ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Controlling the false-discovery rate by procedures adapted to the length bias of RNA-Seq

Tae Young Yang [a,*], Seongmun Jeong [b]

[a] Department of Mathematics, Myongji University, Yongin, Kyonggi 449-728, Republic of Korea
[b] Personalized Genomic Medicine Research Center, Division of Strategic Research Groups Korea Research Institute of Bioscience and Biotechnology, Daejeon, 34141, Republic of Korea

ARTICLE INFO

Article history:
Received 20 September 2016
Accepted 24 August 2017
Available online xxxx

AMS 2000 subject classifications:
primary 62P10
secondary 92B05

Keywords:
Common weight
Individual weight
Length bias
RNA-Seq
Separate procedure
Weighted procedure

ABSTRACT

In RNA-Seq experiments, the number of mapped reads for a given gene is proportional to its expression level and length. Because longer genes contribute more sequencible fragments than do shorter ones, it is expected that even if two genes have the same expression level, the longer gene will have a greater number of total reads. This characteristic creates a length bias such that the proportion of significant genes increases with the gene length. However, genes with a long length are not more biologically meaningful than genes with a short length. Therefore, the length bias should be properly corrected to determine the accurate list of significant genes in RNA-Seq. For this purpose, we proposed two multiple-testing procedures based on a weighted-FDR and a separate-FDR approach. These two methods use prior information on differential gene length while keeping the false-discovery rate (FDR) controlled at α. In the weighted-FDR controlling procedure, we incorporated prior weights for the length of each gene. These weights increase the power when the gene's length is short and decrease the power when its length is long. In the separate-FDR controlling procedure, we sequentially ordered all genes according to their lengths and then split these genes into two subgroups of short and long genes. The adaptive Benjamini–Hochberg procedure was then performed separately for each subgroup. The proposed procedures were compared with existing methods and evaluated in two numerical examples and one simulation study. We concluded that the weighted p-value procedure properly reduced the length bias of RNA-Seq.

1. Introduction

Whereas standard microarray probes only cover about 20% of a gene on average, capturing only a portion of the biologically relevant data, next-generation sequencing technologies have become widely used, enabling researchers to profile and quantify transcripts across the entire transcriptome in a technique called RNA-Seq. RNA-Seq has various advantages over microarray-based technology, including high resolution to identify novel genes (Zheng, Chung, & Zhao, 2011), and it has become a commonly used alternative to microarray-based gene-expression profiling (Li & Tibshirani, 2011).

In an RNA-Seq experiment, purified mRNAs from a tissue are sheared into small fragments, which are then converted into complementary DNA. This complementary DNA library is amplified and sequenced by a high-throughput platform, such as Illumina's Genome Analyzer (Bullard, Purdom, Hansen, & Dudoit, 2010). This process generates millions of short reads, which are then mapped to the genome or transcriptome (Li, Witten, Johnstone, & Tibshirani, 2012). For a set of regions of interest,

* Correspondence to: Myongji University, Department of Mathematics, Yongin, Kyonggi 17058, Republic of Korea.
E-mail address: tyang@mju.ac.kr (T.Y. Yang).

http://dx.doi.org/10.1016/j.jkss.2017.08.001
1226-3192/

such as gene-level or exon-level units, depending on the purpose of the experiment, the number of reads mapped to each region are counted and then used as an expression level of that region (Li & Tibshirani, 2011). In this article, the region of interest is genes.

Gene expressions for microarrays are represented as continuous numbers, and differential expression among different conditions can often be analyzed by a t-test. In contrast, the discrete counts of RNA-Seq arise from either a Poisson density or a negative binomial density. Additionally, the differential expression among different conditions can often be analyzed using Fisher's exact test or a nonparametric test, such as the Wilcoxon's rank sum test. The gene-level p-value of RNA-Seq can be obtained by various R programs, including edgeR (Robinson, McCarthy, & Smyth, 2010), DEGseq (Wang, Feng, Wang, Wang, & Zhang, 2010), DESeq (Anders & Huber, 2010), and baySeq (Hardcastle & Kelly, 2010). Kvam, Liu, and Si (2012) compared the performance of these R programs.

Because longer genes in RNA-Seq contribute more sequencible fragments than shorter ones do, it is expected that even if two genes have the same expression level, the different length will yield different total read numbers (Bullard et al., 2010; Young, Wakefield, Smyth, & Oshlack, 2010). Oshlack and Wakefield (2009) determined that the characteristic of RNA-Seq that maps reads for a given gene depends on the genes expression level and length. Thus, this characteristic creates a length bias, such that the proportion of significant genes increases with the gene length. Because the analysis of other external sources, such as the microarray datasets and the quantitative real-time polymerase chain reaction (qRT-PCR) dataset in Section 3, also shows no evidence of length bias, gene expression is not related to its transcript length. Thus, genes with long length are not more biologically meaningful than genes with short length (Young et al., 2010). The length bias of RNA-Seq should be properly corrected to determine the accurate list of significant genes, which would be used as a base for subsequent analyses such as pathway analysis or gene-set enrichment analysis.

1.1. Standard procedure for controlling the false-discovery rate

A significant gene list was obtained by first calculating each gene's test statistic based on expression levels under different conditions and ranking genes based on their corresponding p-values. Many gene-level tests are carried out so that the statistical significance of each gene's p-value can be adjusted for multiple testing by controlling the FDR, i.e., the percentage of true-negative genes that are incorrectly rejected. The FDR often serves as a target for control in testing high-dimensional problems in genomic data analysis.

Benjamini and Hochberg (1995) originally provided a method, referred to as the BH procedure, for controlling the FDR at a more stringent level. The FDR of the BH procedure at level α is equal to $\pi_0\alpha$, where π_0 represents, in general, the unknown proportion of the null hypothesis. Only when π_0 is close to 1 does the procedure control the FDR at level α.

The adaptive BH method is the standard method, which properly controls the FDR at level α by first estimating π_0 with $\hat{\pi}_0$ (Storey, 2002) and then performing the BH procedure at level $\alpha/\hat{\pi}_0$. Because $\alpha/\hat{\pi}_0$ is larger than the original level α of the BH procedure, the adaptive BH procedure is less conservative and thus more powerful than the original BH procedure, especially when π_0 is far from 1. Because $\hat{\pi}_0$ in the numerical examples of Section 3 are far from 1, the adaptive BH procedure is more appropriate for controlling the FDR at α than the original BH procedure.

When all genes originate from the same condition, the genes are exchangeable (Cai & Sun, 2009). However, an important feature of RNA-Seq is gene-length bias, which renders the genes are no longer exchangeable. The adaptive BH procedure conducts an analysis on all genes under the assumption that genes are exchangeable. Cai and Sun (2009) and Morris (2008) indicated that when exchangeability is not satisfied, the significant gene list of the procedure would not be appropriate. If we ignore the length bias and conduct the procedure as an analysis on all genes, then the resulting significant gene list favors long genes over short genes (Efron, 2008). This is also shown in Sections 3 and 4. Furthermore, because the procedure treats all null hypotheses equally, it cannot handle the prior information regarding the relationship between a gene's significance and its length.

1.2. Appropriate FDR controlling procedures

For multiple hypothesis testing that maintains control of the false discovery rate while incorporating prior information about the hypotheses, many authors including Genovese, Roeder, and Wasserman (2006) and Roeder and Wasserman (2009) have considered using weighted p-values. Particularly, Roeder and Wasserman (2009) provided two p-value weighting methods of the external weighting and the estimated weighting.

We sequentially ordered all genes according to their lengths and then split these genes into two subgroups of short and long genes. Because the proportion of significant genes in the first subgroup was smaller than that in the second subgroup, each gene within the first subgroup was more likely to be non-significant. The genes in the first subgroup are needed to make the rejection more likely. Conversely, the genes in the second subgroup are needed to make the rejection less likely. More appropriate multiple-testing procedures are needed to properly remove the gene-length bias of RNA-Seq while keeping the FDR controlled at α. For this purpose, we provided two procedures: a weighted FDR controlling procedure and a separate FDR controlling procedure.

In the weighted-FDR controlling procedure, we incorporated the prior weight based on the length of each gene, which takes two forms: individual weight and common weight. These weights increase the power when the gene's length is short and decrease the power when its length is long. Thus, the procedure would provide the result that long genes are not expressed significantly more than short genes are.

Please cite this article in press as: Yang, T.Y., Jeong, S., Controlling the false-discovery rate by procedures adapted to the length bias of RNA-Seq. Journal of the Korean Statistical Society (2017), http://dx.doi.org/10.1016/j.jkss.2017.08.001.