Department of
Computer Science
College of Sciences
San Diego State University
5500 Campanile Drive
San Diego CA 92182-7720
Tel: 619 - 594 - 1672
Fax: 619 - 594 - 6746
http://www.cs.sdsu.edu

redwards@mail.sdsu.edu
http://edwards.sdsu.edu/researc

SAN DIEGO STATE UNIVERSITY

April 21st, 2014

Dear Editors:

Please find enclosed our revised manuscript entitled “FOCUS: an Alignment-free Model to Identify Organisms in Metagenomes Using Non-negative Least Squares”.

In the review the editor suggested that “The authors are suggested to test their new method in additional datasets. Also the methods and results part should be enhanced.” We have added additional datasets and included a deeper analysis of the SimHC dataset. In addition, we include a point by point response to the reviewers comments.

We thank the editors and reviewers for their assistance with this manuscript and believe that it is now ready for publishing in PeerJ.

Sincerely,

Robert A Edwards.

Dr. Robert A. Edwards
Associate Professor of Computer Science

Reviewer 1

Basic reporting

Firstly, the literature review is insufficient, as the 16S rRNA based organism detection methods are not even mentioned. The author should at least cite one of these 16S rRNA reference-based methods. Also, for efficiency analysis, the authors are better cite some efficient metagenomic analysis methods such as Parallel-Meta (Su, BMC Bioinformatics, 2012).

This paper focuses on random community metagenomes, and not 16S metagenomes. Identifying the species present in 16S sequence is an entirely different problem, and one that we have not tackled here. We have recently written a tool to extract 16S sequences from random community samples to profile metagenomes (GenomePeek (McNair and Edwards),) and have included that tool. Parallel-meta is another homology based tool that uses GPUs and we have included that tool.

Secondly, the organization of the "Results" section is strange: seems that the authors are better describe results on simulated data, and then on real data.

We start with the focus on simulated metagenomes because we know "who the organisms are" and can therefore discuss the implications of the predictions based on knowing truth. On the other hand, a more detailed comparison cannot be done using a real metagenome due to the lack of information about what is really there, but that is what people will use the tools for, and so we demonstrate the use of these tools with real data..

Thirdly, the annotations (such as "R", and also "m") for formula 1 is not clear.

This is standard mathematical notation but we have attempted to clarify the sentence so it now reads " Given a matrix $A \epsilon \mathbb{R}^{mxn}$ and $b \epsilon \mathbb{R}^{m}$, where $m \geq n$, find a non-negative vector $x \epsilon \mathbb{R}^{n}$ to minimize the function (1)". We believe that it is clearer now.

Finally, there are quite some wording problem for this draft, such as "what those organisms are doing and who they are" (logically not the right order). Another example: there is a missing "." at the end of sub-section "Jackknife resampling of the data".

We apologize for the mistake. We thank the reviewer for the observation, and we have edited the problems

Experimental design

The description of the simulate datasets is blur: how many organisms simulated, and what relative proportions?

This information asked by the reviewer was previously shared in the file "Supplementary_Table_2.xlsx". It can also be found at http://edwards.sdsu.edu/~gueiros/focus/suppl.

Also, how "using different subset" is implemented, and how to avoid bias?These are completely unclear from the experiment design part of this draft.

The different subsets we selected randomly. We have edited "using different subsets" to "using random subsets", and we believe that it is more clear now.

Validity of the findings

The results part lack complete quantitative comparison of the accuracy of various methods, and thus making it difficult to judge the validity of the finding.

We disagree with this reviewer. We have gone to extraordinary measures to quantify the accuracy of each of the methods, starting with synthetic metagenomes discussed above. The validity of the finding was showed using simulated data because we know the proportion of each organism which are present in the data sets comparing FOCUS to RAIphy, GASiC (added) and PhymmBL. Unfortunately we had to remove from the comparison MetaPhlAn, MG-RAST, and Taxy because it is not possible to change their databases, so it would not be a fair comparison.

Reviewer 2

Basic reporting

The methods section is very short and does not seem to cover sufficiently in depth all steps of the methodology. For example, I suspect that several parameters need to be tuned, but there is no mention of them.

We thank the reviewer comment, and we have improved the methods section.

Experimental design

"No Comments"

Validity of the findings

The presented approach to taxonomically profile metagenomes seems interesting and fast. However, it has been applied on very few metagenomes (2 synthetic and 1 real) and it thus fail to convince that FOCUS is a valid alternative with respect to existing approaches. Given that the tool is very fast to run, it should be easy for the authors to

run it on several real and large metagenomes (for example they can use the HMP dataset for which the profiling with other methods - including 16S sequencing - are already publicly available).

Predicting organisms present in metagenomes is a challenging problem, and we agree that FOCUS is fast. For this paper, we have compared FOCUS with other computational tools that perform the same approach. We have considered comparing random community metagenomes and 16S libraries from the same samples as a way to test the accuracy of our tools, but at the moment the variation between sequencing approaches appears to outweigh the variation between computational approaches. In other words, the predictions of the organisms present in the sample are more variable if 16S sequences are compared to random community genomes than if different tools are used. The biases introduced by the different amplification and sequencing approaches skew the data too much.

We don't agree, therefore, that comparing a large dataset like those sequenced by the HMP, to the 16S sequences from the same sample, is a meaningful way of assessing accuracy or specificity of computational approaches, and every time we have tried to do this we end up being mired in a debate about whether 16S sequencing or random community sequencing is the most appropriate way to idenitfy the species present in the sample.

For this paper, we chose to focus on just random community sequencing, and to compare the speed and accuracy of our tools with the best-in-the-field approaches. These computations are demanding, we have consumed thousands of processor hours "just" to generate the comparison of the tools that we have. As we note in the text, some tools take weeks to process a single metagenome.

In addition to a large set of real (and possibly synthetic) metagenomes there are several other points that should be addressed to really validate the method:

- for both synthetic metagenomes accuracies at all taxonomic levels should be presented

We have added in the supplementary figures the comparison between FOCUS and the other tools for the synthetic metagenomes in following taxonomic levels: phylum; order; family, and genus.

- a quantitative value summarizing the accuracy of the tested tools on the synthetic datasets should be given (squared error, correlation with the real values...)

This paper presents FOCUS, a new tool for metagenome analysis, and we have presented the accuracy of FOCUS on the synthetic datasets. The paper is not a comparison of the quality of other tools – they are merely included to demonstrate the superiority of

FOCUS to the task at hand. The computations that we have performed took a lot of computational resources, and to present this data for all of the tools is not only completely outside the scope of this manuscript it would take an extremely long time to compute.

- MetaPhlAn estimates the relative abundance of organisms, the other tools estimate the fraction of reads coming from each organism. They thus differ when the size of the genomes in the metagenome is not constant. Veillonella have very short genomes (~2MB) and thus will have higher genome relative abundance than reads relative abundance. This should be made clear when presenting and commenting Figure 1.

Thank you for this insightful comment. We have added this into the paper.

- The authors trained the system on 2766 genomes. How fast is the training process? Is it scalable to the quickly increasing number of available reference genomes (not at least 10k)?

The training process basically involves identifying all the *k*-mers in the genomes. There is some excellent computer science research pushing the speed and breaking the limitations in *k*-mer counting, and we are leveraging the approaches that other teams develop. Currently FOCUS uses Jellyfish (Marçais and Kingsford 2011), an ultra-fast tool to compute k-mers, to create the training dataset. Because of its innovative non-locking hash design Jellyfish runs in parallel threads and takes a few seconds to run on thousands of genomes, and will easily scale to 10k genomes as computational cores increase. However, we look forward to the next generation of fast and efficient algorithms to count *k*-mers as they will ensure that FOCUS will continue to scale in the future.

Reviewer 3

Basic reporting

It is an important and valuable problem in metagenomics to identify the organisms present in a microbial community and estimate its abundance from unannotated sequencing reads. There have been a lot of methods on it. The authors argues that current methods do not scale well with increasing data volumes and they introduce a composition based approach using nonnegative least squares (NNLS) to estimate the focal organisms present in metagenomic samples and estimate their abundances. Generally, the idea is sound and brief tests have been done to demonstrate its effectiveness.

However, the paper is a bit too sketchy. Many details were not been covered by this manuscript. If possible, I would like to suggest the authors add more analysis and descriptions on their methods and results.

We appreciate the reviewer comments. We believe that the answers given to the previous reviewers covers the comments above.

Moreover, another recent work has adopted a similar mathematical framework for the metagenomic abundance estimation. The authors should clarity the similarity and difference carefully.

Lindner MS, Renard BY. Metagenomic abundance estimation and diagnostic testing on species level. Nucleic Acids Res. 2013 Jan 7;41(1):e10. doi: 10.1093/nar/gks803.

We thank the reviewer for point out GASiC (Lindner and Renard 2013). We did not know about this tool, but we edited the paper and include GASiC in the list of tools to which we compare FOCUS. We downloaded, installed, and tested GASiC, however, we did not compare to the other metagenomes because GASiC generates a large amount of output, and we did not have enough storage to save it.

Experimental design

The experimental design is reasonable, but the description is too sketchy.

Validity of the findings

The test is ok, but more analysis is needed.

We thank the reviewer comment, and the complete manuscript was improved with more details in the methods and, results, and discussions. For example, we have added one more real metagenome in the analysis, we included one more tool in the comparison (GASiC), and we present the results for the comparisons in 4 taxonomic levels.

References

Lindner, M. S., and B. Y. Renard, 2013: Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.*, **41**, e10, doi:10.1093/nar/gks803.

Marçais, G., and C. Kingsford, 2011: A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770, doi:10.1093/bioinformatics/btr011.

McNair, K., and R. Edwards, GenomePeek – A tool for prokaryotic genome and metagenome analysis.

Su, X., J. Xu, and K. Ning, 2012: Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Syst. Biol.*, **6**, S16, doi:10.1186/1752-0509-6-S1-S16.