Improving transcriptome assembly through error correction of high-throughput sequence reads

Matthew D MacManes[1*] and Michael B. Eisen[1,2]

1 UC Berkeley. California Institute of Quantitative Biology, Berkeley, CA, USA
2 Howard Hughes Medical Institute

∗ Corresponding author: macmanes@gmail.com, Twitter: @PeroMHC

Response to Reviewers Comments

1. The use of real versus fake data.

   This is an excellent suggestion. In response, I have done a parallel analysis using a random 30M PE read subset of the Trinity mouse dataset. I choose to use 30M reads rather than the full 50M read dataset to maximize comparability, though moderate differences in coverage still exist.

   With regards to analyses, the paper continues to focus mostly on the simulated data, though I now state that the results of the simulation study have been corroborated by analysis of real data. That the finding are robust to dataset should be interpreted as evidence of general utility of these methods.

2. Use paired end reads.

   I have used 30 million PE reads in this version of the paper. The results are unchanged, though the magnitude of the improvement is reduced. I believe this to be due to the general improvement of the assembly quality secondary to use of PE reads.

3. Deposit the data!.

   The data are now publicly available. The reads are available on my personal Dropbox (https://www.dropbox.com/s/rkl0ihqom28smb2/empiric.reads.tar.gz and https://www.dropbox.com/s/mp8fu0tijox69ki/simulated.reads.tar.gz). They will be moved to Dryad upon acceptance. The assemblies are available at http://dx.doi.org/10.6084/m9.figshare.725715. The code has been moved to a Github Gist https://gist.github.com/macmanes

4. Justify the use of error correction techniques

   The specific error correction techniques were chosen in an attempt to cover the breadth of computational techniques, as well as to include tools commonly used, and recently benchmarked. I make this explicit in the manuscript. Of note, this version removes the program ECHO from the analysis. Given that it would have taken at least 4 weeks to run with the new PE data and empirically derived dataset, I did not have the time, given the constraints on resubmission.

5. Use more than just Illumina data.

   While I agree that expanding the paper to include different types of sequence data (454 and IonTorrent) would be interesting, it is beyond the scope of this current paper, which focuses on the current (and likely future, given the number of deployed machines) most popular sequencing technology. In addition, that each technology has a unique error model, understanding the results would not be easy, given simulated data not accessible.

6. What's going on with splice variants?

   Good question...

7. Variable coverage and miscorrections?

8. The use of real versus fake data.