Biochemical and Biophysical Research Communications 471 (2016) 274–281

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Meta-analysis method for discovering reliable biomarkers by integrating statistical and biological approaches: An application to liver toxicity

Hyeyoung Cho [a], Hyosil Kim [b], Dokyun Na [c], So Youn Kim [d], Deokyeon Jo [d], Doheon Lee [a, *]

[a] Department of Bio and Brain Engineering, KAIST, Yuseong-gu, Daejeon, 34141, Republic of Korea
[b] Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul, 03722, Republic of Korea
[c] School of Integrative Engineering, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul, 06974, Republic of Korea
[d] Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, 16499, Republic of Korea

ARTICLE INFO

Article history:
Received 6 January 2016
Accepted 14 January 2016
Available online 25 January 2016

Keywords:
Meta-analysis
Biomarker discovery
Effect size
Drug liver toxicity

ABSTRACT

Biomarkers that are identified from a single study often appear to be biologically irrelevant or false positives. Meta-analysis techniques allow integrating data from multiple studies that are related but independent in order to identify biomarkers across multiple conditions. However, existing biomarker meta-analysis methods tend to be sensitive to the dataset being analyzed. Here, we propose a meta-analysis method, iMeta, which integrates *t*-statistic and fold change ratio for improved robustness. For evaluation of predictive performance of the biomarkers identified by iMeta, we compare our method with other meta-analysis methods. As a result, iMeta outperforms the other methods in terms of sensitivity and specificity, and especially shows robustness to study variance increase; it consistently shows higher classification accuracy on diverse datasets, while the performance of the others is highly affected by the dataset being analyzed. Application of iMeta to 59 drug-induced liver injury studies identified three key biomarker genes: Zwint, Abcc3, and Ppp1r3b. Experimental evaluation using RT-PCR and qRT-PCR shows that their expressional changes in response to drug toxicity are concordant with the result of our method. iMeta is available at http://imeta.kaist.ac.kr/index.html.

1. Introduction

Tremendous amount of data is piling up in public databases of biological and medical fields. For instance, the Gene Expression Omnibus database has more than 49,800 study records with over 1.2 million samples in 2014, and the overall submission rate continues to grow more and more rapidly [1]. Moreover, a number of studies have been conducted in related topics repetitively and independently. Accordingly, the potential for integrating these invaluable resources has led researchers to advance biological and medical insights that were formerly unrevealed from an individual study [2–4]. In this regard, the development of effective data integration technique should be essential.

Meta-analysis is a statistical technique for systematic integration of experimental outcomes from multiple studies that are related but independent. Its main advantage is to boost power by increasing sample size to be able to catch signals that are small but consistent. In particular, one promising application is in the field of biomarker discovery (biomarker meta-analysis), whose goal is to find robust biomarkers (genes, proteins, or metabolites) from inconsistent measurements among various studies, mainly arising from biological and technical heterogeneities. For instance, when identifying up- or down-regulated expressed genes between two conditions, the amount of change can vary from study to study, or sometimes even their signs (up or down) can be inversed. In this situation, meta-analysis could efficiently mitigate the issue by finding a consensus by means of summarizing the outcomes.

Existing meta-analyses are classified into three categories according to what to combine: *p*-value, *t*-statistic, and fold-change ratio. First, Fisher's inverse chi-square method combines *p*-values across *k* studies by summation of the logarithm of each *p*-value into a single statistic, $\chi^2_{2k} \sim \sum_{i=1}^{k} \ln(p_i)$, which is then compared against chi-square distribution with $2k$ degree of freedom (Fisher)

* Corresponding author.
E-mail address: dhlee@kaist.ac.kr (D. Lee).

http://dx.doi.org/10.1016/j.bbrc.2016.01.082

[5]. Its strengths include that it is very straightforward, easy to implement, and easily able to extend to more than two class comparison. Unfortunately, however, it is unable to estimate the magnitude of effect itself, so called effect size, which is often crucial where we aim to estimate, for instance, how much the amount of changes varies from study to study.

Second, *t*-statistic-based approach combines *t*-statistics from multiple studies usually with weight regarding study variance. A *t*-statistic itself captures a difference of two group means standardized by its pooled standard deviation. The *t*-statistic-based approach might be the most popular effect size-based method, presumably due to its well-established statistical background. However, practically, using *t*-statistic could be easily dominated by outliers, thereby causing misleading estimation. Moreover, inadequate variance estimation could make the outcome (*t*-statistic) distorted. For instance, a comparison of expression levels between two conditions could make a signal (difference) underestimated due to large variance, often occurring in complex diseases such as cancer whose samples usually have a wide range of expression levels (i.e., large variance) [6]. As many researchers point out, *t*-statistic-based approach seems preferred in a statistical sense rather than a biological sense [7,8]. Several meta-analysis methods have been developed in this category. Marot et al. and Wang et al. both use *t*-statistic but differ in their effect size definition and implementation; a moderated *t*-test statistic (Mod_t) [9] and a Bayesian framework (Bayes_t) [10], respectively. Also, Choi et al. use *t*-statistic of Hedge's and Olkin's [11], which corrects for biases due to small sample size. It utilizes so called inverse-variance technique for regarding within- and between-study variance, in which the study weight is inversely proportional to the study variance (GeneMeta) [12].

Third, fold-change-based approach is another effect size-based method, which combines values utilizing fold-change ratio. A fold-change ratio is simply a ratio between two representative values, usually median, from each condition. An increase or decrease of at least two fold, for instance, may be considered significant. Several features make this statistic useful for biomarker discovery: its simplicity, biological intuitiveness, and resistance to outliers. So, many studies using fold-change ratio find that it often offers more reproducible results than using *t*-statistic in microarray analysis [7,8]. However, the major drawback comes from its tendency to generate biased and erroneous outcomes mainly due to the lack of variance consideration [13]. In this category, Hong et al. propose a nonparametric rank product method utilizing the rank of fold-change ratio, and shows that it is highly effective especially in case of small sample sizes and large between-study variance (RankProd).

In this study, we propose a method called iMeta, which combines two types of effect size by integrating fold-change ratio and *t*-statistic. Prior to the integration, fold-change ratios over multiple studies are quantile-normalized against *t*-statistic distribution so as to minimize bias arising from the distributional difference. This approach aims at borrowing virtues of both effect size measurements to achieve robustness. On top of that, study variance, serving as weight for each study in combining multiple studies, is also estimated under two assumptions: fixed effect model (FEM) and random effects model (REM).

To evaluate predictive performance r5of the biomarkers identified, we compare our method with other five existing methods (Fisher, Mod_t, Bayes_t, GeneMeta, and RankProd), using three simulation datasets and two actual datasets of drug-induced liver injury (DILI). Computational and experimental evaluation shows that biomarkers identified by iMeta are more robust over diverse datasets in terms of sensitivity and specificity.

2. Materials and methods

2.1. Definition of iMeta index

We review iMeta method here. The fundamental idea is to integrate fold-change ratio and *t*-statistic over multiple studies with weight regarding each study variance (Fig. S1). It comprises three steps: 1) two independent statistical tests of gene *i* are performed within a single study *j*; 2) the resulting *t*-statistic (*T*) and fold-change ratio (*F*) are integrated in the control of parameter α, whie the *F* distribution is normalized by quantile-normalization function *f* against the *T* distribution so as to minimize bias arising from distribution difference. [Eq. (2)]; and 3) a weighted mean over *k* studies was calculated, whose weight ω_{ij} is an inverse of study *j*'s variance, regarding within-study variance (s_{ij}^2) and between-study variance (τ_i^2) [Eq. (6)]. The iMeta index, *iM*, is calculated as follows:

$$iM_i = \frac{\sum_{j=1}^{k} \omega_{ij} * E_{ij}}{\sum_{j=1}^{k} \omega_{ij}} \tag{1}$$

$$E_{ij} = \alpha * f(F_{ij}) + (1-\alpha) * T_{ij} \tag{2}$$

$$F_{ij} = \log_2 \frac{\hat{x}_{ij}^t}{\hat{x}_{ij}^c} \tag{3}$$

$$T_{ij} = \frac{\bar{x}_{ij}^t - \bar{x}_{ij}^c}{S_p} \tag{4}$$

$$S_p = \sqrt{\frac{(n_c-1)s_c^2 + (n_t-1)s_t^2}{n_c + n_t - 2}} \tag{5}$$

where E_{ij} is an integrative effect size of gene *i* in study *j*, and α is a parameter, called weight-on-fold-change, which determines relative weight of each statistic: higher values of α increase the impact of fold-change ratio. Our method practically determines the value of α from the dataset being analyzed, adjusting in a range between 0.0 and 1.0 by increment of 0.1 in a way to maximize an area under receiver-operating-characteristic curve (AUC) value. $\hat{x}^t$ and $\hat{x}^c$ denotes the median value of expression levels in treatment (t) and control (c) group, respectively, and $\bar{x}^t$ and $\bar{x}^c$ denotes the mean of expression level of each group. S_p denotes an estimated pooled standard deviation, where s_c and s_t are standard deviation and n_c and n_t are the number of samples in each group.

Many methods using *t*-statistic make two statistical model assumptions on study variance: (i) Fixed effect model assumes that true effect sizes are the same in all studies, but differences occur from sampling error alone; (ii) Random-effects model assumes that true effect sizes might be variable from study to study. Our method implements both of the models. The weight ω_{ij}, called weight-on-study-variance, is defined as an inverse of summation of within-study variance s_{ij}^2 and between-study variance τ_i^2 [Eq. (6)];

$$\omega_{ij} = \frac{1}{s_{ij}^2 + \tau_i^2} \tag{6}$$

$$s_{ij}^2 = \left[\frac{n_c + n_t}{n_c n_t} + \frac{E_{ij}^2}{2(n_c + n_t)} \right]^{(1-\alpha)} \tag{7}$$