GLOBAL SCIENCE LIBRARY

RESEARCH

Journal of Cancer Science and Therapeutics

A review on moderated-t methods for differential expression detection

Lianbo Yu*

Department of Biomedical Informatics, The Ohio State University, Columbus Ohio, USA

Abstract

With the advancement of high-throughput technology, identifying differential expression has become an essential task in multiple domains of biomedical research, such as transcriptome, proteome, metabolome. A wide variety of computational methods and statistical approaches were developed for detecting differential expression. Most of these methods were applicable to modeling expression level of the entire set of features simultaneously. In this article, we provide a review emphasizing on moderated-t methods published in last two decades. We compared similarities and differences between them, and also discussed their limitations in applications.

Keywords: *differential expression, moderated t-statistic, variance smoothing, shrinkage estimator*

Introduction

Since microarray technology was introduced two decades ago, gene expression studies utilizing microarray chips allow measuring expression level of thousands of genes simultaneously. In a short period, microarray studies have been broadly adopted in biomedical research and became a powerful tool to provide insights into understanding cellular processes, disease pathogenesis, drug development, disease diagnosis and treatment. One of most common research questions in such gene profiling studies is to detect genes that are differentially expressed between conditions, which can be phenotypes (e.g. cancer, normal) for observational studies or can be interventional groups (e.g. knock-out, wild type) for biological system perturbation. With the recent advancement of next generation sequencing technology, RNA-seq platforms generate digital expression data and provides much richer information than microarray chips. With price dropping dramatically in last few years, RNA-seq is gradually taking place of microarray in gene expression studies. A great amount of expression profiling data have been generated and are in need of comprehensive analysis. Toward this end, a broad range of computation methods and statistical approaches have been developed and applied to every type of expression data generated from different platforms.

Because of practical considerations in many research projects, gene expression studies are often conducted using only a small number of biological replicates. This experimental constraint poses great challenges to statistical methods in order to provide a solution to successfully answer research questions, but it also provides strong motivation to address these challenges in research community that have already responded by developing many novel computational methods and statistical techniques. One of such challenges is the instability in gene variance estimation, especially for experiments with small number of biological replicates, under which ordinary t-statistic has this issue per se. Researchers have proposed different techniques in order to obtain robust variance estimation. Among these techniques, several versions of moderated-t methods were developed by borrowing information across genes in variance estimation [1-8]. These methods perform better in relative to ordinary t-statistic method that treats each gene independently [9]. In this article, we will review these moderated t-methods in great detail and provide systematic comparisons between them.

Methods

Several earlier moderated-t methods were proposed to test differential expression between conditions that do not have multiple measurements of a single gene for a study unit. We label such methods for independent data points. In contrast, many functional genomic studies profiled transcriptomic data of study units with multiple samples of different origins, diverse cell types, and various time points. These measurements of gene expression levels from multiple samples of the same study unit are correlated in nature. To analyze correlated gene expression levels, a few of moderated-t methods were later proposed to address this correlation issue. We label these methods for correlated data points.

Methods for independent data points

SAM: Tusher et al. [1] observed that signal-to-noise ratio decreases with decreasing gene expression and variance of t-statistic can be high at low expression levels because of small sample variances. Therefore they proposed to add a positive constant to the sample standard deviation in a t-statistic in order

to ensure an independence between t-statistic and gene expression. This added positive constant was chosen to minimize coefficient of variation (CV) of the t-statistic. Without assuming particular parametric distributions on gene expression, permutation test is then used to call significance of expression changes between study conditions. However the usage of permutation test requires that studies have a relatively large number of biological replicates in order to achieve more accurate p-values.

Empirical Bayes Analysis: Efron et al. [2] proposed a non-parametric Bayes model on t-statistics without any assumption on gene expression. Similarly as the SAM method, a positive constant (equal to 90^{th} percentile of all sample standard deviation) is added to the sample standard deviation in t-statistic. Density of t-statistics for all genes were modeled as a mixture density of two populations (unaffected or affected genes). Permutations were then performed to estimate the density of null test statistics of unaffected genes. And logistic regression was used to estimate density ratios between null test statistics and observed test statistics. With an upper bound estimation for the proportion of unaffected genes, the posterior of differential expression was derived by Bayes' rule for the two-component mixture model.

CyberT: Microarray data often demonstrate a functional relationship between gene expression level and gene variance. Long et al. [3] proposed the CyberT method that uses a t-statistic with adjusted variance estimates based on a hierarchical Bayesian model. It assumes that genes of similar expression level have similar measurement errors, therefore the variance of a single gene can be estimated by the weighted average of a prior estimate of the variance for that gene. This weighting factor (hyper-parameter) is determined by experimenters and reflects how confident they are that the variance of a closely related set of genes approximates the variance of the gene under consideration. However, the priori value for the prior degrees of freedom and the chosen window size for smoothing the prior variances were not determined in an optimal way. Later on, Fox et al. [4] proposed an extension for the CyberT method. Their method uses a similar moving average approach to estimate prior variances, but the value of prior degrees of freedom is not assigned a priori. Like the CyberT method, Fox and Dimmic's extension method requires users to fix the window size for variance smoothing.

Limma: Smyth [5] generalized the empirical Bayes model of Lonnstedt and Speed [6] into an approach called Limma that can be used for small sample sizes. Limma uses linear models to analyze the entire set of genes rather than per gene comparison between conditions. By assuming a hierarchical Bayes model, it can borrow information across genes to smooth variances and uses posterior variance estimates in the t-statistic that follows a t-distribution with the estimated posterior degrees of freedom. Empirical Bayes methods and smoothing techniques were used to estimate functional relationships and parameters. Other benefits under linear model framework include that experimental factors can be adjusted as covariates and complex comparisons (i.e. interaction effects) are possible. Later in 2014, voom transformation [10] was amended to model functional relationship between gene variances and their corresponding expression levels.

IBMT: Sartor et al. [7] proposed an intensity-based moderated t-statistic (IBMT) as an extension of Limma for differential expression detection. It uses a local regression method [11] to model a relationship between gene variances and gene expression levels. Because IBMT shrinks sample variances towards expression-dependent prior variances, it provides better posterior variance estimates than Limma. As a result, IBMT reduces the variance of gene variances, which is opposite to prior degrees of freedom, thereby power is increased. This improvement on variance estimation are especially noticeable when sample sizes are small. IBMT allows variances to vary across gene expression levels, but it fixes prior degrees of freedom across levels of gene expression, which means that CV of gene variances is constant between genes. This assumption of constant CV for gene variances is often very restrictive for many microarray experiments. In particular, filtering methods that filter out genes with expression levels close to background noise result in variances that are more homogeneous at lower expression levels.

FMT: Yu et al. [8] observed that not only mean of gene variances changes with gene expression levels, but also CV of gene variances are often not constant across gene expression levels, therefore they proposed a t-statistic that more fully moderates the denominator of t-statistic (FMT) than both Limma and IBMT. Using a local regression method [11], they modeled both mean and CV of gene variances as a function of gene expression levels, which leads to improved estimates of prior degrees of freedom as well as gene variances. Consequently, FMT has more accurate estimates of degrees of freedom and better control of false positives and false negatives across different expression levels than both Limma and IBMT, which were demonstrated in both simulations and spike-ins data set. Limma consistently under-estimates the prior degrees of freedom in these situations, which is critical in small sample size experiments. Because of that, when data were generated under either IBMT or FMT models where variances are larger at the lower expression levels, Limma generates most of its false positives at lower expression levels. While the performance of IBMT and FMT were comparable when data were simulated under the IBMT model, FMT was more powerful than IBMT when CV of gene variances was dependent on expression levels. If prior degrees of freedom is increasing with increased expression levels, the power gain for FMT was greatest at higher intensity levels. For experiments with larger sample sizes, these moderated-t methods were more comparable in power, because as residual degrees of freedom increase, the posterior mean of gene variances is pulled more toward the empirical gene-specific variance estimate. And additionally, the increase in degrees of freedom of moderated t-statistic is relatively small when residual degrees of freedom are large.

Methods for correlated data points

Limma: Smyth et al. [12] implements a strategy in Limma for incorporating the fact that observations or samples may be correlated [ref]. This strategy is similar to fitting a linear mixed-effects model (LMM) for each gene, but they are constrained to share the same intra-replicates correlation. The 'duplicateCorrelation' function in R package limma is used to estimate this consensus correlation. The correlation is then incorporated into the linear model fit and into all tests for

differential expression. Originally this idea was used to estimate the correlation between technical replicates of the same probe on a microarray, such that it preserves more information than simply averaging over replicated probes. More generally, the same idea is also applicable to model the correlation between replicated RNA samples, for example, repeated measures over time or multiple RNA samples collected at the same time on the same individual. However one caveat of this approach is that it enforces a common correlation for all genes, which may work well for within-array technical replicates as originally planned, but is risky for between-array biological replicates.

Dream : Dream method by Hoffman et al. [13] tries to build upon the workflow of Limma but solve it in the setting of LMM, which are fitted using R package lme4. Residual degrees of freedom is estimated by either Satterthwaite approximation [14] or Kenward-Roger approximation [15]. Since the integration with Limma workflow, it allows covariate adjustment and provides linear contrasts of fixed effects' coefficients. Unlike Limma, it estimates random effects separately for each gene. The Dream method shrinks residual variances as in Limma, while variance estimates of random effects are scaled to residual variances before and after shrinkage. In another word, variance smoothing of random effects is bundled with residual variances by enforcing the same shrinkage over the range of gene expression levels for both variances. This fact limits its broad applicability.

FMT-LMM: Yu et al. [16] implemented a novel procedure that shrink different variations independently through FMT under LMM. Their method assumes that variances of both residual errors and random effects have different functional relationships over the range of gene expression levels. Moderated t-statistic for detecting differential expression is based on a combined shrinkage estimator of all variances. To estimate degrees of freedom of the moderated t-statistic, they used two approaches: variance-components approximation [17] and Satterthwaite approximation [14]. The differences between them is that Satterthwaite approximation method uses all variance estimates, but variance-components approximation method only accounts for between-subject variation while ignoring within-subject variation in total degrees of freedom. Through simulations, they show that Satterthwaite approximation method inflates type I error and thus fails to maintain control of false positive rate at the nominal level, but variance-components approximation method can maintain a proper control of false positive rate. Compared to Limma and Dream, FMT-LMM gains more power and maintain proper control of false positive rate as demonstrated in both simulations and the case study under a simplified design. Similar to Dream, FMT-LMM can be extended to complex designs with diverse correlation structures, e.g. multi-level designs.

Summary

This comprehensive review focuses on recently published moderated-t methods specifically developed for detecting differential expression. We compared their similarities and differences under different formats of designs with independent or correlated data points. We also discussed pros and cons for their applications under different scenarios.

Acknowledgement

This research work was supported by the National Institute of Health Grant 2P30CA016058-40 to Ohio State Comprehensive Cancer Center. The author declared no potential conflicts of interest with respect to research, authorship, and publication of this article.

.References

1. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U.S.A.* 2001; 98: 5116-5121.
2. Efron B, Tibshirani R, Storey J, Tusher V. Empirical bayes analysis of a microarray experiment. *J Amer Stat Assoc.* 2001; 96: 1151-1160.
3. Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, et al. Improved statistical inference from dna microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J Biol Chem.* 2001; 276: 19937-19944.
4. Fox RJ, Dimmic MW. A two-sample Bayesian t-test for microarray data. BMC *Bioinformatics*. 2006; 7: 126.
5. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat Appl Genet Mol Biol*. 2004; 3: Article 3.
6. Lonnstedt I, Speed TP: Replicated microarray data. *Statistica Sinica*. 2002; 12: 31-46.
7. Sartor AM, Tomlinson RC, Wesselkamper CS, Sivaganesan S, Leikauf DG, et al. Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics.* 2006; 7: 358.
8. Yu L, Gulati P, Fernandez S, Pennell M, Kirschner L, et al. Fully moderated T-statistic for small sample size gene expression arrays. *Stat Appl Genet Mol Biol.* 2011: 10.
9. Cui X, Hwang JT, Qiu J, Blades NJ, Churchill GA. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics.* 2005. 6: 59-75.
10. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 2014; 15: R29.
11. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. *J Amer Stat Assoc.* 1979; 74: 829-836.
12. Smyth GK, Michaud J, Scott HS. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics.* 2005; 21: 2067-75.
13. Hoffman GE, Roussos P. Dream: Powerful differential expression analysis for repeated measures designs. *Biorxiv*. 2018.
14. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bulletin.* 1946; 2: 110-114.
15. Kenward MG, Roger J H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics.* 1997, 53: 983-997.

16. Yu L, Zhang J, Brock G, Fernandez S. Fully moderated t-statistic in linear modeling of mixed effects for differential expression analysis. *BMC Bioinformatics.* 2019; 20: 675.

17. Schluchter MD, Elashoff JD. Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *J Statistical Computation Simulation*. 1990; 37: 69-87.

Correspondence: Lianbo Yu, Department of Biomedical Informatics, The Ohio State University, Columbus Ohio, USA, Tel: +1 614-292-6446; E-mail: lianbo.yu@osumc.edu

Rec: Feb 10, 2020; Acc: Feb 25, 2020; Pub: Feb 28, 2020
J Cancer Sci Therap. 2020;3(2):19
DOI: 10.36879/JCST.20.00019