Dear Editor,

We would like to present the paper titled “Scalable Clustering of Genotype Information using MapReduce” for peer-review to incob2014 as well as BMC Genomics (Methodology).

Full contact details of the corresponding author:

Denis Bauer

Postal address: Locked Bag 17 North Ryde NSW 1670

Street address: Level 4 Building 53, Riverside Life Sciences Centre (RLSC), Riverside Corporate Park, 11 Julius Avenue, North Ryde NSW 2113

Email: [denis.bauer@csiro.au](mailto:denis.bauer@csiro.au)

Telephone +61 2 9325 3174

Fax +61 2 9325 3226

Declaration of the corresponding author:

I, the corresponding author, certify that the work submitted is original and has not been submitted elsewhere for publication. Furthermore, if accepted, I agree to pay in full, the stipulated article processing charges as well as reformat the accepted version of the manuscript in accordance with the journal requirements and upload the final version of the manuscript to EasyChair by the stipulated deadline.

Copyright statement:

We, the Authors collectively and individually declare that to their best ability, they have ensured that the submission made herein, including the main paper, supplementary data, deposited data, database entries, software code, does not contain any plagiarized material, content or ideas, and that all necessary attributions have been appropriately made and all copyright permissions obtained, cited and acknowledged.

Additional questions:

1. What scientific /methodology problem has been addressed?

Big Data analysis: genotype-based clustering using MapReduce Machine learning framework

1. Which published work of the past 12 months is related to the scientific/methodology problem addressed in the submitted paper.

Massie, M., Nothaft, F., Hartl, C., Kozanitis, C., Schumacher, A., Joseph, A.D., Patterson, D.A.: Adam: Genomics formats and processing patterns for cloud scale computing. Technical Report UCB/EECS-2013-207, EECS Department, University of California, Berkeley (Dec 2013). http://www.eecs.berkeley.edu/Pubs/ TechRpts/2013/EECS- 2013- 207.html

1. Which reported findings are crucial to solve the problem?

MapReduce frameworks are designed to scale to large volumes of data, however the processing algorithm needs to be carefully designed to scale with samples not features.

1. Are these findings novel?

In the context of Bioinformatics yes

1. What impact has the reported finding in the subject area as well as among wider  community of biological and biomedical researchers?

MapReduce is a “hot topic”, however there is not much actual research findings presented using these methodologies in bioinformatics. Hence showcasing the performance and discussing limitations and pitfalls is of great interest to the wider community.

Sincerely yours

Denis Bauer

(on behalf of the co-authors)