**Response to Reviewers**

We would like to thank the editor and reviewers for their comments.

**VariantSpark: Population Scale Clustering of Genotype Information**

Aidan R O’Brien, Neil Saunders, Fabian A. Buske, Denis C Bauer

The reviewers’ comments and suggestions are boxed in gray and itemized below, followed by our responses and a description of the consequent revisions to the paper. Text from the manuscript or supplemental materials is italic and changes are “quoted in red”. Please note that in our responses we have used the abbreviations ‘pg’ and ‘para’ for page and paragraph, respectively.

**Reviewer: 1**

Figure 2 and 3 do not add much to the paper as they are poorly labeled and not easily interpretable.

We updated the figures captions to increase the interpretability of the figures

Figure 2

Schematic overview of VariantSpark. The image shows the flow from the input VCF file to the machine learning library and onto the visualization. It highlights the differences between the Hadoop and Spark implementations for converting data in VCF format to a data structure readable by Mahout and MLlib, respectively.

Figure3

Visualisation of VariantSpark predicted clusters. The figure shows the four clusters predicted for the 1000 Genomes data. Individuals from the super-populations AFR, AMR and EAS are accurately grouped into distinct clusters. The fourth cluster contains predominantly EUR + AMR individuals potentially accurately reflecting migrational backgrounds.

We also updated Figure 1 to 1) added binary-conversion for ADAM and 2) corrected labelling errors.

A better description of new features (in addition to a reasonable speed up) need to be discussed

We added the following sentence to the conclusion to address this point:

Utilising MLlib as well as Spark.ML will enable supervised machine learning applications to e.g. identify variants that jointly interact with phenotypes as well as include electronic health record in addition to the genomic feature vector to e.g. capture medical history as well as predispositions for diagnosis and treatment decisions.

**Reviewer 2:**

I wasn't able to find the reference implementations in R and Python.

We update the method section as follows

We also include the R and Python source file as supplementary material.

JADA

As a baseline for performance comparisons it would appear more natural (to me), to use a C++ implementation, which would actually be very easy to do using the bcftools API to read VCF/BCF and mlpack for clustering.

We investigated the suggested option and do not think it is appropriate as the kmeans clustering in mlpack is not multithreaded and hence performing worse than R and python whose kmeans approach is multithreaded. However we implemented a solution in Matlab which uses c++ libraries.

Scalability is an issue and will certainly become ever more important, especially in the light of projects like Genomics England and the Precision Medicine Initiative. So exploring the use concepts like Spark is certainly welcome. However, VariantSpark caters to only one very specific use case, rather than providing a platform to address many problems one faces. In order for it to become widely adopted, it would need to have a wider scope (in my opinion).

We agree that the reviewed version of the paper focuses on only one application we hence elaborated on the options of using the other machine learning algorithms (See Reviewer1 comment 2)

R and Python do have specific packages to parse VCFs (Bioconductor/VariantAnnotation and pysam). I wonder whether using these would result in better performance.

We tried the VariantAnnotation package and found it to actually be faster than R’s built in library for reading delimited files. Therefore, we will use this package in our comparison. We also optimized our custom function that converts VCF variant strings to integers

(I.e. “1|0:99” to 1).

To elaborate, we now parse VCF files to matrix using readGT from the VariantAnnotation package.

vcf <- readGT("input.vcf")

Optimizations to our function include using fixed character matching instead of regex.

strsplit(a, ":"**, fixed = TRUE**)

As before, we apply our function to the matrix using vapply, with a pre-defined return type. We have compared it to other available apply methods, and find it to be the fastest in this case.

We include the R source file as a supplementary document.

To demonstrate scalability to today's data sets, using the present 1000 Genomes release (phase 3) would be appropriate.

We agree with the reviewer and have included the phase 3 data in our paper.

Page 5, para 3

To further demonstrate the scalability of VariantSpark, we also cluster the 1000 Genomes Project phase 3 data, which contains 3000 individuals from 5 superpop- ulations and as a result has over 80 Million variants. The uncompressed size of the phase 3 files is 770GB compared to the 161GB of the phase 1 dataset. Vari- antSpark successfully completes the clustering in just 27 hours (see Table 2) with an ARI of XX