Author's response to reviews

Title: Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors

Authors:

Christopher A Miller (camiller@bcm.edu)
Stephen H Settle (ssettle@mdanderson.org)
Erik P Sulman (epsulman@mdanderson.org)
Kenneth D Aldape (kaldape@mdanderson.org)
Aleksandar Milosavljevic (amilosav@bcm.tmc.edu)

Version: 3 Date: 15 November 2010

Author's response to reviews: see over
We thank the reviewers for their time and their suggestions. We have revised the manuscript and believe that we have now adequately addressed all of the reviewer comments. We have responded point-by-point below. Thank you.

**Reviewer #1**

In constructing a gene network, the preprocessing step was applied to retain only the genes that meet a frequency of recurrence. Is this process necessary? Because I guess the winnow algorithm might filter the matrix to remove all those trivial connections. Have you ever tested it without the preprocessing step? And, as the minimum frequency threshold in the preprocess, what value was used?

The edge scores that come out of winnow are based on exclusivity only, and do not consider recurrence (this is included during the significance calculations). Thus, if we did not remove low-recurrence genes, there would be a great number of edges in the graph with high exclusivity scores, but that would not be part of significant RME patterns. It may help to think of a situation where two genes are each mutated only once in different samples. In this case, the two genes would have perfect exclusivity, but such a pattern does not offer any meaningful insight since it is likely to occur by chance.

Since we use combinatorial search to consider every possible connected module in the graph, reducing the number of edges is critical to making such a search feasible. In addition, removing this step or lowering the threshold results in more false-positives. A comparison between threshold values of 5% and 10% is shown in Figure 2. When applying our data to GBM, we used a recurrence threshold of 10%, which is noted in the Results section under "Application to glioblastoma tumors".

In the last paragraph of page 6, does "using one gene as a classifier and the rest of the mutation array as training data" mean the leave-one-out cross-validation? Usually, "classifier" is called an algorithm or a function or a tool to classify data. I think the usage of the term "classifier" is not correct in that paragraph.

You are correct that classifier refers to a tool for classifying data. Winnow is exactly that kind of
algorithm, which is used to learn a linear classifier from labeled data. In addition to defining a separating hyperplane, Winnow reports weights that indicate which features are most informative. We take these weights and use them as scores for edges in the graph, then remove low-scoring edges. This allows us to significantly reduce the size of the graph which we must search, as shown in the Supplemental Methods. We have rewritten this section of the text so that this point is better explained (pp 6-7).

In the middle of page 7, what do the "overlapping modules" mean? What are the "component genes"? They should be clarified.

We have removed these terms and replaced them with a more clear definition.

In page 7, the authors are arguing the previous approaches based on probabilistic models have computational problems. But I think this statement is too extreme because many probabilistic models can be efficiently computed by various machine learning algorithms. The authors should give citations of the examples which have the computational problems?

Agreed. We now include text in the Results section (page 13) describing how we directly compared our method for establishing significance with two other probabilistic models. (additional details can be found in the Supplement). In short, we show that since permutation is needed to correct for multiple testing, these models require several orders of magnitude more processing time, making their use infeasible for large and highly-connected data sets. We have also pointed out additional advantages such as better suitability of our method for detecting modules involving more than two genes.

Figure 1 is the overview of RME module detection. However, it is not easy to understand the whole procedure of RME module detection from the figure. How about illustrating the proposed RME algorithm step by step?

Thank you for the suggestion. We have added a new figure (Figure 2) that gives an overview of the method.
In page 11, NCBI Entrez Gene source website, "Dai et al. 2005", and "SAS Institute" should move to the reference section.

Thank you for pointing this out. We have corrected it.

Reviewer #2

1. **Methods (Creating a mutation matrix)**: It is somewhat unclear how to go from SNPs and copy number data to the mutation matrix described in the last paragraph of page 5.

   We have added some text on pages 5 and 6 that should help clarify this procedure.

2. The proposed methodology requires a complex series of data manipulations and analytic stages. I think it would add to the clarity and reproducibility of the paper, if the overall analysis framework would be summarized in a diagram or algorithm description.

   We have added a new figure (Figure 2) that summarizes the analysis pipeline.

3. I think that selecting a subgroup of interesting genes and then try to identify functional modules in this reduced gene-set is a reasonable approach. On the other hand a multi-stage analysis is prone to error propagation, which is clearly not accounted for when the authors identify and evaluate functional modules. I understand that a full methodological development to deal with the issue is perhaps beyond the scope of this paper. However, the issue should be clearly assessed and discussed.

   Yes, unfortunately, we are currently limited by the scope of the TCGA pilot project, which assayed only ~1200 genes. Using this limited set of genes undoubtedly means that we are unable to detect some functional modules, but since our modules rely on mutual exclusivity, we expect that the most highly significant patterns that we find will be detectable even when many more genes are added.

   As we show with simulation and discuss in the conclusion, our method is absolutely applicable to
whole-genome data, and we could actually produce better inferences by working with such a data set. We look forward to applying our tool in such a manner in the future. Text has been added in the conclusion that acknowledges and addresses this point.

4. Prognostic significance: It is unclear how the proposed methodology brought the authors to the identification of EP300 as a prognostic gene. Perhaps this process should be described in more detail.

Though we discovered its interaction with TP53 and MDM2 using no prior information, we later found that EP300 has been associated with the p53 pathway and tumorigenesis in other tumor types (which we cite) but not in glioblastoma. This discovery of the role of EP300 in glioblastoma provides strong evidence that our method is capable of finding relevant genes and interactions.

We then decided to further investigate EP300 in order to show how one might go from gene module discovery all the way to clinically useful results. Thus our method allowed us to take a list of aberrations in glioblastoma samples across over one thousand genes, and quickly zero in on EP300, which turned out to have prognostic value in glioblastoma patients.

This section of the results has been revised to more clearly describe our rationale.

Reviewer #3

First, the most problematic issue is the lack of comparison with other methods. Since the problem of identifying mutually exclusive mutational patterns has been tackled by previous studies (as been cited in the manuscript), it is necessary to demonstrate that the new method excels in both accuracy and efficiency. A straightforward comparison is to examine two additional methods: the hyper-geometric p-value of mutual exclusion and the log likelihood ratio between the model that the mutations of two genes are coupled and the model that they are independent (a linkage-disequilibrium like score). What are the sensitivity, specificity and running time of those comparison methods?

We agree with this critique and have performed additional analyses to compare our method for
calculating significance to both a log-likelihood and hypergeometric method (as described in Yeang, et al 2008). An overview of the comparison can be found in the results section, and a full explanation of the methodology and results can be found in the supplementary data.

Second, some part of the algorithmic description is not clear. A better way to present the algorithm is to first summarize it as pseudocodes in a figure then describe each step in detailed texts.

Thank you. We have added a new figure (now Figure 2) that diagrammatically shows the flow of data through our analysis pipeline.

In addition, I am unclear about the following procedures:

(1) The description of the Winnow algorithm is rather vague. Also, I am confused about how it is used in the algorithm. Is it used to generate the modules? If so how is it related to the algorithmic significance scores?

We have edited the Methods section to better explain how Winnow is used to generate a graph of potential interactions. Briefly, this step occurs first, then we search within connected components of the resulting graph for modules and score them using algorithmic significance.

(2) The description of the greedy method to generate modules is also unclear. How did they grow the modules using the algorithmic significance? What are the stopping criteria for further expansion?

We examine every possible module at or below the specified maximum module size. We have added text to the "Identifying candidate modules" section that explains this.

(3) The authors calculated $P(x_{ij})$ conditioned on $a_{ij}$ and $b_{ij}$, the number of unobserved aberrations in the current genes and samples respectively. How come that the still unobserved information can be used to predict $P(x_{ij})$? The only reason I can think of is that they use the information of the entire matrix to encode $P(x_{ij})$ at each $i$ and $j$. If so then the authors should make clear that it is
not in a sequential prediction setup.

(4) Following (3), the score of $d'$ suggests each $x_{ij}$ is treated as independent. Yet the estimation of each $P(x_{ij})$ uses the entire matrix thus is not independent. Can the authors clarify the relations of the compression complexity and likelihood scores?

Compression is analogous to gambling, and in this encoding scheme, we are essentially placing bets on what the next bit in the matrix will be. Information theory tells us that the optimal betting strategy is to bet proportionally to the probability of success. The more information that our algorithm has, the better it can estimate the probability, and the greater the compression will be. The counterweight is that we have to pay a price for each bit of information put into the system, in the form of a penalty. If the entire matrix is provided to the algorithm, the compression will be optimal, but the penalty will be high, resulting in no gain. By providing only the total number of aberrations in each row and column, we provide enough information to allow our encoder to make significantly better bets, while still keeping the penalty score low.

Any information about the data (the matrix) supplied to the compression algorithm prior to compression is accounted for. Specifically, the number of bits to encode the count of aberrations per sample, $m \log^*(k)$, and the number of bits to encode the count of aberrations per gene, $k \log^*(m)$, are subtracted from the number of bits saved by compression.

To answer your second question, the probability calculation is used for the sole purpose of estimating the number of bits of information necessary to encode the observed data. The algorithm uses summary statistics (number of mutations per gene and per sample) that are supplied to it prior to seeing the data (this information is accounted for, as described above). When encoding the specific element in the matrix, the algorithm also uses information in the matrix that it has already seen. The advantage of this algorithmic compression approach is that these multiple sources of information can be used. An equivalent probabilistic formulation will be much more complex exactly because of the complex dependencies involved. The algorithmic compression approach therefore simplifies formulation greatly while still allowing calculation of a discriminative significance score.

Third, it is intriguing that high EP300 expression was associated with improved
survival. However, logically this finding does not necessarily support the functional relation of EP300 with P53 or MDM2.

Though we discovered its interaction with TP53 and MDM2 using no prior information, we later found that EP300 has been linked to the p53 pathway in other studies (which we cite). This includes direct evidence that EP300 acetylates p53 and contributes to tumorigenesis. This rediscovery of a known oncogenic pathway provides strong evidence that our method is finding real interactions.

We then decided to further investigate EP300 in order to demonstrate how one might go from gene module discovery all the way to clinically useful results. By doing so, we show that our method allowed us to take a list of aberrations in glioblastoma samples across over one thousand genes, and quickly zero in on EP300 through identification of recurrently altered functional modules. Further examination of this gene showed that it has prognostic value in glioblastoma patients.

We have added text which clarifies this process in the results.

**Reviewer #5**

1. The algorithm works off of mutation matrix. Construction of this matrix, while not unreasonably done, is prone to error and some ad hoc procedures. It would be useful to see, perhaps via simulation, how sensitive downstream analyses are to incorrectly calling aberrations/non-aberration status and to various parameters (e.g. threshold for copy number calling).

Yes, the quality of the results that we produce is obviously dependent upon the quality of the input data. That said, our algorithm is robust when confronted the most common types of error associated with the input data.

Deep sequencing and accurate algorithms have made SNP calling error rates very low, and the TCGA SNPs that we use were discovered and then confirmed with a round of verification assays. Even if a few SNPs were excluded or miscalled, the net result would be minimal. Given the large number of samples assayed, and the expectation that these random errors would occur in
different genes, the net effect on any given pathway's mutational signature would be very small.

A much more limiting factor in our experience turned out to be the resolution of copy number data. Determining which gene or genes is driving a copy-number aberration is complex, as we describe in the text. The \textit{CYP21B1/CDK4} oncogenic cluster we discuss in our results section is a good example of what happens when these calls are less exact than we might hope. Nonetheless, as GWAS studies have demonstrated many times, even identifying narrow genomic regions can be quite useful and inform future experiments.

2. The proposed algorithm appears to have practical advantages over COSMIC in terms of computational efficiency, but it is not immediately clear that the proposed approach would be much better in terms of accuracy. Comparisons with this existing approach should be done -- even if the competing methods would only work for small data sets.

We agree, and have now performed these comparisons, which we present in the Results section. A complete description of the methodology is found in the supplement.

3. The definition of sensitivity and specificity on paragraph 1 of page 12 do not appear to match standard definitions. These are typically scored with regard to how often the RME module is called significant. The meaning behind the current definitions are somewhat unclear.

Yes, upon review, these definitions were somewhat confusing. We have rewritten this section to make it more clear.

4. Cover page: "Department of Molecular and Huan Genetics" is probably misspelled.

Thank you for pointing out this oversight.

5. Some of the text from the legend of figure 1 should be directly in the
manuscript (e.g. the exclusivity is never defined within the main body of the article).

We have added some of that information to the text, and exclusivity has now been defined in the main text.

6. The description of the winnowing algorithm is somewhat confusing and could be expanded.

Additional explanatory text has been added on pp 6-7 that helps to explain the role of Winnow in our procedure.

7. Page 8, paragraph 2, line 4: parentheses appear mismatched.

Corrected – thank you.

8. Page 12, paragraph 3: Description of the simulations may fit better in the method section (since other simulation descriptions are there). Alternatively, I supposed one could move the simulation description to the results.

Thank you for your input, but we feel as though the results from the simulations belong in the Results section, and the description of the precise parameters used belongs in the Methods.