Assessing Next-Generation Sequence variant detection methods on African populations.

Noëlle van Biljon
University of Cape Town, Cape Town, South Africa.

I. Introduction:

The vast majority of next generation sequencing and downstream analysis of sequence data has been performed using European genetic material only. African genetic data has been shown to have greater gene diversity and the largest number of total and unique alleles compared to European and Asian genetic data(1). These differences in genetic characteristics affect the performance of downstream bioinformatic analysis tools. Thus, when choosing the appropriate software, ancestry must be considered. Here we aim to identify the most appropriate variant calling program for analysis of African genetic data.
Variant Calling involves the identification of polymorphisms in a given alignment file. Polymorphisms are differences between a given sample and a reference sample, that are considered to occur commonly within a population(2). Because African populations are less well studied in terms of genetics, it is difficult to identify which differences are common in these populations - and hence are considered SNPs (single nucleotide polymorphisms). Current variant calling tools use probabilistic approaches to identify polymorphisms(3). These tools produce a false positive (FP) and a false negative (FN) identification rate as they are not 100% accurate(2). Using the FP and FN rates, we can compare the performance of the various variant calling tools and accordingly identify the most accurate program.

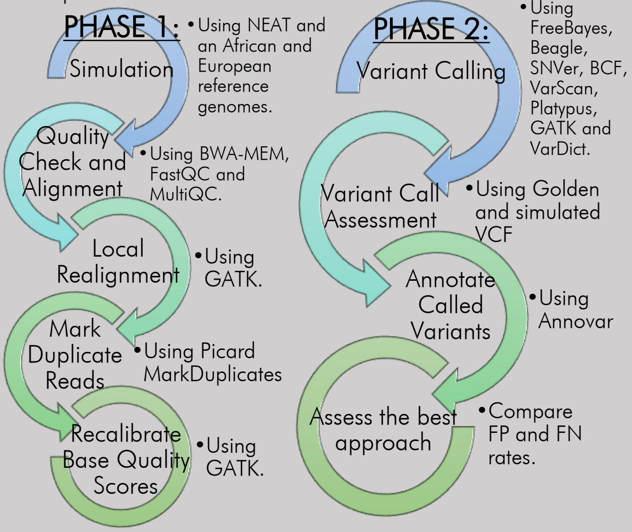
II. Hypothesis: The choice of optimal variant calling tool is affected by ancestry.

III. Aims:

- To identify the best bioinformatic tool for variant calling when working with African genomic data.
- To identify the potential false positive SNPs arising from variant calling tools and to improve the means of SNP identification when using African genetic data.

IV. Methods:

Fig 1: Overview of Simulation and Variant Calling Pipelines



PHASE 1:

Simulation
- Using NEAT and an African and European reference genomes.

Quality Check and Alignment
- Using BWA-MEM, FastQC and MultiQC.

Local Realignment
- Using GATK.

Mark Duplicate Reads
- Using Picard MarkDuplicates

Recalibrate Base Quality Scores
- Using GATK.

PHASE 2:

Variant Calling
- Using FreeBayes, Beagle, SNVer, BCF, VarScan, Platypus, GATK and VarDict.

Variant Call Assessment
- Using Golden and simulated VCF

Annotate Called Variants
- Using Annovar

Assess the best approach
- Compare FP and FN rates.

V. Results:

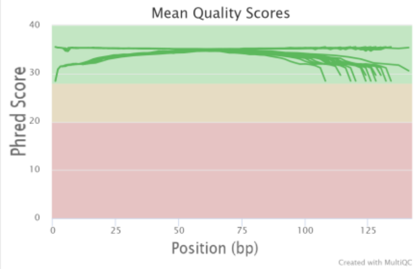


Fig 2: Mean read quality score for 15 simulated African genome sequences.

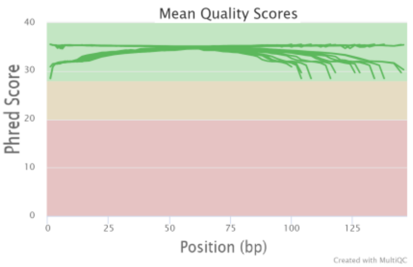


Fig 3: Mean read quality score for 15 simulated European genome sequences.

Figure 2 and 3 show the mean Phred quality scores from simulated sequence runs for 15 African and 15 European genomes. Here it is evident all sequenced reads are of good quality as the average Phred scores for each simulated genome remains above 30. Clearly, African and European genomes have very similar quality scores. For both European and African simulated reads the FastQC reports showed that all elements of read quality are good, the only concern is a low GC content, on average, for both European and African reads.

VI. Conclusion:

We can deduct - from figure 2 and 3 - that the simulation procedure does not favour European genomic data over African data. Hence, the downstream analysis using this simulated data should not be biased by the simulation procedure.
Following the variant calling procedure, we expect to see a general increased FP and FN rate in the variants called from African genomic data compared to that from European. Given this, we also expect to see one variant calling tool with the lowest overall FP and FN rate for Africa data. This will be the optimal choice for variant calling when working with African genomic data.
This decision of optimal variant calling tool will allow future research using African genomic data to receive optimal results. This will also serve as a guide to future researchers working with African genomic data.

VII. Acknowledgements:

I am particularly grateful for the assistance given by my supervisors Prof Nicola Mulder and Dr Emile Chimusa. I would also like to thank the developers of the various software programs used allowing me to complete this research.

VIII. References:

1. Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, et al. The Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data. Am J Hum Genet. 2000 Mar 1;66(3):979–88.
2. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011 Jun;12(6):443–51.
3. Altmann A, Weber P, Bader D, Preuss M, Binder EB, Müller-Myhsok B. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet. 2012 Oct;131(10):1541–54.