Received: 8 May 2018 | Revised: 5 July 2018 | Accepted: 16 July 2018

DOI: 10.1002/ame2.12028

REVIEW ARTICLE

A brief procedure for big data analysis of gene expression

Kewei Wang[1,2,3,4,5,6] | Wenji Wang[1,2,3,4] | Mang Li[1,2,3,4]

[1]Institute of Cell Biotechnology, China and Russia Medical Research Center, Harbin, China

[2]Center for Endemic Disease Control, Chinese Center for Disease Control and Prevention, Harbin, China

[3]Key Laboratory of Etiology and Epidemiology, National Health and Family Planning Commission of the People's Republic of China, Harbin, China

[4]Harbin Medical University, Harbin, China

[5]Departments of Surgery, University of Illinois College of Medicine, Peoria, Illinois

[6]Institute of Laboratory Animal Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Correspondence
Kewei Wang, Institute of Cell Biotechnology, China and Russia Medical Research Center, Harbin, China.
Email: kkww02052015@hotmail.com

Abstract

There are a lot of biological and experimental data from genomics, proteomics, drug screening, medicinal chemistry, etc. A large amount of data must be analyzed by special methods of statistics, bioinformatics, and computer science. Big data analysis is an effective way to build scientific hypothesis and explore internal mechanism. Here, gene expression is taken as an example to illustrate the basic procedure of the big data analysis.

KEYWORDS
big data analysis, cluster analysis, microarray, PCA analysis, regression model

1 | INTRODUCTION

Gene expression data can be originated from different techniques such as quantitative real-time PCR (qRT-PCR), microarray, ChIP assay, ChIP-on-chip, high-throughput ChIP-sequencing, etc. Levels of the gene expression are calculated based on relative quantity of house-keeping genes (ie, GAPDH, β-actin, cyclophilin A1) or absolute magnitude subsequent to a standard curve. The gene expression of special proteins may be typical marker under physiological and pathological conditions. In the field of clinical medicine, biomarkers such as calcitonin for medullary thyroid carcinoma, alpha fetoprotein for hepatocellular carcinoma, glial fibrillary acidic protein for glioma, carcinoma antigen 15-3 for breast cancer, and so on, are often utilized for diagnosis and prognostic evaluation of relevant tumors.[1-4] In addition, F13A1 gene is for screening abnormal bleeding risk.[5] Parkinson's disease may have a high level of tissue transglutaminase.[6] Gene T235 is a marker for the persistent microalbuminuria in children and adolescents with type 1 diabetes mellitus. So far, enough evidence has demonstrated differential levels of gene expression are closely associated with their functionality under pathophysiological status. However, how to form a scientific hypothesis based on existing data? How to choose suitable datasets and further to set up a relationship among different variables? It needs not only professional knowledge and judgment, but also mathematical theory and statistical methods. The present study introduces a brief procedure to perform big data analysis of gene expression.

2 | DATA TYPES OF GENE EXPRESSION

The gene expression can be detected or measured via routine PCR or qRT-PCR. One or more genes are determined at one time, based on designed primers. Relative quantification of the gene expression is usually normalized to the level of housekeeping gene with the

ΔΔCt method. Microarray utilizes single DNA chip to detect a lot of gene expression simultaneously. The principle of DNA microarray is to hybridize unknown DNA sequence with a set of probe oligodeoxynucleotides. The probe nucleotides are immobilized on hard surfaces such as glass in array way, and then were hybridized with fluorescent target genes. The gene expression is determined via fluorescence intensity/position of target nucleotides in the microarray. ChIP-sequencing or ChIP-seq is a technique to analyze the binding sites of DNA-associated nuclear proteins, to investigate interactions between nuclear proteins and specific DNA sequence, and to identify functional role of transcription factors. ChIP-Seq data are derived from the DNA sequencing through a series of operation. Now, the ChIP-Seq has already replaced the previous ChIP-on-chip method for the interactive study between nuclear proteins and genomic DNA. Owing to differences in gene detection methods, there are multiple data sources of gene expression, for example, DNA microarray, Chip-seq data (Figure 1). The first step for gene expression analysis is to cluster gene data with similar characteristics into different groups for further investigation.[7] There are a few terms that explain data types of gene expression in Geo database as follows:

1. GSM (GEO sample) stands for the experimental data of a single sample.
2. GDS (GEO dataset) is a collection of GSM that is arranged manually on a topic. Thus, GSM and GDS share the same platform.
3. GSE (GEO series) includes a multiple experiment in a research project. It may use different platforms.
4. GPL (GEO platform) is a chip platform, such as Affymetrix, Sentrix, Illumina, Aglent, etc.

3 | DATA PREPROCESSING

Levels of gene expression are much varied due to their diversity in nature/function. Raw data from different databases have to be reorganized by gene ID and symbol after having matched microarray annotation table (Table 1). If the raw data are used for direct comparison, it can overemphasize the role of the high abundance genes during the comprehensive analysis, and/or may weaken the function of the low gene-expression levels. Therefore, in order to ensure the reliability of the result, the original data need to be normalized.[8,9] There are many kinds of data normalization methods, including min-max normalization, log function conversion, arc tangent function conversion, *z*-score normalization, and fuzzy quantization. They may be linear type (ie, $M \pm SD$) and curve type (ie, seminormal distribution). Different normalization methods have diverse effects on the evaluation of resultant data.[10-12] Unfortunately, there is no general rule to follow in the selection of data normalization methods. The min-max normalization, also called standardization of deviations, is the linear transformation of the original data, and the result falls to the [0, 1] interval. The log function conversion can be completed via $x^* = \log 10(x)$. A problem is that the result does not necessarily fall on the [0, 1] interval (Table 2). It should be divided by log10 (max). The atan function transformation uses an inverse tangent function to achieve the normalization of data. The data less than 0 will be mapped to the [−1, 0] interval. The most common method is the *z*-score normalization (or zero-mean normalization) (Table 3), which is used as standard method in SPSS package (The IBM SPSS® Software, Armonk, NY, USA). The processed data conforms to the normal distribution, that is, the mean is 0 and the SD is 1. Its conversion function contains μ for the mean value and σ for the SD of all sample data.

4 | MULTIVARIATE DATA ANALYSIS

There are different statistical methods for big data analysis. Several techniques that are generally applied for microarray gene expression are summarized as follows.

4.1 | Correlation analysis

It describes the correlation between two or more than two random variables, for example, the relationship between body height and weight, and relative humidity between the air and rainfall. When abovementioned variables are plotted in the Cartesian coordinate

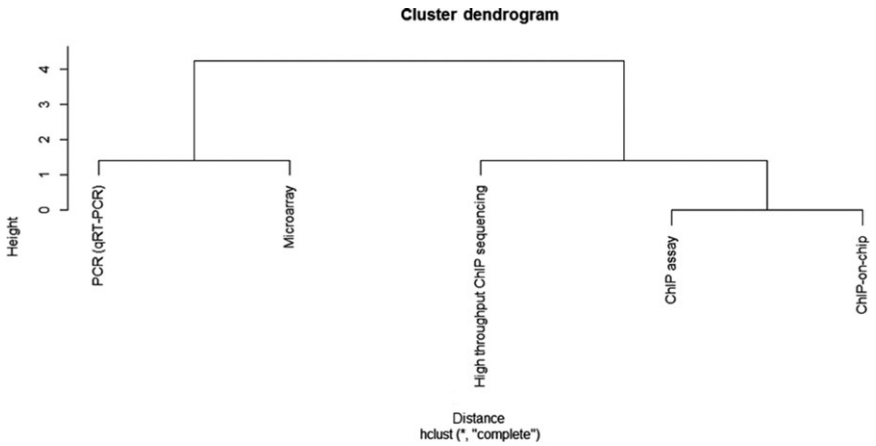

FIGURE 1 Data sources of gene expression derived from different detection methods

system, there is a set of points called "scatter plot." The correlation between the two variables is expressed by correlation coefficient *r*. The correlation coefficient r is any value between −1 and 1.[13,14] The correlation between one variable x_0 and a set of variables (x_1, x_2, … x_n) is measured by multicorrelation coefficient. The range of the multicorrelation coefficient is [0, 1]. In the case of multivariable, the net correlation between two variables is reflected by partial correlation coefficient when the effect of other variables is controlled or considered as a constant.[15,16] For instance, there are three variables X_1, X_2, and X_3. The partial correlation coefficient r13.2 represents the linear correlation between variable X_1 and X_3 after the influence of variable X_2 is thought of as a constant. The partial correlation coefficient can reflect the relationship between the two variables more accurately than the simple linear correlation coefficient.

4.2 | Cluster analysis

The goal of cluster analysis is to collect data on similar basis for classification (Figure 2).[17,18] In different applications, many clustering techniques have been developed, which are used to describe data, to measure the similarity between different data sources, and to classify data sources into different clusters. The statistical methods for the cluster analysis include system clustering, decomposition, addition, dynamic clustering, ordered sample clustering, overlapping clustering, and fuzzy clustering. Traditional clustering algorithms can be classified into five categories: partitioning, hierarchical, density-based, grid based, and model-based. Clustering analysis tools, such as k-mean and k-center algorithm, have been added to many famous packages such as SPSS (The IBM SPSS® Software) and SAS (Cary, NC, USA).

TABLE 1 Different chip categories of GEO microarrays in NCBI database

Datasets	Years	Gene number	Names
1	2007	33297	Aflymetrix Human Gene 1.0 ST array
2	2003	54675	Aflymetrix Human Genome U133 Plus 2.0 Array
3	2002	22283	Aflymetrix Human Genome U133A Array
4	2002	8793	Affymetrix Human HG-Focus Target Array
5	2007	20228	Agilent-012097 Human 1A Microarray (V2) G4110B
6	2006	45220	Agilent-014850 Whole Human Genome Microarray 4×44K G4112F
7	2008	20589	Sentrix Human Ref-8 v2 Expression BeadChip
8	2008	24526	Illumina HumanRef-8 v3.0 expression beadchip

4.3 | Principal components analysis

In many cases, there is a certain correlation between the two variables. It can be explained that the two variables share common data information or have a overlap of datasets. The principal components analysis (PCA) is a technology for analyzing and simplifying datasets (Figure 3).[19,20] There is more information about PCA theory and application in the platforms "Principal Component Analysis (PCA) in R" and "Principal Component Analysis in Python" (website https://datascienceplus.com/principal-component-analysis-pca-in-r/; https://plot.ly/ipython-notebooks/principal-component-analysis/).[19-21] The PCA method is usually used to reduce the dimension of dataset, while maintaining the greatest contribution of the other side in the dataset. This is achieved by preserving low-order principal components and ignoring higher order principal components. Such low-order components often retain the most important aspect of data. The PCA technique can take out as many of the less comprehensive variables as possible to reflect the information of the original variables. However, this is not necessary and depends on the specific application. Moreover, since PCA relies on the given data, the accuracy of the data has a great impact on the analysis results. The most classical approach in PCA analysis is to express the variance of F_1. The F_1 selected in all linear combinations should be the largest variance, so the F_1 is the first principal component. If the first principal component is not enough to represent the information of the original variables, then consider selecting F_2 for the second linear combinations, expressed in the mathematical Cov (F_1, F_2) = 0.

TABLE 2 Logarithmic transformed data (log *e* ratio, reference series: GSE9539)

Gene	4/100 Fold	8/100 Fold	12/100 Fold	24/100 Fold	4/200 Fold	8/200 Fold	12/200 Fold
BAD	0.9330	0.9560	0.9390	0.9930	0.9600	0.9770	0.9800
BAX	1.0710	0.9640	1.0070	0.9390	1.0970	0.9430	0.9870
BCL2	1.0730	0.9490	0.9060	1.0620	1.2360	0.7870	1.2920
BCL2A1	0.9320	1.3620	0.9410	0.9280	0.9090	0.9110	0.8130
BCL2L11	1.0740	0.9790	1.1810	0.9670	0.9880	1.0160	1.1500
BCL2L13	1.1290	0.9720	1.0990	1.0170	1.1080	0.9930	1.0610
BCL2L2	1.1300	1.0480	1.0410	1.0650	1.1120	1.0530	1.0290
BIK	1.0440	0.9560	0.9770	1.0120	0.9920	0.9560	1.0480
BOK	0.9000	0.9350	0.9300	0.9570	0.9620	0.9320	0.8810

TABLE 3 Transformed data by zero-mean normalization

Gere symbol	153-T0-1	153-T0-2	153-T0-3	153-T0.5-1	153-T0.5-2	153-T0.5-3	153-T1.S-1	153-T1.S-1
TNFS F10	−0.5237	−0.6558	−0.4203	−0.5577	−0.4949	−0.6140	2.5225	3.2670
TNF	0.2021	−0.3106	0.0823	−0.2435	0.2932	−0.4927	−0.5215	1.5486
TNFSF 12	0.3596	1.0538	0.1945	0.3850	−0.0340	1.8283	−0.9779	1.9087
FAS	−1.0077	−0.0496	−0.5411	0.0521	−0.8960	−0.1018	−1.3602	−1.1021
TNFRSF 10B	1.8797	1.6844	0.9053	1.9502	1.0238	1.2344	0.0413	−0.6137
TNFRSF 10A	−0.9825	0.0200	−0.6054	−0.5993	−1.0315	−0.6330	0.1672	−0.5839
FADD	0.1915	0.2694	0.8454	−0.5752	−0.4390	−0.5752	−0.8126	−0.9527
TNFRSF 1A	0.0443	0.7225	0.0773	0.0733	0.3874	−0.4379	−0.4959	0.0333
TNFRSF 10D	0.4991	0.6449	1.1089	0.3600	0.6891	0.5102	−0.2631	−0.2520
TNFRSF 10C	−0.5613	−0.0915	1.8041	0.2629	−0.5284	−0.6603	0.6667	1.7877
TRAF2	−0.3673	−0.5960	−0.4435	0.6127	−0.3728	0.8686	−0.4163	−1.6087
TRADD	−0.1146	−0.9489	−0.9365	−1.2715	−1.5972	−0.8590	0.2048	1.0081
TNFRS F11B	−0.8241	−0.4364	−0.5804	−0.5804	−1.1564	−0.5804	0.1729	0.1729
TRAF1	0.0494	0.7638	−0.0038	−1.4858	−1.2730	−1.7290	0.0190	−0.6498

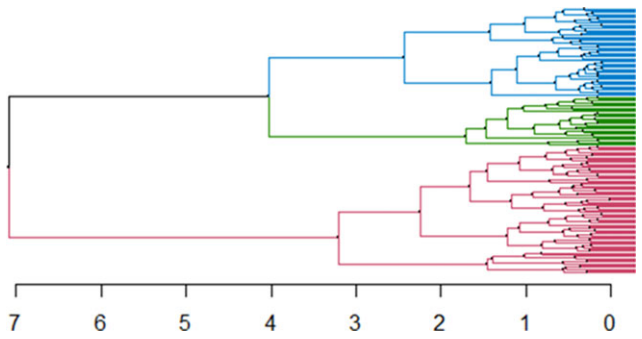

FIGURE 2 Cluster analysis of gene expression. Different colors represent clusters of similar characteristics based on geometric distance

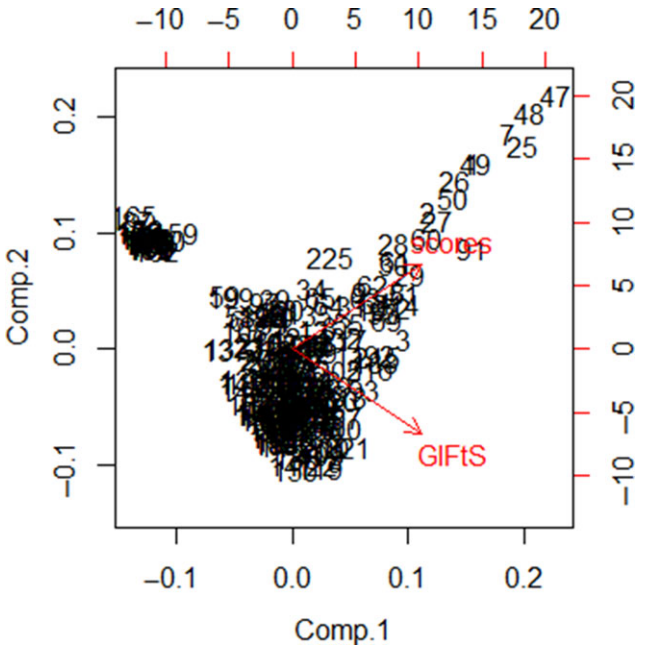

FIGURE 3 PCA analysis of gene expression. The horizontal and vertical coordinates represent two-dimensional distribution of the first two components. The number in the figure stands for different samples

4.4 | Regression model

Representative gene variables are chosen out of above-mentioned original data following a series of statistical procedure. They are utilized to establish mathematical model and to characterize essential features of gene expression (Figure 4). Mathematical modeling is the process of describing actual phenomena in mathematical language.[21,22] The actual phenomena here contain specific natural phenomenon such as free falling, and abstractive phenomenon such as the value tendency of a customer to a certain commodity. The biological mathematical (BioMath) model is a strict language to make the description of all kinds of phenomena into a much scientific, logical, objective, and repeatable structure.[23] The description here includes not only the external form and internal mechanism, but also the prediction, experiment, and explanation of the actual phenomena. Sometimes, we need to do some experiments, but these experiments often use abstract models as a substitute for the actual objects, and the experiment itself is a theoretical substitute for the physical operation. Therefore, the BioMath model is a simplification of real things. The establishment of mathematical models is the process of simplifying and abstracting complex and practical problems into a rational mathematical structure. Mathematical modeling is the bridge linking gene expression and biological/medical problems. Based on the hypothesis, we should collect original data of the gene expression, observe and investigate the inherent characteristics and inherent laws, ascertain the quantitative relationship among variables, to calculate and estimate all the parameters of the model, and then to verify the accuracy, rationality, and applicability of the model.

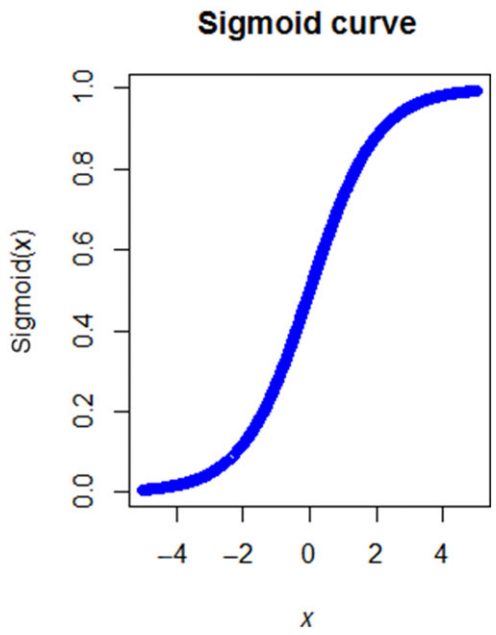

FIGURE 4 Sigmoid curve is a typical pattern of biological response as reflected by possibility range [0, 1]. Logistic regression or logit regression model is able to estimate the probability of a binary response based on one or more independent variables.

In summary, big data accumulated through diverse methods can show different data types. Representative variables are screened out of multiple categories via professional knowledge and statistical methods. The representative variables may be fitted into a suitable mathematical model for the evaluation of pathophysiological process, which can be utilized in clinical diagnosis and treatment.

CONFLICT OF INTEREST

None.

ORCID

Kewei Wang http://orcid.org/0000-0001-5869-2490

REFERENCES

1. Bodenner D, Nalley C, Chien C, Clarke B, Spring PM, Stack BC. Markedly elevated serum calcitonin concentrations associated with initial presentation but not the recurrent presentation of medullary thyroid carcinoma. *Thyroid*. 2010;20(8):927-929.
2. Zhu W, Peng Y, Wang L, et al. Identification of alpha-fetoprotein-specific T cell receptors for hepatocellular carcinoma immunotherapy. *Hepatology*. 2018. https://doi.org/10.1002/hep.29844.
3. Wei P, Zhang W, Yang LS, et al. Serum GFAP autoantibody as an ELISA-detectable glioma marker. *Tumour Biol*. 2013;34(4):2283-2292.
4. Pectasides D, Pavlidis N, Gogou L, Antoniou F, Nicolaides C, Tsikalakis D. Clinical value of CA 15-3, mucin-like carcinoma-associated antigen, tumor polypeptide antigen, and carcinoembryonic antigen in monitoring early breast cancer patients. *Am J Clin Oncol*. 1996;19(5):459-464.
5. Griffin M, Casadio R, Bergamini CM. Transglutaminases: nature's biological glues. *Biochem J*. 2002;368(Pt 2):377-396.
6. Vermes I, Steur EN, Jirikowski GF, Haanen C. Elevated concentration of cerebrospinal fluid tissue transglutaminase in Parkinson's disease indicating apoptosis. *Mov Disord*. 2004;19(10):1252-1254.
7. Guo L, Lu Z. Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data. *Comput Biol Chem*. 2010;34(3):165-171.
8. Franks JM, Cai G, Whitfield ML. Feature Specific Quantile Normalization Enables Cross-Platform Classification of Molecular Subtypes using Gene Expression Data. *Bioinformatics*. 2018;34(11):1868-1874
9. Shaydurov VA, Kasianov A, Bolshakov AP. Analysis of housekeeping genes for accurate normalization of qPCR data during early postnatal brain development. *J Mol Neurosci*. 2018;64(3):431-439.
10. Huang HC, Qin LX. Empirical evaluation of data normalization methods for molecular classification. *PeerJ*. 2018;6:e4584.
11. Larriba Y, Rueda C, Fernandez MA, Peddada SD. A bootstrap based measure robust to the choice of normalization methods for detecting rhythmic features in high dimensional data. *Front Genet*. 2018;9:24.
12. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genom*. 2018;19(1):274.
13. Yuan L, Chen L, Qian K, et al. A novel correlation between ATP5A1 gene expression and progression of human clear cell renal cell carcinoma identified by coexpression analysis. *Oncol Rep*. 2018;39(2):525-536.
14. Ahmed M, Nguyen H, Lai T, Kim DR. miRCancerdb: a database for correlation analysis between microRNA and gene expression in cancer. *BMC Res Notes*. 2018;11(1):103.
15. Erb I, Notredame C. How should we measure proportionality on relative gene expression data? *Theory Biosci*. 2016;135(1–2):21-36.
16. Reverter A, Chan EK. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*. 2008;24(21):2491-2497.
17. Olex AL, Fetrow JS. SC(2)ATmd: a tool for integration of the figure of merit with cluster analysis for gene expression data. *Bioinformatics*. 2011;27(9):1330-1331.
18. Pazos Obregon F, Soto P, Lavin JL, et al. Cluster Locator, online analysis and visualization of gene clustering. *Bioinformatics*. 2018. https://doi.org/10.1093/bioinformatics/bty336.
19. Taguchi YH. Tensor decomposition-based and principal-component-analysis-based unsupervised feature extraction applied to the gene expression and methylation profiles in the brains of social insects with multiple castes. *BMC Bioinform*. 2018;19(Suppl. 4):99.
20. Chen T, Zhang M, Jabbour S, et al. Principal component analysis-based imaging angle determination for 3D motion monitoring using single-slice on-board imaging. *Med Phys*. 2018;45(6):2377-2387.
21. Liu WT, Wang Y, Zhang J, et al. A novel strategy of integrated microarray analysis identifies CENPA, CDK1 and CDC20 as a cluster of diagnostic biomarkers in lung adenocarcinoma. *Cancer Lett*. 2018;425:43-53.
22. Ehler M, Rajapakse VN, Zeeberg BR, et al. Nonlinear gene cluster analysis with labeling for microarray gene expression data in organ development. *BMC Proceed*. 2011;5(Suppl 2):S3.
23. McCarthy GD, Drewell RA, Dresch JM. Global sensitivity analysis of a dynamic model for gene expression in Drosophila embryos. *PeerJ*. 2015;3:e1022.

How to cite this article: Wang KW, Wang WJ, Li M. A brief procedure for big data analysis of gene expression. *Animal Model Exp Med*. 2018;1:189–193.
https://doi.org/10.1002/ame2.12028