Referee: 1

Comments for the Author
Hughey and Butte present a new method for meta-analysis of microarray datasets. This method has two main advantages over the existing methods. First, it uses a multivariate technique to select genes with non-redundant information. This leads to the second advantage of identifying smaller set of genes which are especially useful for diagnostic and prognosis modeling.

All sections of the article is well written and they should be commended for the brevity of the manuscript. I recommend the article be published subject to the following comments.

We very much appreciate your comments. Our responses to each of your concerns are below.

Major Comments:

1) The authors sometimes refer studies by their GSE number and sometimes by the corresponding publication (e.g. Bhattacharjee et al) which I struggle to follow. I also struggle to piece together the characteristics of the other studies. There is no mention what the CLCGP stands for and if it is published or unpublished.

I request a table in the main manuscript listing the characteristics of all eight studies identified here. For each study, you can list the GSE number, corresponding publication and year, sample size (+breakdown for each of the subtype and which samples were excluded here), array platform (e.g. Affymetrix HGU133 plus 2.0 not GPLxxx), whether it was used for discovery or replication here, age range with mean, proportion male, proportion of ex/current/never smokers, whether raw data was available for this study or only processed GEO data. Supplementary Table 1 and Figure 6 can be eliminated.

You are right that our table was missing important information. We have revised and expanded the table according to your suggestions. Because the table now has so many columns, we have put it in an Excel file and kept it as Supplementary Table S1. We believe Figure 6 is still valuable, because it breaks down the subtype frequency by patient sex, information that would make the table too complicated. The Bhattacharjee and CLCGP datasets are not on GEO, which is why we refer to them by name.

2) Why select five studies for discovery and three for replication? What happens if you use seven for discovery and test on the eight. Then repeat for other seven studies (i.e. leave one out).

You make an interesting point. Our original strategy was to keep the number of discovery studies to a minimum, and save the rest for validation, which has

worked well for our lab in previous meta-analyses. That strategy is also easier to explain, because the procedure you describe is actually a nested leave-one-out (leave one study out for validation, then perform leave-one-out cross-validation on the others).

Based on your question, we performed the nested leave-one-out procedure you describe. The performance of our method in this procedure was practically identical to the performance in our original leave-one-study-out cross-validation. The overall accuracy was still 91%, and prediction accuracies for each subtype (95% for AD, 88% for SQ, 76% for SCLC, and 82% for CAR) were also very similar. Because the nested leave-one-out procedure is more complicated and does not alter the performance of the classifier, we do not believe the manuscript would benefit from including this analysis.

3a) I am not familiar with elastic net and I was not able to follow the equations for weight calculation. If there are three studies, then w_i = 3 / n_i according to third equation. And according to second equation sum(w_i) = sum(n_i) which implies 3(1/n1 + 1/n2 + 1/n3) = n1 + n2 + n3 which does not hold for positive values of n_i.

We apologize that our procedure for calculating the sample weights was unclear. We have revised our notation in the manuscript. The key is that w and n are indexed by sample, not by study. Suppose there are three studies, where study A has 2 samples, study B has 3 samples, and study C has 4 samples. So M=3 and N=9. For each sample in study A, n_i=2. For each sample in study B, n_i=3. For each sample in study C, n_i=4. Expressed as a vector, n = [2, 2, 3, 3, 3, 4, 4, 4, 4]. Therefore, w = [3/2, 3/2, 3/3, 3/3, 3/3, 3/4, 3/4, 3/4, 3/4] and sum(w) = 1.5*2 + 1*3 + 0.75*4 = 9. As a result, the weight of a particular sample is inversely proportional to how many samples are in the same study as that sample.

3b) Please specify what ||B||2 and ||B||1 means in first equation.

We apologize for not defining those terms. $||\beta||_2$ refers to the L2-norm and $||\beta||_1$ refers to the L1-norm. We have added this information to the manuscript.

4) It is good that the authors discussed the problem that arises averaging multiple probes that map to one Entrez Gene ID. However, can they discuss the problem of using only the probes that are common to all five discovery datasets?

Here, the authors state they used 7,200 genes but there could easily be 2 - 3 as many genes if one considers the union instead of intersection. The problem becomes even more severe as number of discovery studies goes up. For example, if you have a gene present in 99 of 100 discovery studies, you would exclude it because it is missing in one study. Can you relax your criteria to allow genes present in at least 50% of studies?

Thank you for bringing up this point. We have added a paragraph to the Discussion addressing this issue. One thing to keep in mind is that the size of the intersection is determined not by the total number of datasets, but by the number of unique microarray platforms. Given that the vast majority of publicly available gene expression data is based on a small number of platforms and that any given phenotype is typically associated with the expression of many genes, we believe the current implementation will work well for most meta-analyses. Our deliberate choice of using the Bhattacharjee dataset for discovery represents a near worst-case scenario. Excluding the Bhattacharjee dataset, which was collected on the Affymetrix HGU95Av2 GeneChip, would raise the number of Entrez Gene IDs in the merged discovery data from 7,200 to 13,609 (at the cost of losing valuable samples for SCLC and CAR). In future work, we will explore imputing the expression of genes that are present in some datasets but not in others.

5) You can easily determine the sex information from microarray data and check with the reported sex. There are R packages such as massiR that can do this with a few lines of code. In my own experience, a few samples have misreported of sex information. Your Figure 7 also seems to suggest this is the case.

Thank you for pointing us to this package, which we were not aware of. As you mention, we had guessed that a few samples had misreported patient sex. We have now investigated this further and discovered that of all the genes on the Y chromosome whose expression is measured on the discovery datasets, RPS4Y1 has by far the highest variance in expression and the largest standardized difference in expression between males and females. This explains why the elastic net selects RPS4Y1 as a proxy for patient sex. The prediction of sex using the strategy of massiR is the same as the prediction based on RPS4Y1 alone. We have added a statement about this to the Results section and have added a corresponding Figure in the Supplementary Material.

Minor comments:

6) Can you justify using tSNE for visualization over the more widely used techniques such as PCA?

We have added text explaining why we used t-SNE. In our experience, consistent with the published literature, t-SNE produces superior visualizations of high-dimensional data compared to other techniques such as PCA.

7) Change "639 samples and 7200 genes." to "639 samples and 7200 genes that were present in all five discovery datasets."

Thank you for suggesting this clarification, which we have added to the text.

8) Can you possibly filter out genes that fail detection above background (DABG)

or absence/presence filter in say 50% of samples or less? Could this reduce the problem of signal from one probe being drowned out by the noise?

This is an interesting suggestion. However, when processing raw Affymetrix data, the averaging of multiple probes and probe sets occurs *within* the RMA function, to which we pass the Brainarray custom CDF. Previous work suggests that this procedure improves precision and accuracy of microarray-derived expression (Sandberg and Larsson 2007). Attempting to improve the RMA algorithm seems outside the scope of the present manuscript. It would be feasible to filter out probes from datasets that do not come as raw Affymetrix data, but that would introduce an inconsistency in how different datasets are processed, which we would prefer to avoid.

9) Can you comment on why prediction accuracy is high for some subtypes than others? Is it due to sample size / power issues or are these more heterogeneous cancer subtypes?

We have added a paragraph in the Discussion on this point. In short, both the issues you mention and a few others could be involved.

10) It would be more convincing if they can repeat this procedure with another cancer type (e.g. blood cancer subtypes) to see how their method works. However, I appreciate this is a lot of work.

We are working to apply our method to other diseases and conditions. As you mention, however, this is a lot of work. To keep the manuscript focused, we are only describing one application of the method.

11) Please provide a simple ASCII file with the expression values for 7,200 genes and > 639 samples as supplementary that can be used by other scientists. I applaud you for making the R codes and raw data files available. However the raw expression files were several gigabytes large and took several hours to download before I canceled it.

This is an excellent suggestion. We have added a csv file with the merged gene expression data from the discovery datasets (luca_main_discovery_expr.csv). We have also removed the raw data (which are already publicly available) and instead included RDS files containing the processed data, which total only about 300 MB and enable our meta-analyses to be reproduced without the raw data.

Finally, it is my policy to reveal my identity (Adaikalavan Ramasamy) to the authors when reviewing articles. Thank you.

Referee: 2

Comments for the Author
The authors present an interesting analysis using Elastic Net to combine and use data from multiple studies. The strength of this method is that it was able to successfully combine extract results from eight studies, which is a difficult task. I think the paper was well written, and their analysis approach was complete and well-executed. However I do have some comments/questions regarding this study:

Thank you for your kind words regarding our work. Please see our responses to each of your comments below.

1. Paragraph 2 on the introduction alludes to the fact that there is no general solution for accounting for covariates in meta-analysis problems (which I agree with). However, this lead me to believe that the authors were about to propose a general solution, but then went on to solve a meta-analysis specifically for 8 datasets (not generally). The flow here was a little misleading.

We apologize that the flow was misleading. We do believe that our method offers a general solution (or at least, as general as one could hope for) for dealing with covariates in a meta-analysis, because our method merges the multiple datasets into a single matrix, after which the covariates can be treated simply as additional rows (if rows are features) in the matrix. We show one example of using covariates in our meta-analysis of lung cancer, but there is nothing unique about the covariates in our meta-analysis. We have revised the relevant section of the Results to make this more clear.

2. Along the same lines as #1, in the authors claim to "present a methodological framework for using the elastic net..." whereas in reality, they really only "apply a methodological framework..." I think the distinction is important, because applying a framework will only require methods and software/code used in a specific case, whereas I think the "presentation" of a framework implies one is presenting something more general--which would include the need for software tool for general use on other similar problems.

Thank you for making that distinction. We have improved the documentation of the code and written a detailed procedure for running a meta-analysis (“Running your own meta-analysis.doc” at https://www.copy.com/s/WZnAESCw2ZCilnkQ). We have also changed “present a methodological framework…” to “describe a methodological framework...”

3. The following is a valid statement and an accurate description of the value and contribution of the study: "Our meta-analysis results in a robust and accurate multinomial classifier that distinguishes between four lung cancer subtypes using a small set of genes. Our method also enables us to rigorously demonstrate the

value of a meta-analysis, in that training a classifier on multiple studies improves prediction compared to training a classifier on only one study."

The preceding is an appreciated comment ;-).

4. The authors used "alpha=0.9 for the elastic net penalty". Why was this parameter chosen, and what is the impact on the study result if this was set at a different value? Selection of the lambda parameter via cross-validation was well executed.

We have added a sentence to the Methods section on this point. For this meta-analysis, we found that varying the value of alpha only changed the value of lambda at which the multinomial deviance was at a minimum, but did not change the minimum value of the multinomial deviance. In other words, lower values of alpha led to a classifier with more genes, but with identical performance. Compared to the classifier from alpha=0.9, the classifier from alpha=1 has only five fewer genes. We chose alpha=0.9 to avoid possible issues caused by the lasso's (alpha=1) sensitivity to extreme correlations and inability to select more than n features (from data with n observations and p features).

5. The following statement is not correct: "Importantly, our cross-study normalization does not use the sample metadata (e.g., cancer subtype), so later prediction is unbiased." Note that ComBat with covariates will not introduce prediction bias, but in unbalanced designs will introduce systematic correlation into the adjusted data that may be correlated with the outcome of interest. This can lead to exaggerated confidence or prediction accuracy, but it is not bias. Would be more accurate to say "later prediction is not impacted by the adjustment". Incidentally, not incorporating covariate values can actually introduce (or leave in) bias in the data, but in balanced designs this bias usually results in reduced significance or poorer predictability (but not necessarily the case in unbalanced designs).

Thank you for correcting us on this point. We have revised that statement.

6. The authors should also compare with other methods for generating predictions, including fSVA (others?)

Thank you for referring us to fSVA. We have added text to the Discussion comparing our approach for meta-analysis to fSVA and to other methods. The main difference with fSVA is that fSVA treats the batch information as unknown, whereas in a meta-analysis, the batch information is known.