Reviewer’s report

Title: RNA-sequence data normalization through in silico prediction of reference genes: the bacterial response to DNA damage as case study

Version: 0  Date: 26 Jun 2017

Reviewer: Tom Hampton

Reviewer’s report:

This is a thoughtful and well written manuscript on a topic of vital interest to the field. You present a clear case for your algorithm as an improvement over current approaches by questioning the core assumption of existing approaches, then showing that your mechanism outperforms existing mechanisms. I have two suggestions that I believe would improve this manuscript.

First, though I understand why you have chosen to state that most existing normalization methods assume "the majority of genes is invariant". I would advocate for a more precise articulation of that assumption, because it is central to what makes your method better. I believe it is more accurate to say that the other normalization approaches assume that no systematic biases skew differential expression toward induction or repression. The trimmed mean approach should work fine, for example, even if all genes are "differentially expressed" as long as mean amount of expression is always the same. And that could be the case. At the end of the day, maybe cells always manufacture the same amount of message no matter what is going on. That, I think, is the underlying assumption of the other methods, not that a small fraction of genes are differentially expressed, though, if that is true, then it also follows that mean expression should be constant. I believe the main reason few genes are "differentially expressed" in RNA-seq experiments has more to do with the low power of these experiments, than biology, but that is another discussion.

Second, you are still making assumptions of some sort, and i'd like some discussion of why your assumption ("housekeeping genes exist") is more biologically reasonable than the assumptions that your are replacing. I think yours are more reasonable, but not always. Letting readers know what your theoretical limitations are would make your new method seem more reliable and well thought out, in my opinion. Suppose, for example, one ran Moose on two experiments, and, as you have implied, the invariant genes were highly dependent on the experiment, to the point were the first experiment identified a completely disjoint set of invariant genes from the second. Now, imagine an experimental design in which these two sets of experimental conditions were simultaneously tested. I believe Moose would select yet another set of relatively invariant genes based on the complete data set, and normalizing against this new set would be predicted to introduce artifacts such as spurious gene expression. As one adds more and more conditions to this theoretical experiment, Moose will continue to degrade in its performance, unless universal housekeeping genes exist, and we know that they do not. I think comparing differential gene expression between different bacterial taxa would be a good place to use TMM rather than
Moose, for example. Is Moose then limited to the number of experimental conditions tested or better used for certain kinds of comparisons?

In short, a bit about how your mechanism is more consistent with what we know about biology, and the situations in which your mechanism might do poorly, would add to this manuscript.

**Level of interest**
Please indicate how interesting you found the manuscript:

An exceptional article

**Quality of written English**
Please indicate the quality of language in the manuscript:

Acceptable

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments
which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal

Do you want to get recognition for reviewing this manuscript? Add a record of this review to Publons to track and showcase your reviewing expertise across the world’s journals. Signing up is quick, easy and free!

Yes