Novel integrated systems approach discovers cell specific genetic-transcriptomic-methylation networks and causal pathways underlying diseases.

Zixin Hu[1,*], Shicheng Guo[5], Yun Zhu[3], Panpan Wang[1], David A Bennett[4], Li Jin[1], Momiao Xiong[1,2]

[1*]State Key Laboratory of Genetic Engineering and Ministry of Education, Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai. China
huzixin@fudan.edu.cn
[2]Department of Biostatistics, University of Texas School of Public Health, Houston, TX 77030, USA.
[3]Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA 70112. USA
[4]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL 60612, USA
[5]Department of Bioengineering, University of California, San Diego, CA 92093-0412, USA

Abstract

Most tissues consist of multiple cell types. Global profiling of gene expression and methylation results from the convolution of multiple cell type specific expressions and methylations that lead to heterogeneity of molecular profiles. Heterogeneity of gene expression and methylation data is a major confounding factor in their analysis ignoring data heterogeneity leads to misleading results while recovering molecular components in the original subpopulations can identify the molecular structure of the tissue samples. In the last decade, there has been increasing interest in developing computational methods for deconvolution of gene expression and DNA methylation data. The approaches to deconvolution are classified into reference and reference-free deconvolution. Reference-based molecular profile deconvolution assumes that prior information on the molecular profiles of the subpopulations are known. Reference-free deconvolution requires no prior information on the number and composition of the subpopulations present in mixed samples. Linear regression and nonnegative matrix factorization are widely used computational tools for gene expression and methylation deconvolution. The major application of gene expression deconvolution is to identify cell specific differentially expressed genes. Less attention has been paid to identify cell specific regulatory networks underlying diseases. The traditional paradigm for analyzing biological data is association or correlation analysis. Despite their differences in selection of specific methods for inference, the current correlation and association analysis methods share the same drawback. They cannot efficiently detect, distinguish and characterize the true biological processes. The correlation networks constructed cannot model information flow and rarely contain causal information. Therefore, these approaches generally do not provide clear biological or clinical information relevant to the mechanisms of disease that are yet to be discovered and understood. To overcome these

limitations, we first propose to develop novel nonnegative matrix factorization methods with general loss functions and penalization constraints for gene expression and methylation deconvolution based on both microarray and sequencing data. Then, for each specific cell type, using the de-convolution gene expression and causal inference theory we construct causal gene expression networks and identify the cell specific causal networks underlying diseases. The proposed methods are applied to prefrontal cortex brain region expression dataset with 448 individuals and 4,540 genes in 181 pathways that matched with the UC San Diego reference dataset with expression of 16,242 gene in 4,040 cells. In order to evaluate the potential of the expression deconvolution, we investigated differential expression between Alzheimer' patients and normal individuals in prefrontal cortex brain region and cells. The number of cell types in the prefrontal brain region was 17. We found that the total number of significantly expressed genes in both the tissue and cells were 5 and the number of genes that were differentially expressed only in the specific cells, but not in the tissue were 5. We identified no differentially expressed pathway in the tissue sample data, but we identified three differentially expressed pathways: Circadian rhythm, Systemic lupus erythematosus and Type II diabetes mellitus pathways in at least one cell types. Literatures confirmed that three pathways are associated with Alzheimer's diseases.

Key Words: Gene expression deconvolution, methylation deconvolution, nonnegative matrix decomposition, causal networks, Alzheimer' disease.