Reviewer Report

Title: Transcriptome annotation in the cloud: complexity, best practices and cost.

Version: Original Submission **Date:** 9/14/2020

Reviewer name: Lucas Carey

Reviewer Comments to Author:

Reviewer's report
Title: Transcriptome annotation in the cloud: complexity, best practices and cost. transcriptome data
Reviewers: Qiao Xuanyuan and Lucas B. Carey
Reviewer's report:
The authors provide a perspective on transcriptome annotation in the cloud by presenting a comparative study of the two main public cloud providers, aiming to help the reader determine the proper cloud platform and its utilization in their research. The main strength of this paper is that it addresses a question that little has been studied beforeâ€”the cloud cost estimates and implementation best practices. This is useful not only for labs doing transcriptome annotation, but, because all code is provided and very well commented, it might be of use for labs dealing with big data &amp; distributed computation problems in general.
The manuscript is well written, and I have only a few minor concerns on presentation and the information that is provided.
1. The tested transcriptome data need to be clarified. I read the Data partitioning code for the creation of 20 FASTA files of the different query sizes (Fig 3). The variation in processing time among these files is consistent across clusters, and is therefore presumably due to differences in query sets in each file. This is surprising, as 10,000 is a large number. Are the differences because the sequence records were created sequentially from the input fasta file, with some genes having large numbers of hits? Would subsetting transcripts randomly result in more uniform processing times? There is almost two-fold variation in processing time between query files.
2. Please include a figure showing the distribution of transcript lengths for Opuntia streptacantha, and write that it is the prickly pear cactus. Presumably timing depends on the transcript lengths and on the number of BLAST hits.
3. It is difficult to draw the conclusions from the way the data are plotted in Figure 4. The author concluded that
"Reducing the number of transcripts per input file will reduce the total running time but will also increase the cost of the project as more instances will be in used."
In addition to the raw data boxplot, it is better to show how the time and cost scale with query size, or with the number of instances. (The number of instances equals query size divided by total transcripts).
Controlling the total transcripts and making instances as variable may be helpful in balancing the interpretation and data.
The plot below shows the relationship between the number of CPUs and time, and query size and time. (data collected using https://automeris.io/WebPlotDigitizer/). It also provides direct-viewing evidence to

support the author's conclusion "AWS is more suitable for large data analysis groups to establish a set of queues and compute environments for multiple pipelines." Unlike the boxplot in the current manuscript, this figure also shows the differences in scaling between 16, 32 &amp; 64 vCPU nodes. Please add graphs showing query file sizes vs time (as below) and query file sizes vs cost. As well as cost vs time. These are the important take-home messages from the manuscript, but it is difficult to extract this information from the current figures.

Optional (not necessary) suggestions:
The figures could be improved to give a clearer visualization of the data. For Figure 3a, will adding a secondary Y-axis regarding money and draw a line plot be better to show the relationship between total time and cost? For Figure 3b&amp;c&amp;d, drawing a component bar chart could allow the readers to compare the job-dependent time among various configurations and demonstrates the proportionality at the same time.

Level of Interest

Please indicate how interesting you found the manuscript: Choose an item.

Quality of Written English

Please indicate the quality of language in the manuscript: Choose an item.

Declaration of Competing Interests

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

- 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?
- 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?
- Do you hold or are you currently applying for any patents relating to the content of the manuscript?
- 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?
- Do you have any other financial competing interests?
- 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.

Choose an item.

To further support our reviewers, we have joined with Publons, where you can gain additional credit to further highlight your hard work (see: https://publons.com/journal/530/gigascience). On publication of this paper, your review will be automatically added to Publons, you can then choose whether or not to claim your Publons credit. I understand this statement.

Yes Choose an item.