ASSESSING NEXT-GENERATION SEQUENCE VARIANT DETECTION METHODS ON AFRICAN POPULATIONS.

NOËLLE VAN BILJON

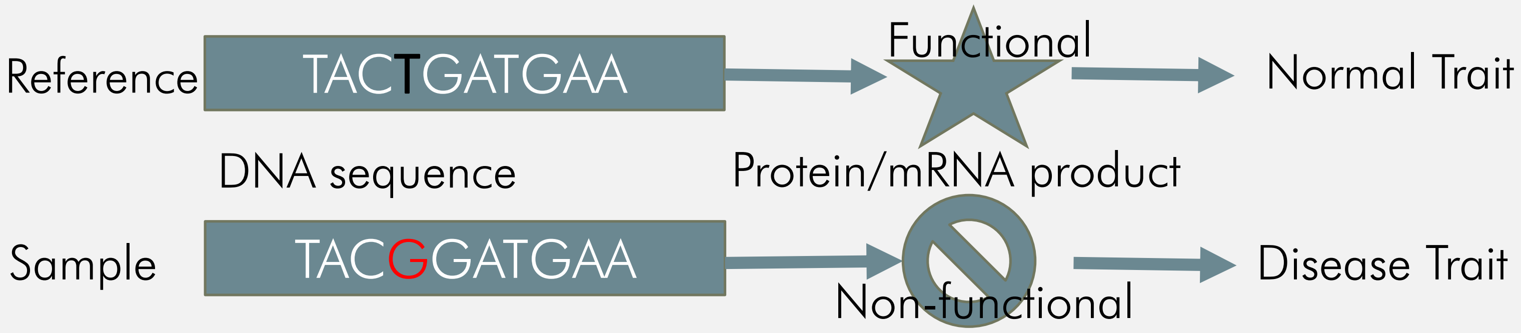
SUPERVISORS: PROF NICOLA MULDER & DR EMILE CHIMUSA

UNIVERSITY OF CAPE TOWN
COMPUTATIONAL BIOLOGY DIVISION

VARIANT CALLING (VC)

- Variants: Variants are differences between a given sample and a reference sample, that are considered to occur commonly within a population (1).

- Variant Calling: The process whereby variants (SNPs, Indels, SV, CNV) are identified in a genome.



- Why VC?: Deviation from the reference allele may affect cellular activities.
  - Find variants associated with a specific trait, phenotype or condition. (2)

- Sensitivity(FN) and specificity(FP) equally important (3).

VC TOOLS

- VC Methods:
  - Heuristic -filtering and quality cut off values to identify an initial set of genotypes from which SNPs are inferred (1, 4, 5)
  - Statistical – based on likelihood of observing a specific outcome given all prior known information (1)
- Types of variants: Germline, Somatic, Structural Variants and Copy Number Variants.
- Different VC tools – we focus on germline tools.

THE PROBLEM

- African genomic data is less studied relative to European (7,8).
- African data has different characteristics:
  - Greater genetic diversity and the largest number of total and unique alleles (9).
- Data characteristics affect bioinformatic tool performance.
- Inappropriate reference increases FP and FN.

VC -TOOLS

Table 1: VC tools with defining methodological feature.

GATK Haplotype Caller	**Uses a Bayesian approach and local realignment for variant calling.**
BCFtool	Performs multiallelic variant calling
Samtools	Performs SNP and genotype calling as well as computation of genotype likelihoods.
VarScan2	Uses a heuristic/ statistical approach to identify somatic and germline variants.
VarDict	Uses local realignment to improve indel calling.
SNVer	Statistical framework to find rare and common variants. Can be used on pooled or individual data.
Lofreq	Uses a Poisson- binomial distribution model to identify variants.
Platypus	Haplotype based variant caller that uses a Bayesian approach.
Freebayes	Identifies MNVs and complex variant events.

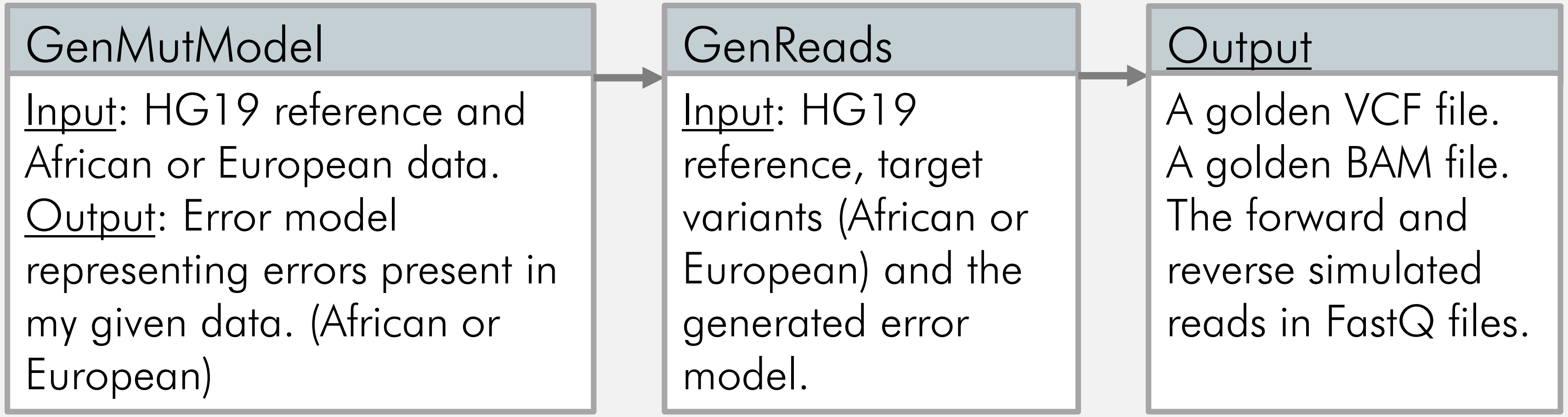
- VC tools have different methods = different results (1,3,6,10,11).

- Have been assessed previously.

AIMS

- Aim: Assess performance of VC tools and identify best VC tool for African data.
- Objectives:
  - Simulate data that represents African and European data,
  - Align simulated data and perform QC,
  - VC using simulated data and nine VC tools,
  - Identify %FP and %FN for each tool,
  - Assess tool accuracy using African and European data.

METHODS - SIMULATION



GenMutModel

Input: HG19 reference and African or European data.
Output: Error model representing errors present in my given data. (African or European)

GenReads

Input: HG19 reference, target variants (African or European) and the generated error model.

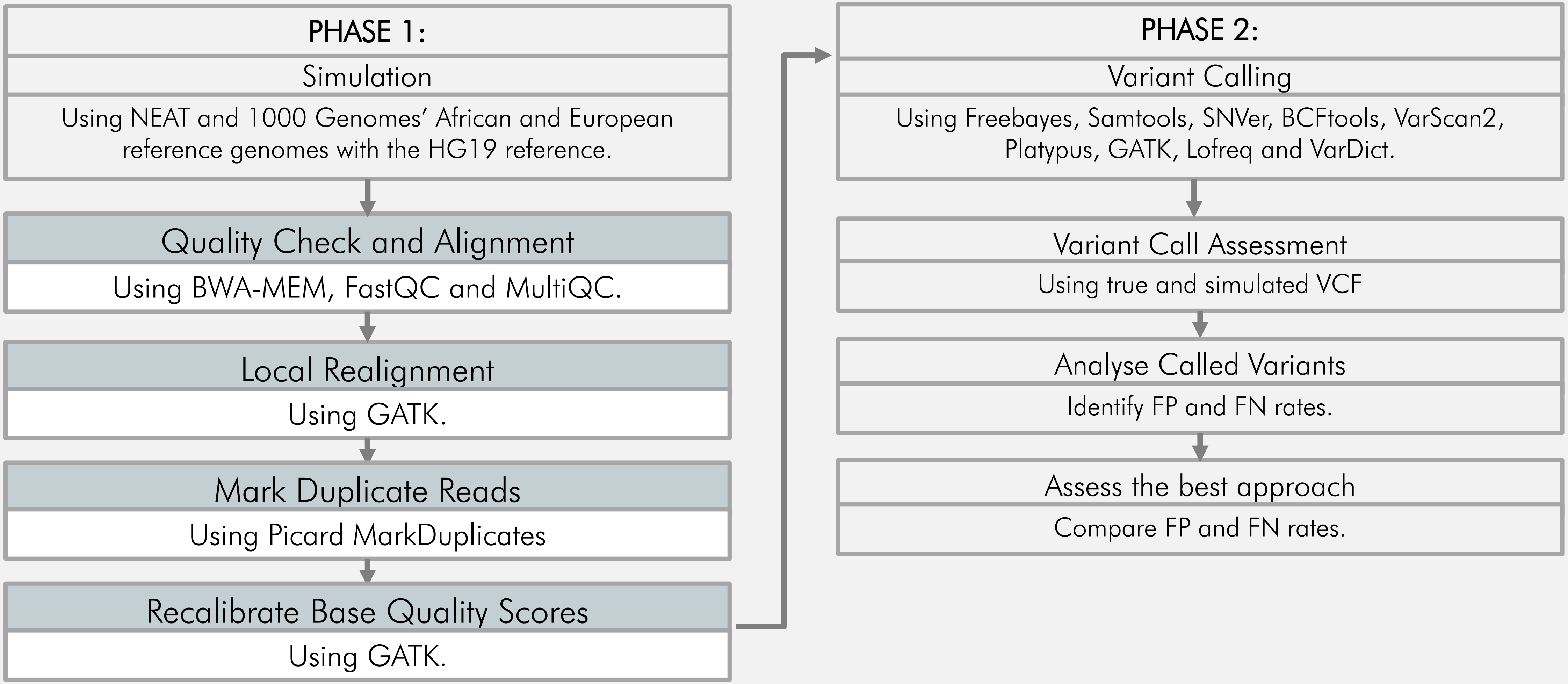
Output

A golden VCF file.
A golden BAM file.
The forward and reverse simulated reads in FastQ files.

Figure 1: Overview of the NEAT simulation procedure.

- Gives empirical dataset
- First: Mutation rate 0
- Second: Mutation rate 0.1 (12)

METHODS – QC AND ALIGNMENT



PHASE 1:

Simulation

Using NEAT and 1000 Genomes' African and European reference genomes with the HG19 reference.

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, Samtools, SNVer, BCFtools, VarScan2, Platypus, GATK, Lofreq and VarDict.

Variant Call Assessment

Using true and simulated VCF

Analyse Called Variants

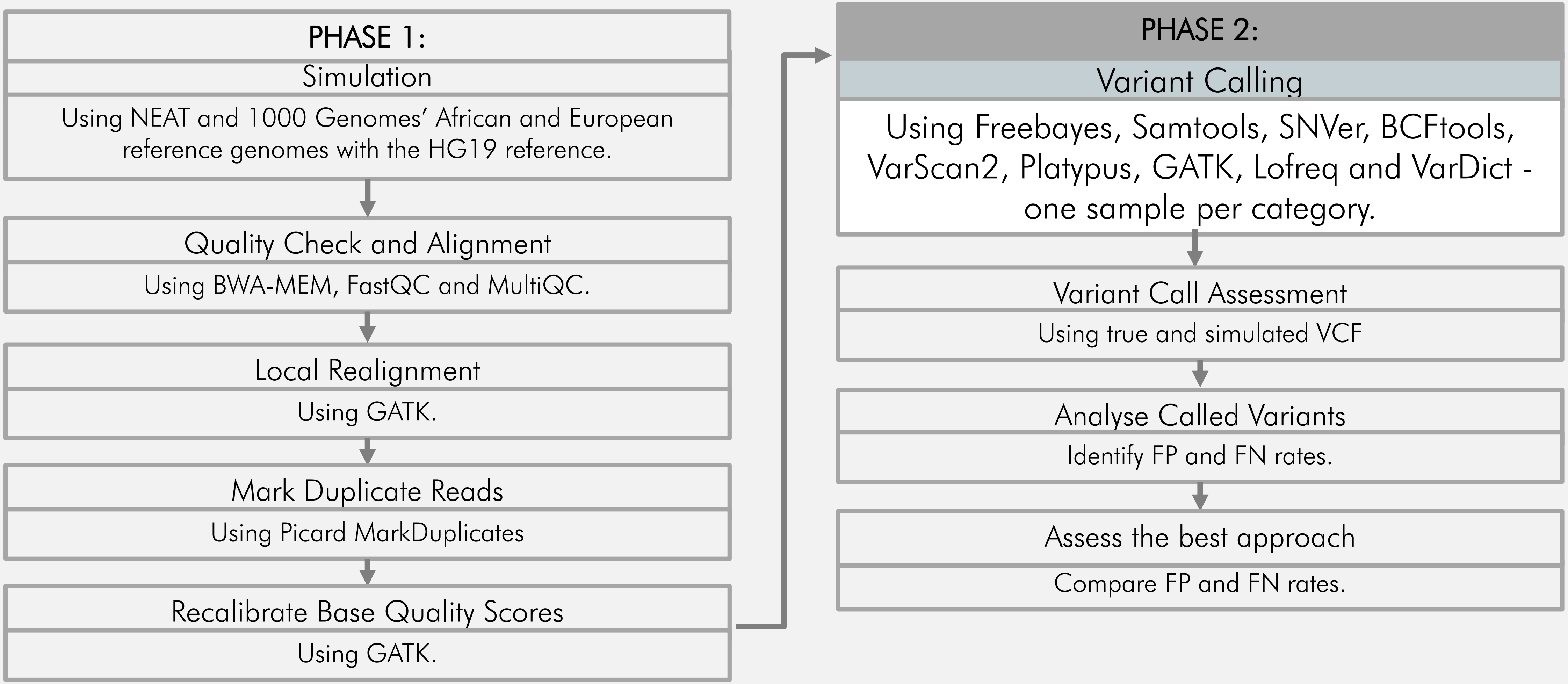
Identify FP and FN rates.

Assess the best approach

Compare FP and FN rates.

Figure 2: Overview of the entire experimental pipeline. Phase One: Involves data simulation and data preparation. Phase Two: Involves Variant calling and analysis of called variants.

METHODS - VC



PHASE 1:

Simulation

Using NEAT and 1000 Genomes' African and European reference genomes with the HG19 reference.

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, Samtools, SNVer, BCFtools, VarScan2, Platypus, GATK, Lofreq and VarDict - one sample per category.

Variant Call Assessment

Using true and simulated VCF

Analyse Called Variants

Identify FP and FN rates.

Assess the best approach

Compare FP and FN rates.

Figure 2: Overview of the entire experimental pipeline. Phase One: Involves data simulation and data preparation. Phase Two: Involves Variant calling and analysis of called variants. (4,5,13-18)

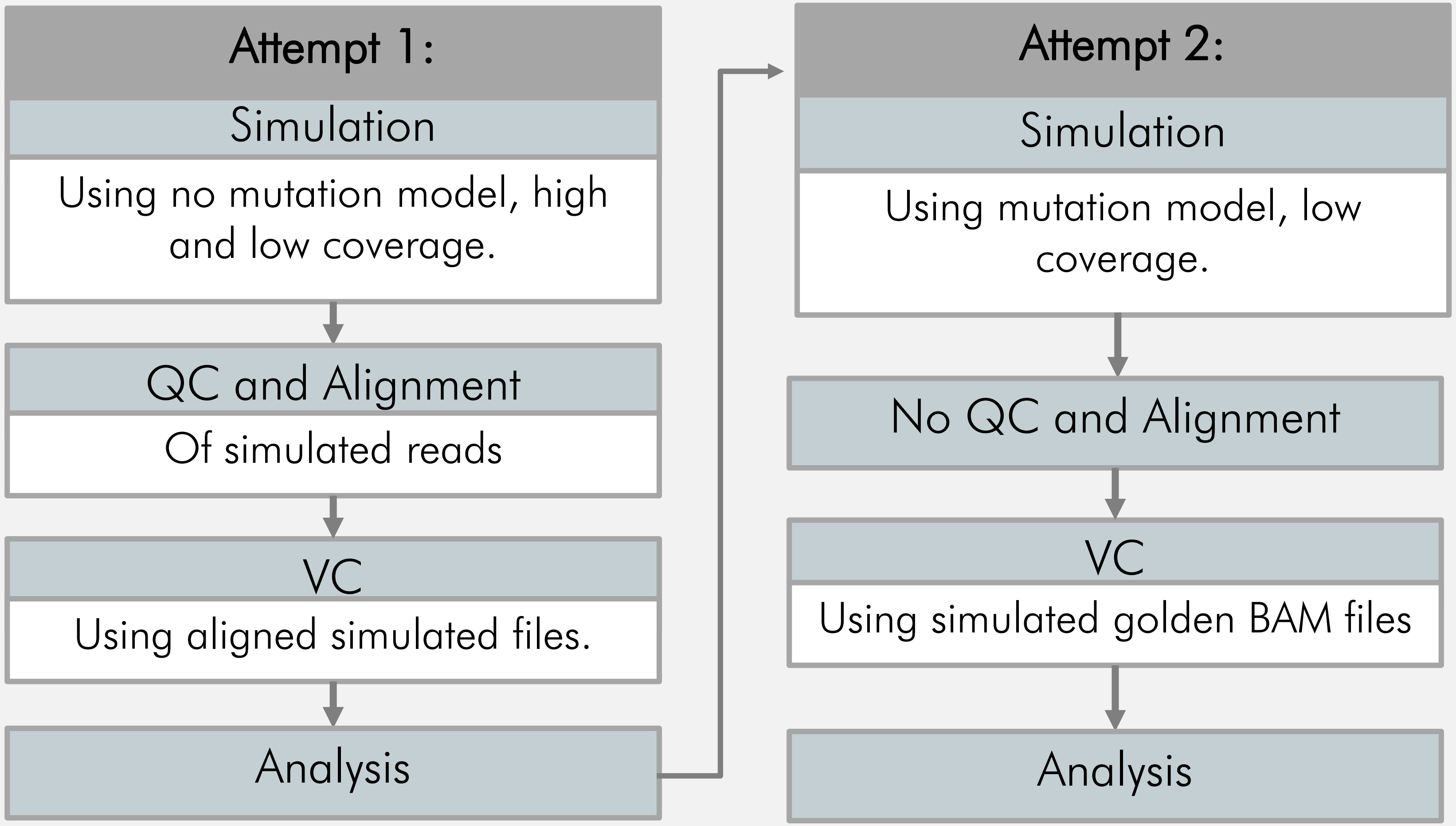
METHODS - ANALYSIS

- Venn Diagrams created using R.
- Calculate % FP and %FN.

$$\%\ Positions\ FP = \frac{Number\ of\ FP\ positions\ called}{Total\ number\ of\ positions\ called} \times 100$$

$$\%\ Positions\ FN = \frac{Number\ of\ FN\ positions\ called}{Total\ number\ of\ variant\ positons} \times 100$$

METHODS - OVERVIEW



Attempt 1:

Simulation

Using no mutation model, high and low coverage.

QC and Alignment

Of simulated reads

VC

Using aligned simulated files.

Analysis

Erroneous Results

Attempt 2:

Simulation

Using mutation model, low coverage.

No QC and Alignment

VC

Using simulated golden BAM files

Analysis

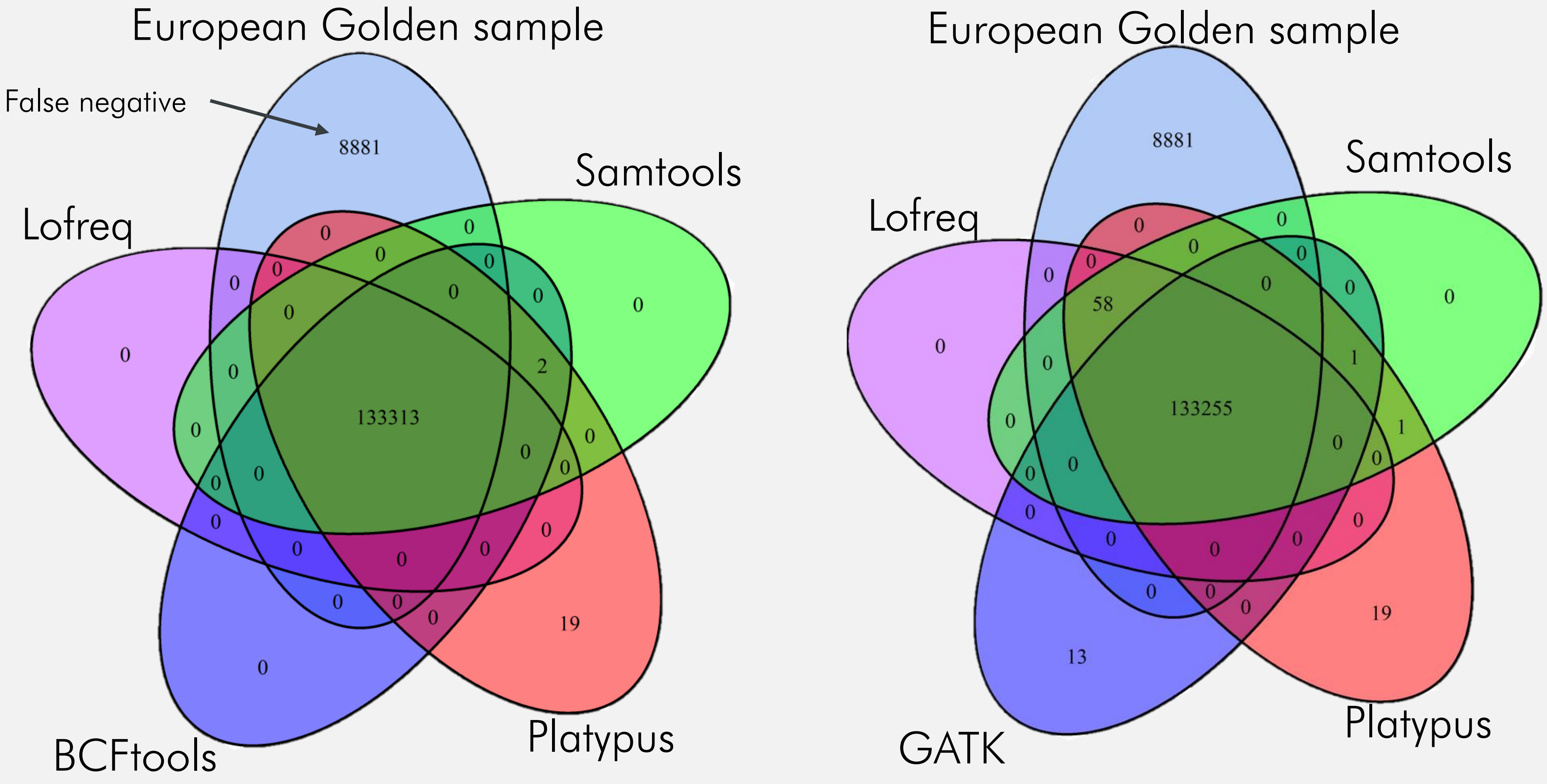
Figure 3: True pipeline with differences between attempts.

RESULTS - SIMULATION

Table 2: Total number of variants present in the golden variant call files produced by the simulation processes.

	Number of SNPs present in the African Golden VCF	Number of SNPs present in the European Golden VCF
First Simulation Run	399737	246551
Second Simulation Run	236037	263188

RESULTS – SECOND VC APPROACH



Note: Not all VC tools results are listed as some produced unfeasible results.

Figure 4: A Venn diagram representing the variant positions identified by respective VC tools for European samples

RESULTS – SECOND VC APPROACH

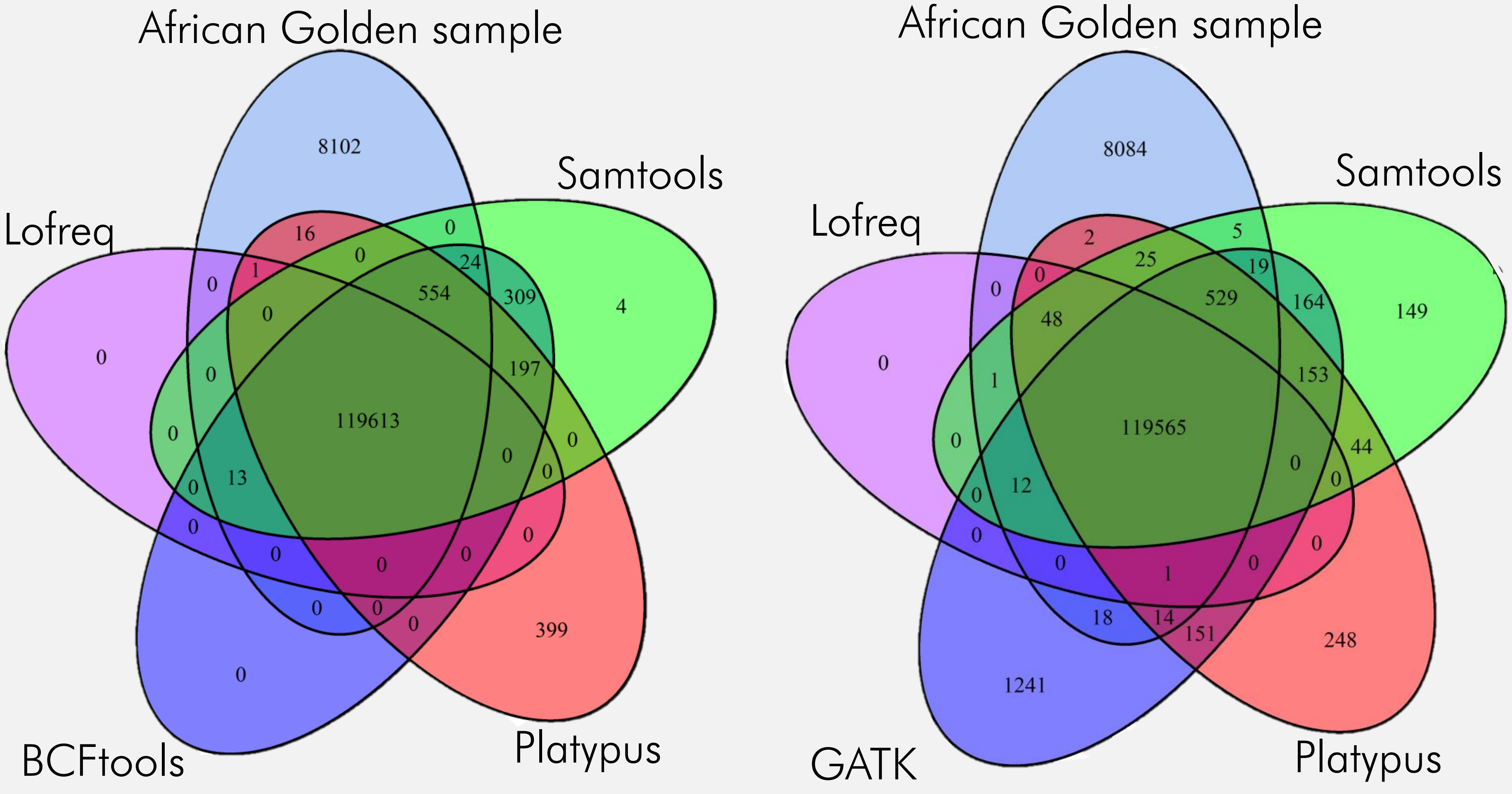


Figure 5: A Venn diagram representing the variant positions identified by respective VC tools for African samples

RESULTS - SECOND VC APPROACH

Table 3: %FP and FN positions called by VC tools using African data from the second simulation approach.

	% FP positions	% FN positions
Lofreq	0	6.7766
BCFtool	0.4192	6.3270
Samtools	0.4225	6.3270
Platypus	0.4935	6.3426
GATK	1.4023	6.3613
Average	**0.5475**	**6.4269**

Table 4: %FP and FN positions called by VC tools using European data from the second simulation approach.

	% FP positions	% FN positions
Lofreq	0	6.2457
BCFtool	0.0015	6.2457
Samtools	0.0015	6.2457
Platypus	0.0143	6.2457
GATK	0.0105	6.2865
Average	**0.0056**	**6.2539**

DISCUSSION

- GenMutModel does not accurately model African genetic variation,
- All VC tools favour specificity over sensitivity,
- The average FN% and average FP% is lower using European that African,
- Tools better suited to European data,
- Lofreq most accurate VC tool using European data,
- Optimal VC tool for African data overall is BCFtools.

LIMITATIONS AND FUTURE WORK

- Limitations:
  - Simulation output formatting was not standard,
  - Unable to assess all tools and all data simulated,
  - Could not resolve all issues,
- Future Work:
  - Use Golden BAM as a control for effect of QC and alignment pipeline,
  - Combinations of VC tools and alignment tools (19,21),
  - Incorporate de novo assembly based VC tools (20).

CONCLUSION

- Encountered setbacks and were not able to reach all of our goals,
- VC tools are less accurate using African data,
- Simulation tool does not accurately represent African variation,
- BCFtools best choice for African data currently,
- VC tools should be adapted and improved to account for different genomic characteristics.

ACKNOWLEDGEMENTS

- Supervisors: Prof Nicola Mulder and Dr Emile Chimusa
- CBIO Lab
- CHPC

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad •

SPES BONA

CBIO

Computational Biology @ UCT

CHPC

CENTRE FOR HIGH PERFORMANCE COMPUTING

QUESTIONS?

REFERENCES:

1.Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).

2.The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

3.Sandmann, S. *et al.* Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci. Rep.* **7**, (2017).

4.Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinforma. Oxf. Engl.* **25**, 2283–2285 (2009).

5.Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *ArXiv12073907 Q-Bio* (2012).

6.Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **15**, 256–278 (2014).

7.Campbell, M. C. & Tishkoff, S. A. AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annu. Rev. Genomics Hum. Genet.* **9**, 403–433 (2008).

8.Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nat. News* **538**, 161 (2016).

9.Jorde, L. B. *et al.* The Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data. *Am. J. Hum. Genet.* **66**, 979–988 (2000).

10.Sandmann, S., Graaf, A. O. de, Reijden, B. A. van der, Jansen, J. H. & Dugas, M. GLM-based optimization of NGS data analysis: A case study of Roche 454, Ion Torrent PGM and Illumina NextSeq sequencing data. *PLoS ONE* **12**, (2017).

11.Pirooznia, M. *et al.* Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum. Genomics* **8**, 14 (2014).

12.Glenn, T. C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11**, 759–769 (2011).

13.Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

14.Lai, Z. *et al.* VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).

15.Wei, Z., Wang, W., Hu, P., Lyon, G. J. & Hakonarson, H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* **39**, e132 (2011).

16.Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).

17.DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

18.Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).

19.Smith, H. E. & Yun, S. Evaluating alignment and variant-calling software for mutation identification in C. elegans by whole-genome sequencing. *PLOS ONE* **12**, e0174446 (2017).

20.Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 18 (2012).

21.Wu, L., Yavas, G., Hong, H., Tong, W. & Xiao, W. Direct comparison of performance of single nucleotide variant calling in human genome with alignment-based and assembly-based approaches. *Sci. Rep.* **7**, (2017).

Term	Abbreviation	Definition
Single nucleotide polymorphism (variant)	SNP	Single base pair mutation that occurs at a specific position in a genome and varies across individuals.
Copy Number Variant	CNV	Sections of the genome are repeated – the number of times these sections are repeated varies among individuals.
Structural Variant	SV	These are large variants in the genome (generally 1kb or larger) that are made up of inversions, translocations or indels.
Indel	-	Insertion or deletion of bases in the genome of an individual.
Multiple nucleotide variant	MNV	Several consecutive variant sites where variation is seen across individuals.
Heuristic	-	Offers a practical method to problem solving that is not guaranteed to be optimal but will be faster than more impractical solutions.
Bayesian	-	Method of inference whereby probability for a hypothesis is updated as more information becomes available.
Sequence Coverage (Depth)	-	The number of reads that contain a given nucleotide base in the reconstructed sequence.
Variant Calling	VC	The process whereby polymorphisms (SNPs, Indels, SV, CNV) are identified in a genome.
Quality Control	QC	The process whereby sequence reads are assessed and modified if low quality data is present.

Term	Abbreviation	Definition
False Positive	FP	When we reject the null hypothesis given it was true.
False Negative	FN	When we do not reject the null hypothesis given it was false.
Allele	-	Variant forms of a base position in a genome.
Variant Call File	VCF	The file produced by the variant calling process that contains the variants identified from a sample.
Phenotype	-	An observable characteristic of an organism that results from an interaction between genotype and environment.
Genome wide association studies	GWAS	An experiment whereby a genome wide set of genetic markers are compared between individuals to identify a potential association with a trait.
Genotype	-	The genetic make-up of an organism with respect to one or multiple traits.
Linkage disequilibrium	-	A non-random association of alleles at different locations in a genome.
Haplotype callers	-	Short haplotypes are read from input data and thus do not call variants based on only one position at a time.
Variant Call file	VCF	A file containing identified variant sites along with genotype information.
Binary alignment Map	BAM	A compressed version of a SAM file.
Sequence alignment map	SAM	File format containing aligned sequence reads.

IN DEPTH VC TOOL DETAILS

- VarScan2 (Koboldt et al., 2012) has been designed to use a heuristic and statistical approach to identify variants and may also be used to identify somatic mutations (Sandmann et al., 2017a)
- The Samtools package consists of two different variant calling tools –Samtools and BCFtools. Samtools and BCFtools both have Bayesian underling processes and do not require any genotyping assumptions to call variants (Li, 2011; Li et al., 2009). The key difference between these tools is that BCFtools performs VC with a multiallelic calling model while Samtools uses a consensus calling model (Li, 2011).
- SNVer uses a binomial-binomial model to test the significance of observed allele frequencies against the sequencing error rates (Wei et al., 2011).
- GATK Haplotype Caller also uses a Bayesian approach to identify variants, however this is under the assumption of uniform copy numbers (Garrison and Marth, 2012). GATK does not involve genotype calling to inform variant identification (DePristo et al., 2011). GATK incorporates "technical covariates, known sites of variation, genotypes for individuals, linkage disequilibrium, and family and population structure" (DePristo et al., 2011) into its variant calling approach to separate out true variants from machine artefacts. As GATK has been developed as a toolkit it also offers the ability for local realignment and base quality score recalibration to eliminate FP variants (DePristo et al., 2011).
- LoFreq uses a Poisson-binomial distribution (Sandmann et al., 2017a) to model sequencer run specific error rates and as a result can call rare variants(Wilm et al., 2012).
- Freebayes is a haplotype based variant calling tool that uses a Bayesian statistical framework that can model multiallelic loci in a set of individuals with non-uniform copy numbers (Garrison and Marth, 2012).
- Platypus is also a multi-sample haplotype based variant caller. Platypus integrates various approaches into one to perform variant calling – mapping based assembly and reference free assembly are incorporated into a Bayesian framework to perform variant calling (Rimmer et al., 2014).
- VarDict is the newest of the tools chosen to compare, it uses two types of local realignment to improve estimated allele frequencies and is able to call complex combinations of variants simultaneously (Lai et al., 2016).

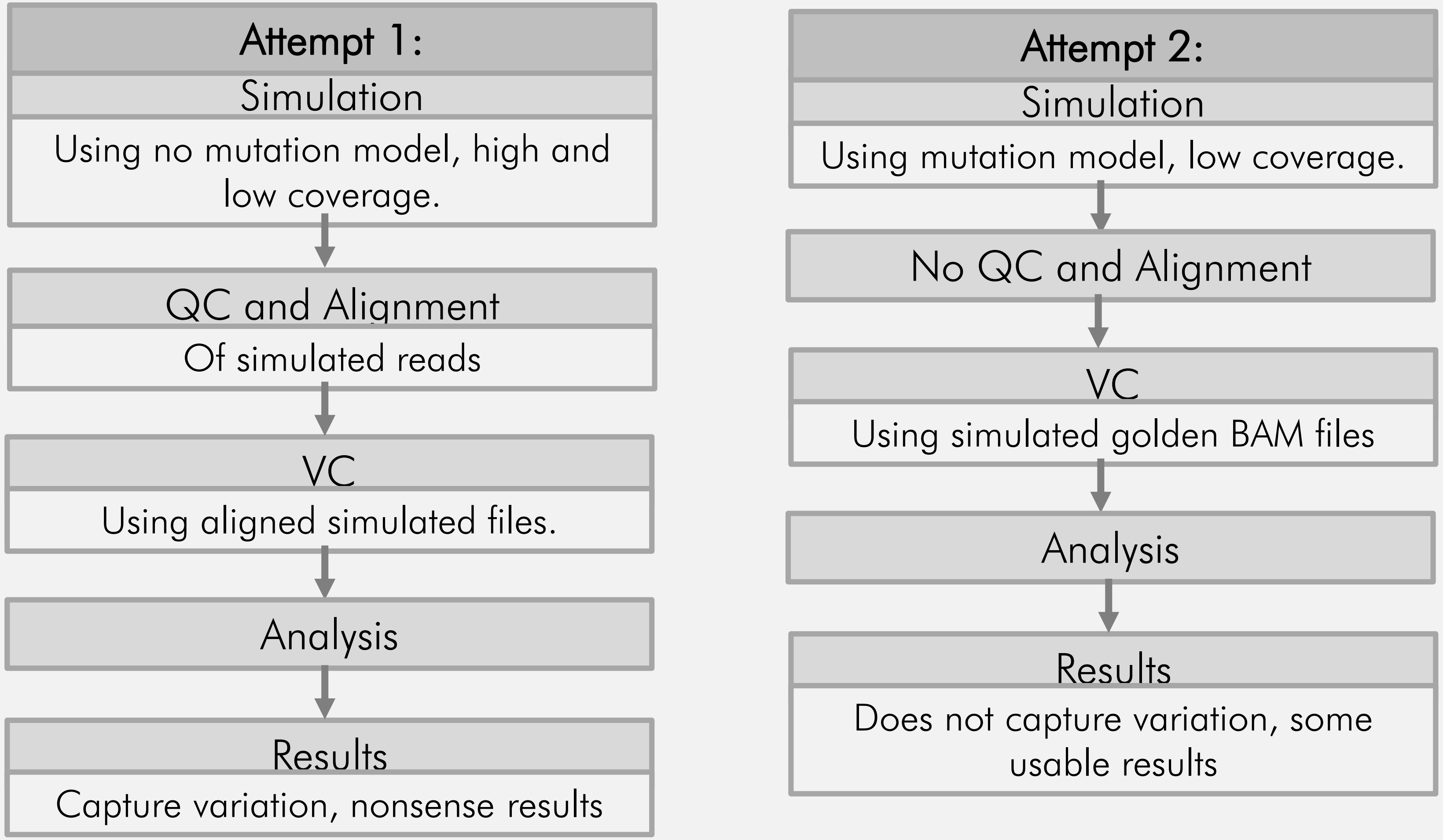
ALIGNMENT, QUALITY CONTROL AND DATA PREPARATION:

First, simulated FastQ files were quality checked using FastQC (http://www.bioinformatic .babraham.ac.uk/projects/fastqc/). FastQC reports were aggregated using MutiQC (Ewels et al., 2016).

Using the Burrows Wheeler Alignment tool (Li and Durbin, 2009), the simulated FastQ reads were aligned to create a SAM file. This was done using the HG19 human reference.

Picard SortSam was used to sort the SAM files. Picard http://picard.sourceforge.net.) MarkDuplicates and BuildBamIndex were used to mark duplicates (these are due to PCR artefacts) in the BAM files and index the BAM file, respectively. Picard AddOrReplaceReadGroups, BuildBamIndex and SortSam, were used to add read group names to the simulated samples and once again index and sort these files. GATK (McKenna et al., 2010) was used to perform realignment on these BAM files. Picard FixMateInformation was used to verify and correct any mate pair information. GATK (McKenna et al., 2010) was used for a final step of read recalibration and realignment.

RESULTS – SECOND VS FIRST



Attempt 1:

Simulation

Using no mutation model, high and low coverage.

QC and Alignment

Of simulated reads

VC

Using aligned simulated files.

Analysis

Results

Capture variation, nonsense results

Attempt 2:

Simulation

Using mutation model, low coverage.

No QC and Alignment

VC

Using simulated golden BAM files

Analysis

Results

Does not capture variation, some usable results

QC RESULTS

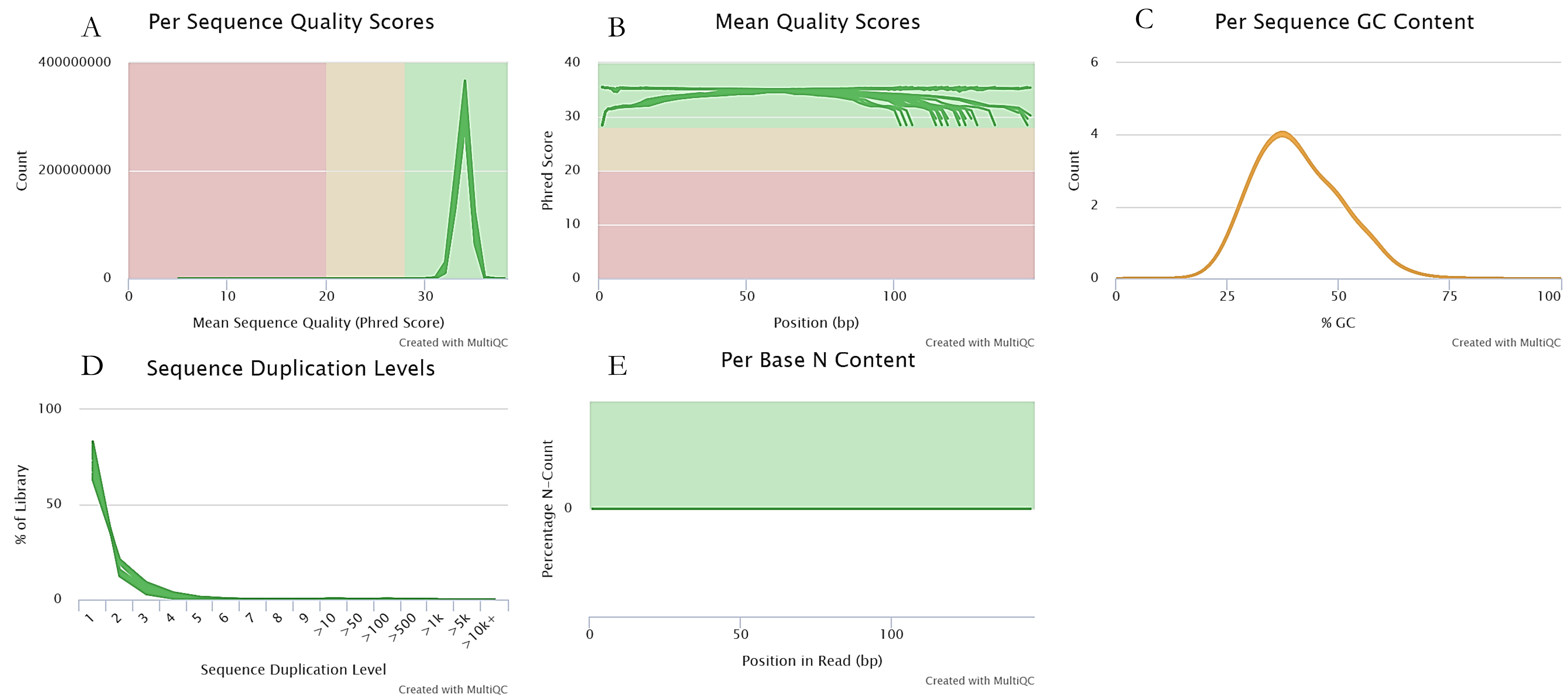


Figure 4: Aggregated FastQC report output for all Simulated High Coverage AFR samples. FastQC was used to test quality of simulated reads and the produced reports were aggregated using MultiQC. FastQC reports the (A) Per sequence quality scores, (B) Mean quality scores, (C) per sequence GC content, (D) sequence duplication levels and (E) the per base N content.

RESULTS – FIRST VC APPROACH

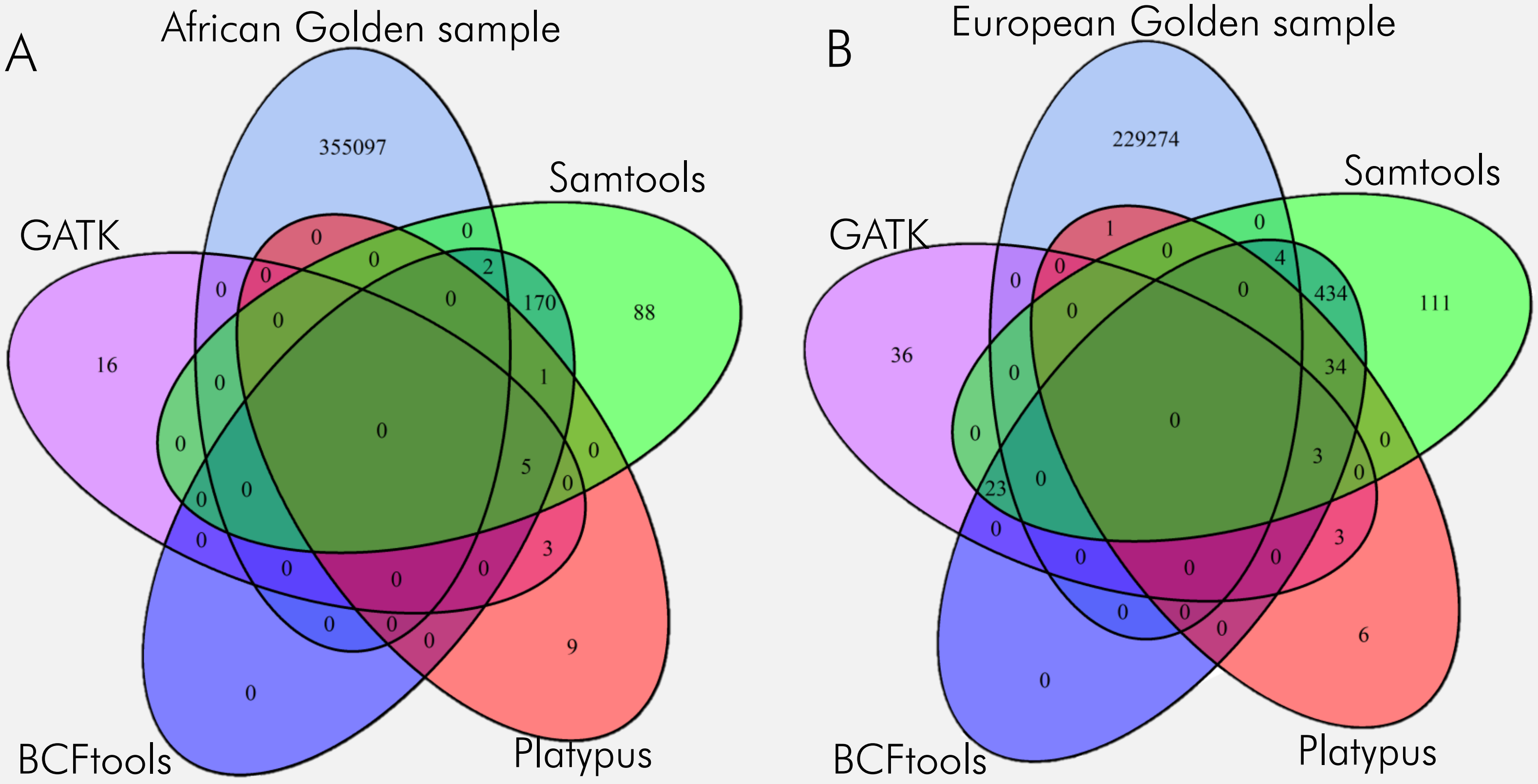


Figure 8: A Venn diagram representing the variant positions identified by respective VC tools for (A) African and (B) European high coverage data.

RESULTS – SECOND VC APPROACH

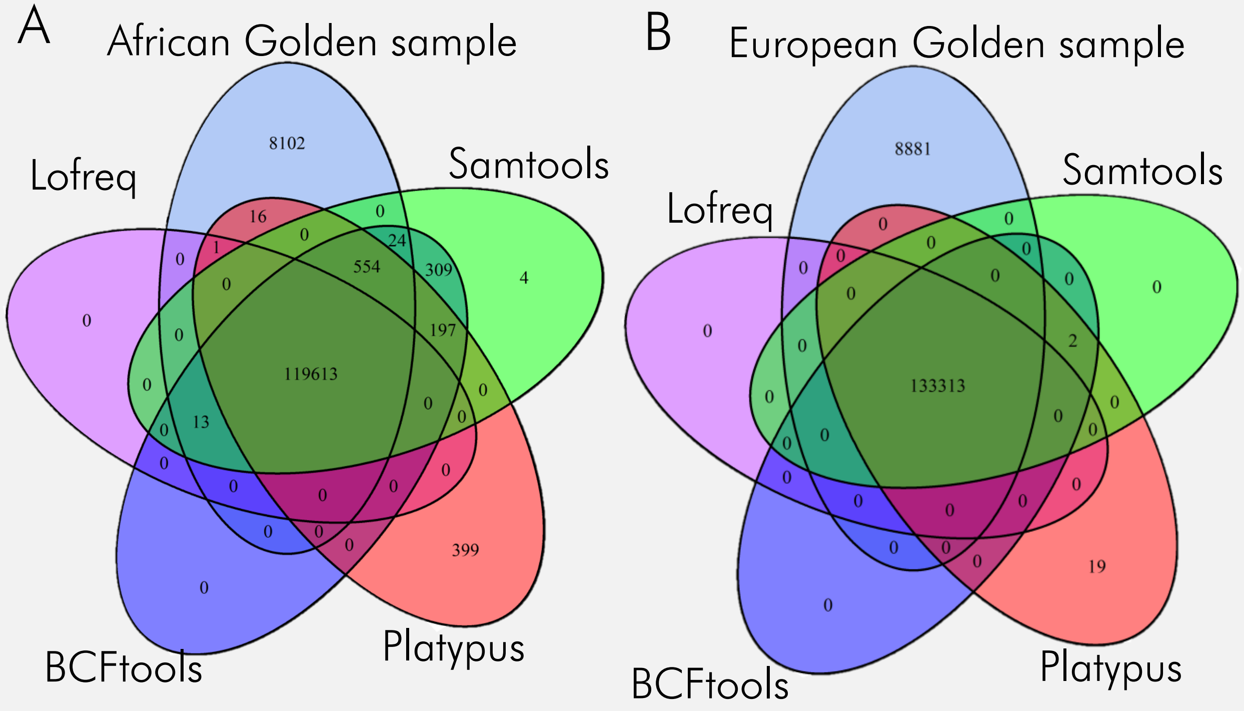


Figure 10: A Venn diagram representing the variant positions identified by respective VC tools for (A) African and (B) European data.

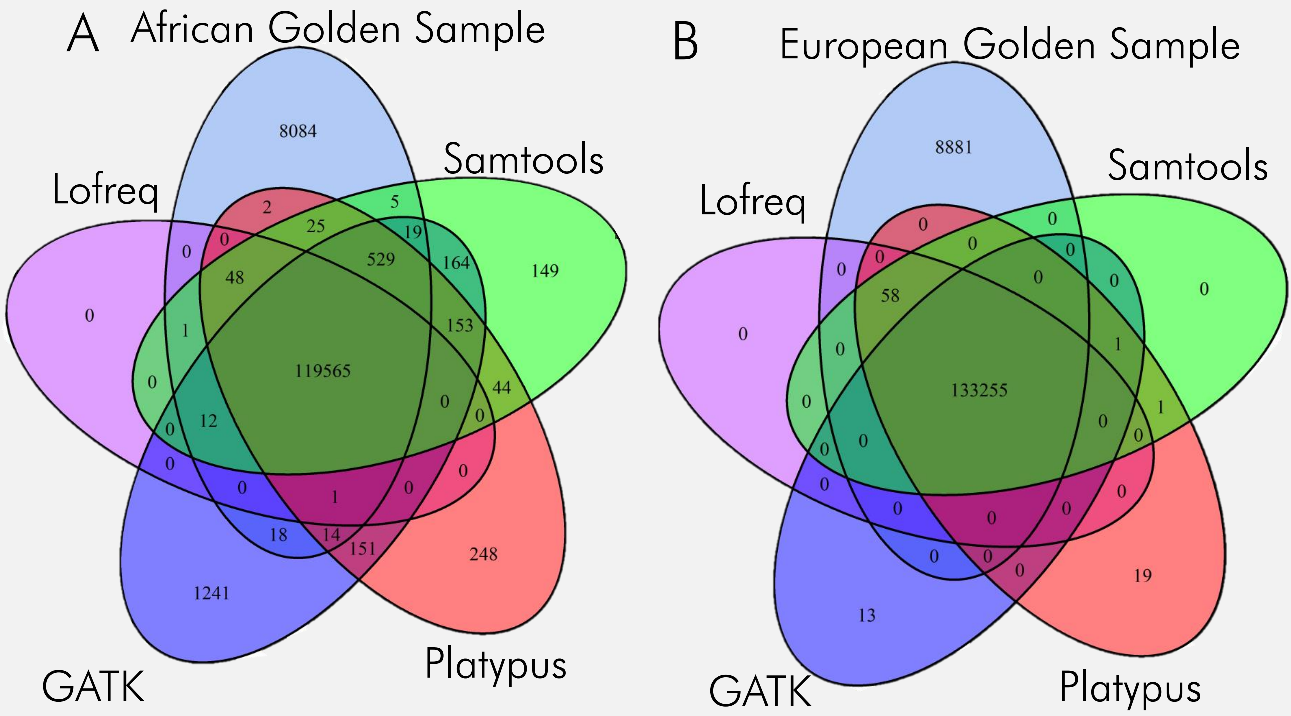


Figure 11: A Venn diagram representing the variant positions identified by respective VC tools for (A) African and (B) European data.

Table 3: Variable positions called by VC tools using African data from the second simulation approach. These values were obtained using data obtained from the created Venn diagrams. This depicts the sensitivity and specificity of the VC tools seen by the #FP and #FN respectively.

	True # variant sites	# Variant positions called	# FP positions	# FN positions	% FP positions	% FN positions
Lofreq	128323	119627.0	0.0	8696.0	0.0000	6.7766
BCF	128323	120710.0	506.0	8119.0	0.4192	6.3270
Samtools	128323	120714.0	510.0	8119.0	0.4225	6.3270
Platypus	128323	120780.0	596.0	8139.0	0.4935	6.3426
GATK	128323	121867.0	1709.0	8163.0	1.4023	6.3613
Average	-	120739.6	664.2	8247.2	0.5475	6.4269

Table 4: Variable positions called by VC tools using European data from the second simulation approach. These values were obtained using data obtained from the created Venn diagrams.

	True # variant sites	# Variant positions called	# FP positions	# FN positions	% FP positions	% FN positions
Lofreq	142194	133313	0	8881	0.0000	6.2457
BCF	142194	133315	2	8881	0.0015	6.2457
Samtools	142194	133315	2	8881	0.0015	6.2457
Platypus	142194	133332	19	8881	0.0143	6.2457
GATK	142194	133269	14	8939	0.0105	6.2865
Average	-	133318.75	7.4	8892.6	0.0056	6.2539