Reviewer Report

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

Version: Original Submission **Date:** 7/27/2020

Reviewer name: Tazro Ohta

Reviewer Comments to Author:

In this manuscript, the authors describe the usability and the performance of public cloud computing services and their batch job execution services for the transcriptome annotation pipeline. The authors also compare the services of two major public cloud vendors, Google Cloud Platform (GCP), and Amazon Web Service (AWS). The results show that the two cloud providers can provide a similar experience on its operation, workflow execution time, and payment for execution. The performed experiments follow the modern best practice of data analysis using tools and frameworks for better reproducibility, including Docker container, the Common Workflow Language (CWL), and Jupyter Notebook. This article can be a good example of a reproducible study with open data and open-source software.
This report is very significant with the practical statistics, which can be a helpful reference for all the cloud use cases in biomedical data analysis. The title states that this study focuses on the transcriptome annotation, but the output provides insight for all the cloud use cases. I suggest the authors change the title because the current one looks the best practice valid only for the transcriptome annotation in the wide variety of genomic data analysis.
Below are minor comments for the manuscript.
Page 3, the last paragraph:
The authors claim as "little has been published describing cloud costs and implementation best practices". However, there is a study implemented software to monitor the runtime metrics of a given workflow, and support the cost estimation for the executions on the cloud (https://doi.org/10.1093/gigascience/giz052). This article describes the cost for the normal EC2 instance and does not mention the batch execution services, yet it provides additional information to the readers. Please consider introducing this in the background section as a related study. (Disclaimer: I am the first author of this article)
Page 12, Best practices
The authors describe here that they recommend CWL as a workflow language. I think the authors should introduce the reason they chose CWL at the Method section where they first mention CWL (Page 4, Transcriptome Annotation Workflow). The authors also should provide a more practical reason to choose CWL because CWL is not only one framework that has portability and scalability. For example, having multiple workflow runners or its syntax easily parsed and connected to the resources such as runtime metrics can be a reason to choose CWL in this use case.
Workflow dependencies
Both AWS batch and Google LifeScience allow users to specify only one container per one batch job. This means the user needs to install the workflow runner (in this article cwltool) and the tools to process the data (e.g., BLAST) in a single container. However, the current best practice for bioinformatics is to

separate the containers for each tool. Thus, in most cases, users cannot reuse the containers they use for a normal computing environment to the cloud batch system. The authors need to mention the limitation that one needs to create a container for cloud batch systems. Another possible solution would be to run sibling containers from the main batch job container, though I do not know if it is feasible with AWS and GCP.

Software and framework related to the cloud batch services

Many workflow languages and runners are supporting AWS batch and GCP. For example, Nextflow (https://www.nextflow.io/docs/latest/awscloud.html), Cromwell (https://cromwell.readthedocs.io/en/stable/backends/Google/), or Snakemake (https://snakemake.readthedocs.io/en/stable/executing/cloud.html#executing-a-snakemake-workflow-via-tibanna-on-amazon-web-services) can run the workflows via AWS batch or GCP. Task Execution Service (TES) of the Global Alliance for Genomics and Health (GA4GH) Cloud working group is also a framework to utilize the cloud batch system (https://github.com/ga4gh/task-execution-schemas). Some tools that run workflows on the cloud using the ETL framework (https://doi.org/10.21105/joss.01069) or cloud batch services (https://github.com/DataBiosphere/dsub) are also available. These may not be directly relevant to this study, but it would be helpful for readers to understand how the cloud batch services are being used by the researchers.

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.