BRIGHAM AND WOMEN'S HOSPITAL

HARVARD MEDICAL SCHOOL

77 Avenue Louis Pasteur, Room 466D
Boston, Massachusetts 02115-6195
Tel: 617.525.4725, Fax: 617.525.4705
Email: mlbulyk@rascal.med.harvard.edu

Martha L. Bulyk, Ph.D.
Assistant Professor of Medicine, Pathology, and Health Sciences & Technology (HST)
Division of Genetics, Department of Medicine
Brigham & Women's Hospital,
Harvard Medical School
Department of Pathology
Brigham & Women's Hospital,
Harvard Medical School
Harvard-MIT Division of Health Sciences & Technology (HST),
Harvard Medical School

Dr. Peter Newmark, Editor-in-Chief
BMC Bioinformatics
BioMed Central Ltd
Middlesex House
34-42 Cleveland Street
London W1T 4LB, UK

February 13, 2006

Dear Dr. Newmark,

I have enclosed a revised manuscript in *BMC Bioinformatics* **Research Article** format, entitled **"Meta-Analysis Discovery of Tissue-Specific DNA Sequence Motifs from Mammalian Gene Expression Data"**. In this revised manuscript, we have attempted to address all of the reviewers' comments, as described in detail in the Response to Reviewer Comments on the following pages. In addition, although our manuscript is a Research Article, as part of our revised manuscript we are now making available the software that we developed for this study.

As described in my cover letter accompanying our original manuscript submission, I believe that this manuscript fits very well within the scope of *BMC Bioinformatics*, as it covers a number of computational approaches and problems that are central to the analysis of genomic datasets and for the understanding of transcriptional gene regulation. In addition, I think that this manuscript should be considered as a **Research Article**, because of our findings on the occurrences of known transcription factor binding site motifs within tissue-specific gene expression clusters, and how that relates to whether the given transcription factor is detected as being expressed in those particular tissues. These findings may be of particular interest to researchers focusing on particular tissue types, and such researchers may miss these results if this manuscript were to be published as a Methodology article or Software paper.

Sincerely,

Martha L. Bulyk

Martha L. Bulyk, Ph.D.

Response to reviewer comments for:
Title: Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data
Authors: Bertrand R. Huber and Martha L. Bulyk
MS ID: 3515251788298314
Journal: BMC Bioinformatics
Article type: Research article

We thank the anonymous reviewers for their interest in our manuscript, and for all of their helpful comments. Provided below is a point-by-point response describing our attempts to address, and when possible incorporate, all of their requested revisions in our manuscript.

Reviewer #1:

Major Compulsory Revision #1: We agree with the reviewer that it would valuable to be able to estimate the false discovery rate (FDR) for the discovered motifs. However, the issue of calculating a FDR is a difficult one, particularly because of our observation that once we considered five separate sets of matched randoms, the number of highly group specific motifs resulting from the matched randoms far exceeded the number of comparably scoring motifs resulting from the skeletal muscle CRMs (see also p. 11-12 of the revised manuscript). Indeed, over all expression clusters that we examined, on average we discovered 3.9 novel motifs and 2.9 motif matching TRANSFAC motifs that scored 2 SDs above the geometric mean of the matched randoms, whereas over all the corresponding size-matched randoms, on average we discovered 4.6 novel motifs and 3.5 motif matching TRANSFAC motifs that scored 2 SDs above the geometric mean of the matched randoms. At a significance threshold of 1 SD above the geometric mean of the matched randoms, these numbers are 9.0 novel motifs and 6.6 TRANSFAC motifs for the expression clusters, and 14 novel motifs and 11 TRANSFAC motifs for the matched randoms. We believe this occurs for the following reasons: (1) when strong motifs are present in an input sequence set, the motif finders tend to keep finding variant versions of those same motifs over and over, and (2) for practical reasons, we capped the number of motifs to be discovered by each motif finder at 30 per run. Thus, for the expression clusters, we see that the same motifs are being found over and over, whereas for the size-matched randoms, a larger number of distinct motifs are discovered. It is possible that with more stringent block filtering criteria, so many distinct motifs would not be discovered for the matched randoms as compared to for the expression clusters.

To sum up, it's not clear how or if one could use the size-matched randoms to derive a FDR for the motif discovery. Nevertheless, the above finding suggests to us that our comparison of group specificity scores for motifs resulting from an input tissue-specific query set versus motifs resulting from five separate sets of matched randoms (i.e., "SDs above the geometric mean of the matched randoms") is likely a conservative threshold for assigning statistical significance. One could use the distribution of scores from the matched randoms to calculate a rough approximation of the false discovery rate. This is why we provide in Additional Data File 3 the geometric means and standard deviations of the group specificity scores for motifs resulting from the five sets size-matched randoms for each expression cluster (separately for block-filtered and non-block-filtered), and it is this information that we used to generate the motif heat maps shown in Figure 7a,b.

Regarding the discordance between our set of discovered motifs and the motifs discovered by Xie et al., we offer some possible explanations for this discrepancy on p. 5 and on p. 21-22 of the revised manuscript. The following sentence on p. 22 highlights one key difference in our motif search strategies:

> "Since our study was focused on identifying tissue-specific motifs within 18 expression clusters, it is quite possible that in effect we were searching for less common motifs, that would have been unlikely to have been found in Xie *et al.*'s genome-wide search for over-represented *k*-mers."

Moreover, in identifying tissue-specific motifs, we only considered those motifs whose group specificity scores were greater than that of the geometric mean of the matched randoms. Thus, if a number of the Xie et al. motifs are found associated with genes that are members of a number of different expression clusters, then we would likely miss them by limiting ourselves to tissue-specific motifs. In addition, we limited ourselves to just the regions conserved between human and mouse within 1 kb upstream of transcription start site, while Xie et al. looked within 2 kb upstream of transcription start site. Finally, as described on p. 5 of the revised manuscript, Xie et al. employed a motif conservation score that relies upon the assumption that the position of a motif occurrence has been highly conserved in mammalian genomes, but in reality motifs can be in diffferent positions in the genomes under consideration.

Minor Essential Revision #1: We have performed the suggested GO term over-representation analysis, and modified that section of the Results and Discussion accordingly:

> "It is important to note that motif matches do not necessarily indicate direct regulation by the indicated TF, but rather simply indicate sequence matches beyond the similarity threshold. For example, the NKX2.2 motif was also discovered in the skeletal muscle, adipocyte, and immune gene expression clusters, although in these tissues NKX2.2 was present at an AD value below 200. Given that muscle and adipose tissue have previously been shown to be involved in glucose homeostasis [44,45], we were curious whether NKX2.2 might have an as yet undescribed role in these tissues in regulating genes involved in glucose homeostasis. However, these expression clusters did not have an over-representation of Gene Ontology annotation terms pertaining to glucose homeostasis. Therefore, some other TF that binds a motif similar to the NKX2.2 motif might actually regulate genes in these expression clusters. Indeed, a Pfam search indicates that there are 277 homeobox proteins in the human genome (data not shown). It is quite possible that TFs of the same structural class and with a high degree of sequence similarity in their DNA binding domains might potentially have similar DNA binding site specificities."

We believe that this is actually an important point to make, since it relates to conclusions one might be tempted to make when doing any motif matches or *de novo* motif searches in a genome that contains families of highly similar transcription factors.

Minor Essential Revision #2: Columns 4, 5 and 6 are identical because the original Wasserman analysis only examined the first 1 kb of sequence upstream of the of the transcriptional start site. The sites listed here are only those that were listed in prior studies by Wasserman, and correspond to experimentally verified binding sites found in the literature. Columns 5 and 6 could be removed, but have been included for completeness.

Minor Essential Revision #3: We have added the missing word.

Discretionary Revision #1: We ran each motif finder with its own default parameters. For merging of similar motifs by the previously published program Tree, we used a Pearson correlation coefficient threshold of 0.6 for human motifs and 0.7 for yeast motifs. Our goal in incorporating motif finders into MultiFinder was to get a set of dissimilar search algorithms that would maximize the number of different motif finding algorithms used. The motif finders also had to output a motif profile that could be used to generate position weight matrices. Finally, the motif finders had to be readily available so that others would be able to download and use them locally. We have added a brief explanation of this selection

strategy to the 1st paragraph under the "MultiFinder.pl" heading in the Methods section, on p. 28-29 of the revised manuscript.

The methods of motif finding that were selected were expectation maximization (MEME), Gibbs sampling (AlignACE, BioProspector) and word enumeration (MDscan). A graph traversing search method that output a motif profile and could be installed locally was not to our knowledge available at the time this work started. AlignACE and BioProspector differed in that BioProspector used a 3rd order representation of the background while AlignACE used only base frequency for the background. Of these four motif finders, only MEME and AlignACE were tested in combination by Tompa and colleagues and showed a modest increase in the number of overall motifs found. Other combinations of motif finders were better at increasing the number of motifs found. A future direction of this work will be to include more motif finders in the set of motif finders that are compatible with MultiFinder.

Discretionary Revision #2: We have expanded the description of the motif merging process in the 5th paragraph under the "MultiFinder.pl" heading in the Methods section.

Discretionary Revision #3: We have shortened the abstract by eliminating points (a) through (d) of the Conclusions, as suggested.

Reviewer #2:

Major Compulsory Revision #1: We have shortened the manuscript. The main body of the manuscript text, excluding the title page, references, figure legends and tables, was previously 51 double-spaced pages. The main body of the manuscript text is now 40 double-spaced pages. We have also moved Table 1 to Additional Table 1, which removes an additional 4 double-spaced pages from the manuscript.

Major Compulsory Revision #2: We have uploaded to the *BMC Bioinformatics* website a compressed folder containing SequenceExtractor, MultiFinder, and BlockFilter, along with sample input and output files. Upon publication we will also make this package available freely to academic and non-profit users on our lab website; the academic license will allow us to maintain an email list of users whom we should contact in case of updates to the software. We have also added to the supplementary information for this manuscript an Additional Methods file, which is a PDF file in which we provide instructions for the installation and usage of the MultiFinder, SequenceExtractor and BlockFilter programs.