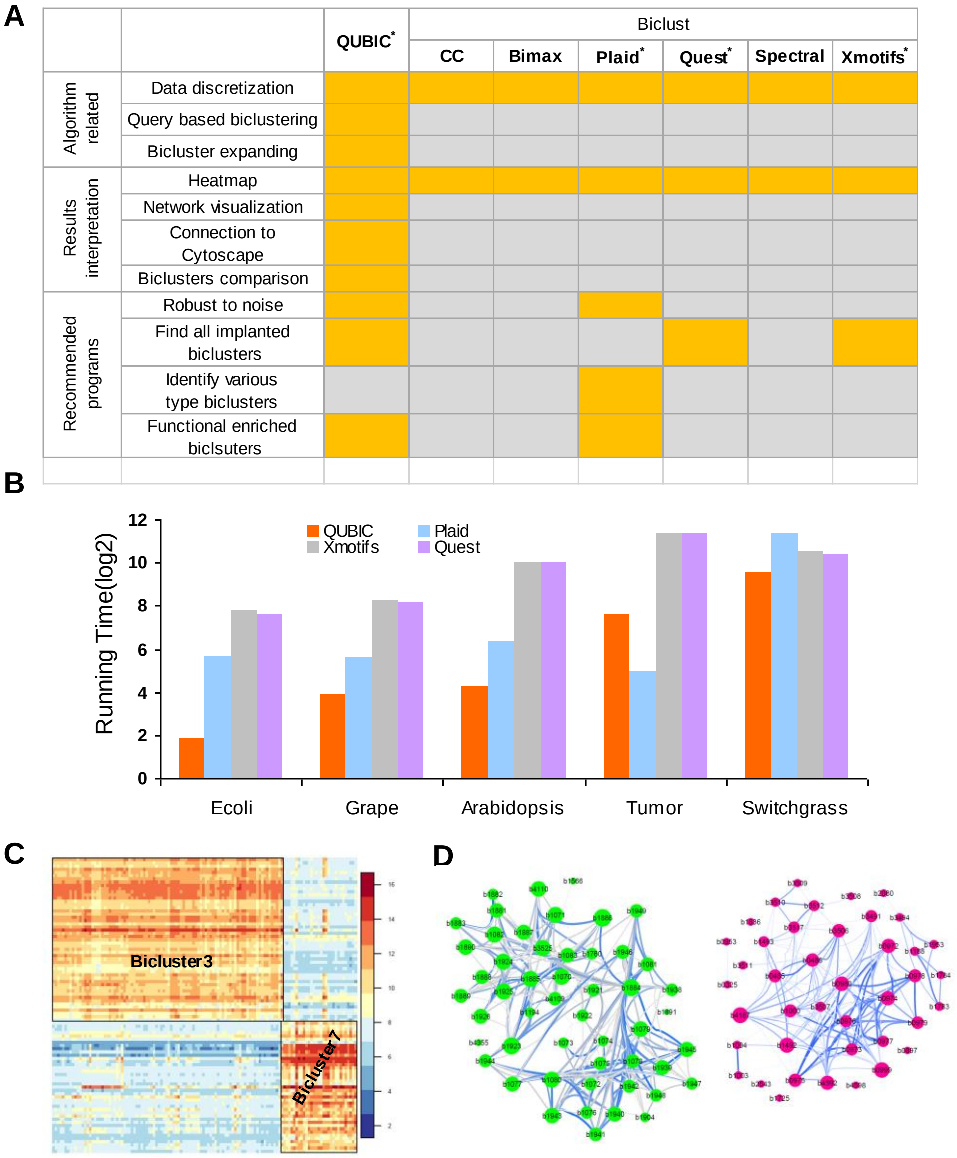


Figure 10. Performance of QUBIC2, QUBIC, two decomposition methods and two clustering methods in term of F score on a human scRNA-Seq data.

CHAPTER 3: QUBICR- A Biconductor Package for Qualitative Biclustering Analysis of Gene Co-expression Data

Biclustering is a widely accepted approach for gene expression data mining. Several biclustering algorithms have been published in the past two decades, and QUBIC has been reviewed as one of the best programs by several review studies. To enable the biclustering users lacking comprehensive computational background, a web server of QUBIC was developed in 2012 [23]. Since gene expression datasets keep increasing in scale, we developed this user requested R package of QUBIC (QUBIC-R for short), to provide an efficient optimized implementation and to eliminate large-scale data submission to a webserver.

The unique features of QUBIC-R include: (i) biclustering is integrated with analyses functions, i.e., data discretization, query-based biclustering, bicluster expanding, biclusters comparison, heatmap visualization and co-expression network elucidation (**Figure 11A**); (ii) the QUBIC source code is optimized and converted from GNU C to C++, thus has better memory control and is more efficient than the original QUBIC (an average 82.4% saving of running time); (iii) on five large-scale datasets, QUBIC-R consistently performs the best among four popular tools according to the running time (**Figure 11B**). In the following part, I will present the main features of QUBICR.



		QUBIC*	Biclust					
			CC	Bimax	Plaid*	Quest*	Spectral	Xmotifs*
Algorithm related	Data discretization	🟨	🟨	🟨	🟨	🟨	🟨	🟨
	Query based biclustering	🟨	⬜	⬜	⬜	⬜	⬜	⬜
	Bicluster expanding	🟨	⬜	⬜	⬜	⬜	⬜	⬜
Results interpretation	Heatmap	🟨	🟨	🟨	🟨	🟨	🟨	🟨
	Network visualization	🟨	⬜	⬜	⬜	⬜	⬜	⬜
	Connection to Cytoscape	🟨	⬜	⬜	⬜	⬜	⬜	⬜
	Biclusters comparison	🟨	⬜	⬜	⬜	⬜	⬜	⬜
Recommended programs	Robust to noise	🟨	⬜	⬜	🟨	⬜	⬜	⬜
	Find all implanted biclusters	🟨	⬜	⬜	⬜	🟨	⬜	🟨
	Identify various type biclusters	⬜	⬜	⬜	🟨	⬜	⬜	⬜
	Functional enriched biclsuters	🟨	⬜	⬜	🟨	⬜	⬜	⬜

Figure 11. A. Comparison of QUBIC-R and 6 R packages in biclust. Yellow color indicates that a package provides the function or is recommended in a specific biclustering application and gray color represents the opposite; B. Comparison of running time among four recommended programs, annotated with asterisks in Figure 11A; C. Heatmap visualization of two biclusters identified in *E. coli* data; D. Co-expression networks of Figure 11C biclusters. Green nodes represent bicluster #3 and red nodes represent bicluster #7. The larger the size of a node, the higher its degree of presence; and the thicker an edge the heavier its co-expression value is.

3.1 Implementation

QUBIC-R package [115] is developed for the R statistical computing environment, and is freely available at http://bioconductor.org/packages/release/bioc/html/QUBIC.html. It depends on the *biclust* package developed by Kaiser et al. [48] to be compatible with the *biclust* output. Its output format can also be used by network analysis software, such as Cytoscape [116].

The original QUBIC program, written in GNU C with POSIX library, is limited in its portability. A memory leak may occur if the primary functions are called more than once. This problem was addressed by refactoring the C source code and transforming it into C++. Specifically, to avoid memory leak, we changed the majority of data structures and replaced C pointers by STL containers. We also optimized core function structures to facilitate future package updates and developments. The program efficiency has been significantly increased with the same predicting results (**Figure 11A**). An input data as large as 30,000×30,000 can be finished within half an hour (detailed limits test is in **Figure 12**). All the computational experiments were conducted on a computer with Windows 7 x64, Memory 48G, Intel Core i7-6700 3.4G.

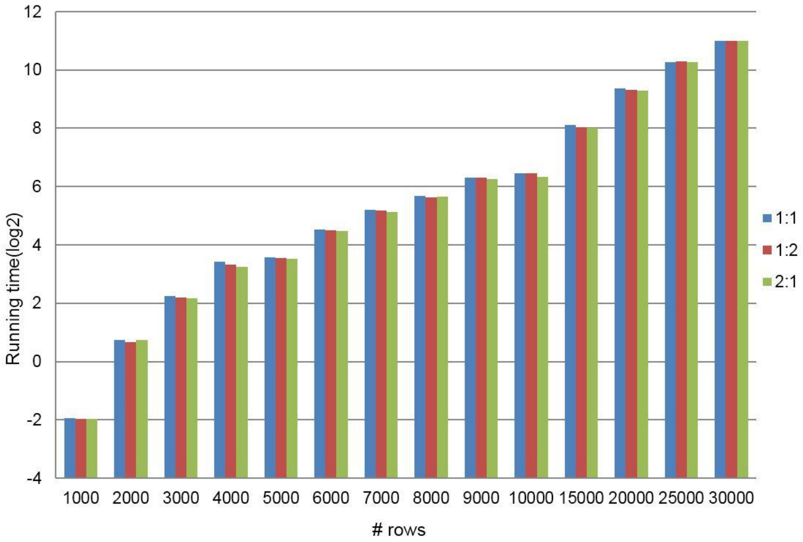


Figure 12. Data limit test of QUBICR on simulated datasets. In this test, n-by-n matrixes were generated with increasing number of n (1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, 15,000, 20,000, 25,000, 30,000). Each matrix was planted with several non-overlapped 10-by-10, 10-by-20, or 20-by-10 biclusters (corresponding to the 1:1,1:2 or 2:1 row-column-ratio, respectively). Default parameters of QUBICR were applied (c=0.95, r=1, f=1, q=0.06, o=100, k=max(ncol(x)%/%20,2)), and the running time in y-axis is in its log2 scale.

3.2 Functions

(i) ***qudiscretize*** is useful to obtain discrete gene expression matrix. This matrix can be used in other biclustering program, where -1 represents lowly express, 0 represents normally express, and 1 represents highly express. For example:

```
library(QUBIC)
matrix1 <- ecoli[1:3, 1:4]
matrix1
##       dinI_U_N0025 dinP_U_N0025 lexA_U_N0025 lon_U_N0025
## b4634     9.077693     9.225537     9.138900    9.114353
## b3241     7.122300     7.195453     7.051193    7.124200
## b3240     7.184417     7.336610     7.283377    7.188263

matrix2 <- qudiscretize(matrix1)
matrix2
##       dinI_U_N0025 dinP_U_N0025 lexA_U_N0025 lon_U_N0025
## b4634           -1            1            0           0
## b3241            0            1           -1           0
## b3240           -1            1            0           0
```

(ii) ***BCQU*** and (iii) ***BCQUD*** are used to perform biclustering for continuous and discretized gene expression data, respectively:

```
# QUBic-R on continuous data
res <- biclust(ecoli, method = BCQU(), f = 0.25,verbose=F)
res
##
## Number of Clusters found:  19
##
## Cluster sizes:
##                    BC 1 BC 2 BC 3 BC 4 BC 5
## Number of Rows:     437  121   51  108  103
## Number of Columns:   29   45   94   44   38

# QUBIC-R on discrete data
res1 <- biclust(x = qudiscretize(ecoli), method = BCQUD(), f = 0.25, ve
rbose=F)
res1
# QUBIC algorithm can be called independently via qubiclust and qubiclu
st_d for both continuous and discrete data, respectively:

res2 <- qubiclust(x = ecoli, f = 0.25, verbose=F)
res2
res3 <- qubiclust_d(x = qudiscretize(ecoli), f = 0.25)
res3
# note that res, res1, res2 and res3 are the same
```

(iv) Using the parameter ***weight***, a user can conduct a query-based biclustering, with additional biological information.

Specifically, a user can input additional biological information and utilize that information to guide the biclustering progress in QUBICR, using the newly-added parameter ***weight***. This kind of function is so-called query-based biclustering and has been widely applied in bioinformatics [24, 117] . The format of this input file should be supported by ***igraph***, e.g., a file with three columns with column #1 and #2 representing the gene names and column #3 being the score of the two genes. QUBICR will (step 1) rank all the gene pairs in this additional input file, according to the corresponding biological information (e.g., protein-protein interaction information or co-regulation relationship), in an increasing order; (step 2) rank all the gene pair in original gene expression data, according to their co-expression similarity trained from the QUBIC algorithm, in an increasing order; and (step 3) add the two ranks together for each gene pair. Then the summed ranks will be used as a new weight for each gene pair for all the following biclustering procedures. It is noteworthy that if a gene pair appears in this additional input file but not in original gene expression file (or in the opposite situation), its rank in step 1 will be assigned as 0.

In this example, the instance file “511145.protein.links.v10.txt” was downloaded from string (http://stringdb.org/download/protein.links.v10/511145.protein.links.v10.txt.gz). Note that after using the ***weight*** parameter, the output biclusters changed.

```
# Conduct a query-based biclustering by adding the weight parameter
```

```
library(QUBIC)
library(QUBICdata)
data("ecoli")
library(igraph)
file = "511145.protein.links.v10.txt ";
graph = read.graph(file, format = "ncol")
get.edgelist(graph, names = TRUE)
E(graph)$weight
weight <- get.adjacency(graph, attr = "weight")
res0 <- biclust(ecoli, method = BCQU(),verbose = F)
res0

res4<- biclust(ecoli, method = BCQU(), weight = weight, verbose = F)
res4
```

(v) Using the ***seedbicluster*** parameter, a user can expand existing biclusters by recruiting more genes according to specified consistency level. The existing biclusters can be any biclustering results obtained from QUBICR or from any other algorithms in the

biclust package. This function has been successfully applied in [118] and a flowchart of this function can be found in Figure 1 of [118].

In the following example, we expand previously obtained biclusters by QUBICR (res). Note that number of genes in some biclusters increase after expanding (e.g., 437 genes in BC1 from res vs 593 genes in BC1 from res5).

```
res5 <- biclust(x = ecoli, method = BCQU(), seedbicluster = res, f = 0.
25,verbose = F)
res5
##
## Number of Clusters found:  19
##
## Cluster sizes:
##                   BC 1 BC 2 BC 3 BC 4 BC 5
## Number of Rows:    593  151   51  110  117
## Number of Columns:  29   45   94   44   38
```

(vi) Using the parameter ***showinfo***, the biclustering results from different algorithms or from a same algorithm with different combinations of parameter can be compared. Specifically, we can compare the number of detected biclusters, the row number and column number of the first bicluster, the area of the first biclusr, the overlap of first two biclusters , and so on.

```
test <-ecoli [1:50,]
res6 <-biclust(test, method = BCQU(), verbose = F)
res7 <- biclust (test, method = BCCC())
res8 <- biclust(test, method = BCBimax())
showinfo (test, c(res6, res7, res8))
```

(vii) The function ***quheatmap*** can visualize the identified biclusters using heatmap in support of overall expression pattern analysis, either for a single bicluster or for two biclusters.

```
# heatmap for single bicluster
```

```
par(mar = c(5, 4, 3, 5) , cex.lab = 1.1, cex.axis = 0.5, cex.main = 1.1
)
quheatmap(ecoli, res, number = 4)
```

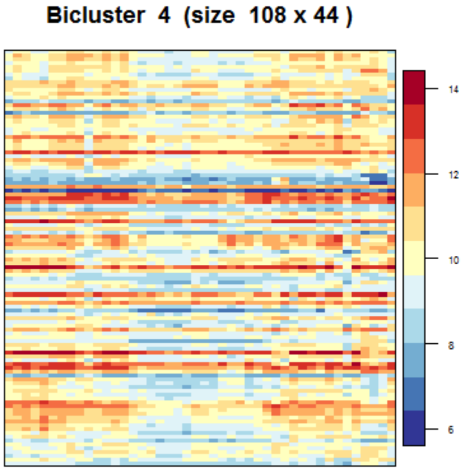


Figure 13. Heatmap for the 4th bicluster identified in the E. coli data.

```
# heatmap for two biclusters
par(mar = c(5, 4, 3, 5) , cex.lab = 1.1, cex.axis = 0.5, cex.main = 1.1
)
quheatmap(ecoli, res, number = c(3,7))
```

(viii) We can construct and visualize network for the identified biclusters, using the function ***qunetwork***, either for a single bicluster or for two biclusters.

In the gene co-expression network, each node represents a gene, and a pair of nodes is connected with an edge if they have a significant correlation (with the cutoff as 0.6 in default). Specifically, for a single bicluster with m genes and n conditions, we used the m-*by*-n expression matrix to calculate the correlation between each pair of genes in the network. For two given biclusters, whose gene sets are $\{m_1\}$ and $\{m_2\}$ and condition sets are $\{n_1\}$ and $\{n_2\}$, we used the expression matrix, with genes $|\{m_1 \cup m_2\}|$ and conditions $|\{n_1 \cup n_2\}|$, to generate the correlation coefficient scores among genes. QUBICR provides three methods to calculate the correlation, i.e., Pearson, Kendall and Spearman, to facilitate different preference in practical application.

```
# Construct the network for the 4th identified bicluster in the E.coli
data
library(qgraph)
net1 <- qunetwork(ecoli, res, number = 4, group = 4, method = "spearman
")
qgraph(net1[[1]], groups = net1[[2]], layout = "spring", minimum = 0.6,

color = cbind(rainbow(length(net1[[2]]) -  1), "gray"), edge.label = F)
```

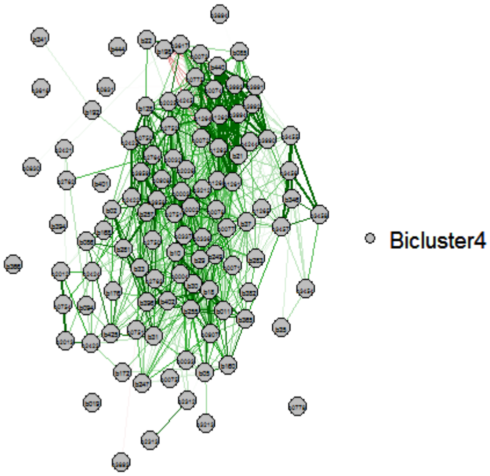


Figure 14. Network for the 4th bicluster identified in the *E. coli* data.

```
# Construct the network for the 3th and 7th bicluster in the E.coli dat
a
net2 <- qunetwork(ecoli, res, number = c(3, 7), group = c(3, 7), method
 = "spearman")
qgraph(net2[[1]], groups = net2[[2]], layout = "spring", minimum = 0.6,
 legend.cex = 0.5, color = c("red", "blue", "gold", "gray"), edge.label
 = FALSE)
```

(ix) The function ***qunet2xml*** can convert the constructed networks into XGMML format, facilitating further functional enrichment analysis (e.g. DAVID) and advanced network visualization (e.g. Cytoscape, Biomax and JNets)

```
# Output overlapping heatmap XML, could be used in other software such
# as Cytoscape, Biomax or JNets
sink("tempnetworkresult.gr")
qunet2xml(net2, minimum = 0.6, color = c("red", "blue", "gold", "gray")
)
sink()
# We can use Cytoscape, Biomax or JNets open file named
# tempnetworkresult.gr
```

3.3 Summary

Biclustering algorithms facilitate researchers in the identification of co-expressed gene subsets in their gene expression dataset and have become a useful approach for the interpretation of gene expression profile data. Our R package implements a well-cited biclustering algorithm, QUBIC. It provides more efficient source code and fully integrated functions to identify and analyze biclusters and visualize identified biclusters and corresponding co-expression networks. This package is a powerful tool for gene expression data mining and co-expression network modeling.

CHAPTER 4: Application of Biclustering on Biological and Biomedical Data

The advent of much-improved biotechnology and the decreased associated costs have generated a massive amount of biological and biomedical data. NGS allows for rapid generation of larger volumes of biological information than ever before. Also, large amounts patient clinical data are generated through NGS and Electronic Health Record (EHR), which presents significant opportunities for knowledge discoveries in biomedical research [119]. These complex and large volumes of data, collected from different sources, have changed the way biological and biomedical research is conducted [120, 121]. 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.

Furthermore, with the advancement of informatics technology, EHR contains sufficient information that can be transformed into disease phenotypes [122]. 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 [122], 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 [123-126], and (ii) genotype biclusters [122].

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 to 2017 (**Figure 15**). 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 is reflecting absolute gene expression with continuous fluorescence intensity values [127], while RNA-Seq data measures the relative expression level using discrete, positive, and highly skewed read counts [88, 128-130]. 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 scRNA-Seq data [82, 131]. 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 unique design in 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 [38, 39, 41, 42]. 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 are presented in Figure 15. The numbers of biclustering studies on algorithm design and application were similar at the earliest stage when few tools were available. The proportion of application related studies decreased relative to algorithm design until 2010. In the 1,650 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 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.

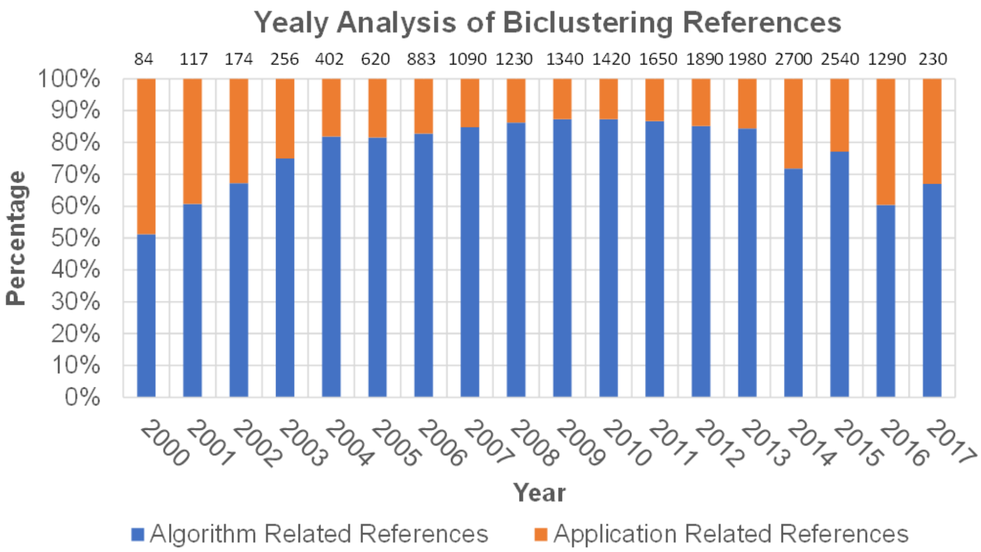


Figure 15. Yearly comparison of biclustering algorithm development and algorithm application related studies. The references in 2017 were collected as of 03/26/2017. The overall annual reference numbers that shown on the top of each bar were collected by searching the keyword "biclustering" on google scholar, and proportion of algorithm development shown in blue was captured by adding the keyword "algorithm," and the rest are considered as application related, which were shown in orange.

4.1 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 [132, 133]. Biotechnology has developed very fast in the last two decades, from traditional arrays (e.g., microarray and tilling array) to NGS (e.g., DNA-Seq, RNA-Seq, and 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 towards the goal of generating a system-level understanding.

4.1.1 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 [134]. 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 [135], and KEGG [101] are three dominant representatives of such comprehensive databases; RegulonDB is one of the most widely-used resources for *E. coli K-12* gene regulation [102]; The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) offers genomics, epigenomic and proteomic data for thousands of tumor samples across more than 20 types of cancer; and PlantTFDB provides comprehensive genomic transcriptional factor (TF) repertoires of green plants [136]. 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 [137, 138] and co-evolution [139, 140]. 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 [141]. 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 [99]. 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 [142, 143]. 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 [144].

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 [145]. Thus, experimental validation of the predictions is needed. However, the percentage of unannotated genes is very high even in well-studied model organisms [93] (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 [142], and in most cases, they rely on computational validation (e.g., cross-validation [146] and random forest [143]) and published literature support.

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 [147]. Recently, researchers proposed using genome-scale ChIP-Seq data for the validation of the prediction of CEMs [144]. Table 5 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.

Table 5. Case studies of Functional annotation of unclassified genes

Data	**Methods**	**Tools/Databases**	**Outcomes**	**Refs**
Functional annotation of Yeast				
Microarray (6,200 ORFs under 515 conditions)	• Biclustering for gene classification	SAMBA	2,406 biclusters; 196 annotations of unknown genes;	[146]
	• Functionally assign the unannotated genes in biclusters to the most abundant class;	SGD [148]		
	• Cross-validation for annotation assessment.	-		
Functional annotation of plant genomes				
Microarray (21,031 genes of Arabidopsis under 351 conditions)	• Biclustering on known PCW genes;	QUBIC	417 seed biclusters; 2,438 candidate PCW genes co-expressed with 349 PCW genes.	[147]
	• Expand biclusters to include additional genes;	QUBIC		
	• Construct co-expression network;	Cytoscape		
	• Predict and annotate motifs in promoter regions of co-expressed genes in each module.	WeederTFBS;		
		MotifSampler		
		CompariMotif		
		PLACE		

		AGRIS		

Microarray
(122,973 probes of Switchgrass, 94 conditions)
	• Homologous mapping of identified CW genes;	Tblastn	991 homologs CW genes;
			104 clusters of co-expressed genes;
			823 new PCW genes;
			112 new genes.
				[144]

	• Assign mapped genes to CW-associated functions;	DAVID		
	• Biclustering of mapped genes and expand for new candidates;	QUBIC		
	• Identify motifs for each bicluster;	-		
	• Validate prediction by annotated Arabidopsis CW genes	PCWGD *		
Functional annotation of Human and Mouse				
A correlation matrix with associations among mouse lincRNA, protein-coding genes, and lincRNAs	• Identify lincRNA;	ChIP-Seq	Sets of lincRNAs associated with a diverse range of functions including cell proliferation, immune surveillance, muscle development, etc..	[142]
	• Create association matrix of lincRNA and protein-coding genes;	GSEA		
	• Biclustering to identify functional modules consisting of lincRNAs and protein-coding genes;	SAMBA		
	• Assign putative functions to each lincRNA;	-		
	• Validate inferred biological functions for lincRNAs.	-		
65 human microarray datasets and GO function categories	• Discover network patterns based on frequent item sets and biclustering;	-	1,126 functions assigned to 895 genes (779 knowns and 116 unknowns).	[143]
	• Design network topology statistic based on graph random walk;	-		
	• Assess functional annotation by a random forest method.	-		

Note: - denotes for no specific existed tools and this also applies to all the following tables.

* Purdue Cell-Wall-Genomics Database (https://cellwall.genomics. purdue.edu)

4.1.2 Modularity Analysis

Compared to 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 [133]. Increasing evidence indicates that biological systems are inherently modular [149-

151], therefore, modularity analysis has been widely applied to investigate the organization and dynamics of biological systems at different levels, i.e., module identification, module dynamic 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., miRNA sponge modules in [152] and miRNA-mRNA modules in [153]), functionally related genes/proteins or any other manually defined clusters [154]. Depending on the target modules, different inputs and strategies are needed. For example, scRNA-Seq gene expression data was utilized to identify molecularly distinct subtypes of cells that contribute different brain functions [155]; and an integrated correlation matrix was derived from expression data with target site information to predict miRNA-mRNA functional modules [153]; time series expression data provides valuable information regarding the cellular dynamic activity, thus it is often utilized to identify temporal transcriptional modules that consist of activated genes at consecutive time points [38]. 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 [61]. Similar strategies have been implemented with the integration of other supporting data types (e.g., operon prediction, ChIP-Seq data, and network connections) [53].

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 [152, 153, 156]. Where expression profiles are available in multiple evolutionarily correlated species, researchers can conduct inter-specific comparisons and investigate the underlying evolutionary story. For example, Waltman *et al*. performed biclustering of multiple-species data and then used a conservation score to identify conserved modules among these species [157]. Based on co-regulation modules, Yang *et al.* 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 [158]. With temporal modules, the dynamic regulatory interaction can be explored. Gonçalves *et al*. [159] 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 [160]. Table 6 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.

Table 6. Case studies of Modularity analysis.

Data	**Methods**	**Tools/Databases**	**Outcomes**	**Refs**
	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	[152]
	• 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	[153]
	• Biclustering to identify functional modules;	BUBBLE		
	• Visualize and evaluate modules.	miRMAP		
SC-RNA-Seq (3,005 mouse cortical cells)	• Biclustering	BackSPIN	47 distinct cell subclasses	[155]
Regulatory modules				
Microarray data (*S. cerevisiae* under 2,200 conditions); upstream and downstream Sequences.	• Biclustering	COALESCE	450 regulatory modules	[61]
Microarray (*M. tuberculosis under 2,325 measurements*); and 154 TFs ChIP-Seq data	• Biclustering	cMonkey2	600 modules	[53]
Time series expression for 2,884 genes of *Saccharomyces cerevisiae* in response to heat stress under five time-points	• Biclustering	CCC-Biclustering	167 biclusters; Regulatory snapshots of documented regulators at each time point	[38, 159]
	• Ranking the prioritize prominent regulators targeting each the modules at each time point	Regulatory Snapshots		
	• Graphically depict the regulatory activity in a module	Baiacu; BiGGEsTs		
Evolutionary study				
Three normalized expression matrixes (*B. subtilis*, *B. anthracis, and L. monocytogenes*); upstream Sequences; metabolic and signaling pathways, co-membership in an operon and phylogenetic profile networks	• Biclustering on expression data;	FD-MSCM	150 biclusters	[157]
	• Evaluate the conservation between biclusters	-		
Microarray (4117 orthologs in 15, 14, and	• Biclustering to predict co-regulated modules;	ISA	1,181 modules	[158]

17 tissue groups in rice, maize, and Arabidopsis, respectively)	• 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	[160]
	• Construct bicluster distance matrix;	-		
	• Construct neighbor-joining tree and designate host subsystem	-		

4.1.3 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 lncRNAs) and targets, and edges are regulatory interaction directing from regulators to targets. Analyzing these networks provides systematic views and novel insights in understanding underlying mechanisms controlling cellular processes. Table 7 shows some examples in network analysis, mainly focus on network inference and network decomposition.

Compared with random networks, one distinct characteristic of the biological networks is modularity, forming dense subgraphs [161, 162]. Several computational approaches have utilized 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[132]. This approach has been successfully applied in several other studies [163-165]. On the other hand, Tanay *et al.* [150] 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[163-165]; and Tanay *et al.* combined expression data, various interactions, and phenotypes [150].

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[166-168]. Henriques and Madeira [35] 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.

Table 7. Case studies of Biological networks elucidation.

Inputs	Methods	Tools/Databases	Outputs	Refs
Yeast transcriptional network				
Nearly 1,000 *Saccharomyces cerevisiae* expression profiles; 110 TF binding location profiles; 30 growth profiles; 1,031 protein interaction; 4,177 complex interactions and 1,175 known interactions from MIPS	• Modeling genomic information as weighted graph	-	665 significant modules; Global Yeast molecular network	[150]
	• Biclustering	SAMBA		
	• Generate module graph and explore associations between modules			
Methanogenesis regulatory network				

Microarray (1,661 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;	cMonkey	166 biclusters;
			GRN model including a set of 1,227 EF and TF regulatory influences that inter-link the regulation of 1,661 genes
				[163]

	• Construct GRN to infer transcriptional influences of each bicluster;	Inferelator		
	• Visualize GRN;	Cytoscape		
		Gaggle		
	• Use TF knockout experiment and extra data and to validate the GRN model	-		
Mycobacterium tuberculosis regulatory network				

Microarray data (*Mycobacterium tuberculosis* genes under 2,325 conditions);
Upstream regions of all genes;
~5000 Operon prediction from MicrobesOnline;
~250,000 protein interactions from String
	• Biclustering to identify co-regulated gene subsets;	cMonkey	598 biclusters;
			A global regulatory network covering 98% of MTB genes
				[164]

	• Construct GRN model to infer transcriptional influences of each bicluster;	Inferelator		
	• Validate the GRN model using new datasets; Visualize Network.	BioTapestry		
Phaeodactylum tricornutum regulatory network				
RNA-Seq (1,214 Phaeodactylum tricornutum genes from 179 samples); Genome annotation, Chloroplastic and mitochondrial genomic information, functional annotation, Protein-protein interactions	• Biclustering to identify putatively co-regulated genes;	cMonkey2	121 biclusters covering 1,214 metabolic genes and TFs	[165]
	• Construct regulatory network to infer regulatory influences;	Inferelator		
	• GO enrichment analysis to identify potential biological processes carried out by the co-regulated genes	-		

Biological network decomposition				
Two Gene interaction networks for yeast; Two PPIs from *E. coli* and human	• Biclustering	BicNET	modules with heightened biological significance	[35]
	• Assess biological significance of retrieved modules	GOrilla		

4.2 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 [169]. Besides, understanding the association between above genomic information and specific diseases has led to the discovery of new drugs [170].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 [170]. 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 detecting disease subtype, identifying biomarker and gene signatures of disease and gene–drug association.

4.2.1 Disease Subtype Identification

Disease subtype could provide a framework for the development of more accurate biomarkers by stratification of patient populations [171]. It can be defined by related molecular characteristics or clinical features [172]. 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 [49]. 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 [31, 173]. Hence, *denovo* 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 [63]. 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 8 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 [174, 175]. 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 ($r = number\ of\ rows, c = number\ of\ columns$), 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, E represents the expected frequency under the null hypothesis, which is computed by:

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

The test statistics will be compared to 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 Example S1). Meanwhile, interpretation of the identified biclusters in gene dimension can be carried out, more details of biomarker and gene signatures detection can be found in the next section.

Table 8. Case studies of disease subtype identification.

Data	Methods	Tools/Databases	Outcomes	Refs
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	[31]
Gastric cancer				

Microarray data for 80 paired gastric cancer and reference tissues from non-treated patients	• Biclustering on gene expression data for bicluster identification;	QUBIC [31];	Pathways associated with cancer development; identified gastric cancer subtypes	[176]
	• Pathway enrichment analysis.	DAVID [177]		
		KOBAS [178]		
		HPID [179]		
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 non-adjuvant treated tumor data set;	ComBat [180]	Similar clinical features associated with tumor within the same cluster	[181]
	• Biclustering to identify molecular-based tumor subgroup;	cMonkey [52]		
	• Determine molecular classifiers for each bicluster;	PAM [182]		
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;	PageRank [183]	Cancer subtypes	[63]
	• Weighted biclustering algorithm based on a semi-nonnegative matrix tri-factorization.	NCIS [63]		
Colon and lung cancers				
290 colon cancer samples, each has 384 methylation probes covering 151 cancer-specific differentially methylated region (cDMRs); Expression levels of 12,625 genes in 56 patients having lung cancer	• Heterogeneous sparse singular value decomposition (HSSVD) based Biclustering	-	Variance biclusters of methylation data in cancer versus normal patients using colon cancer data; cancer subtype patterns using lung cancer data	[173]

4.2.2 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 [184, 185]. Biclustering has been used with human gene expression data to identify cancer subtype patterns [31, 63, 173, 181, 186], metabolic pathways highly related to cancer progression[176], marker genes of a specific cancer type/subtype [8, 155], and clinical risk factors of cancer [187]. Also, studies of common or rare diseases have used biclustering of human gene expression data to identify phenotype-genotype associations [188, 189], dysregulated transcription modules [190], and genetic risk variants [191]. 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 [176]. 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 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 [187].

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 [188, 189, 191]. With the developments of RNA-Seq, whole transcriptomic data are becoming available to characterize and quantify gene expression [192]. 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 [8]. Using biclustering, researchers can not only group cells into subpopulations but also identify biologically important gene signatures for each class simultaneously [193]. For example, Zeisel *et al.* [155] 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.* [8] 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. Meanwhile, new biclustering applications are emerging, such as detecting disease marker genera from gut biome [194]. 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 [194].

These identified biclusters were subjected to downstream analysis of functional gene annotation [186, 188], gene network inference [188] or phenomic analysis [188, 189, 191]. Most of the gene functional annotations were done through the UCSC Genome Browser [195]. 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 9 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 compared with those in basic biological applications, regarding the data sources, data preprocessing methods and downstream statistical analyses.

Table 9. Case studies of Biomarker and gene signatures detection.

Data	**Methods**	**Tools/Databases**	**Outcomes**	**Refs**
Breast cancer				
Association matrix of 1,008 gene expression microarray profiles of primary breast tumors	• Biclustering binary data matrix.	iBBiG	Modules associated with clinical covariates in breast cancer	[187]
Matrix of normalized miRNA Sequencing expression profiles	• Biclustering to evaluate miRNA deregulation;	ISA[100]	12 different miRNA clusters	[186]
	• Validate each bicluster by an external repository of different groups of miRNAs in human species;	MetaMirClust [196]		
		UCSC [195]		
	• Compare results with a different biclustering algorithm.	SAMBA [146]		
Osteoporosis				
Regression coefficients matrix of 1,109 unique SNPs associated with 23 studied traits from the GWAS data of the Framingham Osteoporosis Study	• GWAS database mining;	Tagger [197]	SNP-phenotype connections; Highly genetically correlated traits; Candidate genes identified for	[188]
	• Biclustering on matrix of SNPs against phenotypes;	Bayesian biclustering [198]		
	• Gene annotation and identification of enriched canonical pathway and gene network inference.	UCSC [199]		
		IPA		

			multiple bone traits	
Williams-Beuren syndrome				
Normalized skin fibroblast microarray dataset including 9,329 probe sets and 96 samples	• Identify transcriptional modules;	ISA[100]	72 dysregulated modules were found	[190]
	• Test modules containing at least ten genes for dysregulation using hypergeometric distribution.	-		
Schizophrenia				
8,023 subjects, 4,196 patients, and 3,827 controls, with 2,891 SNPs in each subject	• Perform biclustering for both phenotype and genotype data;	bioNMF [62]	Causally cohesive genotype-phenotype relations	[189]
	• Cross-correlate phenotype and genotype biclusters;	-		
	• Organize and encode relations into topologically organized networks;	PGMRA [189]		
	• Estimate genotype associated disease risk.	SKAT[200]		
Complex diseases				
p-value matrix of 466423 SNPs in 32 independent diseases/traits	• Identify biclusters of diseases/traits and SNPs	SparseBC [201]	Genetic risk variants for complex diseases	[191]
		LAS [57]		
		SSVD [202]		
	• Map detected SNPs to genes	-		

4.2.3 Gene-drug Association

In drug development, it is vital to understand the complicated responses in the human body to various drug treatments [203, 204]. 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 [205]. 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 [206]. 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 10 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 [206]. 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 [206, 207]. Enrichment of drug annotation terms can be assessed by various databases, such as STRING [208] and DAVID [177], for identification of transcriptional factors linked to these biclusters [206, 209, 210].

Table 10. Case studies of gene-drug association.

Data	**Methods**	**Tools/Databases**	**Outcomes**	**Refs**
		Drug-gene associations		
NCI-60 cancer cell line in drug response; Gene expression data	• Identify co-modules of drugs and genes;	PPA [207]	859 co-modules were identified, and drug-gene associations were predicted more accurately than other algorithms	[207]
	• Test drug-gene association.	DrugBank [211]		
		Connectivity Map [212]		
		Drug-Induced Transcriptional Modules		
6,100 gene expression profiles of human cancer cell treated with 1,309 small	• Biclustering drug-induced gene expression profiles [100];	ISA [50]	Drug-induced transcriptional modules	[206]
	• Hypergeometric test for significance assessment of overlaps among gene members;	-		

molecules from CMap **[212]**; 1,743 expression profiles from liver tissues of drug-treated rats **[213]**.	• Predict novel gene functions by comparing modules of human cancer and rat liver cell lines;	STRING [208]		
	• Test enriched gene functions and identified biological themes among transcriptional modules.	DAVID [177]		
Transcriptional factors (TFs) for drug-associated gene modules				
7,056 genome-wide expression profiles of five different human cell lines treated with 1,309 chemical agents at different dosages from CMap **[212]**	• Identify drug-gene modules by biclustering method;	FABIA [32]	Links between 28 modules with 12 TFs were detected	[209]
	• Indicate Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) information associated with genes in modules;	DAVID [110]		
	• Use cumulative hypergeometric test to evaluate drug target enrichment.	-		
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;	-	Transcriptional effects of compounds	[210]
	• Identify transcriptional modules;	FABIA [32]		
	• Identify transcriptional modules related to the desired effect using target-related bioassay measurements.	PSVM [214]		

4.3 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 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 16, 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. 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 results [31, 181, 215-217]. The connections between data and methods offer model analysis paths for researchers to use depending on the characteristics of their data.

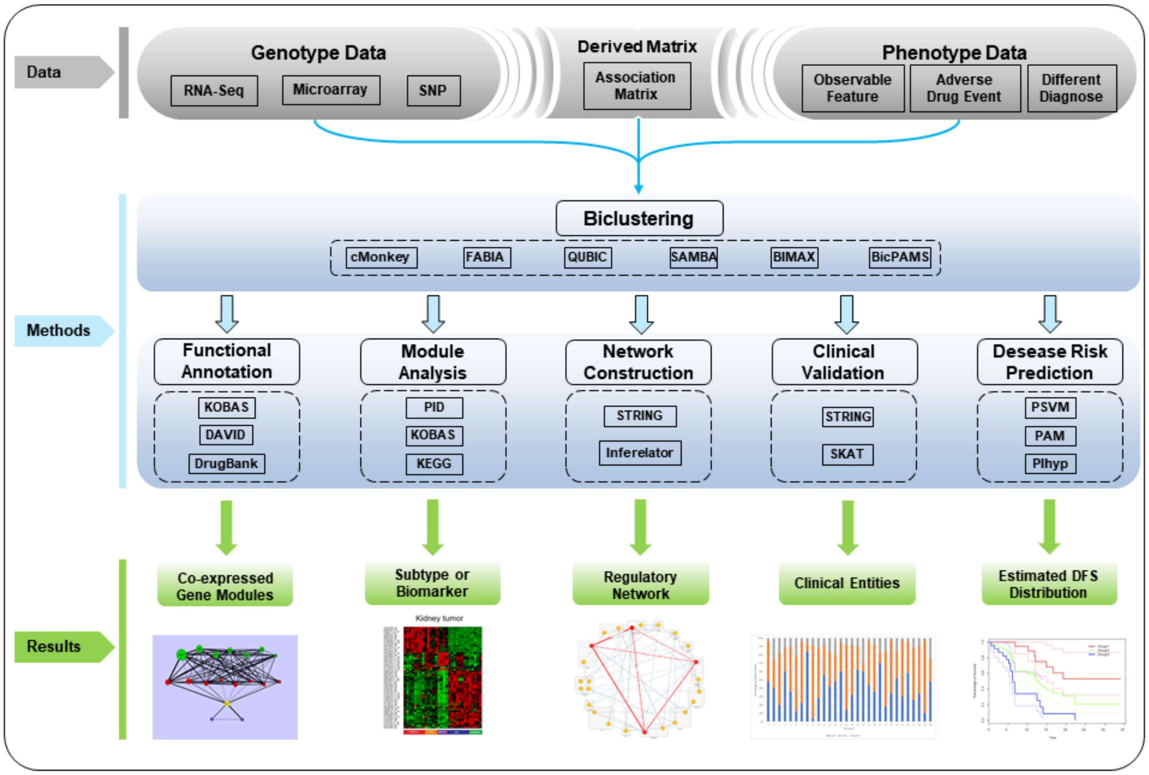


Figure 16. The overall workflow of biclustering application mechanism related to upstream and downstream process. Three layers are shown to provide the path from raw data, appropriate analytical methods/tools to various cases of the result. The power of biclustering is illustrated by the ability to generate co-expressed gene modules, subtype or biomarker, regulatory networks, clinical entities and estimated disease-free survival (DFS) distribution.

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 the characteristics of biological and biomedical data into account, are still needed to facilitate larger-scale applications of biclustering.

REFERENCES

1. Selvaraj, S. and J. Natarajan, *Microarray Data Analysis and Mining Tools.* Bioinformation, 2011. **6**(3): p. 95-99.
2. Marioni, J.C., et al., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome research, 2008. **18**(9): p. 1509-1517.
3. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. **320**(5881): p. 1344-1349.
4. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nature reviews genetics, 2011. **12**(2): p. 87-98.
5. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews genetics, 2009. **10**(1): p. 57-63.
6. Garber, M., et al., *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nature methods, 2011. **8**(6): p. 469-477.
7. Saliba, A.-E., et al., *Single-cell RNA-seq: advances and future challenges.* Nucleic Acids Research, 2014. **42**(14): p. 8845-8860.
8. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nat Methods, 2017. **14**(5): p. 483-486.
9. Aibar, S., et al., *SCENIC: single-cell regulatory network inference and clustering.* Nat Methods, 2017. **14**(11): p. 1083-1086.
10. Prince, M.E., et al., *Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma.* Proc Natl Acad Sci U S A, 2007. **104**(3): p. 973-8.
11. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.
12. Xu, X., et al., *Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor.* Cell, 2012. **148**(5): p. 886-95.
13. Ulitsky, I., et al., *Expander: from expression microarrays to networks and functions.* Nat Protoc, 2010. **5**(2): p. 303-22.
14. Hartigan, J.A., *Direct clustering of a data matrix.* Journal of the american statistical association, 1972. **67**(337): p. 123-129.
15. Cheng, Y. and G.M. Church. *Biclustering of expression data*. in *Ismb*. 2000.
16. Lazzeroni, L. and A. Owen, *Plaid models for gene expression data.* Statistica sinica, 2002: p. 61-86.
17. Jaitin, D.A., et al., *Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types.* Science, 2014. **343**(6172): p. 776-779.
18. Shekhar, K., et al., *Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics.* Cell, 2016. **166**(5): p. 1308-1323.e30.
19. Kluger, Y., et al., *Spectral biclustering of microarray data: coclustering genes and conditions.* Genome Res, 2003. **13**(4): p. 703-16.
20. Murali, T. and S. Kasif. *Extracting conserved gene expression motifs from gene expression data*. in *Pacific symposium on biocomputing*. 2003.
21. Gu, J. and J.S. Liu, *Bayesian biclustering of gene expression data.* BMC Genomics, 2008. **9 Suppl 1**(1): p. S4.

22. Chen, Y., et al., *Genome-wide discovery of missing genes in biological pathways of prokaryotes.* BMC Bioinformatics, 2011. **12 Suppl 1**: p. S1.
23. Zhou, F., et al., *QServer: a biclustering server for prediction and assessment of co-expressed gene clusters.* PLoS One, 2012. **7**(3): p. e32660.
24. Dhollander, T., et al., *Query-driven module discovery in microarray data.* Bioinformatics, 2007. **23**(19): p. 2573-80.
25. De Smet, R. and K. Marchal, *An ensemble biclustering approach for querying gene expression compendia with experimental lists.* Bioinformatics, 2011. **27**(14): p. 1948-56.
26. Zhao, H., et al., *Query-based biclustering of gene expression data using Probabilistic Relational Models.* BMC Bioinformatics, 2011. **12 Suppl 1**: p. S37.
27. Madeira, S.C. and A.L. Oliveira, *A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series.* Algorithms Mol Biol, 2009. **4**: p. 8.
28. Tanay, A., R. Sharan, and R. Shamir, *Discovering statistically significant biclusters in gene expression data.* Bioinformatics, 2002. **18 Suppl 1**: p. S136-44.
29. Bergmann, S., J. Ihmels, and N. Barkai, *Iterative signature algorithm for the analysis of large-scale gene expression data.* Physical review E, 2003. **67**(3): p. 031902.
30. Prelić, A., et al., *A systematic comparison and evaluation of biclustering methods for gene expression data.* Bioinformatics, 2006. **22**(9): p. 1122-1129.
31. Li, G., et al., *QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.* Nucleic Acids Res, 2009. **37**(15): p. e101.
32. Hochreiter, S., et al., *FABIA: factor analysis for bicluster acquisition.* Bioinformatics, 2010. **26**(12): p. 1520-1527.
33. Henriques, R. and S.C. Madeira, *BiC2PAM: constraint-guided biclustering for biological data analysis with domain knowledge.* Algorithms Mol Biol, 2016. **11**: p. 23.
34. Bunte, K., et al., *Sparse group factor analysis for biclustering of multiple data sources.* Bioinformatics, 2016. **32**(16): p. 2457-63.
35. Henriques, R. and S.C. Madeira, *BicNET: Flexible module discovery in large-scale biological networks using biclustering.* Algorithms Mol Biol, 2016. **11**: p. 14.
36. Alzahrani, M., et al., *Gracob: a novel graph-based constant-column biclustering method for mining growth phenotype data.* Bioinformatics, 2017: p. btx199.
37. Hochreiter, S., et al., *FABIA: factor analysis for bicluster acquisition.* Bioinformatics, 2010. **26**(12): p. 1520-7.
38. Madeira, S.C., 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): p. 153-65.
39. Gonçalves, J.P., S.C. Madeira, and A.L. Oliveira, *BiGGEsTS: integrated environment for biclustering analysis of time series gene expression data.* BMC research notes, 2009. **2**(1): p. 124.
40. Medina, I., et al., *Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W210-3.

41. Henriques, R. and S.C. Madeira, *BicPAM: Pattern-based biclustering for biomedical data analysis.* Algorithms for Molecular Biology, 2014. **9**(1): p. 27.
42. Henriques, R., F.L. Ferreira, and S.C. Madeira, *BicPAMS: software for biological data analysis with pattern-based biclustering.* BMC Bioinformatics, 2017. **18**(1): p. 82.
43. Bentham, R.B., K. Bryson, and G. Szabadkai, *MCbiclust: a novel algorithm to discover large-scale functionally related gene sets from massive transcriptomics data collections.* Nucleic Acids Res, 2017. **45**(15): p. 8712-8730.
44. Barkow, S., et al., *BicAT: a biclustering analysis toolbox.* Bioinformatics, 2006. **22**(10): p. 1282-3.
45. Cheng, K.O., et al., *BiVisu: software tool for bicluster detection and visualization.* Bioinformatics, 2007. **23**(17): p. 2342-4.
46. Santamaria, R., R. Theron, and L. Quintales, *BicOverlapper 2.0: visual analysis for gene expression.* Bioinformatics, 2014. **30**(12): p. 1785-6.
47. Wu, C.J. and S. Kasif, *GEMS: a web server for biclustering analysis of expression data.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W596-9.
48. Kaiser, S., et al., *biclust: Bicluster algorithms.* R package version 0.7, 2009. **2**.
49. Zhang, Y., et al., *QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data.* Bioinformatics, 2016: p. btw635.
50. Csardi, G., Z. Kutalik, and S. Bergmann, *Modular analysis of gene expression data with R.* Bioinformatics, 2010. **26**(10): p. 1376-7.
51. Kluger, Y., et al., *Spectral biclustering of microarray data: coclustering genes and conditions.* Genome research, 2003. **13**(4): p. 703-716.
52. Reiss, D.J., N.S. Baliga, and R. Bonneau, *Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks.* BMC Bioinformatics, 2006. **7**(1): p. 280.
53. Reiss, D.J., et al., *cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism.* Nucleic Acids Res, 2015. **43**(13): p. e87.
54. Lee, M., et al., *Biclustering via sparse singular value decomposition.* Biometrics, 2010. **66**(4): p. 1087-1095.
55. Gu, J. and J.S. Liu, *Bayesian biclustering of gene expression data.* BMC genomics, 2008. **9**(1): p. 1.
56. Bozdağ, D., J.D. Parvin, and U.V. Catalyurek. *A Biclustering Method to Discover Co-regulated Genes Using Diverse Gene Expression Datasets*. in *Bioinformatics and Computational Biology*. 2009. Berlin, Heidelberg: Springer Berlin Heidelberg.
57. Shabalin, A.A., et al., *Finding large average submatrices in high dimensional data.* The Annals of Applied Statistics, 2009: p. 985-1012.
58. Zeisel, A., et al., *Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science, 2015. **347**(6226): p. 1138-1142.
59. Kutalik, Z., J.S. Beckmann, and S. Bergmann, *A modular approach for integrative analysis of large-scale gene-expression and drug-response data.* Nature biotechnology, 2008. **26**(5): p. 531-539.
60. Madeira, S.C., et al., *Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm.* IEEE/ACM

Transactions on Computational Biology and Bioinformatics (TCBB), 2010. **7**(1): p. 153-165.

61. Huttenhower, C., et al., *Detailing regulatory networks through large scale data integration.* Bioinformatics, 2009. **25**(24): p. 3267-74.
62. Pascual-Montano, A., et al., *bioNMF: a versatile tool for non-negative matrix factorization in biology*. BMC bioinformatics, 2006. **7**(1): p. 366.
63. Liu, Y., et al., *A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression.* BMC Bioinformatics, 2014. **15**: p. 37.
64. Waltman, P., et al., *Multi-species integrative biclustering.* Genome biology, 2010. **11**(9): p. R96.
65. Gusenleitner, D., et al., *iBBiG: iterative binary bi-clustering of gene sets.* Bioinformatics, 2012. **28**(19): p. 2484-2492.
66. Bryan, K. and P. Cunningham. *Bottom-Up Biclustering of Expression Data*. in *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*. 2006.
67. Tan, K.M. and D.M. Witten, *Sparse biclustering of transposable data.* Journal of Computational and Graphical Statistics, 2014. **23**(4): p. 985-1008.
68. Bentham, R.B., K. Bryson, and G. Szabadkai, *MCbiclust: a novel algorithm to discover large-scale functionally related gene sets from massive transcriptomics data collections.* Nucleic Acids Research, 2017. **45**(15): p. 8712-8730.
69. Eren, K., et al., *A comparative analysis of biclustering algorithms for gene expression data.* Briefings in bioinformatics, 2013. **14**(3): p. 279-292.
70. Saelens, W., R. Cannoodt, and Y. Saeys, *A comprehensive evaluation of module detection methods for gene expression data.* Nature Communications, 2018. **9**(1): p. 1090.
71. Madeira, S.C. and A.L. Oliveira, *Biclustering Algorithms for Biological Data Analysis: A Survey.* IEEE/ACM Trans. Comput. Biol. Bioinformatics, 2004. **1**(1): p. 24-45.
72. Bozdağ, D., A.S. Kumar, and U.V. Catalyurek. *Comparative analysis of biclustering algorithms*. in *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*. 2010. ACM.
73. Chia, B.K.H. and R.K.M. Karuturi, *Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms.* Algorithms for molecular biology, 2010. **5**(1): p. 23.
74. Padilha, V.A. and R.J. Campello, *A systematic comparative evaluation of biclustering techniques.* BMC Bioinformatics, 2017. **18**(1): p. 55.
75. Li, L., et al., *A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data.* BioData Mining, 2012. **5**(1): p. 8.
76. Pontes, B., R. Giraldez, and J.S. Aguilar-Ruiz, *Biclustering on expression data: A review.* J Biomed Inform, 2015. **57**: p. 163-80.
77. Busygin, S., O. Prokopyev, and P.M. Pardalos, *Biclustering in data mining.* Computers & Operations Research, 2008. **35**(9): p. 2964-2987.
78. Eren, K., et al., *A comparative analysis of biclustering algorithms for gene expression data.* Brief Bioinform, 2013. **14**(3): p. 279-92.

79. Kasim, A., et al., *Applied Biclustering Methods for Big and High-Dimensional Data Using R*. 2016: Chapman & Hall/CRC. 433.
80. Li, G., et al., *A new framework for identifying cis-regulatory motifs in prokaryotes.* Nucleic acids research, 2010. **39**(7): p. e42-e42.
81. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic acids research, 2009. **37**(suppl_2): p. W202-W208.
82. Lun, A.T., K. Bach, and J.C. Marioni, *Pooling across cells to normalize single-cell RNA sequencing data with many zero counts.* Genome Biol, 2016. **17**(1): p. 75.
83. Bacher, R. and C. Kendziorski, *Design and computational analysis of single-cell RNA-sequencing experiments.* Genome biology, 2016. **17**(1): p. 63.
84. Bengtsson, M., et al., *Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels.* Genome Res, 2005. **15**(10): p. 1388-92.
85. Lu, C. and R.D. King, *An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems.* Bioinformatics, 2009. **25**(16): p. 2020-2027.
86. Hebenstreit, D., et al., *RNA sequencing reveals two major classes of gene expression levels in metazoan cells.* Mol Syst Biol, 2011. **7**: p. 497.
87. Glaus, P., A. Honkela, and M. Rattray, *Identifying differentially expressed transcripts from RNA-seq data with biological variation.* Bioinformatics, 2012. **28**(13): p. 1721-1728.
88. Rau, A. and C. Maugis-Rabusseau, *Transformation and model choice for RNA-seq co-expression analysis.* Brief Bioinform, 2017.
89. Reuter, J.A., et al., *Simul-seq: combined DNA and RNA sequencing for whole-genome and transcriptome profiling.* Nature Methods, 2016.
90. Cohen, A.C., *Simplified estimators for the normal distribution when samples are singly censored or truncated.* Technometrics, 1959. **1**(3): p. 217-237.
91. Stegle, O., S.A. Teichmann, and J.C. Marioni, *Computational and analytical challenges in single-cell transcriptomics.* Nat Rev Genet, 2015. **16**(3): p. 133-45.
92. Sha Cao, T.S., Xin Chen, Qin Ma, Chi Zhang, *A probabilistic model-based bi-clustering method for single-cell transcriptomic data analysis.* bioRxiv, 2017.
93. Monk, J., J. Nogales, and B.O. Palsson, *Optimizing genome-scale network reconstructions.* Nat Biotechnol, 2014. **32**(5): p. 447-52.
94. Sun, X. and A.B. Nobel, *On the size and recovery of submatrices of ones in a random binary matrix.* Journal of Machine Learning Research, 2008. **9**(Nov): p. 2431-2453.
95. Kiselev, V.Y., et al., *SC3: consensus clustering of single-cell RNA-seq data.* Nature methods, 2017.
96. Overbeek, R., et al., *The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes.* Nucleic Acids Research, 2005. **33**(17): p. 5691-5702.
97. Keseler, I.M., et al., *The EcoCyc database: reflecting new knowledge about Escherichia coli K-12.* Nucleic Acids Research, 2017. **45**(Database issue): p. D543-D550.

98. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.* Science, 2016. **352**(6282): p. 189-96.
99. Castillo-Davis, C.I. and D.L. Hartl, *GeneMerge—post-genomic analysis, data mining, and hypothesis testing.* Bioinformatics, 2003. **19**(7): p. 891-892.
100. Bergmann, S., J. Ihmels, and N. Barkai, *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): p. 031902.
101. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res, 2017. **45**(D1): p. D353-d361.
102. Gama-Castro, S., et al., *RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond.* Nucleic Acids Res, 2016. **44**(D1): p. D133-43.
103. Yan, L., et al., *Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells.* Nat Struct Mol Biol, 2013. **20**(9): p. 1131-9.
104. Guo, M., et al., *SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis.* PLOS Computational Biology, 2015. **11**(11): p. e1004575.
105. Xu, C. and Z. Su, *Identification of cell types from single-cell transcriptomes using a novel clustering method.* Bioinformatics, 2015. **31**(12): p. 1974-80.
106. Ståhl, P.L., et al., *Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.* Science, 2016. **353**(6294): p. 78-82.
107. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update.* Nucleic acids research, 2013. **41**(D1): p. D991-D995.
108. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-1111.
109. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature biotechnology, 2010. **28**(5): p. 511-515.
110. Dennis, G., et al., *DAVID: database for annotation, visualization, and integrated discovery.* Genome biology, 2003. **4**(9): p. R60.
111. Shalek, A.K., et al., *Single-cell RNA-seq reveals dynamic paracrine control of cellular variation.* Nature, 2014. **510**(7505): p. 363-9.
112. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-1740.
113. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.
114. Zechel, S., et al., *Topographical transcriptome mapping of the mouse medial ganglionic eminence by spatially resolved RNA-seq.* Genome Biol, 2014. **15**(10): p. 486.
115. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics.* Genome Biol, 2004. **5**(10): p. R80.
116. Smoot, M.E., et al., *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics, 2011. **27**(3): p. 431-432.
117. Alqadah, F., et al. *Query-based Biclustering using Formal Concept Analysis*. in *SDM*. 2012. SIAM.

118. Wang, S., et al., *Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis.* BMC Plant Biology, 2012. **12**(1): p. 1-12.
119. Luo, J., et al., *Big Data Application in Biomedical Research and Health Care: A Literature Review.* Biomed Inform Insights, 2016. **8**: p. 1-10.
120. Wu, X., et al., *Data mining with big data.* IEEE transactions on knowledge and data engineering, 2014. **26**(1): p. 97-107.
121. Swan, M., *The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery.* Big Data, 2013. **1**(2): p. 85-99.
122. Hripcsak, G. and D.J. Albers, *Next-generation phenotyping of electronic health records.* J Am Med Inform Assoc, 2013. **20**(1): p. 117-21.
123. Burgel, P.R., J.L. Paillasseur, and N. Roche, *Identification of clinical phenotypes using cluster analyses in COPD patients with multiple comorbidities.* Biomed Res Int, 2014. **2014**: p. 420134.
124. Han, M.K., et al., *Chronic obstructive pulmonary disease phenotypes: the future of COPD.* Am J Respir Crit Care Med, 2010. **182**(5): p. 598-604.
125. Henriques, R., C. Antunes, and S.C. Madeira, *A structured view on pattern mining-based biclustering.* Pattern Recognition, 2015. **48**(12): p. 3941-3958.
126. Carreiro, A.V., et al., *Prognostic prediction through biclustering-based classification of clinical gene expression time series.* J Integr Bioinform, 2011. **8**(3): p. 175.
127. Yeung, K.Y., et al., *Model-based clustering and data transformations for gene expression data.* Bioinformatics, 2001. **17**(10): p. 977-87.
128. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nature biotechnology, 2016. **34**(5): p. 525-527.
129. Pachter, L., *Models for transcript quantification from RNA-Seq.* arXiv preprint arXiv:1104.3889, 2011.
130. Rau, A., et al., *Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models.* Bioinformatics, 2015. **31**(9): p. 1420-7.
131. Bacher, R. and C. Kendziorski, *Design and computational analysis of single-cell RNA-sequencing experiments.* Genome Biol, 2016. **17**(1): p. 63.
132. Babu, M.M., et al., *Structure and evolution of transcriptional regulatory networks.* Curr Opin Struct Biol, 2004. **14**(3): p. 283-91.
133. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
134. Gillis, J. and P. Pavlidis, *"Guilt by association" is the exception rather than the rule in gene networks.* PLoS Comput Biol, 2012. **8**(3): p. e1002444.
135. Consortium, G.O., *Gene ontology consortium: going forward.* Nucleic acids research, 2015. **43**(D1): p. D1049-D1056.
136. Jin, J., et al., *PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants.* Nucleic Acids Res, 2017. **45**(D1): p. D1040-d1045.
137. Obayashi, T., 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**(Database issue): p. D863-9.

138. Yang, M.Q., L.M. Koehly, and L.L. Elnitski, *Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes.* PLoS Comput Biol, 2007. **3**(4): p. e72.
139. Oldham, M.C., S. Horvath, and D.H. Geschwind, *Conservation and evolution of gene coexpression networks in human and chimpanzee brains.* Proc Natl Acad Sci U S A, 2006. **103**(47): p. 17973-8.
140. Mezey, J.G., et al., *Coordinated evolution of co-expressed gene clusters in the Drosophila transcriptome.* BMC Evol Biol, 2008. **8**(1): p. 2.
141. Ma, Q., et al., *Computational analyses of transcriptomic data reveal the dynamic organization of the Escherichia coli chromosome under different conditions.* Nucleic Acids Research, 2013. **41**(11): p. 5594-5603.
142. Guttman, M., et al., *Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.* Nature, 2009. **458**(7235): p. 223-7.
143. Huang, Y., et al., *Systematic discovery of functional modules and context-specific functional annotation of human genome.* Bioinformatics, 2007. **23**(13): p. i222-9.
144. Chen, X., et al., *Genome-scale identification of cell-wall-related genes in switchgrass through comparative genomics and computational analyses of transcriptomic data.* BioEnergy Research, 2016. **9**(1): p. 172-180.
145. Horan, K., et al., *Annotating genes of known and unknown function by large-scale coexpression analysis.* Plant Physiol, 2008. **147**(1): p. 41-57.
146. Tanay, A., R. Sharan, and R. Shamir, *Discovering statistically significant biclusters in gene expression data.* Bioinformatics, 2002. **18 Suppl 1**(suppl 1): p. S136-44.
147. Wang, S., et al., *Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis.* BMC Plant Biol, 2012. **12**(1): p. 138.
148. Cherry, J.M., et al., *SGD: Saccharomyces Genome Database.* Nucleic Acids Res, 1998. **26**(1): p. 73-9.
149. Wagner, G.P., M. Pavlicev, and J.M. Cheverud, *The road to modularity.* Nat Rev Genet, 2007. **8**(12): p. 921-31.
150. Tanay, A., et al., *Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 2981-6.
151. Purnick, P.E. and R. Weiss, *The second wave of synthetic biology: from modules to systems.* Nat Rev Mol Cell Biol, 2009. **10**(6): p. 410-22.
152. Zhang, J., et al., *Identifying miRNA sponge modules using biclustering and regulatory scores.* BMC bioinformatics, 2017. **18**(3): p. 44.
153. Bryan, K., et al., *Discovery and visualization of miRNA-mRNA functional modules within integrated data using bicluster analysis.* Nucleic Acids Res, 2014. **42**(3): p. e17.
154. Wilson, C.M., et al., *Clostridium thermocellum transcriptomic profiles after exposure to furfural or heat stress.* Biotechnology for Biofuels, 2013. **6**(1): p. 131.
155. Zeisel, A., et al., *Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science, 2015. **347**(6226): p. 1138-42.

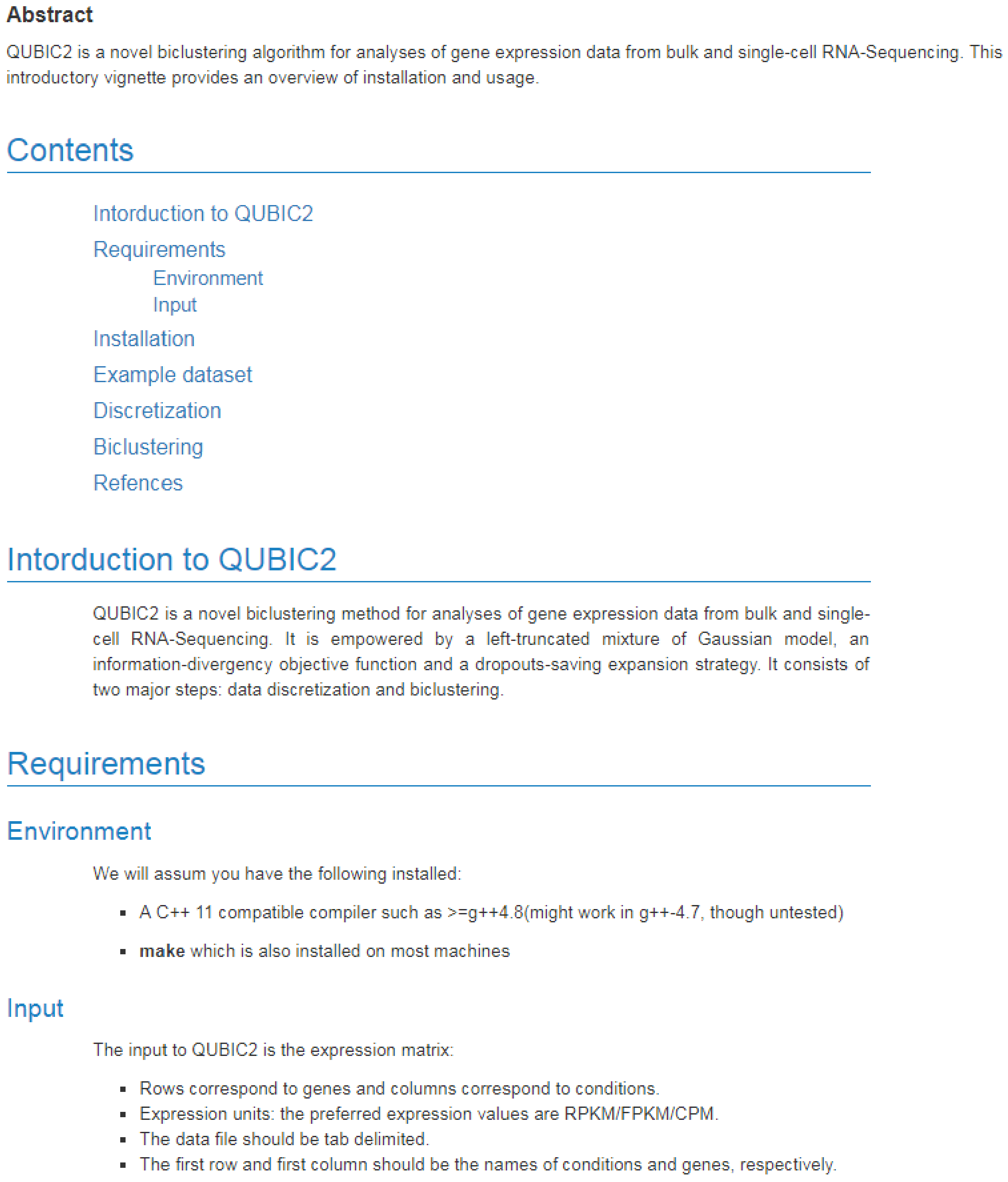
156. Yang, J., E. Worley, and Q. Ma, *Nitrogen remobilization and conservation, and underlying senescence-associated gene expression in the perennial switchgrass Panicum virgatum.* 2016. **211**(1): p. 75-89.
157. Waltman, P., et al., *Multi-species integrative biclustering.* Genome Biol, 2010. **11**(9): p. R96.
158. Yang, R. and X. Wang, *Organ evolution in angiosperms driven by correlated divergences of gene sequences and expression patterns.* Plant Cell, 2013. **25**(1): p. 71-82.
159. Gonçalves, J.P., et al., *Regulatory Snapshots: Integrative Mining of Regulatory Modules from Expression Time Series and Regulatory Networks.* PLOS ONE, 2012. **7**(5): p. e35977.
160. MacPherson, J.I., et al., *Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems.* PLoS Comput Biol, 2010. **6**(7): p. e1000863.
161. De Smet, R. and K. Marchal, *Advantages and limitations of current network inference methods.* Nat Rev Microbiol, 2010. **8**(10): p. 717-29.
162. Wang, X., et al., *Gene-module level analysis: Identification to Networks and Dynamics.* Current opinion in biotechnology, 2008. **19**(5): p. 482-491.
163. Yoon, S.H., et al., *A systems level predictive model for global gene regulation of methanogenesis in a hydrogenotrophic methanogen.* Genome Res, 2013. **23**(11): p. 1839-51.
164. Peterson, E.J., et al., *A high-resolution network model for global gene regulation in Mycobacterium tuberculosis.* Nucleic Acids Res, 2014. **42**(18): p. 11291-303.
165. Levering, J., et al., *Integrated Regulatory and Metabolic Networks of the Marine Diatom Phaeodactylum tricornutum Predict the Response to Rising CO2 Levels.* mSystems, 2017. **2**(1): p. e00142-16.
166. Sharan, R., I. Ulitsky, and R. Shamir, *Network-based prediction of protein function.* Mol Syst Biol, 2007. **3**: p. 88.
167. Liu, G., et al., *Functional diversity of topological modules in human protein-protein interaction networks.* Scientific Reports, 2017. **7**: p. 16199.
168. Zhang, Y., et al., *Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data.* BMC Bioinformatics, 2008. **9**(1): p. 203.
169. Lewis, C.M. and J. Knight, *Introduction to genetic association studies.* Cold Spring Harb Protoc, 2012. **2012**(3): p. 297-306.
170. Chen, B. and A.J. Butte, *Leveraging big data to transform target selection and drug discovery.* Clin Pharmacol Ther, 2016. **99**(3): p. 285-97.
171. Starmans, M.H. and P.C. Boutros, *Biomarkers and subtypes of cancer.* Aging (Albany NY), 2015. **7**(5): p. 280-1.
172. Wang, M., et al., *Statistical methods for studying disease subtype heterogeneity.* Stat Med, 2016. **35**(5): p. 782-800.
173. Chen, G., P.F. Sullivan, and M.R. Kosorok, *Biclustering with heterogeneous variance.* Proc Natl Acad Sci U S A, 2013. **110**(30): p. 12253-8.
174. Yeoh, E.-J., et al., *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling.* Cancer Cell, 2002. **1**(2): p. 133-143.

175. Parise, C.A., 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): p. 593-602.
176. Cui, J., et al., *An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer.* Nucleic Acids Res, 2011. **39**(4): p. 1197-207.
177. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature protocols, 2009. **4**(1): p. 44-57.
178. Wu, J., et al., *KOBAS server: a web-based platform for automated annotation and pathway identification.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W720-4.
179. Schaefer, C.F., et al., *PID: the Pathway Interaction Database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D674-9.
180. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-127.
181. Wang, Y.K., C.G. Print, and E.J. Crampin, *Biclustering reveals breast cancer tumour subgroups with common clinical features and improves prediction of disease recurrence.* BMC Genomics, 2013. **14**(1): p. 102.
182. Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proc Natl Acad Sci U S A, 2002. **99**(10): p. 6567-72.
183. Brin, S. and L. Page, *The anatomy of a large-scale hypertextual web search engine.* Computer networks and ISDN systems, 1998. **30**(1): p. 107-117.
184. Swanton, C., *Intratumor heterogeneity: evolution through space and time.* Cancer Res, 2012. **72**(19): p. 4875-82.
185. Bedard, P.L., et al., *Tumour heterogeneity in the clinic.* Nature, 2013. **501**(7467): p. 355-64.
186. Fiannaca, A., et al., *Analysis of miRNA expression profiles in breast cancer using biclustering.* BMC Bioinformatics, 2015. **16 Suppl 4**(4): p. S7.
187. Gusenleitner, D., et al., *iBBiG: iterative binary bi-clustering of gene sets.* Bioinformatics, 2012. **28**(19): p. 2484-92.
188. Gupta, M., et al., *Identification of homogeneous genetic architecture of multiple genetically correlated traits by block clustering of genome ‐ wide associations.* Journal of Bone and Mineral Research, 2011. **26**(6): p. 1261-1271.
189. Arnedo, J., et al., *PGMRA: a web server for (phenotype x genotype) many-to-many relation analysis in GWAS.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W142-9.
190. Henrichsen, C.N., et al., *Using transcription modules to identify expression clusters perturbed in Williams-Beuren syndrome.* PLoS Comput Biol, 2011. **7**(1): p. e1001054.
191. Teng, B., et al., *Exploring the Genetic Patterns of Complex Diseases via the Integrative Genome-Wide Approach.* IEEE/ACM Trans Comput Biol Bioinform, 2016. **13**(3): p. 557-64.

192. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
193. Shi, F. and H. Huang, *Identifying Cell Subpopulations and Their Genetic Drivers from Single-Cell RNA-Seq Data Using a Biclustering Approach.* J Comput Biol, 2017. **24**(7): p. 663-674.
194. Falony, G., et al., *Population-level analysis of gut microbiome variation.* Science, 2016. **352**(6285): p. 560-4.
195. Fujita, P.A., et al., *The UCSC Genome Browser database: update 2011.* Nucleic Acids Res, 2011. **39**(Database issue): p. D876-82.
196. Chan, W.-C., et al., *MetaMirClust: discovery of miRNA cluster patterns using a data-mining approach.* Genomics, 2012. **100**(3): p. 141-148.
197. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.
198. Liang, F. and W.H. Wong, *Evolutionary Monte Carlo: Applications to C p model sampling and change point problem.* Statistica sinica, 2000: p. 317-342.
199. Kent, W.J., et al., *The human genome browser at UCSC.* Genome research, 2002. **12**(6): p. 996-1006.
200. Wu, M.C., et al., *Powerful SNP-set analysis for case-control genome-wide association studies.* Am J Hum Genet, 2010. **86**(6): p. 929-42.
201. Tan, K.M. and D.M. Witten, *Sparse Biclustering of Transposable Data.* J Comput Graph Stat, 2014. **23**(4): p. 985-1008.
202. Lee, M., et al., *Biclustering via sparse singular value decomposition.* Biometrics, 2010. **66**(4): p. 1087-95.
203. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-4.
204. Evans, W.E. and H.L. McLeod, *Pharmacogenomics--drug disposition, drug targets, and side effects.* N Engl J Med, 2003. **348**(6): p. 538-49.
205. Rutherford, K.D., G.K. Mazandu, and N.J. Mulder, *A systems-level analysis of drug-target-disease associations for drug repositioning.* Brief Funct Genomics, 2017.
206. Iskar, M., et al., *Characterization of drug‐induced transcriptional modules: towards drug repositioning and functional understanding.* Molecular systems biology, 2013. **9**(1): p. 662.
207. Kutalik, Z., J.S. Beckmann, and S. Bergmann, *A modular approach for integrative analysis of large-scale gene-expression and drug-response data.* Nat Biotechnol, 2008. **26**(5): p. 531-9.
208. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.
209. Xiong, M., et al., *Identification of transcription factors for drug-associated gene modules and biomedical implications.* Bioinformatics, 2014. **30**(3): p. 305-9.
210. Verbist, B., et al., *Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project.* Drug discovery today, 2015. **20**(5): p. 505-513.

211. Wishart, D.S., et al., *DrugBank: a comprehensive resource for in silico drug discovery and exploration.* Nucleic Acids Res, 2006. **34**(Database issue): p. D668-72.
212. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.* Science, 2006. **313**(5795): p. 1929-35.
213. Natsoulis, G., et al., *The liver pharmacological and xenobiotic gene response repertoire.* Mol Syst Biol, 2008. **4**: p. 175.
214. Hochreiter, S. and K. Obermayer, *Support vector machines for dyadic data.* Neural Computation, 2006. **18**(6): p. 1472-1510.
215. Wang, S., et al., *Genome-scale identification of cell-wall related genes in Arabidopsis based on co-expression network analysis.* BMC plant biology, 2012. **12**(1): p. 138.
216. Yang, J., et al., *DMINDA 2.0: integrated and systematic views of regulatory DNA motif identification and analyses.* Bioinformatics, 2017. **33**(16): p. 2586-2588.
217. Liu, B., et al., *Bacterial regulon modeling and prediction based on systematic cis regulatory motif analyses.* Sci Rep, 2016. **6**: p. 23030.

APPENDIX 1: QUBIC2 tutorial



Abstract

QUBIC2 is a novel biclustering algorithm for analyses of gene expression data from bulk and single-cell RNA-Sequencing. This introductory vignette provides an overview of installation and usage.

Contents

- Intorduction to QUBIC2
- Requirements
  - Environment
  - Input
- Installation
- Example dataset
- Discretization
- Biclustering
- Refences

Intorduction to QUBIC2

QUBIC2 is a novel biclustering method for analyses of gene expression data from bulk and single-cell RNA-Sequencing. It is empowered by a left-truncated mixture of Gaussian model, an information-divergency objective function and a dropouts-saving expansion strategy. It consists of two major steps: data discretization and biclustering.

Requirements

Environment

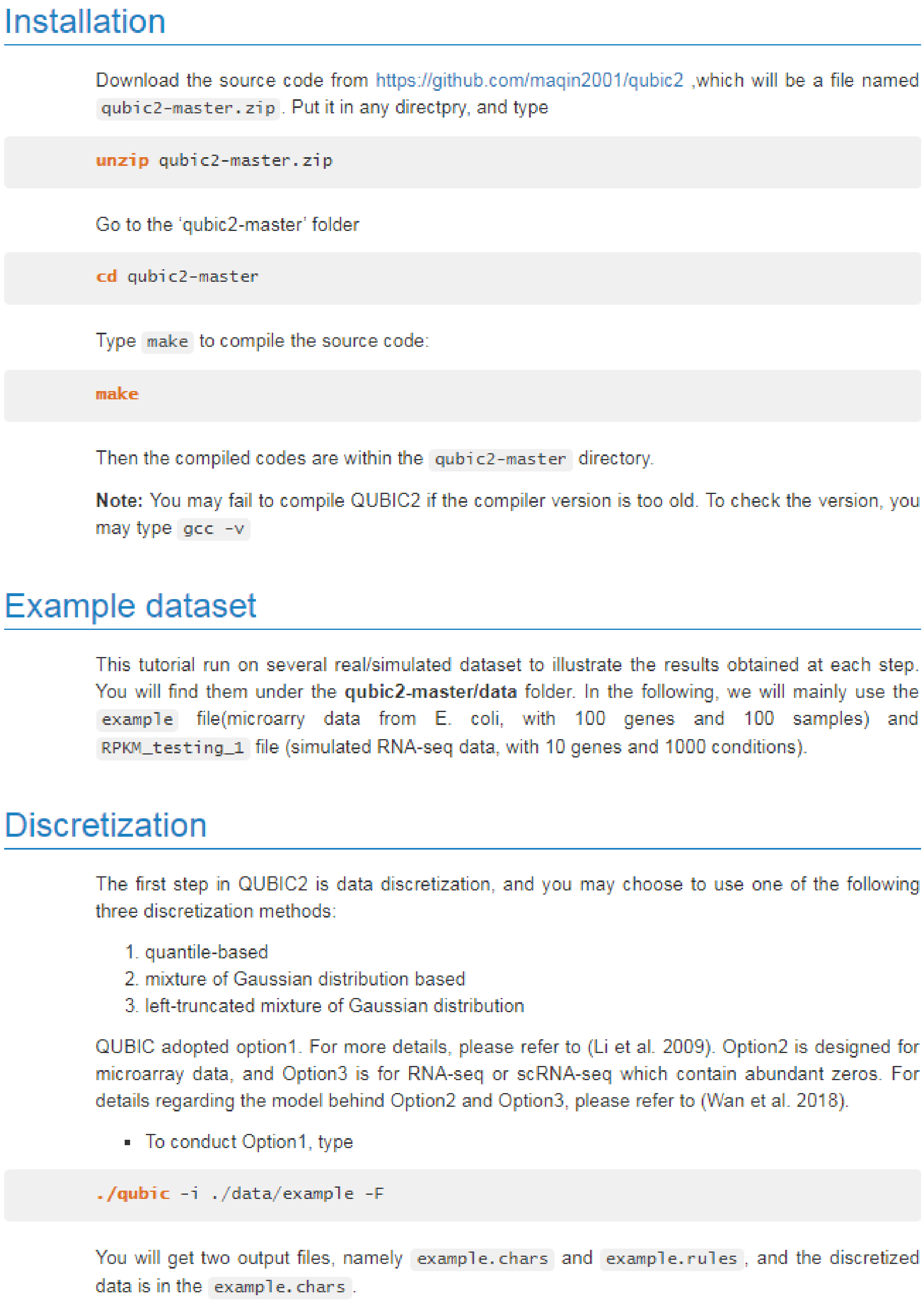
We will assum you have the following installed:

- A C++ 11 compatible compiler such as >=g++4.8(might work in g++-4.7, though untested)
- **make** which is also installed on most machines

Input

The input to QUBIC2 is the expression matrix:

- Rows correspond to genes and columns correspond to conditions.
- Expression units: the preferred expression values are RPKM/FPKM/CPM.
- The data file should be tab delimited.
- The first row and first column should be the names of conditions and genes, respectively.



Installation

Download the source code from https://github.com/maqin2001/qubic2 ,which will be a file named `qubic2-master.zip` . Put it in any directpry, and type

```
unzip qubic2-master.zip
```

Go to the 'qubic2-master' folder

```
cd qubic2-master
```

Type `make` to compile the source code:

```
make
```

Then the compiled codes are within the `qubic2-master` directory.

Note: You may fail to compile QUBIC2 if the compiler version is too old. To check the version, you may type `gcc -v`

Example dataset

This tutorial run on several real/simulated dataset to illustrate the results obtained at each step. You will find them under the **qubic2-master/data** folder. In the following, we will mainly use the `example` file(microarry data from E. coli, with 100 genes and 100 samples) and `RPKM_testing_1` file (simulated RNA-seq data, with 10 genes and 1000 conditions).

Discretization

The first step in QUBIC2 is data discretization, and you may choose to use one of the following three discretization methods:

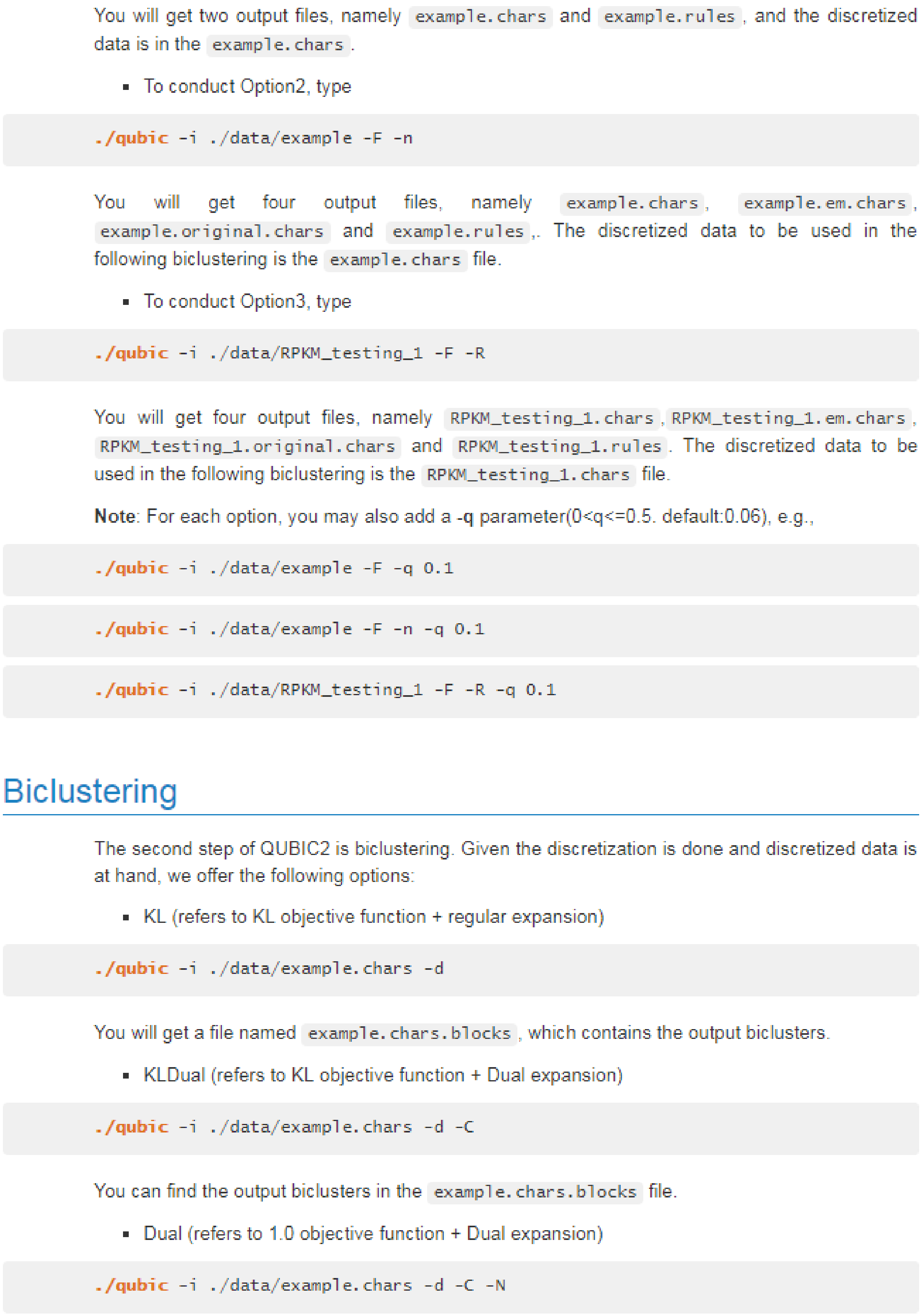
1. quantile-based
2. mixture of Gaussian distribution based
3. left-truncated mixture of Gaussian distribution

QUBIC adopted option1. For more details, please refer to (Li et al. 2009). Option2 is designed for microarray data, and Option3 is for RNA-seq or scRNA-seq which contain abundant zeros. For details regarding the model behind Option2 and Option3, please refer to (Wan et al. 2018).

- To conduct Option1, type

```
./qubic -i ./data/example -F
```

You will get two output files, namely `example.chars` and `example.rules`, and the discretized data is in the `example.chars`.



You will get two output files, namely `example.chars` and `example.rules`, and the discretized data is in the `example.chars`.

- To conduct Option2, type

```
./qubic -i ./data/example -F -n
```

You will get four output files, namely `example.chars`, `example.em.chars`, `example.original.chars` and `example.rules`,. The discretized data to be used in the following biclustering is the `example.chars` file.

- To conduct Option3, type

```
./qubic -i ./data/RPKM_testing_1 -F -R
```

You will get four output files, namely `RPKM_testing_1.chars`, `RPKM_testing_1.em.chars`, `RPKM_testing_1.original.chars` and `RPKM_testing_1.rules`. The discretized data to be used in the following biclustering is the `RPKM_testing_1.chars` file.

Note: For each option, you may also add a **-q** parameter(0<q<=0.5. default:0.06), e.g.,

```
./qubic -i ./data/example -F -q 0.1
```

```
./qubic -i ./data/example -F -n -q 0.1
```

```
./qubic -i ./data/RPKM_testing_1 -F -R -q 0.1
```

Biclustering

The second step of QUBIC2 is biclustering. Given the discretization is done and discretized data is at hand, we offer the following options:

- KL (refers to KL objective function + regular expansion)

```
./qubic -i ./data/example.chars -d
```

You will get a file named `example.chars.blocks`, which contains the output biclusters.

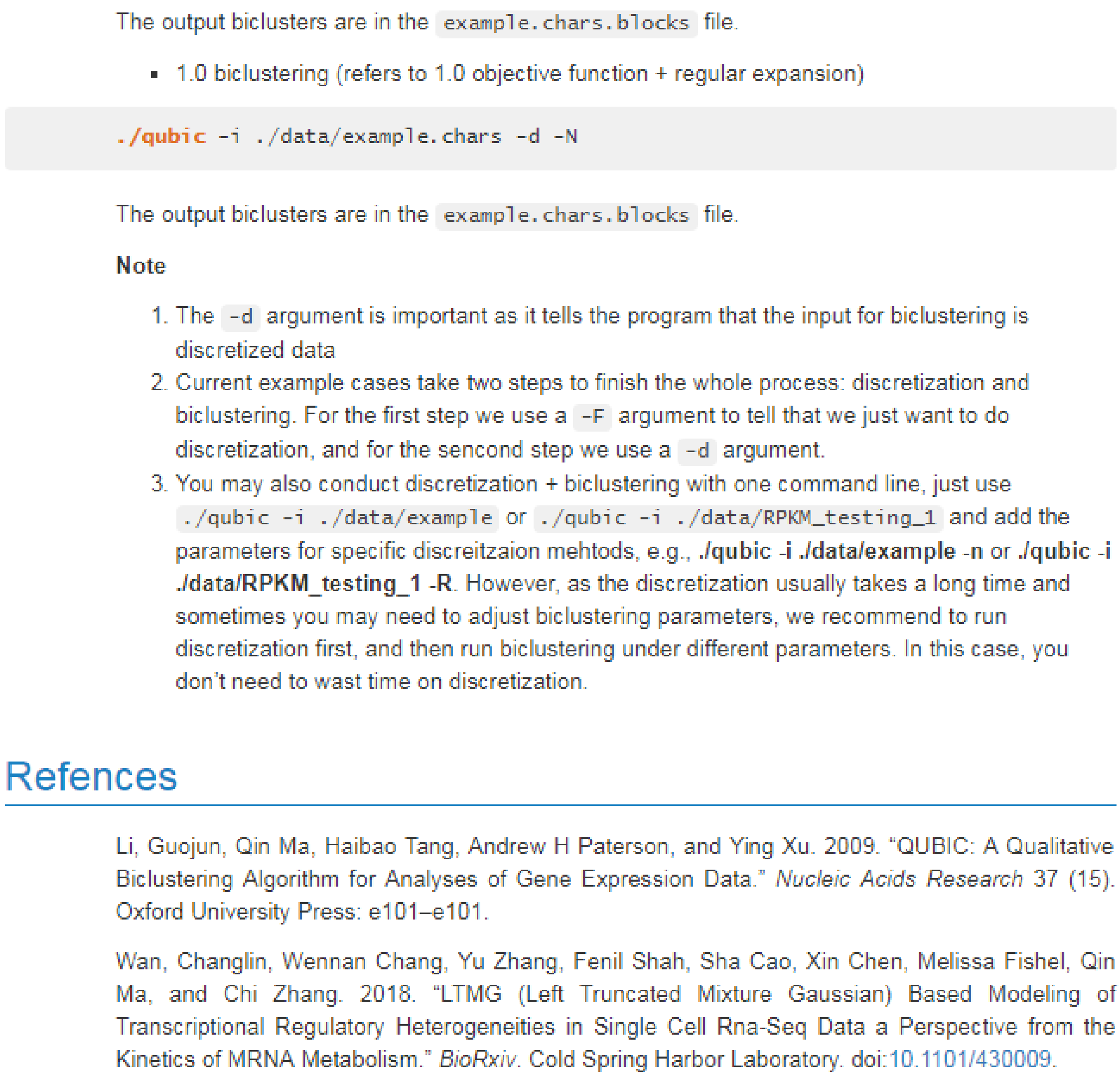
- KLDual (refers to KL objective function + Dual expansion)

```
./qubic -i ./data/example.chars -d -C
```

You can find the output biclusters in the `example.chars.blocks` file.

- Dual (refers to 1.0 objective function + Dual expansion)

```
./qubic -i ./data/example.chars -d -C -N
```



The output biclusters are in the `example.chars.blocks` file.

- 1.0 biclustering (refers to 1.0 objective function + regular expansion)

```
./qubic -i ./data/example.chars -d -N
```

The output biclusters are in the `example.chars.blocks` file.

Note

1. The `-d` argument is important as it tells the program that the input for biclustering is discretized data
2. Current example cases take two steps to finish the whole process: discretization and biclustering. For the first step we use a `-F` argument to tell that we just want to do discretization, and for the sencond step we use a `-d` argument.
3. You may also conduct discretization + biclustering with one command line, just use `./qubic -i ./data/example` or `./qubic -i ./data/RPKM_testing_1` and add the parameters for specific discreitzaion mehtods, e.g., **./qubic -i ./data/example -n** or **./qubic -i ./data/RPKM_testing_1 -R**. However, as the discretization usually takes a long time and sometimes you may need to adjust biclustering parameters, we recommend to run discretization first, and then run biclustering under different parameters. In this case, you don't need to wast time on discretization.

Refences

Li, Guojun, Qin Ma, Haibao Tang, Andrew H Paterson, and Ying Xu. 2009. "QUBIC: A Qualitative Biclustering Algorithm for Analyses of Gene Expression Data." *Nucleic Acids Research* 37 (15). Oxford University Press: e101–e101.

Wan, Changlin, Wennan Chang, Yu Zhang, Fenil Shah, Sha Cao, Xin Chen, Melissa Fishel, Qin Ma, and Chi Zhang. 2018. "LTMG (Left Truncated Mixture Gaussian) Based Modeling of Transcriptional Regulatory Heterogeneities in Single Cell Rna-Seq Data a Perspective from the Kinetics of MRNA Metabolism." *BioRxiv*. Cold Spring Harbor Laboratory. doi:10.1101/430009.

APPENDIX 2: Data Links

The link for all the datasets used in the thesis is provided in **Table11**.

Table 11**.** Data links

Data	**Link**	**Note**
E. coli RNA-Seq	http://bmbl.sdstate.edu/downloadFiles/E.coli%20RNA-seq/	Include expression matrix and five sets of pathways
Simulation data	http://bmbl.sdstate.edu/downloadFiles/simulation/	Include expression matrix and ten groups of modules
TCGA data	https://zenodo.org/record/1157938#.W489C_ZFwiQ	
scRNA-Seq data	https://scrnaSeq-public-datasets.s3.amazonaws.com/manual-data/yan/nsmb.2660-S2.csv	Used in 2.3 and 2.5.
GSE52583	http://www.cs.cmu.edu/~jund/scdiff/download/data/treutlein2014	
mouse olfactory bulb scRNA-Seq data	http://www.spatialtranscriptomicsresearch.org/wp-content/uploads/2016/07/Rep2_MOB_count_matrix-1.tsv	MOB Replicate2
GSE 48968	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE48968	
GSE60402	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60402	

APPENDIX 3: Citation Map for QUBIC

The citation map for QUBIC (including QUBIC algorithm, Qserver and QUBICR) is provided in Figure 17.

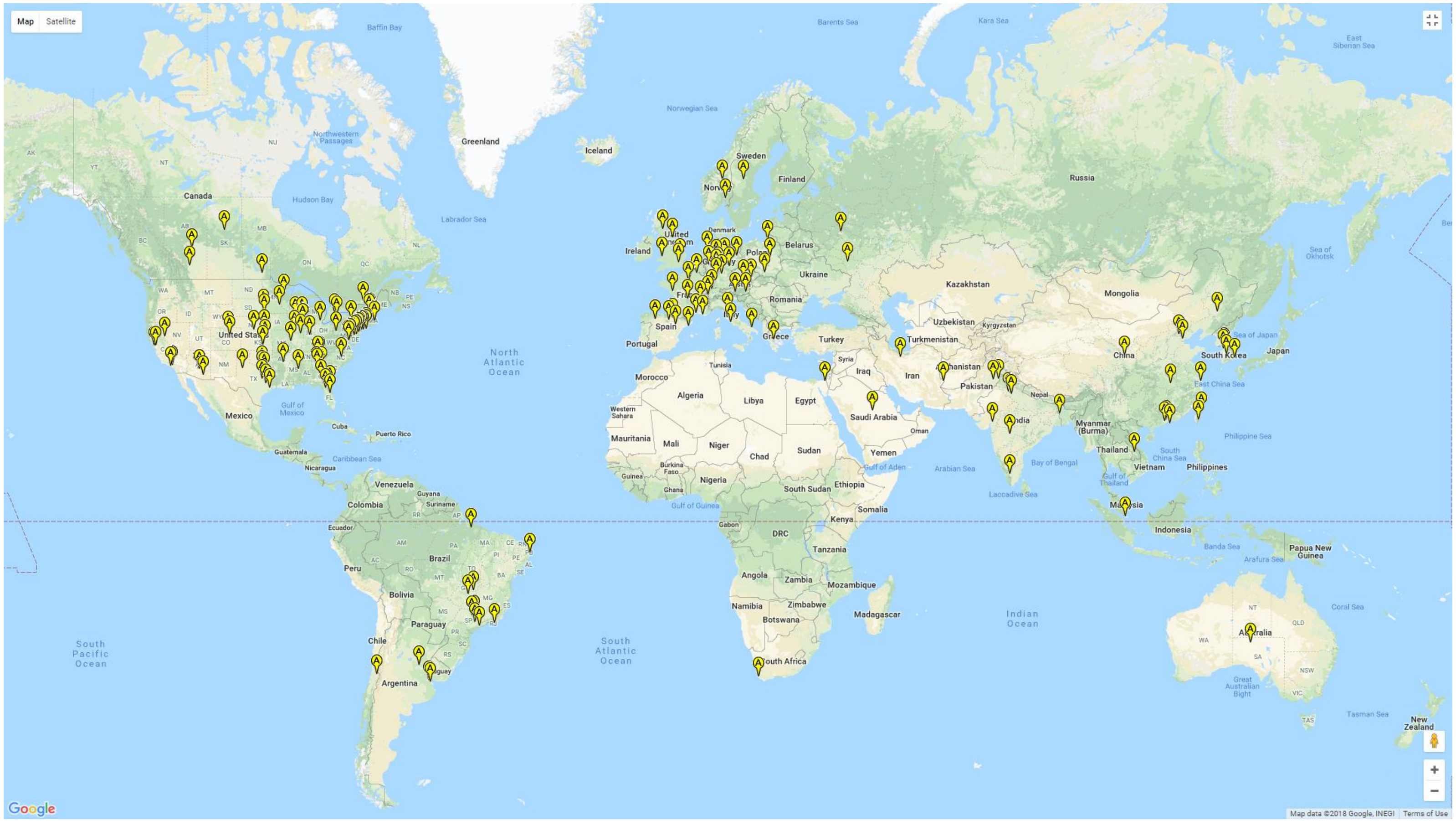
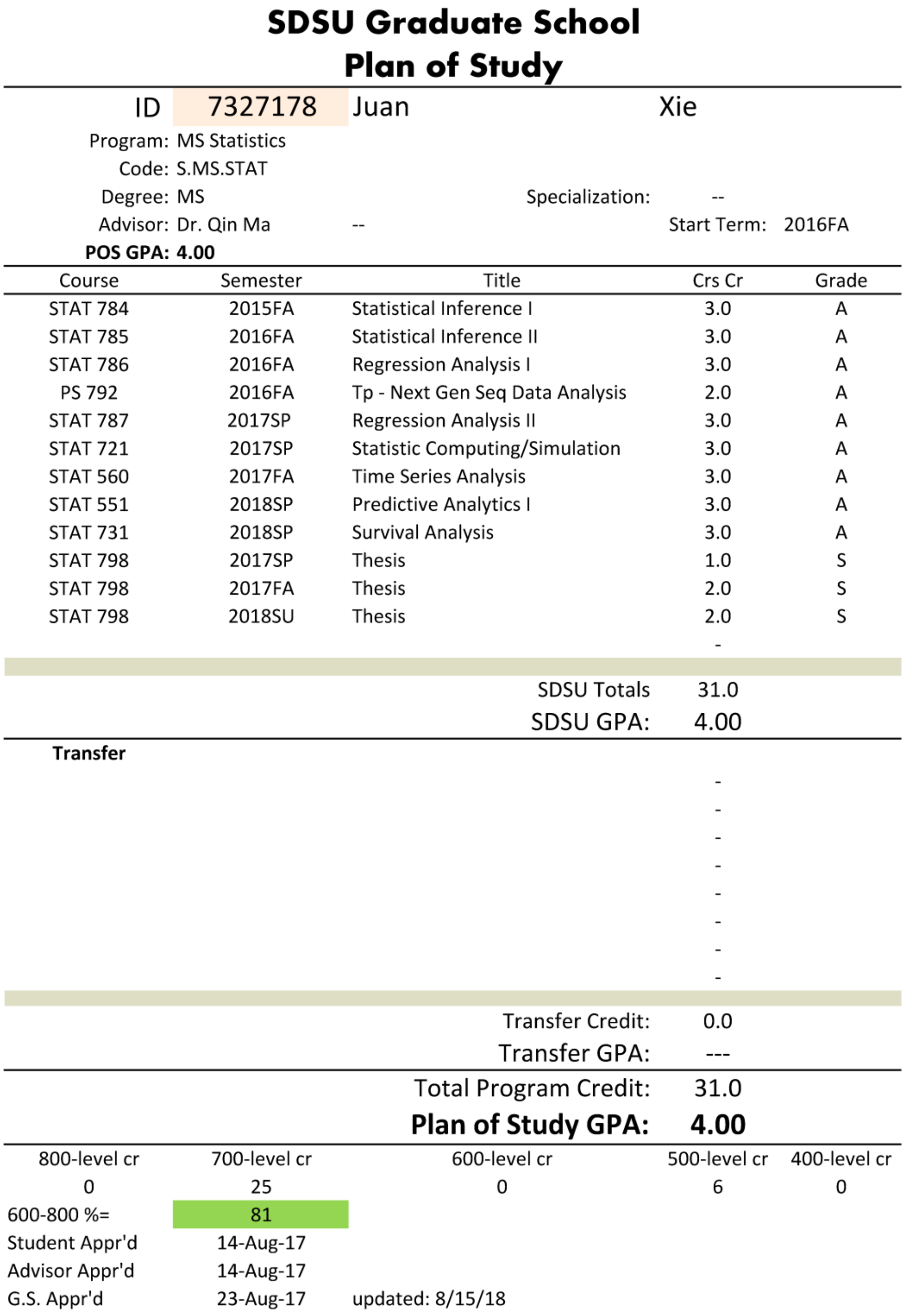


Figure 17. Citation map for QUBIC

APPENDIX 4: Plan of Study



SDSU Graduate School
Plan of Study

ID	7327178	Juan	Xie	
Program:	MS Statistics			
Code:	S.MS.STAT			
Degree:	MS		Specialization:	--
Advisor:	Dr. Qin Ma	--	Start Term:	2016FA
POS GPA:	**4.00**			

Course	Semester	Title	Crs Cr	Grade
STAT 784	2015FA	Statistical Inference I	3.0	A
STAT 785	2016FA	Statistical Inference II	3.0	A
STAT 786	2016FA	Regression Analysis I	3.0	A
PS 792	2016FA	Tp - Next Gen Seq Data Analysis	2.0	A
STAT 787	2017SP	Regression Analysis II	3.0	A
STAT 721	2017SP	Statistic Computing/Simulation	3.0	A
STAT 560	2017FA	Time Series Analysis	3.0	A
STAT 551	2018SP	Predictive Analytics I	3.0	A
STAT 731	2018SP	Survival Analysis	3.0	A
STAT 798	2017SP	Thesis	1.0	S
STAT 798	2017FA	Thesis	2.0	S
STAT 798	2018SU	Thesis	2.0	S
			-	
		SDSU Totals	31.0	
		SDSU GPA:	4.00	

Transfer

			-	
			-	
			-	
			-	
			-	
			-	
			-	
			-	
		Transfer Credit:	0.0	
		Transfer GPA:	---	
		Total Program Credit:	31.0	
		Plan of Study GPA:	**4.00**	

800-level cr	700-level cr	600-level cr	500-level cr	400-level cr
0	25	0	6	0
600-800 %=	81			
Student Appr'd	14-Aug-17			
Advisor Appr'd	14-Aug-17			
G.S. Appr'd	23-Aug-17	updated: 8/15/18		

APPENDIX 5: Curriculum Vitae

Juan Xie

605-690-3056 | https://www.linkedin.com/in/juan-xie/ | juanxie.84@gmail.com

EDUCATION

M.S. in Statistics (GPA: 4.0) **August 2016-Sep 2018**

South Dakota State University, Brookings, SD

- Coursework: Statistical Inference, Regression analysis, Time series data analysis, Next Generation Sequencing Data Analysis, Survival analysis, Predictive analysis

M.S. in Environmental Science

University of Chinese Academy of Sciences, Beijing, China

B.S. in Food Science and Engineering

Hua Zhong Agricultural University, Wuhan, China

EMPLOYMENT

Graduate Research Assistant **Aug 2016--Present**

Bioinformatics and Mathematical Biosciences Lab, South Dakota State University

- Develop biclustering algorithm and tools for large scale gene expression data
- Analyze NGS data from bacteria and plant
- Maintain lab’s high-performance computing resource
- Write local and national level research proposals

PUBLICATIONS

1. Yu Zhang, Juan Xie, Jinyu Yang, et al. QUBIC: a Bioconductor package for qualitative biclustering analysis of gene co-expression data. ***Bioinformatics***, btw635,2016.
2. Juan Xie, Anjun Ma, Anne Fennell, Jing Zhao, Qin Ma, 2018, It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. ***Briefings in Bioinformatics***, DOI:10.1093/bib/bby014
3. Adam McDermaid, Xin Chen, Yiran Zhang, Cankun Wang, Shaopeng Gu, Juan Xie, Qin Ma. A new machine learning-based framework for mapping uncertainty

analysis in RNA-Seq read alignment and gene expression estimation. ***Frontiers in Genetics***. doi: 10.3389/fgene.2018.00313

4. Ying Li, Shi Xiaohu, Liang Yanchun, Juan Xie, Yu Zhang, Qin Ma. RNA-TVcurve: A Web Server for RNA Secondary Structure Comparison based on a Multi-Scale Similarity of its Triple Vector Representation. ***BMC Bioinformatics,*** 2017, doi: 10.1155/2017/5652041
5. Juan Xie, Anjun Ma, Yu Zhang, Bingqiang Liu, Changlin Wan, Sha Cao, Chi Zhang, Qin Ma. QUBIC2: A novel biclustering algorithm for large-scale bulk RNA-sequencing and single-cell RNA-sequencing data analysis. (***bioRxiv,*** https://www.biorxiv.org/content/early/2018/09/07/409961)
6. Suresh Damodaran, Amélie Dubois, Juan Xie, Qin Ma, Valérie Hindié, Senthil Subramanian. GmZPR3d interacts with GmHD-ZIP III proteins and regulates soybean root and nodule vascular development. (Under review in ***Frontiers in Plant Science***)

PRESENTATIONS

- Hypothesis-driven and discovery-driven analysis of Grapevine expression data. January 16, 2018. Plant and Animal Genome Conference, Jan. 14-18, San Diego, CA, USA. (Oral presentation)
- Biclustering in Big Biological Data Analysis. 8th Annual Avera SDSU Research Symposium, Oct 26, 2016, Brookings, SD. (Poster presentation)
- QUBIC: An R package of qualitative biclustering for gene co-expression analysis. The 101st Annual Meeting of the South Dakota Academy of Science, April 8-9, 2016, Sioux Falls, SD. (Poster presentation)

SKILLS

- R programming
- Bash scripting
- SAS, C, C++, SPSS experience
- Next-Generation Sequencing data analysis