Statistica Sinica **36** (2026), 1-22
doi:https://doi.org/10.5705/ss.202023.0109

REGULATION-INCORPORATED GENE EXPRESSION NETWORK-BASED HETEROGENEITY ANALYSIS

Rong Li[1], Qingzhao Zhang[2] and Shuangge Ma*[1]

[1] *Yale University and* [2] *Xiamen University*

Abstract: Gene expression-based heterogeneity analysis has been extensively conducted. In recent studies, it has been shown that network-based analysis, which takes a system perspective and accommodates the interconnections among genes, can be more informative than that based on simpler statistics. Gene expressions are regulated. Incorporating regulations in analysis can better delineate the "sources" of gene expression effects. Although conditional network analysis can somewhat serve this purpose, it does not render enough attention to the regulation relationships. In this article, significantly advancing from the existing heterogeneity analyses based only on gene expression networks, conditional gene expression network analyses, and regression-based heterogeneity analyses, we propose heterogeneity analysis based on gene expression networks (after accounting for or "removing" regulation effects) as well as regulations of gene expressions. A high-dimensional penalized fusion approach is proposed, which can determine the number of sample groups and parameter values in a single step. An effective computational algorithm is proposed. It is rigorously proved that the proposed approach enjoys the estimation, selection, and grouping consistency properties. Extensive simulations demonstrate its practical superiority over closely related alternatives. In the analysis of two breast cancer datasets, the proposed approach identifies heterogeneity and gene network structures different from the alternatives and with sound biological implications.

Key words and phrases: Gene expression network, heterogeneity analysis, penalization, regulation.

1. Introduction

Many complex diseases are intrinsically heterogeneous, with samples having the same disease diagnosis behaving differently. In early studies, heterogeneity analysis is often based on low-dimensional clinical and demographic measurements. With the development of high-throughput profiling, omics measurements, which may more informatively capture disease biology, have been increasingly used in heterogeneity analysis (Lee, Park and Kim, 2021). Among the various omics measurements, gene expressions have drawn special attention because of important biological implications, broad availability of data, and promising empirical results. Through a series of studies (Church, Williams and Mar, 2019;

*Corresponding author. E-mail: shuangge.ma@yale.edu