Author’s response to reviews

Title: A novel joint analysis framework improves identification of differentially expressed genes in cross disease transcriptomic analysis

Authors:
Wenyi Qin (wqin2@uic.edu)
Hui Lu (huilu@uic.edu)

Version: 1 Date: 24 Jan 2018

Author’s response to reviews:

Reviewer reports:

Reviewer #1: In a typical transcriptomic dataset, the sample size is small. This makes detection of Differentially Expressed (DE) genes very difficult. One way of addressing this problem is to integrate multiple datasets of the same disease. This does have some improvement, but in many cases, the power is still limited. For two similar diseases, there could be overlap between the sets of DE genes. Under this assumption, the authors proposed a statistically principled framework to boost the detection power, where multiple datasets from similar diseases are analyzed jointly, and disease-shared and -specific DE genes are identified. By extensive simulations and two real data applications, the authors demonstrated the superiority of the new method over a single dataset method and a popular joint analysis method.

Overall, this paper is well-written. I have one major comment. In the proposed procedure, EM algorithm is used to estimate the prior probability. This can be computationally intensive, especially when the number of datasets considered is large. Does the prior probability have a large impact on the performance? Please discuss on this point. I suggest acceptance once this issue is addressed.

We thank the reviewer’s comment and added the following paragraphs in the “Conclusion and Discussion” part of the manuscript based on the reviewer’s suggestion. In the new paragraphs, we discussed the importance of prior probability in the proposed joint analysis framework and some potential solutions when the number of data sets is large. The new paragraph is shown as follows:

“The prior probability is the most essential quantity in the proposed joint analysis framework and has a large impact on the performance of the method because similarity between diseases are directly determined by this quantity. This has been demonstrated through both simulation study and real data application. In simulation studies, we observed that when jointly analyzed with diseases with higher similarity, which was realized by adjusting prior probability value among
diseases, the target data set gained more statistical power than less similar diseases. In real data application, more DE genes were identified among similar cancers than dissimilar ones where similarity among cancers were computed through estimated prior probability. In short, prior probabilities among different diseases could determine if the proposed joint analysis framework would be effective or not.

There would be several improvements for the proposed joint framework in the future. The first issue to be addressed is how to jointly analyze more disease data sets. As mentioned by one reviewer, the estimation of the prior probability in the proposed framework here is computationally intensive when the number of diseases to be jointly analyzed is large (~2N, where N is the total number of diseases). The estimation of prior probability would become infeasible when the number reaches 20 or more. Some potential solution to this problem has been proposed in a recent paper (Lai et.al, 2017). The basic idea is to assume special structures about the prior probability such that the number of prior probability to be estimated could be significantly reduced, thus incorporating more disease data sets becomes available. Another improvement would be to design a disease similarity test so that researchers could determine if two diseases are similar enough to be jointly analyzed. A similar idea has been proposed by Chung et.al (Chung et.al, 2014) where a likelihood test was designed to evaluate if two diseases contain similar SNPs. Finally, next generation sequencing support is expected to be added to current framework such that microarray and sequencing data could be analyzed simultaneously.”

We hope this addition addresses the reviewer’s concern.

Reviewer #2: In the paper "A novel joint analysis framework improves identification of differentially expressed genes in cross disease transcriptomic analysis", the authors proposed a Bayesian method to borrow information from other gene expression dataset to increase test power. The paper is well organized the explains the method clearly. After reading the paper, I have a few comments and concerns:

In equations, the authors use both Pr and f to show probability. What is the difference.

We use two different notations to obtain a better representation for the distribution of a categorical variable and continuous variables respectively. The Pr is used to represent the probability of a categorical variable, for example, a gene in DE or non-DE status; f is used to represent the density distribution of continuous variables such as the conditional density distribution of Z-scores. We hope this explanation answers the reviewer’s question.

In Figure 2B, FDR increases at 0.8 and 0.9 for 2 datasets. Is there an explanation for that?

We appreciate the reviewer to point that out. To address the reviewer’s concern, we reran experiments with same simulation setup 50 times by setting shared percentage at 0.6, 0.7, 0.8 and
0.9 and examined the average sensitivity and average FDR. We find that in this run the average sensitivity is still positively correlated to shared percentage, similar to what was observed in Figure 2A while average FDR in this experiment varies in certain range i.e. 0.00294 for 0.6; 0.00263 for 0.7; 0.0027 for 0.8; 0.00277. Based on this observation, we conclude that the observed increase of FDR where shared percentage is set to 0.8 and 0.9 in Figure 2B is not really an increase but a random fluctuation of FDR which is caused by the varying number of non-DE genes with a relatively strong signal (called “noisy genes” hereafter) in the target data set in each random run.

To explain the variation of FDR observed in two data sets setup in Figure 2B, note that our simulation randomly sets genes with different means and variance and will lead to different number of noisy genes for each generated data set. If there happens to be many noisy genes in a data set, the FDR at a specified cutoff will be larger because these noisy genes will tend to be identified as DE genes with limited sample size under a two data sets scenario for either single data set analysis or joint analysis. The different noise level in the target data set could also be observed when 4 and 6 data sets are jointly analyzed together in Figure 2B, i.e. the slight increase when shared percentage is set to 0.6 and 0.7 under four data sets scenario and the slight increase from 0.7 to 0.8 under six data sets scenario. On the other hand, the observed trend of decreased FDR with increasing shared percentage in four and six data sets setup is mainly because although noisy genes still exist in the target data set, true positives in the target data set could borrow information from more data sets when shared percentage increases which leads to more true positives above the specified cutoff than that in two data set scenario.

Nevertheless, the different noise level in each generated data set will not change the conclusion that with increasing shared percentage between target data sets and other ones, test power of joint analysis will increase.

Figure 5B shows the big correlation between colorectal cancer and pancreatic cancer of Z value. However, in table 3, the correlation between the two cancers is small. Why?

We thank the reviewer for pointing that out. This is a misplacement of correlation coefficient and we have fixed the figure (now Figure 6 as we added a new figure) accordingly.

Does the sample size for each data set have influence to the result?

Based on the reviewer’s suggestion, we ran another two-data set simulation. We first fix sample size of data set 1 (target data set) at 15 samples of controls and 15 samples of diseases and then change sample size of controls and diseases samples with 5, 10 and 15 in data set 2 respectively. The shared percentage value between two data sets is set to 0.6. We fixed the mean and variance
of each gene in each data set and repeated the experiment 100 times. We then examined the average sensitivity and false discovery rate and present the result in a new Figure 4. The description of results is also added to the manuscript which is shown as follows:

“Influence of Sample Size of a Similar Disease

Finally, the influence of sample size of a similar disease to be borrowed from is evaluated. To achieve this purpose, we first fixed the target data set with 15 disease and 15 control samples. Then, we generate the second similar (60% similarity) disease data sets with different sample sizes, each of which contains 5, 10 and 15 disease and control samples respectively. The mean and variance for each gene in each data set is fixed in this simulation. This simulation procedure is then repeated 100 times for each sample size parameter. After that, we apply both single and joint analysis on the simulated data sets and record the average sensitivity and FDR at specified cutoff=0.95 for each sample size parameter. The result is shown in Figure 4. As expected, the average sensitivity increases as the sample size increases. The average FDR is well controlled and only shows very small fluctuation due to sampling error in generating expression values for each gene. In conclusion, the simulation results demonstrate that the proposed joint analysis framework could borrow more information from a similar disease of a larger sample size.”

Did the authors try a simulation that there is no correlation between datasets? In that case, the result should be similar to single test.

Based on the reviewer’s suggestion, we generated 100 replicated data sets with 2,4 and 6 diseases with no overlapped DE genes between target disease data and other disease data sets from which information would be borrowed. We then calculated sensitivity and false discovery rate by setting the cutoff at 0.95 and added the result to Figure 2. The new Figure 2 showed that the result of integration of diseases with no overlapping DE genes is similar to that of integration of random diseases and both results of two negative controls are not different from the single data set analysis. The paragraph in the “Result” part is also revised accordingly.