BIOINFORMATICS

Vol. 00 no. 00 2009
Pages 1–4

Correctly identifying tissue-specific genes: The Intersection-Union Test and a Bayesian approach

K. Van Deun [a,*], H. Hoijtink, [b], L. Thorrez [c], F. Schuit, [d], I. Van Mechelen [e]

[a,e]SymBioSys, Catholic University of Leuven, 3000 Leuven, Belgium, [c]Department of Methodology and Statistics, University of Utrecht, 3508 TC Utrecht, The Netherlands [c,d]Gene Expression Unit, Catholic University of Leuven, 3000 Leuven,Belgium,

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation Finding genes that are selectively expressed in a particular tissue or condition is a problem that cannot be solved by standard statistical testing procedures. A relatively unknown procedure that can be used is the intersection union test (IUT). However, two disadvantages of the IUT are that it is conservative and that it is a discrete method that does not allow to distinguish in the degree of tissue-specificity.

Results We propose a Bayesian procedure that quantifies how much evidence there is in the expression profile for tissue-specificity. In a small simulation study, it is shown that the proposed method outperforms the IUT when it comes to finding specifically expressed genes. The method is illustrated to find muscle-specific expressed genes in publicly available data.

Availability Both MATLAB and R code that implement the IUT and the Bayesian procedure in an efficient way, can be downloaded at http://ppw.kuleuven.be/software/....

Contact katrijn.vandeun@psy.kuleuven.be

1 INTRODUCTION

Gene expression is often profiled for several normal tissues to get insights into gene function and regulation (Su et al., 2004; Dezso et al., 2008; Liu et al., 2008). In this respect, finding genes that are selectively expressed in a particular tissue is of utmost importance (Su et al., 2004; Liang et al., 2006) and many papers have appeared that report about genes specifically expressed in a targeted tissue. Although such genes are claimed to have higher expression in the target tissue than in *each of* the other tissues, often the statistics used are appropriate to find those genes that are significantly higher expressed in the target compared to *at least* one of the other tissues. A common strategy is to test each of the target - other tissue pairs (e.g., using t-tests) and to account for the problem of multiple testing with a multiple comparison procedure like Tukey-Kramer (see for example Liang et al. (2006)): A gene for which each of the pairwise comparisons leads to a significant result is then called tissue-specific. To understand why such an approach is erroneous, observe that multiple comparison procedures control the chance of wrongfully rejecting the null hypothesis of no significantly different target - other tissue pair against the alternative of *at least one* significantly different pair. What is needed, is a test that controls the chance of wrongfully rejecting the null hypothesis against the alternative of *all* significantly different target - other tissue pairs.

A test that would have the desired alternative hypothesis of significantly higher expression in the target tissue compared to each of the other tissues, was proposed by Berger (1982) and is known as the intersection-union test (Berger and Hsu, 1996). However, the intersection-union (IUT) test has two disadvantages. First, it is conservative implying that many tissue-specific genes would be missed; and second, the obtained results are discrete as the test only indicates whether a gene is tissue-specific or not without any distinction in the degree of tissue-specificity (see also Allison et al. (2006)). As an alternative, we would like to propose a Bayesian procedure.

In this paper we first briefly introduce the intersection union test and we introduce a Bayesian alternative. Subsequently both procedures are compared in a simulation study and they are used to find muscle-specific genes.

2 METHODS

2.1 Intersection Union Test

A gene is tissue selective if the null hypothesis H_{0j} that the gene is expressed equally or higher in tissue j than in the target tissue can be rejected against the alternative H_{1j} of higher expression in the target *for each tissue*. Using formal notation, the set of compound hypotheses is composed by $H_0 : \bigcup H_{0j}$ and $H_1 : \bigcap H_{1j}$. Note that this is different from multiple comparison procedures where the underlying set of compound hypotheses is that all partial null hypotheses hold ($H_0 : \bigcap H_{0j}$) against the alternative that at least one partial null hypothesis can be rejected ($H_1 : \bigcup H_{1j}$). Berger (1982) introduced a procedure to test the composite null hypothesis $H_0 : \bigcap H_{0j}$ against the composite alternative $H_1 : \bigcup H_{1j}$ (see also Berger and Hsu (1996)). A result is declared significant by this test at level α if it holds that each partial null hypothesis H_{0j} can be rejected at level α. As proven by Berger, the significance level of the compound test is less or equal than the significance level used for each of its partial tests.

The IUT can be applied to the problem of finding tissue-specific genes as follows. For a particular gene, test each target tissue - other tissue pair at the desired significance level (e.g., .05) using a suitable

*to whom correspondence should be addressed

© Oxford University Press 2009.

test-statistic like the t-test. Only when all pairs yield a significant result, the gene is declared tissue-specific. Note that to account for the problem of testing multiple genes, the significance level used can be adapted using Bonferonni's or Sidak's correction. An efficient implementation of this approach in MATLAB or R can be found online: http://ppw.kuleuven.be/software/...

Often the significance level of the IUT is (much) less then α such that the procedure is conservative: a gene will not be easily declared to be tissue-specific. This yields a very low false discovery rate, however at the cost of many false negatives. Another disadvantage of the method is that it indicates whether a gene is tissue-specific or not, but not to which degree. No distinction is made, for example, at the .05 level of significance between a tissue-specific gene that has p-values of .03 for each of the target - other tissue pairs and a gene that has $p-$values of .0001.

2.2 Bayesian Evaluation of the Constrained Hypothesis

The expression profile for a particular gene consists of $i = 1, \ldots, N$ expression levels y_i for $j = 1, \ldots, J$ tissues. The total sample size is equal to $N = \sum_j N_j$. The model for the expression level is

$$y_i = \mu_1 d_{1i} + \ldots + \mu_J d_{Ji} + \epsilon_i, \tag{1}$$

where μ_j denotes the population mean of tissue j, d_{ji} is 1 if an array is measured for tissue j and 0 otherwise, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The density of the data for this ANOVA model is

$$f(y \mid d_1, \ldots, d_J, \mu_1, \ldots, \mu_J, \sigma^2) =$$

$$\prod_j \prod_{i \in j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2} \frac{(y_i - \mu_1 d_{1i} - \ldots - \mu_J d_{Ji})}{\sigma^2}, \tag{2}$$

where $y = [y_1, \ldots, y_N]$ and $d_j = [d_{j1}, \ldots, d_{jN}]$.

The goal is to determine the support in the data for two hypotheses;

$$H_1 : \mu_1 > \{\mu_2, \ldots, \mu_J\}, \tag{3}$$

which states that μ_1 is larger than each of the means in the set $\{\mu_2, \ldots, \mu_J\}$, and

$$H_2 : \text{not } H_1. \tag{4}$$

Support in the data will be quantified using the Bayes factor (Kass and Raftery, 1995). Using Chib's approach (Chib, 1995) the Bayes factor of H_1 versus H_2 can be written as:

$$\mathrm{BF}_{12} = \frac{f(y \mid \theta, \sigma^2) h(\theta, \sigma^2 \mid H_1) / g(\theta, \sigma^2 \mid y, H_1)}{f(y \mid \theta, \sigma^2) h(\theta, \sigma^2 \mid H_2) / g(\theta, \sigma^2 \mid y, H_2)} \tag{5}$$

where $\theta = [\mu_1, \ldots, \mu_J]$, $h(\cdot)$ denotes the prior distribution of the parameters for the hypothesis indicated, and $g(\cdot)$ the posterior distribution. Note that the dependence of $f(\cdot)$ and $g(\cdot)$ on $d_1, \ldots, d_J$ is left implicit.

It is convenient to write

$$\mathrm{BF}_{12} = \mathrm{BF}_{1u} / \mathrm{BF}_{2u}, \tag{6}$$

where u refers to $H_u : \mu_1, \ldots, \mu_J$, that is, an unconstrained model. Once a prior distribution has been specified for H_u the prior distribution of H_m for $m = 1, 2$ is obtained via

$$h(\theta, \sigma^2 \mid H_m) = \frac{h(\theta, \sigma^2 \mid H_u) I_{\theta \in m}}{\int_\theta, \sigma^2 h(\theta, \sigma^2 \mid H_u) I_{\theta \in m} d\theta d\sigma^2}$$

$$= 1/c_m h(\theta, \sigma^2 \mid H_u) I_{\theta \in m}, \tag{7}$$

where $I_{\theta \in m}$ equals 1 if θ is in agreement with the constraints of model m and zero otherwise, and c_m is the proportion of H_u in agreement with H_m.

The prior distribution of H_u is

$$h(\theta, \sigma^2 \mid H_u) = \prod_j \mathcal{N}(\mu_j \mid \mu_0, \tau_0) \text{Inv-}\chi^2(\sigma^2 \mid \nu, \sigma_0^2), \tag{8}$$

that is, the prior distribution of each μ_j is the same. With this specification, and independent of the choice of μ_0, τ_0, ν_0 and σ_0^2, $c_1 = 1/J$ and $c_2 = (J-1)/J$, which makes sense because H_2 consists of a set of $J-1$ models: $\{\mu_j > \{\mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_J\}$, for $j = 2, \ldots, J\}$ and H_1 is the only model not contained in H_2. Using (2.2) and (8) BF_{1u} can, for any value of θ in agreement with H_1, be written as:

$$\mathrm{BF}_{1u} = \frac{f(y \mid \theta, \sigma^2) h(\theta, \sigma^2 \mid H_1) / g(\theta, \sigma^2 \mid y, H_1)}{f(y \mid \theta, \sigma^2) h(\theta, \sigma^2 \mid H_u) / g(\theta, \sigma^2 \mid y, H_u)}$$

$$= \frac{J h(\theta, \sigma^2 \mid H_u) / ((1/f_1) g(\theta, \sigma^2 \mid y, H_u))}{h(\theta, \sigma^2 \mid H_u) / g(\theta, \sigma^2 \mid y, H_u)}$$

$$= J f_1, \tag{9}$$

where f_1 is the proportion of the unconstrained posterior distribution in agreement with the constraints of H_1 because:

$$g(\theta, \sigma^2 \mid y, H_m) = \frac{g(\theta, \sigma^2 \mid y, H_u) I_{\theta \in m}}{\int_\theta g(\theta, \sigma^2 \mid y, H_u) I_{\theta \in m} d\theta}$$

$$= 1/f_m g(\theta, \sigma^2 \mid y, H_u) I_{\theta \in m}. \tag{10}$$

Using a similar derivation for BF_{2u}, it is obtained that

$$\mathrm{BF}_{12} = \mathrm{BF}_{1u} / \mathrm{BF}_{2u} = \frac{J f_1}{\frac{J}{J-1} f_2}. \tag{11}$$

See Klugkist and Hoijtink (2007) for a more elaborate discussion of the derivation of the Bayes factor for inequality constrained hypotheses in the context of ANOVA.

The only question remaining is the estimation of f_1 because $f_2 = 1 - f_1$. Using a very large (but finite) number for τ_0, any number for μ_0, $\nu = -2$ and $\sigma_0^2 = 0$, that is very uninformative priors, the following algorithm renders a sample from the unconstrained posterior distribution of $\mu_1, \ldots, \mu_J, \sigma^2$ and an estimate of f_1:

- Step 1: Assign initial values: $\mu_j = \overline{y}_j$ for $j = 1, \ldots, J$ and $\sigma^2 = 1/N \sum_i (y_i - \overline{y}_1 d_{1i} - \ldots - \overline{y}_J d_{Ji})^2)$, where, $\overline{y}_j$ denotes the sample average for tissue j.
- Step 2: For $j = 1, \ldots, J$ sample μ_j from $g(\mu_j \mid \sigma^2, y)$ which is a normal distribution with mean $\overline{y}_j$ and variance σ^2 / N_j.
- Step 3: Verify whether or not the current values of $\mu_1, \ldots, \mu_J$ are in agreement with the constraints of H_1.

Table 1. Proportion of genes for which the IUT rejected H_0 in favor of H_1 ('the gene is tissue-specific'). The different rows correspond to different amounts of support of H_1 while the different columns correspond to different effect sizes.

	$\delta = 0.5s$	$\delta = 2s$
Complete support of H_1	0.002	0.278
One tissue not supporting H_1	0.002	0.000
No support at all of H_1	0.000	0.000

- Step 4: Sample σ^2 from $g(\sigma^2 \mid \mu_1, \ldots, \mu_J, y)$ which is a scaled inverse chi-square distribution with degrees of freedom $N-2$ and scale parameter $1/(N-2)\sum_i (y_i - \mu_1 d_{1i} - \ldots - \mu_J d_{Ji})^2$).
- Iterate Steps 2 through 4. The proportion of vectors $\mu_1, \ldots, \mu_J$ sampled in Step 3 in agreement with the constraints of H_1 is an estimate of f_1.

An efficient implementation of this algorithm in MATLAB or R can be found online: http://ppw.kuleuven.be/software/...

3 RESULTS

3.1 Simulation

Here we compare the performance of the two testing procedures in a controlled setting using simulated data. In the simulation we mimicked the expression data: for each gene, 64 expression values were generated divided over 20 tissues with minimum three and maximum five replications for a tissue. Each of these values was sampled from a normal distribution with standard deviation s equal to 0.04. The mean for the target tissue, μ_T was set equal to seven and the remaining means were defined relatively to the target as follows: For a tissue supporting the hypothesis that the gene is selectively expressed in the target, $\mu_j = \mu_T - \delta$, while for a tissue not supporting the hypothesis $\mu_j = \mu_T + \delta$. δ is a positive number taking two values, namely $\delta = 0.5s = 0.02$ and $\delta = 2s = 0.08$. In this way we could manipulate the amount of support for tissue specificity (complete support by all 19 tissues, one tissue not supporting and 18 supporting, no support all by all 19 tissues), both for moderate ($\delta = 0.02$) and large ($\delta = 0.08$) effect sizes.

The results for the IUT are summarized in Table 1 that reports the proportion of genes for which the null hypothesis was rejected in favor of the alternative H_1 ('higher expression in target than in each other tissue') for the three different amounts of support. For the Bayesian procedure, the results are summarized by means of the histograms in Figure 1: these are histograms of the log transformed Bayes factor[1]: a value above zero indicates that there is more support for H_1 in the data then for not H_1. Clearly the IUT is conservative. Even in the most favorable case (complete support, large effect size), only 28 percent of the genes are declared to be tissue-specific[2]. Compare this with the corresponding results for the Bayesian procedure (upper panels of Figure 1): The majority of the

[1] Bayes factors below 0.0001 were set equal to 0.0001 while Bayes factors over 10000 were set equal to 10000.

[2] This number drops to 0 when accounting for the multiple testing associated to the testing of 5000 genes, results not shown.

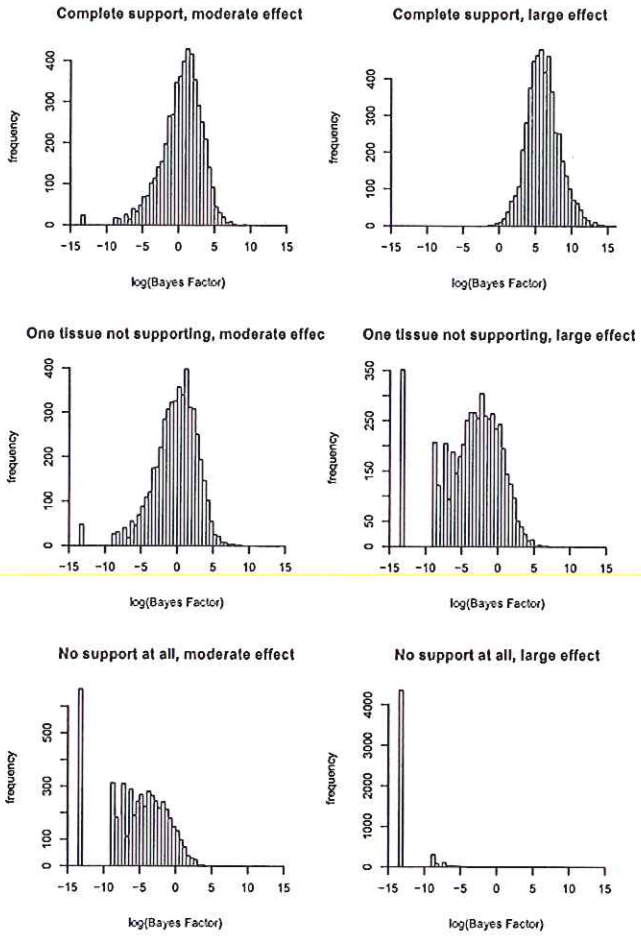


Fig. 1. Histograms of the $\log_2$ transformed Bayes factors. (a) Complete support of $H1: \mu_j = \mu_T - \delta$ for all i=1,...,19 (b) One group not supporting H_1: $\mu_j = \mu_T + \delta$ for one i and $\mu_j = \mu_T - \delta$ for the 18 remaining i's (c) No support at all of H_1: $\mu_j = \mu_T + \delta$ for all i=1,...,19. Panels at the left are obtained with a medium effect size, at the right with a large effect size.

Bayes factors indicate that H_1 is supported, even when the effect size is small. This more liberal character of the Bayesian approach can lead to more false positive results (middle and lower panels of Figure 1) but this can be solved by requiring that the log Bayes factor should be larger than seven.

3.2 Muscle-specific genes

We used a publicly available microarray dataset (http://www.ncbi.nlm.nih.gov/geo/, accession number GSE9954; see alsoThorrez et al. (2008)) consisting of 22 different murine tissues, with 3-5 replicates for each tissue. Here, we are interested in finding genes selectively expressed in muscle. Because heart and diaphragm contain a lot of muscle, these tissues (6 arrays) were removed from the data. The data were RMA pre-processed and log2 transformed. In total, 761 unique genes were found that have a Bayes factor larger than 150 ($\log_2 150 = 7.2$) of which 457

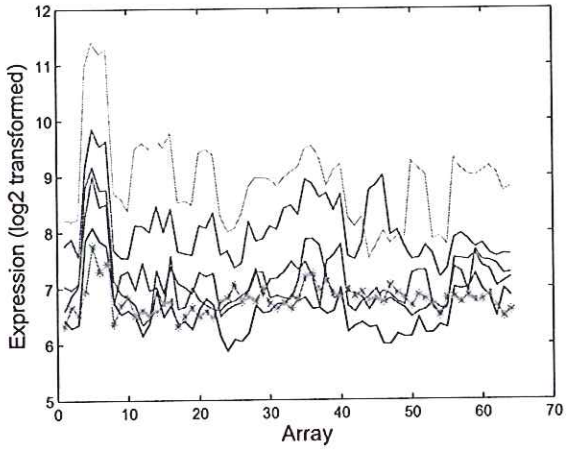


Fig. 2. Expression profile for all probesets matching CLIP1; the dotted profile was found to be expressed selectively in muscle by the Bayesian procedure but not by the IUT.

consistently on each probeset matching the gene[3]. Using the IUT, only about halve (254 to be exact) of these genes were found. An example of a gene with many matching probesets that was detected with the Bayesian procedure but not with the IUT, is CLIP1. Its expression profiles are plotted in Figure 2.

Something about genes also found for human??

1. public data to which we apply the Bayesian procedure
2. using TIGER, an online tool to find tissue-specific genes

REFERENCES

D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7:55–65, 2006.

R. L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24:295–300, 1982.

R. L. Berger and J. C. Hsu. Bioequivalence trials, intersecion-union tests, and equivalence confidence sets. *Statistical Science*, 11:283–319, 1996.

Z. Dezso, Y. Nikolsky, E. Sviridov, W. Shi, T. Serebriyskaya, D. Dosymbekov, A. Bugrim, E. Rakhmatulin, R. Brennan, A. Guryanov, K. Li, J. Blake, R. Samaha, and T. Nikolskaya. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biology*, 6(1):49, 2008.

Shuang Liang, Yizheng Li, Xiaobing Be, Steve Howes, and Wei Liu. Detecting and profiling tissue-selective genes. *Physiol. Genomics*, 26(2):158–162, 2006.

Xiong Liu, Xueping Yu, Donald Zack, Heng Zhu, and Jiang Qian. Tiger: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9(1):271, 2008.

Andrew I. Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A. Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P. Cooke, John R. Walker, and John B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6062–6067, 2004.

L. Thorrez, K. Van Deun, L. Tranchevent, L. Van Lommel, K. Engelen, K. Marchal, Y. Moreau, I. Van Mechelen, and F. Schuit. Using ribosomal protein genes as reference: A tale of caution. *PLoS ONE*, 3(3)(e1854), 2008.

[3] Probesets with expression in the background, this is expression below $\log_2 150$ for each tissue, not included