

Dear Dr. Dudoit:

We are grateful to you and the anonymous reviewer for your truly helpful comments. These comments have driven us to review once again our understanding of modern state of the art in this extremely challenging sub-area of biostatistics. We did our best to be receptive when revising our paper. Before providing our detailed response to the points raised in the review, we would like to make a general comment.

Every scientist working in the field of microarray analysis has his/her own point of view on different aspects of the problem. This is because there are still many hypotheses and beliefs in this field which have no rigorous proof. We are no exception, and some of our convictions differ, sometimes drastically, from what has become a common belief in the literature. These are exemplified by but not limited to the following issues:

1. The idea that genes can be ranked by the magnitude of test-statistics (or p -values) originates from the belief that this criterion is biologically sensible. We do not think so. This does not mean that we rely entirely on biologists to make the outright decision. We only think that other sources of information (probably on between-gene interactions) should be invoked to build an adequate decision-making procedure. As of now, nobody can prove or disprove such a conjecture and the studies on gene ranking are perfectly legitimate. However, the burden of proof should be on those researchers who propose purely statistical ranking procedures rather than on those who dissent from the whole concept. This is the reason why we avoid discussing this issue directly in our paper which is focused on a different methodological problem. We have additional comments on this subject when responding to specific items below.

2. People who attempt to make sense out of 3-4 biological replicates truly believe that this is achievable. With the traditional statistical approach, however, such a belief definitely appears suspicious. This is probably the reason why the idea of pooling across genes has become so popular. In our paper, we show that the price for this way of overcoming sample size limitations can be enormous. In doing so, it is necessary to use a large data base but it is clear that our conjecture is even more valid for smaller sample sizes. Even a single counter-example typically disproves a concept, or at least indicates a serious problem. This is exactly what we did in our paper. Furthermore, it is our firm belief that no reliable statistical inference can be made with samples of arrays as small as 3-4 in this setting. We find it quite amazing that some traditional testing procedures controlling the FWER seem to do a good job (in terms of the stability of gene selection and overall power) with such moderate sample sizes as 40-50 (arrays) per group. The reason is that the hypotheses to be tested are formulated in terms of marginal distributions of gene expression levels and not in terms of joint distributions. As to the performance of FDR-based procedures, they appear to be more sensitive to correlation in the data even when the sample size is relatively large. We are currently submitting another paper that addresses these issues. As it stands, the microarray analysis produces too many artefacts to be truly useful to biologists. We believe that statisticians should be much tougher on the sample size issue if we want to make this technology work and produce biologically significant results. Therefore, it is unrealistic to expect us to provide a recommendation on what to do in situations where only 3-4 replicates are available for analysis. We simply think that nothing good can be done in such situations. Other investigators may feel differently, but it is their responsibility to prove

otherwise.

3. By the same token, we should not be held responsible for fixing the nonparametric empirical Bayes methodology because it is unclear whether it can ever be fixed. The best we can do is to verify some beliefs in the foundation of this method which pertain to many works that resort to pooling information across genes. For example, Dr. Storey believes that there is a clumpy dependence in microarray data and that such clumps are small. In the present paper, we provide evidence that this assumption is probably far from reality. A more direct evidence was provided in our previous publication (Qiu et al., 2005) where we showed that such clumps can be very big so that even consistency of associated estimators becomes questionable. However, Dr. Storey may still be able to find situations where the consequences of this assumption are much less disastrous. This is a normal course of events in science.

It is always good to offer an alternative when uncovering problems with existing methodologies. However, any alternative calls for a serious justification which implies extensive evaluation of its performance and, if necessary, validation of the underlying hypotheses. This cannot be done in a single paper. As a matter of fact, we do have an alternative in mind which I can informally present here. It has the following components:

- (i) A substantial increase of the sample size.
- (ii) The use of nonparametric testing based on the samples of arrays without any pooling across genes. We feel that the t -test is not good for this purpose and should be gradually replaced with distribution-free tests.
- (iii) A search for better concepts of familywise error rate in the spirit of your and Mark van der Laan's work in the field. In particular, the generalized FWER, denoted by $\text{gFWER}(k)$, is definitely a big step in the right direction.
- (iv) The second generation of methods should be designed to more fully utilize the information on interactions between genes. This is an absolutely different dimension in microarray data analysis and we are currently working in this direction.

You will agree that this kind of an alternative strategy hardly merits discussion in our paper which has a very specific focus.

I apologize for this lengthy preamble but it is necessary to show what kind of psychological difficulties we experience when trying to improve our paper along the lines of the review.

Specific comments.

The Reviewer

1. This is a valid point and we have changed the title accordingly. Regarding the parametric empirical Bayes methodology, it is obvious that it must suffer from the same problem. The problem is even exacerbated here because of the way the likelihood function is constructed. Hierarchical modeling that attempts to fit sample correlations does not remedy this difficulty because such correlations are also dependent variables. The actual magnitude of the resultant bias (which can only be bigger than for the NEMB) and variability is of minor importance as far as the proof (or rather disproof) of principle is concerned. We suggest a tool to be used by the authors of various parametric Bayesian approaches for making their cases.

2. The code is pretty simplistic and its inclusion in this paper is probably not warranted. The

code can be obtained from the authors on request. We have made note of that in the revised paper.

3. This is a delicate matter. In computer simulations, we do not need any special methods for estimating the FDR because it can be estimated nonparametrically from simulation runs. It is, so to say, the true FDR and we provide its values in relevant tables. However, the situation is not the same in real data analysis where we have no choice but resorting to some indirect methods proposed by other authors. These methods have their own flaws that are difficult to separate from those inherent in the NEBM. We present these methods only for illustration. However, their performance invites a special investigation which has just been completed. We plan to publish this material in another paper which is now in preparation. When discussing computer simulations it is more advantageous to provide direct estimates rather than those based on additional assumptions. We have added clarity to this issue in the revised version of our paper.

4. We agree. The number of figures related to different estimation methods has been reduced. The idea behind these figures was to demonstrate that it would be hopeless to fix the problem by just choosing a different density estimation method.

5. Figures and Tables in our paper represent supporting evidence so that the paper can be read without looking at them. We did our best to reduce the illustrative material by cutting 5 figures and 5 tables. We do not know what else can be done because all the evidence is important to prove our claims.

6. Unfortunately, this is not feasible. To the best of our knowledge, no working algorithms are currently available to generate multivariate normal variables with an arbitrary covariance matrix of such a high dimension. This is the reason why all papers dealing with dependencies in microarray data present simulated data with exchangeable correlation structure. We would like to emphasize that our simulation studies play a subsidiary role; their only purpose is to explain the effects observed in real data. By no means can such simulations be considered as an attempt to model the actual correlation structure which is clearly much more complex as shown in our previous paper (Qiu et al., 2005). We have made this clearer in the revised paper.

7. It is a large sample size that made it possible to carry out our study. In this sense, the St. Jude data set is unique in providing great scope for the use of resampling techniques. We are unaware of any other data set that can provide similar advantages. This approach is definitely impossible to implement wherever the sample size is small. In particular, the Apo AI data do not seem to be adequate for this purpose. We cannot produce concrete recommendations regarding sample size limitations because no theory is currently available to support any claim. This is a general problem in resampling methodology that remains unresolved. The problem we point out in our paper is of fundamental nature; it pertains to the methodology based on pooling across genes rather than to a specific data set. One of the main messages we are trying to convey is that extreme caution has to be exercised when dealing with dependent variables (test-statistics) in the analysis of microarray data. We think that this message is far more important than an attempt to show that the reported effect varies in magnitude among different data sets.

8. This heterogeneity cannot not affect our inference because we use a pivotal test-statistic within the nonparametric framework. The heterogeneity of subjects would certainly have been

a real concern had we performed a goodness-of-fit test since this would require specifying the distribution to be tested. In two-sample settings with pivotal test-statistics, it may only affect the power which is of minor concern in our study. The choice of the t -statistic was suggested by common practice although we believe that this is not the best choice. However, we had to implement the NEMB in exactly the same way as other authors do in order to prove our assertions.

9. It is a direct consequence of our analysis that the NEMB-based ranking of genes is also affected. Whenever the variability of the number of selected genes is high, the stability of membership in the list of candidate genes is expected to be low. This obviously has a strong effect on the ranking of candidate genes based on purely statistical criteria such as the magnitude of associated test-statistics or estimated posterior probabilities. It is clear that the reverse is not true. If the variance of the total number of selected genes is low, there still can be tangible variations in the stability of selection for individual genes, thereby affecting the composition of the resultant list of candidate genes. However, we do not need to look into the latter possibility because we observe huge variations in the number of genes declared differentially expressed. We have included additional explanations and the suggested references in the manuscript.

10. Done.

Minor revisions

A reference or at least further explanation is needed for the statement about how with some distribution free statistics, the needed null distributions can be derived theoretically, thereby obviating the use of observations altogether.

We have added an appropriate reference. Our special thanks for this comment because the part of our sentence that reads “thereby obviating the use of observations altogether” is misleading. What we meant was that one does not need to do permutations or bootstrap resampling to obtain quantiles of the null distribution. However, one still needs observations to compute the observed test-statistic and the corresponding p -value.

Page 4 line 4.

Done.

Page 5 line 2.

Done.

The Editor

Section 1:

This section provides good motivation for the problem addressed in the manuscript and a simple introduction to empirical Bayes methods. However, it is not clear from this section (p. 4-5) how the FDR is estimated and where exactly the key assumption of independent gene expression measures comes into play. Please clarify these issues.

Done. See also Item # 3 in the Reviewer’s section.

As mentioned in the referee report, the distinction between gene selection and gene ranking

should be addressed.

Done. See also Item # 9 in the Reviewer's section.

Section 2:

The reviewer suggests analyzing a number of other microarray datasets.

See Items # 7 and 8 in the Reviewer's section.

Other simulation models should be considered, e.g., with variable gene correlation ρ , ρ derived from actual microarray data. It seems from the description of the simulation study design, in Section 2.2, that only one dataset (i.e., one 1255×40 matrix) is generated for each model (i.e., SIMU00, SIMU02, SIMU04, SIMU06). Why not generate multiple datasets for each model? This would provide the sampling distributions of estimators of various quantities of interest in the NEBM, such as f , f_0 , π_0 , Z , and FDR.

See Item # 6 in the Reviewer's section.

Sections 3-6:

The sheer number of tables and figures is overwhelming and somewhat confusing. A subset of tables and figures should be selected for presentation.

We did our best to address this concern by cutting 5 figures and 5 tables. It was thought that space was not an issue, and the more specific evidence the better. See also Items # 4 and 5 in the Reviewer's section.

The table and figure captions are not always very clear. The captions should clearly state which quantity is being investigated (i.e., estimators of f , f_0 , π_0 , Z , or FDR) and whether the results pertain to microarray or simulated data. E.g. It is not clear what results are presented in Table 11. Mean and SD refer to which variable?

Done.

Boxplots may provide better graphical displays of the results for estimators of π_0 , Z , and FDR, and could perhaps replace a number of tables.

This is a good idea but we suggest a different display instead. Our suggestion is to graphically display the mean and variance as they are the quantities of primary interest. The estimated true FDR can be displayed numerically in the same figures. We did this to reduce the number of tables.

It would be useful to have a table summarizing the main results. Such a table could have rows corresponding to different estimators (estimators of f , f_0 , π_0 , Z , and FDR) and columns to various model parameters (e.g., ρ). Cells would indicate the behavior of an estimator in response to variations in a model parameter.

We are not sure that this is feasible because it would require multiple entries in each cell. Please note that our simulations serve only explanatory purposes and specific numerical values of per-

formance indicators are of no special interest.

The impact of NEBM assumption violations should be examined for both gene selection and gene ranking.

We discuss this issue in the revised manuscript. See also Item # 9 in the Reviewer's section.

Section 7:

The discussion of multiple testing methods for dependent data should refer to the recent work of Dudoit et al. (2004a,b), van der Laan et al. (2004a,b), and Pollard and van der Laan (2004).
Done.

Once again, let me express our gratitude for your thoughtful review of our article.

Yours sincerely,

Andrei Yakovlev