Chapter 21
Class Discovery, Comparison, and Prediction Methods for RNA-Seq Data

Ahu Cephe
Erciyes University, Turkey

Necla Koçhan
https://orcid.org/0000-0003-2355-4826
İzmir Biomedicine and Genome Center, Turkey

Gözde Ertürk Zararsız
Erciyes University, Turkey

Vahap Eldem
İstanbul University, Turkey

Gökmen Zararsız
Erciyes University, Turkey

ABSTRACT

Gene-expression studies have been studied using microarray data for many years, and numerous methods have been developed for these data. However, microarray technology is old technology and has some limitations. RNA-sequencing (RNA-seq) is a new transcriptomics technique capable of coping with these limitations, using the capabilities of new generation sequencing technologies, and performing operations quickly and cheaply based on the principle of high-throughput sequencing technology. Compared to microarrays, RNA-seq offers several advantages: (1) having less noisy data, (2) being able to detect new transcripts and coding regions, (3) not requiring pre-determination of the transcriptomes of interest. However, RNA-seq data has several features that pose statistical challenges. Thus, one cannot directly use methods developed for microarray analyses, which has a discrete and overdispersed nature of data, quite different from the continuous data structure of microarrays. This article aims to provide an overview and practical guidance to researchers working with RNA-seq data for different purposes.

DOI: 10.4018/979-8-3693-3026-5.ch021

INTRODUCTION

Measuring gene-expression plays a vital role in life sciences such as cancer genomics. It enables us to quantify the level at which a particular gene is expressed within a cell, tissue or organism, thereby providing a tremendous amount of information (Alberts et al., 2002). There are different technologies (i.e., microarray and next-generation technologies) that can measure gene-expression levels. Microarray technology is an outdated technology with some limitations and lost its popularity with the advent of next-generation technologies. On the other hand, RNA-seq is one of the next-generation technologies capable of coping with these limitations, using the capabilities of next generation sequencing technologies, and performing operations quickly and cheaply based on the principle of high-throughput sequencing technology. Moreover, compared to microarrays, RNA-seq offers several advantages: (i) having less noisy data, (ii) being able to detect new transcripts and coding regions, (iii) not requiring pre-determination of the transcriptomes of interest.

RNA-seq technology allows measuring the expression levels of thousands of genes in cells simultaneously, leading to high dimensional data to be further analyzed. The information stored in these high dimensional data can be used for different purposes: (i) identifying "biomarker" genes that can characterize different disease subclasses, that is, class comparison; (ii) identifying new subclasses for a particular disease, that is, class discovery and (iii) assigning samples into known disease classes, that is, class prediction (Dudoit et al., 2002; Weigelt et al., 2010).

Class comparison is known as differential analysis or analysis of differential-expression. In these studies, gene-expression profiles of samples, which are predefined groups, are compared to identify differentially expressed genes between groups. Differentially expressed genes are identified in cells from different tissues, different patients, or cells exposed to different experimental conditions. For example, comparing treated and untreated cells to detect the effect of a new drug on gene-expression levels; comparisons between healthy tissue and diseased tissue to identify genes with altered expression; comparing gene-expression in tumor tissue for patients responding to a particular treatment versus gene-expression in patients with the same cancer diagnosis who do not respond to treatment. Such studies yield lists of genes that were significantly altered between groups. The aim is to provide insight into the underlying biological mechanisms and perhaps identify potential therapeutic targets.

In class prediction studies, as in class comparison studies, genes that differ between predefined classes are tried to be determined. However, in class prediction studies, gene-expression values are explanatory variables rather than outcome variables. Moreover, the purpose of the analysis of class prediction studies is to identify a small set of genes that can accurately distinguish between different classes rather than identify all genes that differ. Classes are defined beforehand in class predictions, and the aim is to create a classifier that can distinguish between these classes based on the gene-expression profiles of the samples and can be applied to the expression profiles of a new sample. For example, a classifier that distinguishes between 2 different disease states; a classifier that distinguishes short-term survivors from long-term survivors; a classifier can be created that predicts whether a patient will respond to a particular drug. In class comparison studies, whether a new patient will react to treatment can be predicted based on gene-expression profiles.

Class discovery differs from class comparison and class prediction studies in that classes are not predefined. The purpose of these studies is to determine whether subsets of samples with apparently homogeneous phenotypes can be distinguished based on differences in gene-expression profiles. For example, there are many diseases in which individuals with apparently similar phenotypes have sig-