Leveraging Whole Genome Sequences to Compare Mutational Mechanism and Identify Medically Relevant Variation in African versus Non-African Descend Populations

By

Shatha Mobarak Alosaimi
ALSSHA003

Dissertation submitted to the University of Cape Town
in fulfilment of the requirements for the degree of Master of Science (MSc) in Human Genetics.

Faculty of Health Sciences
University of Cape Twon

Date of Submission: 10/02/2020

Supervised by: **Assoc/Prof. Emile R. Chimusa**
Division of Human genetics
Department of Pathology
University of Cape Town, SA

Declaration

I, Shatha Alosaimi, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever

Signature:

Signed by candidate

Date: 01/02/2020.

Publications

I confirm that I have been granted permission by the University of Cape Town's Master's Degrees Board to include the following publications in my thesis, and where co-authorships are involved, my co-authors have agreed that I may include the publications.

1. Shatha Alosaimi, Armand Bandiang, Noelle van Biljon, Denis Awany, Prisca K Thami, Milaine S S Tchamga, Anmol Kiran, Olfa Messaoud, Radia Ismaeel Mohammed Hassan, Jacquiline Mugo, Azza Ahmed, Christian D Bope, Imane Allali, Gaston K Mazandu, Nicola J Mulder, Emile R Chimusa., 2019. A broad survey of DNA sequence data simulation tools. Briefings in Functional Genomics. DOI 10.1093/$bfgp/elz$033.

2. Shatha Alosaimi, Noelle van Biljon, Denis Awany, Prisca K Thami, Imane Allali, Gaston K Mazandu, Nicola J Mulder, Emile R Chimusa (2020). Leveraging sequencing models in the simulation of African Versus Non-African low/high coverage whole genome sequence data to Assess variant calling approaches. In preparation.

3. Shatha Alosaimi, Noelle van Biljon, Denis Awany, Prisca K Thami, Imane Allali, Gaston K Mazandu, Nicola J Mulder, Emile R Chimusa (2020) Dissecting Genetic Mutations and Secondary Finding from Twenty World-wide Ethnics Group. In preparation.

Abstract

Whole-Genome Sequencing (WGS) is ushering a new era in healthcare and research in identifying genetic variation in all populations. However, the African populations are still under-represented. Since African populations are being the most genetically diverse with high heterogeneity rate, we need to benchmark the Whole Genome Sequence (WGS) analysis pipeline to ensure reliable mutation detection. Therefore, it is essential to ensure that all steps of WGS downstream analysis are accurate, mainly the variant calling (VC). Current VC tools may produce false-positive/negative results; such result may produce misleading conclusions in prioritisation of mutation, clinical relevancy and actionability of genes. With such many VC tools, two questions have arisen. Firstly, which tool has a high rate of sensitivity and precision in low either high coverage African sequences, given they have high genetic diversity and heterogeneity? Secondly, does the improvement of the VC result will advance the accuracy of detecting mutation and incidental finding (actionable genes) in African populations?

In this project, a total of 100 DNA sequence samples was simulated (of which every 50 samples mimicked the genetics background of African and European, respectively) at different coverage (high and low). In particular, the sensitivity to discover polymorphisms was done by nine different VC tools. These tools were assessed in term of false positive/negative call rate given the simulated golden variants. Combining our result on sensitivity and positive predictive value (PPV). Lofreq performs best in African population data (sens=0.85, PPV=0.983, F-score=0.91) on high/low coverage data; as a result, we chose Lofreq to perform variant calling, and Gene-based annotation is performed to conduct in-sillico predication of mutation on publicly available data (the African Genome Variation and 1000 Genome Project). In doing so, we have leveraged WGS to examine and validate four of burden diseases in the African content, such as communicable diseases: HIV/AIDS, Malaria, Tuberculosis (TB), and Non-communicable diseases: such as Sickle cell disease, these diseases have uniquely shaped ethnic-specific and continental genomics variation and therefore provides unprecedented opportunities to map disease genes across the African continent. Moreover, the current actionable gene recommended by The American College of Medical Genetics and Genomics (ACMG) in the African population and update on additional African-specific actionable genes.

Our result suggests African and African diaspora ethnic groups, particularly Bantu and Khoesan ethnics have gene diversity, high proportion of derived allele at low minor allele frequency $(0.0 - 01)$ and the highest proportion of pathogenic variants within HIV, TB, Malaria, Sickle-Cell disease, while non-African ethnic groups including Latin America, Afro-Asiatic European related ethnic groups have high proportion of pathogenic variants within current actionable gene list.

Overall, given the observed highest genetic diversity found in African ethnics and African diaspora related ethnics at these four Africa burden diseases and current actionable gene associated, our results support (1) the use of personalised medicine as beneficial to both African continent and worldwide; (2) a recommendation for African-specific actionable list of genes to further improve African and diaspora healthcare.

Acknowledgments

First and Foremost praise and all glory to the Almighty Allah, second all my gratitude goes to my beloved husband Marwan Alosaimi for believing in me, encouraging me and all his never-ending support, and my most profound gratitude goes to my dearest, precious son Jaber.

I would also like to express my sincere and deep gratitude to my supervisor Assoc/Prof. Emile Chimusa for his assistance and valuable guidance throughout this project and for giving me the opportunity to continue my post-graduate studies at the University of Cape Town, and for introducing me to this amazing field of bioinformatics, genetics and allowing me to discover the beauty and the complexity of the African population genetics.

Additionally, there are no proper words to convey my deep recognition to my family, especially my father and my mother, for their prayers, support and love. Seeing my parents succeeded their path in life has helped me to get through mine. Furthermore, I am most grateful to the University of Cape Town for support and resources as well as, The Center for High-Performance Computing (CHPC) in the Republic of South Africa is gratefully acknowledged with thanks.

Finally, i genuinely thank Dr Gloudi Agenbag for teaching me laboratory techniques, also a special Thanks to Noëlle van Biljon, and Loratoeng Mpolokeng, for cooperating under the same project, and i am also thankful to my all friends, colleagues and all the staff of the department of human genetics at the university of cape town. Thank you all.

Contents

Abbreviations xi

1 Introduction, Background and Literature Review 1
- 1.1 Introduction . . . 1
- 1.2 Problem Statement . . . 3
- 1.3 Research Questions . . . 4
- 1.4 Research Aims and Objectives . . . 4
  - 1.4.1 Overall hypothesis . . . 4
  - 1.4.2 Project objective . . . 4
- 1.5 Dissertation Outline . . . 5
- 1.6 Overview on Next Generation Sequencing Downstream Analysis . . . 6
- 1.7 Variant Calling . . . 8
  - 1.7.1 Algorithms used to call the variant . . . 8
    - 1.7.1.1 Heuristic approach . . . 8
    - 1.7.1.2 Statistical approach . . . 9
- 1.8 Previous Studies Comparing Variant Calling Tools . . . 9
- 1.9 Evaluation Metric of Variant Calling . . . 11
- 1.10 Overview of Mutations in The African Populations . . . 16
  - 1.10.1 Communicable Diseases: Susceptibility of Infectious Diseases in Africa . . . 16
    - 1.10.1.1 Tuberculosis . . . 16
    - 1.10.1.2 Malaria . . . 17
    - 1.10.1.3 HIV/AIDS in Africa . . . 17
  - 1.10.2 Non-communicable Diseases in Africa . . . 18
    - 1.10.2.1 Sickle Cell Disease . . . 18
- 1.11 Secondary Finding (Actionable Genes) . . . 19

2 DNA Sequencing Simulation 20
- 2.1 Introduction . . . 20
  - 2.1.1 Overview on Different NGS Simulation Tools . . . 20
  - 2.1.2 General process of DNA read simulation . . . 21
- 2.2 Materials and Methods . . . 23
  - 2.2.1 Data Description . . . 24
    - 2.2.1.1 Mutation Model . . . 24
    - 2.2.1.2 Sequencing Error Model . . . 24
  - 2.2.2 NEAT-GenReads . . . 24
- 2.3 Results . . . 25
  - 2.3.1 Assessing and Examining Simulation Outputs . . . 25
    - 2.3.1.1 Quality Control Check on the Simulated Forward and Reverse FastQ files . . . 25
    - 2.3.1.2 Golden BAM files . . . 28
    - 2.3.1.3 Golden VCF files . . . 28

2.4 Brief Discussion and Chapter Summary 29

3 Assessment of Nine Different Variant Calling Tools on African Versus Non-African Populations 30
3.1 Introduction 30
3.2 Characteristics and Specifications of Variant Calling tools 31
3.3 Materials and Methods 36
3.3.1 Data Description 36
3.3.2 Data Generation and Processing 36
3.3.3 Performing Variant Calling 37
3.4 Results 38
3.5 Discussion and Overall Chapter Summary 46

4 Dissecting Genetic Mutations and Secondary Finding from Twenty Worldwide Ethnics Groups 48
4.1 Introduction 48
4.2 Methods and Material 49
4.2.1 Data Description and Quality Check 49
4.2.2 Variants Discovery Analysis 50
4.2.3 Variant Annotation 51
4.2.4 Phased and Haplotypes Inference 51
4.2.5 Disease- and Actionable Gene-specific Population Structure 51
4.2.6 Proportion of Ancestral/Derived Alleles among Risk conferring Alleles . . 52
4.2.7 Distribution of Minor Allele Frequency and Gene-specific in SNP Frequencies 53
4.2.8 Aggregating SNPs Summary Statistics at the Gene level 53
4.3 Results 53
4.3.1 Disease- and Actionable Gene-specific Population Structure 53
4.3.2 Pathogenic Mutation at Polymorphisms within Disease Related Associated Genes 56
4.3.3 Distribution of Minor Allele Frequency and Gene-specific in SNPs Frequencies 62
4.3.4 Gene-specific in Derived Allele Proportion and Relationship between Derived and Ethic-specific Minor Allele 65
4.3.5 Genetic Diversity: Observed and Expected Heterozygosity 74
4.4 Discussion and Chapter Summary 75

5 General Discussion and Conclusion 77

Appendices 80

A Supplementary Information 81
References 85

List of Figures

1.1 An overview on the WGS Data Analysis Pipeline. 7
1.2 Worldwide Number of Newborns with Sickle Cell Anemia from (Piel, Steinberg, & Rees, 2017). 18

2.1 Overview of the general Simulation process. 22
2.2 Sketch of the method used in NEAT-GenReads. 23
2.3 Aggregated Report from MultiQC of all the FastQC reports of the simulated African- High coverage samples. 26
2.4 Aggregated Report from MultiQC of all the FastQC reports of the simulated European- High coverage samples. 26
2.5 Aggregated Report from MultiQC of all the FastQC reports of the simulated African- Low coverage samples. 27
2.6 Aggregated Report from MultiQC of all the FastQC reports of the simulated European- Low coverage samples. 27
2.7 An example of Golden BAM header and flagstat. 28

3.1 Overview of the variant calling analysis pipeline. 36
3.2 Relation between Positive Predictive Value (PPV) and Sensitivity in case of comparing variant calling tools on the African and European Populations regarding different coverage. 41
3.3 An UpSetR diagram visualising the intersections of the variant positions produced by Samtools, BCFtools, SNVer, GATK , Platypus, VarScan, Lofreq and Vardict, on African population- High coverage data. 42
3.4 An UpSetR diagram visualising the intersections of the variant positions produced by Samtools, BCFtools, SNVer, GATK , Platypus, VarScan, Lofreq and Vardict, on European population- High coverage data. 43
3.5 An UpSetR diagram visualising the intersections of the variant positions produced by Samtools, BCFtools, SNVer, GATK , Platypus, VarScan, Lofreq and Vardict, on African population- Low coverage data. 44
3.6 An UpSetR diagram visualising the intersections of the variant positions produced by Samtools, BCFtools, SNVer, GATK , Platypus, VarScan, Lofreq and Vardict, on European population- Low coverage data. 45

4.1 Principal Component Analysis (PCA) of the actionable genes, , plot of the first and the second eigenvectors for all populations. 54
4.2 Principal Component Analysis (PCA) of genes associated with HIV, plot of the first and the second eigenvectors for all populations. 54
4.3 Principal Component Analysis (PCA) of genes associated with Tuberculosis, plot of the first and the second eigenvectors for all populations. 55

4.4 Principal Component Analysis (PCA) of genes associated with Sickle Cell Disease, plot of the first and the second eigenvectors for all populations. . . . 55
4.5 Principal Component Analysis (PCA) of genes associated with Sickle Cell Disease, plot of the first and the second eigenvectors for all populations. . . . 56
4.6 The proportion of pathogenic variants within ACG-specific (Actionable Genes) genes among all 20 ethnic groups. . . . 57
4.7 The proportion of pathogenic variants within HIV-specific genes among all 20 ethnic groups. . . . 58
4.8 The proportion of pathogenic variants within Malaria-specific genes among 20 world-wide ethnic groups. . . . 59
4.9 The proportion of pathogenic variants within Sickle Cell Disease-specific genes among 20 world-wide ethnic groups. . . . 60
4.10 The proportion of pathogenic variants within Tuberculosis-specific genes among 20 world-wide ethnic groups. . . . 61
4.11 The distribution of the minor allele frequency giving a gene level (Actionable Genes) among all ethnic groups. . . . 62
4.12 The distribution of the minor allele frequency giving a gene level (HIV) among all ethnic groups. . . . 63
4.13 The distribution of the minor allele frequency giving a gene level (Malaria) among all ethnic groups. . . . 63
4.14 The distribution of the minor allele frequency giving a gene level (Sickle Cell Disease) among all ethnic groups. . . . 64
4.15 The distribution of the minor allele frequency giving a gene level (Tuberculosis) among all ethnic groups. . . . 64
4.16 The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding Actionable Genes. . . . 65
4.17 The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding HIV genes. . . . 66
4.18 The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding Malaria genes. . . . 66
4.19 The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding Sickle cells diseases genes. . . . 67
4.20 The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding Tuberculosis genes. . . . 67
4.21 ACG Gene-specific proportion of derived allele among 20 world-wide ethnic groups. . . . 69
4.22 HIV Gene-specific proportion of derived allele among 20 world-wide ethnic groups. . . . 70
4.23 Malaria Gene-specific proportion of derived allele among 20 world-wide ethnic groups. . . . 71
4.24 Sickle-Cell Disease Gene-specific proportion of derived allele among 20 world-wide ethnic groups. . . . 72

4.25 **TB Gene-specific proportion of derived allele among 20 world-wide ethnic groups.** 73
4.26 **Plot Expected heterozygosity as a function of observed heterozygosity per genes of specific diseases within each population.** 74

List of Tables

1.1 **Comparisons of previous studies for different variant calling tools.** . . . 12

2.1 **Main characteristics of NGS technologies and examples of simulation tools representing these platforms (Escalona, Rocha, & Posada, 2017).** 21
2.2 **Individuals from 1000 Genome Human Project, used for generating target variants.** . . . 24
2.3 **Total variants number present in the golden VCF files gnerated by NEAT-GenReads for African and European Populations.** . . . 28

3.1 **The Characteristics of the Variant Calling tools used for this study.** . . . 33
3.2 **Data generated by NEAT-genReads, used to analyse the performance of variant calling tools.** . . . 36
3.3 **List of the variant calling tools used to detect SNPs from simulated WGS data.** . . . 37
3.4 **Summary of the performance metric regarding eight variant calling tools evaluated from simulated data representing African and European Population.** The samples represent simulated data of different coverages. True Positive (TP), False Positive (FP) and False Negative were used to calculate the performance metric of each variant calling tool. . . . 39

4.1 **Data obtained from 1000 Genomes Project (1KGP) (Consortium et al., 2012) and the African Genome Variation Project (AGVP) (Gurdasani et al., 2015) and used for analysis.** . . . 49
4.2 **The Number of SNPs after Quality Control (QC) in each group of genes associated with (HIV, TB, SCD, Malaria and actionable genes.)** 52

A.1 **Lists of gene-disease pairs of HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and Actionable genes.** . . . 81
A.2 **Output files and information of the nine variant calling tools investigated.** . . . 83

Abbreviations

1KGP	The 1000 Genomes Project
454	Roche's 454
ACG	Actionable Genome Consortium
ACGS	The Association for Clinical Genetic Science
ACMG	The American College Of Medical Genetics And Genomics
AFR	African
AGVP	African Genome Variation Project
AIDS	Acquired Immune Deficiency Syndrome
AMR	Ad Mixed American
APR	Precision-Recall Curve
BAM	Binary Alignment Map
bp	Base Pair
BWT	Burrows-Wheeler Transform
CCR5	Chemokine (CC motif) receptor 5
CHPC	The Center For High-Performance Computing
chr	Chromosome
ClinGen	The Clinical Genome Resource
CMDS	Correlation Matrix Diagonal Segmentation
CNA	Copy Number Alteration
CNV	Copy Number Variations
CPU	Central Processing Unit
DAF	Derived allele-frequency
dbsnps	The Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic Acid
EAS	East Asian
EM	Expectation- Maximisation
EUR	European
EXAC	Exome Aggregation Consortium3
FDR	False Discovery Rate
FN	False-Negative
FP	False-Positive
FPR	False-Positive Rate
GATK-HC	Genome Analysis Tool Kit– Haplotypecaller
GATK-UG	Genome Analysis Tool Kit– Unifiedgenotyper
GIAB	Genome In A Bottle
GRCh38	Genome Research Consortium Human Build-38
GUI	Graphical User Interface
GWAS	Genome-Wide Association Studies
H3africa	The Human Heredity And Health In Africa

HBB	Haemoglobin Beta
HbC	Hemoglobin c
HbF	Fetal haemoglobin
HbS	Sickle Hemoglobin
hg38	Human Genome Build 38
HGMD	Human Gene Mutation Database
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigens
HTML	Hypertext Markup Language
IFs	Incidental Findings
Indels	Insertions And Deletions
LD	Linkage Disequilibrium
LOH	Loss Of Heterozygosity
Lowqual	Low Quality
MAF	Minor Allele Frequency
MCC	Matthew's Correlation Coefficient
MNPs	Multi-Nucleotide Polymorphisms
MP	Mate Pair
MQ	Mapping Quality
NGS	Next-Generation Sequencing
NIST	National Instate Of Standerds And Technology
OS	Operation System
PacBio	Pacific Bioscience
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PE	Paired End
PPV	Positive Predictive Values
QC	Quality Control
QD	Qual By Depth
RG	Read Group
ROC	Receiver Operator Characteristic
SAC	South African Colored
SAM	Sequence Alignment Map
SAS	South Asian
SCD	Sickle Cell Disease
SE	Single End
sens	Sensitivity
SFF	Standard Flowgram Format
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
SSMP	Singapore Sequencing Malay Project
STR	Short Tandem Repeat
SV	Structural Variants
TB	Tuberculosis
Ti/Tv	Transition/Transversion
TLR	Toll-Like Receptor
TN	True Negative
TP	True Positive
TVC	Torrent Variant Caller
VC	Variant Calling

VCF	Variant Call Format
VUS	Variant of Uncertain Significance
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
WHO	World Health Organization

Chapter 1

Introduction, Background and Literature Review

1.1 Introduction

"Ex Africa surgit semper aliquid novi", it means there is always something new rises out of Africa, this quote was written nearly 2000 years ago by the Roman scholar and a naturalist philosopher Pliny the Elder. This quote applies perfectly to the genomics studies of the African populations, as there are always new findings of genetic variations (Sirugo et al., 2008) from their genomes.
These discoveries in the African genetics markedly increased alongside the massive evolvement of the sequencing methods over the past few decades. Next-Generation Sequencing (NGS) has allowed for a vast increase in data outputs in a shorter time with decreasing costs.

The impact of NGS on modern biological science is second to none. It is a revolution that could be compared with the one that followed Sanger sequencing 40 years ago (Mielczarek & Szyda, 2016). As a result, the field of application for such data has expanded; it has shed light on the path of genomics, epigenomics, pharmacogenomics, and personalise medicine (Shen, de Stadt, Yeat, & Lin, 2015). Furthermore, this revolution has led to understanding diseases etiology, diagnosis, management and treatment planning for both communicable (e.g. HIV/AIDS, Malaria and Tuberculosis) and non-communicable diseases (e.g. Sickle cell disease), and these diseases are a burden on the African continent with the highest prevalence rate in the whole world.

Moreover, alongside with findings resulted from Whole-genome sequencing/ whole-exome sequencing, it is a necessity to investigate and return secondary finding (Actionable genes) as advised by The American College of Medical Genetics and Genomics (ACMG) (Green et al., 2013). Such a development in NGS has an impact on developing the downstream analysis tools, accordingly, increase the demands on developing deoxyribonucleic acid (DNA) sequencing simulations tools to validate these tools. However, despite the advancement of the NGS and the downstream analysis tools, yet the African populations are still under-represented (Retshabile et al., 2018; Bope et al., 2019). The complex demographic history of the African populations, ethnic diversity, migrations have altered the pattern of genetics variations, accordingly, the African genomes harbor the highest genetics diversity alongside with low level of linkage disequilibrium (LD) when compared to other populations (Sirugo et al., 2008; C. N. Rotimi et al., 2017; Bope et al., 2019). Regardless of the challenges, it is important to understand and investigates this diversity in the genomes of the African populations.

To discover these variations, we must focus on the downstream analysis step, particularly, variant calling. Major advantage of NGS is to detect genetic variant like Single Nucleotide Polymorphism/Variation (SNP/SNV), Insertions and Deletions (Indels), Structural Variants (SV), e.g. Copy Number Variation/Alteration (CNV/CNA). SNPs are single base pair mutation

that may affect nucleic acid or protein binding and activity. Hence, we can identify these SNPs to investigate genetic alterations and then determine if these variations may be the cause of a specific phenotype or trait condition. There are various variant calling approaches available which use NGS data to identify SNPs (E. R. Martin et al., 2010; Durtschi, Margraf, Coonrod, Mallempati, & Voelkerding, 2013; Lai et al., 2016; Fang et al., 2017; Pipek et al., 2017; Sandmann, de Graaf, et al., 2017) and each approach implements a different algorithm with different assumptions. Consequently, all do not always provide the same output (Pabinger et al., 2014), either same SNPs discovery along the genome.

Although, NGS has developed to produce massive amount of parallelism and a huge amount of data, with reduced sequencing time and coast, yet the downstream analysis process is still an important bottleneck (Escalona et al., 2017). Particularly, it is challenging when dealing with the African genomes as it has the highest genetic diversity and low level of linkage disequilibrium, as it may lead to a high rate of false-positive/negative results (Bope et al., 2019). Most of the current NGS downstream analysis tools have been performed and tested on European data (Campbell & Tishkoff, 2008; Popejoy & Fullerton, 2016; A. R. Martin, Teferra, Möller, Hoal, & Daly, 2018) known of high range of haplotypes, however to best to our knowledge, there are no variant calling tools been tested or either designed to deal with the complexity of the African genomes. While the world is moving toward precision medicine, it is vital to develop bioinformatics tools includes variant calling tools with high sensitivity and precision tailored to populations characterised by high genetic variations and low linkage disequilibrium. As a result, these developed tools will contribute in improving the overall health care and scientific research and most important precision medicine in Africa. Moreover, these developed tools will enhance the diagnosis and treatment of Africa's most common diseases such as Malaria, HIV/AIDS, Tuberculosis (TB) and non-communicable disease such as Sickle cell disease.

Previous studies have primarily focused on detecting the best variant calling tools, such as (Bauer & Bauer, 2011; Zook et al., 2014; Huang, Mullikin, & Hansen, 2015) and others. Yet, to the best of our knowledge, none of these studies have focused on the African genomes. These studies have evaluated variant calling tools on different parameters, including sensitivity and low rate of false-positive and false-negative, etc. False-positive (FPs) may potentially arise through the use of an inappropriate reference genome. The current reference may not be suitable for all populations, notably for the African genomes, which are known to be very high diverged with low level of linkage disequilibrium. Current variant calling methods have differing advantages and disadvantages. Thus, it is imperative to evaluate these tools and determine which method is the least error-prone on either low or high sequences coverage of the African genomes. These concerns are shared with the consortium of The Human Heredity and Health in Africa (H3Africa), by developing policies and guidance to return genetic feedback of findings to improve genetic research in the African continent (H3 Africa Working Group on Ethics, 2017).

Further, the burden of communicable and non-communicable diseases in the African continent has uniquely shaped ethnic-specific and continental genomic variation and therefore provides unprecedented opportunities to map disease genes across the continent. As a result, it is important to leverage the availability of genotypes and WGS data from newly and previously studied African populations to understand the spectrum of medical and clinical implications of genome variation from the African populations (Mpye et al., 2017). Furthermore, (Bope et al., 2019) recommended to develop African reference panel and benchmark best variant calling tool using African sequences. In doing so, researchers will provide better classification of the pathogenic and actionable variants (also known as secondary/accidental finding)(Green et al., 2013). Towards this end, our research project consists of three parts, following similar approaches as (Sandmann, De Graaf, et al., 2017) and (Bope et al., 2019).

The overall objectives of this research is (1) to detect best variant calling tool from using simulated DNA sequences data based on benchmarking nine state-of-the-art variant calling tools;

(2) apply the obtained best variant calling tool to perform variant calling on 20 world-wide ethic groups from real data (publicly available data as 1000 Genome project and the African Genome Variation Project); (3) conduct gene-based annotation and in silico prediction of mutation using the obtained variant callings data sets based on known associated genes of four selected disorders of relevance to Africa, include communicable diseases (HIV/AIDS, Malaria, Tuberculosis), and Non-communicable diseases (Sickle cell disease) and a list of current known actionable genes as recommended by the American College of Medical Genetics (ACMG) (Green et al., 2013).

In this Chapter, previous studies regarding variant calling will be reviewed, alongside a brief overview of the genes that associated with diseases especially, the diseases that considered as burden in the African continent such as communicable diseases: HIV/AIDS, Malaria, Tuberculosis (TB), and Non-communicable diseases: such as Sickle cell disease and investigating the actionable genes.

1.2 Problem Statement

Globally, human populations show structured genetic diversity as a result of geographical dispersion, admixture, selection and drift. Understanding this genetic variation can provide insights from our human origins into clinical applications. In these contexts, Africa represents the ancestral birthplace of modern humans. Populations from Africa have the highest levels of genetic diversity and less Linkage disequilibrium (LD)

The complex history of the African populations remains a challenge to unravel based on the present documented records of the interaction and movement among populations. Although, there has been a remarkable growth of African genomics data, the sequencing of individuals within Africa is limited. Critically, through the development of high-density microarrays and next-generation sequencing technologies, the past decade has seen a considerable movement towards the generation of high-resolution genomics data, which has contributed to the identification and fine-mapping of complex disease loci, mostly in European populations.

Advances in high-throughput sequencing have facilitated the development of a range of statistical genomics approaches and applications. These advances in high-throughput technologies have led to an unprecedented increase in the computational complexity of downstream data analysis. An obstacle to validating and bench-marking methods for genome analysis is that there are few reference data sets available for which the "ground truth" about the mutational landscape of the sample genome is known and fully validated. Furthermore, accuracy, effectiveness and performance assessments of different analytical methods used to analyse next generation sequence data are important aspects of medical population genetics.

Variant calling (VC) is an important aspect of genomics studies as polymorphism information can be used to influence important clinical decisions. However, currently most variant calling tools have been designed to leverage populations of long-range haplotypes such as European populations. Differences in genetic characteristics as mentioned above can significantly affect the performance of not only the variant discovery tools, but also downstream bioinformatics analysis tools. Another concern with variant calling is the use of inappropriate reference samples -which leads to an increased rate of false positive (FP) and false negative (FN) SNPs. Therefore, it is critical to detect the true and accurate mutation mainly for rare Mendelian diseases using the Whole Genome Sequencing (WGS) analysis and furthermore secondary findings, without mistaken it with false-positive variants or lose tracing it in false-negative results during variant calling filtering processes. Particularly, when detecting mutations in the African genome data.

1.3 Research Questions

Reflecting to the above problem statement in previous section, a number of questions have arisen:

(a) Which variant calling tool has the least error pore with a high rate of sensitivity and precision, regarding dealing with African genome?

(b) How can we prevent false-positive variants through the process of mutation identification?

(c) Does the improvement of the VC result will advance the accuracy of detecting pathogenic mutation, actionable variants relevant to clinical applications?

1.4 Research Aims and Objectives

1.4.1 Overall hypothesis

We hypothesise that choosing the best variant calling tool, which can accurately call true variant with high sensitivity and can handle the high variation diversity presents in the African genome. Alongside, applying this variant calling tool on leveraging whole genome sequence of African versus non-African will elucidate evolutionary variation in causal mutation patterns and genes actionability to improve the spectrum of medical and clinical implications.

1.4.2 Project objective

This project aimed to:

1. To perform a join variant calling on simulated data (of which every 50 samples mimicked the genetics background of African and European, respectively) to benchmark nine different variant calling tool, and detecting best tool performed best on low/high coverage of African genome data.

2. To systematically assess and identify the false-positive SNPs, from variant calling data analysis, and to improve the mean of SNP identification. This aim will enable us to investigate the evolutionary variation of mutation across 20 world-wide population ethic groups.

3. To apply the best variant calling tool on publicly available data, the African Genome Variation and 1000 Genome Project and examine the evolutionary variation of pathogenic mutation based on selected known disease-genes from four big African burden diseases include HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and a set of known actionable genes across 20 world-wide population ethic groups.

4. To perform disease-genes population structure from these known disease-genes (HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and a set of known actionable genes) among 20 world-wide ethnic-specific data.

5. To examine the heterozygosity ratio, the proportion of ancestral/Derived alleles, and the distribution of minor allele frequencies based on these selected disease-genes from HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and a set of known actionable genes across 20 world-wide ethnic-specific data.

1.5 Dissertation Outline

In this current chapter, we continue reviewing relevant literature. This project is divided into three important parts represented by each chapter as follow:

Chapter 2 covers the simulation of a total of 100 DNA sequence samples (of which each set of 50 samples of low and high sequence coverage mimicked the genetics background of African and European, respectively) at different coverage. This chpater provides also a brief overview of current DNA sequence simulations tools and the general simulation processes.

Chapter 3, assesses and evaluates nine different state-of-the-art variant calling tools on the African and European simulated DNA sequence data resulted from chapter 2. This chapter illustrates the characteristics and specifications of each variant calling tools used in this project. Furthermore, the evaluation metrics of these tools are also discussed.

Chapter 4, discusses the application of the suggested best variant calling tool as per the obtained result from chapter 3, on WGS public data grouped into 20 different ethnics. This chapter presents downstream analyses of the variants discovery from WGS and discusses the secondary finding of 20 world-wide ethnics groups. Furthermore, it presents the analysis of genetics diversity, heterozygosity ratio, proportion of ancestral/Derived alleles, and the distribution of minor allele frequencies based on sets of selected known disease-associated variants from four top African burden diseases include HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and a set of known actionable gene in 20 world-wide ethnic population groups.

Finally, the overall discussion and conclusion are in Chapter 5.

1.6 Overview on Next Generation Sequencing Downstream Analysis

Next Generations Sequencing has revolutionised genetics by massive parallelisation, resulting in a tremendous amount of short DNA fragment in a short time (Mardis, 2008; Metzker, 2010). Once the library preparation, cluster amplification, and DNA sequencing are done, the researchers are confronted with an enormous amount of raw data (Pabinger et al., 2014). Analysing NGS raw data will help us to provide a molecular diagnosis of diseases, and it consists of five explicit steps below and **Figure** 1.1 illustrates the overall process:

1. **Quality assessment of NGS**: it is the first step of ensuring that the raw FASTQ files which are generated from the sequencing platforms are in an excellent quality prior alignment to a reference genome. The FASTQ files that do not meet the defined standard should either be removed, trimmed or filtered. There are various quality control tools to assess NGS reads quality; one of these tools is FastQC (Andrews et al., 2012), which provides diagnosis reports and summary graphs and export the results to an HTML report. FastQC is able to calculate the Phred quality score that is distributed along with the reads, and calculating the mean of GC content distribution, read length and duplication level. FastQC allows detecting possible over-represented reads, which may be caused by contamination in the primer or adopter (Bao et al., 2014).

2. **Read alignment to a reference genome**: It is necessary to ensure the accuracy of the reading alignment as it plays a crucial role in identifying variations. Reads are either aligned to a reference genome or without a reference which is known as de novo assembly. There are numerous alignment algorithms, such as hash tables, suffix/prefix tree and Burrows-Wheeler Transform (BWT) algorithm (H. Li & Homer, 2010). In this project BWA alignment tool will be used, it is one of the most widely used aligner based on suffix/prefix tree (H. Li & Durbin, 2009, 2010).

3. **Variant identifications**: It is one of the most important steps in the downstream analysis; it is the process of identifying variants that are different from the reference data. These variants may cause mutations, so it is essential to call the true positive variant by using an accurate variant calling tool. This step is considered as a challenge since there are many variants calling tools with different calling algorithms such as Heuristic approach and statistical approach. There are two different variant calling tools, somatic/tumor or germline (inherited) caller. We will expatiate on variant calling step further in much more details in Section 1.7.

4. **Annotations and Prioritising mutations**: After variant calls are generated, we need to perform mutation prediction and prioritisation to be able to understand the biological and the functional insight within the generated data (H. Yang & Wang, 2015). Annotation tools have been evolved to aid in filtering and prioritise different kinds of variants such as, SNPs, Indels and CNV in-order to predict potential mutation that causes diseases (Pabinger et al., 2014) and characterise their biological functions.

In like manner, annotations tools integrate detail and information such as minor allele frequency (MAF), clinically significant variants, deleterious prediction of variant function to gather more information about variants (Bao et al., 2014) from public databases,. Most studies focus on non-synonymous SNVs, Indels in the protein-coding regions, as it is the cause of 85% of the discovered disease causing mutation (Botstein & Risch, 2003; Rabbani, Tekin, & Mahdieh, 2013; Bao et al., 2014).
ANNOVAR is one of the annotation tools that prioritise candidate genes. It is a fast and

efficient tool that can perform gene-based, region-based and filter-based annotation of the functional consequences of genetic variant (K. Wang, Li, & Hakonarson, 2010). ANNOVAR has the ability to calculate the score for all popular annotation software such as SIFT (Ng & Henikoff, 2003), Polyohen2 (Adzhubei et al., 2010), MutationTaster (Schwarz, Rödelsperger, Schuelke, & Seelow, 2010), MutationAssessor (Reva, Antipin, & Sander, 2011), FATHMM (Shihab, Gough, Cooper, Day, & Gaunt, 2013), VEST (Carter, Douville, Stenson, Cooper, & Karchin, 2013), CADD (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2018), and GERP++ (Davydov et al., 2010). Each step of the downstream analysis are essential to ensure getting the true and accurate result to detect mutations. However, it relies mainly on the sensitivity and the accuracy of discovering the variant from VC tools. However, in our project, we aim to assess and focus on variant identifications on the African versus European populations.

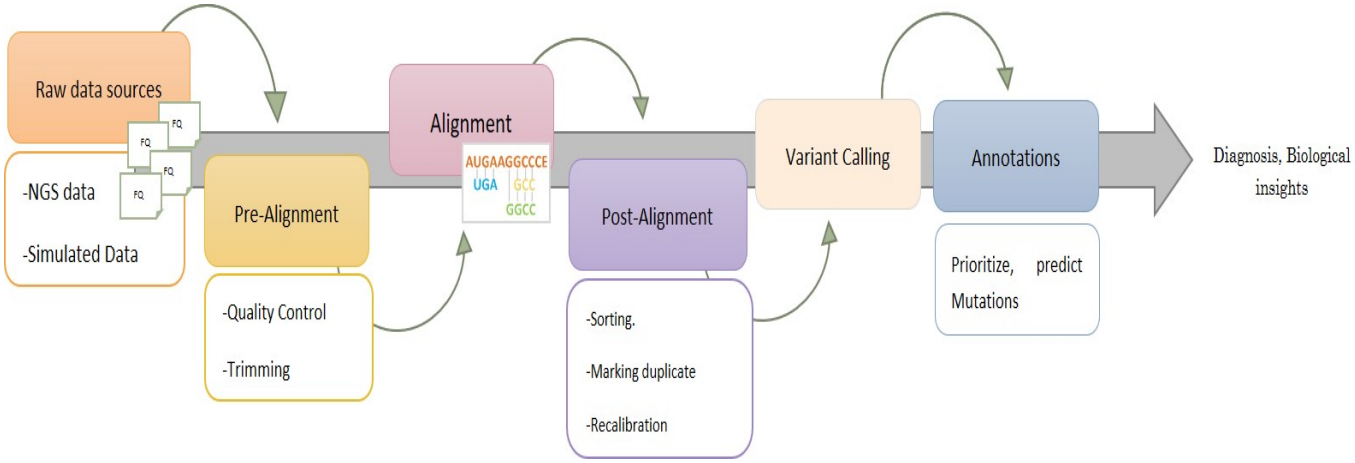


Figure 1.1: An overview on the WGS Data Analysis Pipeline.

Figure 1.1 presents a simple and general overview on the process of analysing the generated raw data, it may differ with some researchers whether to include more steps or less, first step is pre-alignment which include quality control and reads trimming, the second is alignment to a reference genome, third is after-alignment which include: sorting the reads, marking duplicate, and recalibration, after this reads are ready for variant calling, last but not least comes the annotation and finally the diagnosis and result.

Despite the major genetics differences and variations between various populations, the African populations yield the highest level of genetics variations among them, yet, it is still under-represented. While most of the next-generation sequencing and downstream data analysis tools are using European genetics data, this may affect the identification of population-specific variants associated with diseases or variable traits (Campbell & Tishkoff, 2008). Additionally, there are many variant calling tools and considering that majority of NGS downstream analysis tools has been performed and tested in European sequence data (Campbell & Tishkoff, 2008), one can be confused to which one of these current tools can have the highest accuracy with low false positive and false negative rate, especially when dealing with African sequence data sets. However, it is critical to note that discovering population-specific variants associated with diseases in the African populations would enable a range of applications from medical population genetics to precision medicine.

A number of previous studies has evaluated variant calling tools on specific populations, yet rare or little evaluated them on African populations. Here in this project, we will evaluate and assess nine different variant calling tools on both African and European populations at

different coverage depth. According to the review done by (Mielczarek & Szyda, 2016), that an accurate variant calling requires high sequencing depth. In addition to this, it enables to utilise the information on the quality of individual reads, therefore providing measures of uncertainty on every predicted variant.
In this project we have chosen the most known and open-source variant calling tools, which are VarScan2 (Koboldt et al., 2012), SAMtools (H. Li et al., 2009), GATK-HaplotypeCaller (McKenna et al., 2010), SNVer (Wei, Wang, Hu, Lyon, & Hakonarson, 2011), BCFtools (H. Li, 2011), FreeBayers (Garrison & Marth, 2012), Lofreq (Wilm et al., 2012), PlatyPus (Rimmer, Phan, Mathieson, Iqbal, & Twigg, 2016) and VarDict (Lai et al., 2016), see (**Table** 3.1), we excluded other variant calling tools as we follow (Sandmann, De Graaf, et al., 2017) exclusion criteria which is either they are using same tool or they required matched sample or calling Indels. we also excluded somatic variant calling tools.

1.7 Variant Calling

After finishing with Next Generation Sequencing, researchers and bioinformaticians have faced challenges with enormous amount of raw data (FASTQ file), which contain the DNA sequence of individuals. The FASTQ file needs to be aligned to a reference genome file which will produce SAM/BAM file. After Alignment, the SAM/BAM file will contain genetic differences that need to be identified; this step is called variant calling (VC), which is represented in the Variant Call Format (VCF) file.

The genetic variations and polymorphisms are identified by variant calling tools, and these variations are classified into different categories, such as somatic variants, germline (inherited) variants, structure variants (SVs) and copy number variants (CNVs). Moreover, these variations differ and rapidly evolve beyond single nucleotide polymorphisms (SNPs), to another different kind of complexity such as short insertions and deletions (Indels), short tandem repeats (STRs) and multi-nucleotide polymorphisms (MNPs) (Tan, Abecasis, & Kang, 2015). If these genetics variations are well defined and accurately annotated, they can help identifying mutations of which information can improve clinical diagnosis. of course, not all genetics variations are diseased or fatal; some of these variations make individuals unique and different from one to another.

1.7.1 Algorithms used to call the variant

Numerous Algorithms have been designed to detect and discover variants. There are two main approaches applied by current tools to discover variants, the heuristic approach and the statistical approach.

1.7.1.1 Heuristic approach

The heuristic algorithm can resolve the genotypes of the normal and tumor samples depending on reads quality, coverage depth and allele frequency, all along with the statistical significance to detect variants (Koboldt et al., 2012). One of the variant calling tools applying heuristic methods is VarScan2, along with Fisher's exact test can call somatic variants and germline variants (Koboldt et al., 2012). Varscan2 can differentiate between somatic variants and germline variants by comparing between normal and tumor genotypes. if the variants are in both genotypes are called somatic, while if it heterozygous in normal and homozygous in tumor are called as Loss of heterozygosity (LOH), and if the variants are shared among samples are called as germline

variants (Koboldt et al., 2012). The heuristic method requires high computational demands, in effect, it not usually applied by other variant calling tools (Mielczarek & Szyda, 2016).

1.7.1.2 Statistical approach

Genotypes calling by the probabilistic method is the calculation of the statistical uncertainty for the called genotype, and it adopts Bayes' theorem (Mielczarek & Szyda, 2016). This algorithm is implemented by GATK variant caller (McKenna et al., 2010). It calculates raw genotypes likelihoods using a Bayesian model (Equation 1.1) used in GATK tool, showing the probability of a candidate genotype $P(R_i|G_l)$, for a diploid genotype G_i, consist of one copy of allele A_1, and one copy of allele A_2, it indicates the mean read likelihoods for alleles in specified genotype (Poplin et al., 2017).

$$P(R_i|G_l) = \prod_i (\frac{P(R_i|A_1)}{2} + \frac{P(R_i|A_2)}{2}) \tag{1.1}$$

Furthermore, to calculate the posterior probability of each candidate genotype $P(G_l|R_i)$ is by using Bayesian mode to marginalise the raw genotyps likelihoods $P(R_i|G_l)$ (Equation 1.2)

$$P(G_l|R_i) = \frac{P(G)P(R_i|G_l)}{\Sigma_l P(R_i|G_l)P(G_l)} \tag{1.2}$$

The numerator consists of the product of the prior probability of a genotype $P(G)$ and a raw genotype likelihood divided by the sum of the likelihoods of all the possible genotypes for the set of alleles called in the variant (McKenna et al., 2010).

1.8 Previous Studies Comparing Variant Calling Tools

There have been several studies that investigated and compared different variant calling tools on their own parameters and population, but, none of them has used African sequence data set or related simulation data set in doing so. Some of these benchmark studies differ in characteristics (whole genome, whole exome), (somatic variant or germline variants) and evaluation parameter but, most of these studies agreed on detecting true variants with high accuracy and specificity.

Furthermore, many authors compared their tools on different depth of coverage reads, as these are important parameters to detect variant from NGS reads. The higher the read depth coverage, the more confident base calls, and more distinguishable from the sequencing error (Cheng, Teo, & Ong, 2014), accordingly, many studies used high coverage read sequence depth as it improves the accuracy of variants calling. Other studies consider variant filtering to be a suggested step as it could improve the specificity and sensitivity, and reduce false-positive rate (Spencer et al., 2014).

Numerous studies such as (Zook et al., 2014; Laurie et al., 2016; Kumaran, Subramanian, & Devarajan, 2019) and others have bench-marked variant calling tools using the set of NA12878 Genome in a Bottle (GiaB) high confidence GRCh37 variants as a gold standard reference set, whereas other previous studies include (Liu, Han, Wang, Gelernter, & Yang, 2013), has used WES data to benchmark VC tools, while other studies used simulated data as a gold stander. While some authors evaluated VC tools with different coverage to see which tool performs best, (Stead, Sutton, Taylor, Quirke, & Rabbitts, 2013) suggested that VaraScan2 perform very well with different sequence depths, as they were assessing sequence depth that is required to detect low-allelic variants.

(Pabinger et al., 2014) have surveyed more than 60 variant calling tools and compared 9 VC tools, somatic and germline callers respectively, they suggest to use several variants calling tools for a better resul. (Bao et al., 2014) recommended to apply multiple tools to call the variant in order to reduce false-positive and increase sensitivity. Additionally, (O'Rawe et al., 2013) have evaluated VC tools on both whole exome and wholes genome sequence data, suggesting to study larger multi-generational families in order to increase the accuracy of detecting *de-novo* variants.

Several studies have focused on detecting the somatic variants (known as low-frequency variants) to discover cancer mutations, they have compared different somatic variant calling tools such as VarScan2 (Koboldt et al., 2012), GATK haplotype caller(McKenna et al., 2010) (Q. Wang et al., 2013; Roberts et al., 2013; Spencer et al., 2014; H. Xu, DiCarlo, Satya, Peng, & Wang, 2014; Alioto et al., 2015; C. Xu, 2018) (they increase sequence depth up to 100 and apply PCR free methods which show significant benefits).
Apart from this, (Pirooznia et al., 2014) have compared two variant tools with realignment and recalibration steps, they suggested realignment/recalibration increase further the accuracy of the call.
Similar to other studies such as (Yu & Sun, 2013; Liu et al., 2013; Cornish & Guda, 2015; Laurie et al., 2016) founding that GATK UG has yielded high-quality variant calls outperforming others. In contrast, studies in (Yi et al., 2014; Pirooznia et al., 2014; Warden, Adamson, Neuhausen, & Wu, 2014) suggested that GATK HC was the best. Despite several studies suggested to use GATK; however, the unfavorable thing with GATK is the long runtime (Warden et al., 2014; Talwalkar et al., 2014; Huang et al., 2015; Laurie et al., 2016; Sandmann, De Graaf, et al., 2017; Z. Li, Wang, & Wang, 2018). Also, another downside with GATK is the inability to detect low-frequency variants (Cheng et al., 2014) of which might characterised populations with high diversity and with low level of linkage disequilibrium. In contrast, (Hwang, Kim, Lee, & Marcotte, 2015) suggested to use Samtools when dealing with SNP and GATK HC when calling Indels. Among all these studies, (Huang et al., 2015) have evaluated different tools by using pooled samples and they conclude that LoFreq(Wei et al., 2011) had high sensitivity and low false-positive variants.
Alternatively, other studies such as (Bao et al., 2014; Pantano, 2016; Laurie et al., 2016; Said Mohammed et al., 2018) recommended to use FreeBayers. Whilst, others recommend Vardict such as (Sandmann, De Graaf, et al., 2017). The same study (Sandmann, De Graaf, et al., 2017) concluded that LoFreq, VarDict, FreeBayers (Garrison & Marth, 2012) and SNVer have had detected variants with low allele frequency.

Overall, large data such as Whole-genome sequencing may be a challenge to filter out false positive, one of the new emerging trend to handle such a huge size is applying machine learning algorithms (Zook et al., 2014), several recent new tools used this algorithm such as decision-tree-based methods such as FUWA (Z. Li et al., 2018), random forests such as SNooPer (Spinella et al., 2016) and many others, which can learn the excellent way to filter out genotypes error. Finally, among all these previous studies, and to our best of knowledge, there is limited or no studies have compared different variant calling tools on the African populations. Study as (Cheng et al., 2014) who did Southeast Asian Malays population-based sample to analyse and assess VC tools, are a good example to follow and apply on the African populations. Since there no agreement on which tool to use, and with so many recommendations, one has confused. In this study, we will illustrate and evaluate nine different variant calling tools on African population whole-genome data. In our project, we excluded indels, somatic and structural variants since it required a different set of algorithms.

1.9 Evaluation Metric of Variant Calling

As suggested and agreed by many studies that sensitivity and specificity are one of the key metrics of evaluating variant calling tools. There are different evaluating metrics such as sensitivity, positive predictive value (PPV), false-positive rate (FPR), F-score, and the Transition/transversion (Ti/Tv) ratio.

False Negative and False Positive

A variant calling tool can classify variants whether it is positive or negative depending on the caller, and depending on the benchmark validation dataset can confirm if it is true-positive/negative or false-positive/negative variants (Talwalkar et al., 2014).
The sensitivity (also known as recall) of the variant calling tool, is the measurement of the actual true positive in the total called variants, and as shown in Equation 1.3 sensitively is calculated by the proportion of true positives (TP) divided by condition positive (TP+FN).

$$Sensitivity(Recall) = \frac{TP}{TP + FN} \tag{1.3}$$

Many techniques and steps are important to increase the sensitivity from pre-data processing such as increasing sequence depth, till post- variant calling step as filtering variants. Furthermore, as suggested by (Depristo et al., 2011) who proposed GATK-Best Practices Workflows, recalibration, local realignment and marking reduplicate are equally essential. Whereas Positive Predictive Value (known as precision), as shown in Equation 1.4, is calculated by dividing TP by the total of TP and FP. Moreover, a higher rate of false-positive and false-negative may be caused by Short Reads sequencing (SRS)(Caspar et al., 2018).

$$PPV(PositivePredictiveValue) = \frac{TP}{TP + FP} \tag{1.4}$$

Other metrics are as equally important and can be calculated as fellow:
False Positive Rate (FDR), is the ratio of false-positives to all total variant call as shown in Equation 1.5, researchers as (Farrer, Henk, MacLean, Studholme, & Fisher, 2013) have benchmarked the FDR to compare SNPs result.

$$FPR(FalsePositiveRate) = \frac{FP}{FP + TN} * 10^6 \tag{1.5}$$

F-score is known as the harmonic mean of sensitivity and the positive predictive value, the higher the F-score, the higher the accuracy. It can be calculated as shown in Equation 1.6.

$$F - score = \frac{2TP}{(2TP + FP + FN)} \tag{1.6}$$

One of the Post-Alignment steps is to remove duplicate reads, by filtering alignment from PCR amplifications that introduce duplicate reads which may lead to call false positive variants (Hintzsche, Robinson, & Tan, 2016). Furthermore, (Farrer et al., 2013) concluded that read-trimming has decreased the per cent of false-positive SNPs (Durtschi et al., 2013).

Transition/transversion ratio

Almost all of these studies have compared the tools by evaluating their sensitivity, specificity and Ti/Tv ratio. The higher the ratio of Ti/Tv, the higher the accuracy of variant call tool, moreover Ti/Tv ratio may help in evaluating novel SNPs (Wei et al., 2011).

Table 1.1: Comparisons of previous studies for different variant calling tools.

Study	Data set	Sequencing Platforms	Sequence Read Aligner	Variant Caller	Benchmark Methods	Reported Best Tool or pipeline
(Bauer & Bauer, 2011)	Yoruba trio from 1000 Genome Project	Illumina Hiseq	BWA CASAVA1.8	GATK CASAVA1.8	Novel SNP rate, Ti/Tv rate , het/hom ratio	CASAVA1.8
(O'Rawe et al., 2013)	Families of human research precipitants ascertained in clinic at the university of Utah	Illumina Hiseq2000 MiSeq	BWA SOAP2	GATK-UG SAMtools SNVer SOAPsnp GNUMAP pipeline	Ti/Tv ratio, Sensitivity, Specificity	Multiple pipeline
(Stead et al., 2013)	Simulated and Real Data from Horizon Discovery, Cambridge, UK	Illumina	BWA	LoFreq VarScan2 SNVer Bcftools GATK-UG Atlas2	Matthew's correlation coefficient (MCC) Receiver Operator Characteristic (ROC)	VarScan2
(Yu & Sun, 2013)	WGS from pilot1 -1000Genome Project	454 sequence data	BWA SOAP2	SOAPsnp Atlas-SNP2 SAMtools GATK-UG	Coverage cutoff, empirical positive calling rate, sensitivity	GATK-UG
(Liu et al., 2013)	Infinium HumanExome v1.1 Beadchip	Illumina Sanger sequencing for validation	BWA	Atlas-SNP2 glftools SAMtools GATK-UG	The number of SNPs, Rediscovery rate, Specificity, sensitivity , Ti/Tv ratio	GATK-UG
(Zook et al., 2014)	Pilot Genome in a bottle Consortium (GiAB) , CEU from 1KHG from National instate of standerds and Technology (NIST) ,Complete Genomic , X prize, Broad, illumine, Life Technologies	Illumina Gallx 454 SOLiD 4 Complete Genomic Illumina HiSeq Ion Torrent	BWA Ssaha2 Novoalign CASAVA Lifescope CGTools Tmap BWA-MEM	GATK-HC GATK-UG FreeBayes SAMtools	Ti/Tv ratio, Sensitivity, Specificity , False Positive Rate, Receiver Operator Characteristic (ROC)	-

Continued from previous page

Table 1.1 – Continued from previous page

Study	Data set	Sequencing Platforms	Sequence Read Aligner	Variant Caller	Benchmark Methods	Reported Best Tool or pipeline
(Cheng et al., 2014)	Real WES data from Singapore Sequencing Malay Project (SSMP)	Illumina HiSeq 2000	BWA	GATK-UG, SAMtools, Consensus Assessment of Sequence and Variation (CASAVA), VarScan, glfTools SOAPsnp	Sensitivity, False-Positive, Genotype calls concordance, Variant Accuracy	CASAVA
(Pabinger et al., 2014)	Real WES data from KTS (KohlschütterTönz Syndrome), Simulated data	Illumina HiSeq 2000	BWA	CRISP GATK-UG SAMtools SNVer VarScan2 SliderII	Number of SNPs,True Positive	Advising using serval VC tools
(Yi et al., 2014)	Illumina exome-seq dataset	Illumina Genome Analyser IIx	Eland BWA	GATK-UG GATK-HC SAMtools VarScan2 CASAVA CLCBio	Ti/Tv ratios, specificity and sensitivity	GATK-HC
(Pirooznia et al., 2014)	WES from Bipolar disorder from National institute of mental health (NIMH)	illumina Hiseq-2000, Sanger sequencing for validation	BWA	SAMtools GATK-HC GATK-UG	specificity and sensitivity	GATK-HC

Continued from previous page

Table 1.1 – Continued from previous page

Study	Data set	Sequencing Platforms	Sequence Read Aligner	Variant Caller	Benchmark Methods	Reported Best Tool or pipeline
(Bao et al., 2014)	Genome in a bottle Consortium (GiAB), Simulated Data	illumina	BWA Nonoalign Bowtie2	GATK-UG SAMtools Atlas2 FreeBayers	specificity and sensitivity and Precision rate	Novoalign + FreeBayes Pipeline
(Warden et al., 2014)	European Nucleotide Archive, SRP019719 exome data, 1000 Genome Project	Illumina GAII	BWA	GATK-HC GATK-UG VarScan2	Sensitivity, Recovery rates, false discovery rate(FDR), Accuracy	GATK-HC
(Highnam et al., 2015)	Genome in a bottle Consortium (GiAB), Simulated Data	Illumina	BWA BWA-MEM Nonoalign Bowtie2	SAMtools GATK-HC GATK-UG Isaac	Receiver Operator Characteristic (ROC), Positive Predictive Value(PPV)	Nonoalign+ GATK-HC Pipeline
(Huang et al., 2015)	Simulated Data from ClinSeq and the 1000 Genome Project	Illumina Hiseq	Nonoalign BWA	GATK-UG CRISP LoFreq VarScan2 SNVer	Receiver Operator Characteristic (ROC), Sensitivity, False Positive Rate, Accuracy, Singleton sensitivity rate.	LoFreq
(Hwang et al., 2015)	Genome in a bottle Consortium (GiAB)	Illumina2000 Illumina2500 Ion Proton	BWA-MEM Nonoalign Bowtie2	SAMtools GATK-HC FreeBayers Torrent Variant Caller (TVC)	Precision-Recall curve, area under the precision-recall curve(APR) score	BWA-MEM + SAMtools

Continued from previous page

Table 1.1 – Continued from previous page

Study	Data set	Sequencing Platforms	Sequence Read Aligner	Variant Caller	Benchmark Methods	Reported Best Tool or pipeline
(Cornish & Guda, 2015)	National instate of standerds and Technology (NIST) Genome in a bottle Consortium (GiAB)	Illumina Hiseq 2000	BWA BWA-MEM Nonoalign Bowtie2 MOSAIK CUSHAW3	SAMtools GATK-HC GATK-UG FreeBayers SNPSVM	Positive Predictive Value (PPV), sensitivity	Nonoalign+ GATK-UG Pipeline
(Laurie et al., 2016)	National instate of standerds and Technology (NIST) Genome in a bottle Consortium (GiAB)	Illumina Hiseq 2000	BWA GEM3	SAMtools GATK-HC FreeBayers	specificity and sensitivity and F1 score	GATK-HC FreeBayers
(Sandmann, De Graaf, et al., 2017)	Real amplicon-based targeted sequence data, TrueSight DNA amplicon sequencing panel , Simulated Data	Illumina Hiseq Illumina NextSeq	BWA-MEM	SAMtools GATK-HC FreeBayers Platypus VarScan LoFreq SNVer VarDict	Positive Predictive Value (PPV), sensitivity and F1 score	VarDict
(Said Mohammed et al., 2018)	Simulated Data	Illumina	BWA-MEM	FreeBayers VarScan2 LoFreq VarDict	Sensitivity, specificity, Precision, false positive rate, Accuracy	FreeBayers
(Kumaran et al., 2019)	HapMap/1000 Genome Project, Genome in a bottle Consortium (GiAB), Simulated Data	Illumina Hiseq 2000	Novoalign BWA Bowtie2 MOSAIK SOAP	GATK SAMtools FreeBayes DeepVariant	F-Score,Receiver Operator Characteristic (ROC), false discovery rate (FDR), Calculatin TP,FP and FN, Ti/Tv ratios	Novoalign , BWA+ DeepVariant, SAMtools

1.10 Overview of Mutations in The African Populations

Africa considered the utmost sources of modern humans, and have the greatest genetics variation in the world. For this reason, it is crucial to study and investigate the admixture event and the pattern of ancestral migrations that led to genetic variations, by characterising these genetic variations in the African populations (Choudhury, Aron, Sengupta, & Hazelhurst, 2018). This will provide a fundamental understanding of how genes contribute to phenotypic variations, response to the pharmaceutical drugs, susceptibility of infectious diseases such as Tuberculosis (TB), malaria, HIV/AIDS that are considered as a major burden in Africa.

Likewise, the importance of understanding human historical background, biology and the differing distribution of diseases frequency by ancestry and geography (C. N. Rotimi et al., 2017), since the human demographical history, migrations, adaptation, population admixture and expansion, shape the genetic variation and disease susceptibility in Africa .
Furthermore, the urbanisation expansion in Africa has increased the prevalence of infectious diseases alongside with non-communicable disease in low- and middle-income African countries (Oni et al., 2015).

Moreover, the achievement of linkage studies and genetic association studies such as genome-wide association studies (GWAS) have shed light on the diversity of human genome. While, in Africa, research projects such as the Human Hereditary and Health in Africa (H3Africa) has prompted the studies of genetics and genomics, yet, current knowledge on African genetics and genomics is still at infant level. Because not only the discovery of the structural complexity of the genetic mutations in the African populations is critical, but also, it is important to return the secondary findings that are known as actionable genes as recommended by the American College of Medical Genetics and Genomics (ACMG). Therefore, it requires more investigations and studies to improve public health and diminish the heavy burden of genetic diseases in the African continent (Choudhury et al., 2018; Mboowa, 2019).

Here we will illustrate genes that are associated with each different diseases. In this section, we will give a brief background on each disease such as infectious diseases (ID): HIV/AIDS, Malaria, Tuberculosis, and non-infectious diseases: sickle cell disease and its related genes reported by previous researched studies, and we will also investigate previous studies on the secondary (incidental) finding (actionable genes) specifically for African population.

1.10.1 Communicable Diseases: Susceptibility of Infectious Diseases in Africa

Understanding the genetics of diseases susceptibility and diseases resistant is important in order to improve diagnosis, treatment, disease prevention, and health in general. As such an example, (Sirugo et al., 2008) is one of the previous genetic studies of Africa that reviewed most of diseases susceptibility and response to the vaccine and therapeutic, the authors have illustrated the genetic association and the genetic background of the most common infectious diseases in Africa.

1.10.1.1 Tuberculosis

Tuberculosis(TB) is an infectious disease caused by the bacteria called *Mycobacterium tuberculosis*, and it mainly affects the lung. TB is considered as one of the devastating diseases in the world, and according to the World Health Organisation (WHO) 10.0 million individuals are affected with TB worldwide, and the majority of them are African https://www.afro.who.int/health-topics/tuberculosis-tb. Consequently, Africa has the highest TB incident per capita, with high rate of HIV/TB co-infection (Dye et al., 1999; Wood et al., 2010). The co-morbidity of HIV and TB in Africa is common, which make it difficult to analyse TB in Africa, according to this it is important to investigate the genetics susceptibility of TB considering co-infection

with HIV. From many genes known to be associated with TB, we have collected 136 genes (supplementary information in the Appendix) from different literature mostly from (Sirugo et al., 2008) and from other gene-diseases databases as will be described in next method section. Genes like *SLC11A1* (Søborg et al., 2007), vitamin D receptor gene are associated with TB susceptibility, while other genetic association such as a novel in chromosome 11p13 causes resistance to TB (Thye et al., 2012), this was confirmed by GWAS study done by (Chimusa et al., 2014) on South African Colored (SAC) population. Furthermore, previous studies such as (Davila et al., 2008; Salie et al., 2015; Schurz, Daya, Möller, Hoal, & Salie, 2015; Schurz et al., 2018) have investigated the influence and association of toll-like receptor (TLR) family, which has an immune response against invading pathogens. Further, according to (Möller & Hoal, 2010), African were twice as likely to be infected with M. tuberculosis than individuals with European ancestry, which demonstrates the necessity for well-integrated investigations of African genomics information and variations.

1.10.1.2 Malaria

Malaria is caused by Plasmodium species parasite, and four species of Plasmodium infect humans *P. falciparum, P. malariae, P. ovale, P. vivax* and more recently *P. knowlesi*. In 2018, 93% of malaria cases occurred in Africa, sub-Saharan mostly`https://www.afro.who.int/health-topics/malaria`. Most of malaria mortalities are children(Kwiatkowski, 2005). Several genetic variations causes malaria resistance in Africa, such an example, the mutation in Hbs of the β globin gene, which leads to sickle cell disease, another is Duffy gene (Campbell & Tishkoff, 2008). Studies by (Sirugo et al., 2008; Jallow et al., 2009) have reviewed the genetic of malaria susceptibility genes association. For example, *CD40* ligand, *CD36*, Haptoglobin genes cause severe malaria in the African populations (Sirugo et al., 2008). Thus, it is crucial to investigate malaria infection susceptibility and resistance from an inter-ethnic African population groups. We have collected a list of genes associated with susceptibility and resistance of malaria disease and they are presented in (**Table** A.1) in the supplementary information in the Appendix.

1.10.1.3 HIV/AIDS in Africa

Acquired immunodeficiency syndrome (AIDS) is a lethal life-threatening disease caused by the human immunodeficiency virus (HIV). AIDS affected more than 35 million individuals worldwide, majority of infected region are in Africa `https://www.afro.who.int/health-topics/hivaids`. Many studies have investigated the genetic association of the susceptibility and resistance of HIV/AIDS (Sirugo et al., 2008; Joubert et al., 2010; Vannberg, Chapman, & Hill, 2011; Peer, 2015). Further (Picton, Paximadis, & Tiemessen, 2010) studied the association of polymorphic variation of Chemokine (CC motif) receptor 5 (CCR5) gene in HIV-1 infected/uninfected South African population. We have collected a list of genes associated with susceptibility and resistance of HIV/AIDS shown in (**Table** A.1) in the Appendix.

1.10.2 Non-communicable Diseases in Africa

1.10.2.1 Sickle Cell Disease

According to WHO https://www.afro.who.int/health-topics/sickle-cell-disease, Sickle Cell Disease (SCD) is considered as a major and common cause of illness and mortality in the world (Makani, Williams, & Marsh, 2007). Furthermore, the majority of newborns affected with SCD in the world are in Africa; as a result, it is considered as the main birthplace of sickle mutations (Diallo & Tchernia, 2002; Mboowa, 2019), as illustrated in **Figure** 1.2 obtained from (Piel et al., 2017).

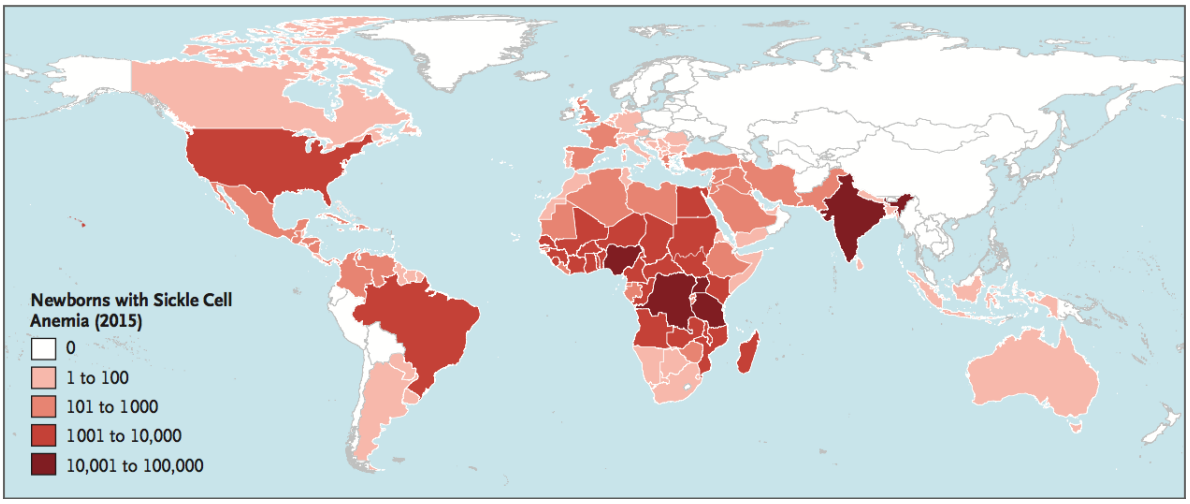


Figure 1.2: Worldwide Number of Newborns with Sickle Cell Anemia from (Piel et al., 2017).

SCD is a single gene disorder caused by a single nucleotide mutation ($HBB : c.20A > T$) in the human β globin gene (*HBB*) (located on chromosome 11p15.5), which gives rise to a hemoglobin structural variant (HbS) (David et al., 2018). The mutation in HbC and HbS, cause a substitution of glutamic by lysine or valine, respectively (Agarwal et al., 2000). The homozygosity for the globin S gene mutation (*HbSS*), is the most common genotype leading to SCD in Africa (Makani et al., 2007), occurs on five "classical" β haplotype backgrounds in ethnic groups of African ancestry. A number of previous studies observed the association of HbF with the *BCL11A* locus on chromosome 2p, and with a broad region around the c-Myb locus (called *HBS1L-MYB*) and within the β globin (Orkin & Bauer, 2019; Wonkam et al., 2019). Furthermore, the human leukocyte antigens *HLA DRB1*1302* haplotype and *HLA-B53* are associated with protection from both forms of the severe disease (Agarwal et al., 2000). Fetal haemoglobin (HbF) is an heritable trait that influences the clinical course of sickle cell disease. Study by (Wonkam et al., 2014) had investigated the relationship between HbF levels and the relevant genetic loci in 610 African patients with SCD. Furthermore, the causes of death in affected children are poorly documented, despite the high mortality associated with SCD in Africa, (Rees, Williams, & Gladwin, 2010). Retrospective design done by (Macharia et al., 2018), described the clinical epidemiology of SCA within a malaria-exposed among African populations, suggesting attention to be payed on SCD care and treatment to reduce high child-mortality. We have collected a list of genes associated with SCD and are displayed in the supplementary information in the Appendix section (**Table** A.1).

1.11 Secondary Finding (Actionable Genes)

In 2013, The American College of Medical Genetics and Genomics (ACMG) provided a recommendation and policy to return secondary finding in 56 genes associated with medically actionable conditions (Green et al., 2013; ACMG, 2015). Secondary finding (also known as accidental finding) provides information about genes unrelated to the primary cause of testing. Several studies such as (Berg et al., 2013) investigated actionable secondary findings following ACMG policy, and have been carried out on different populations. A study by (Dorschner et al., 2013) have identified actionable finding in 1,000 individuals (500 African and European, respectively), they return a result of 114 genes reported as actionable genes. They have classified the variants into four criteria, (1) pathogenic, (2) likely pathogenic, (3) variant of uncertain significance (VUS), (4) likely benign VUS. Furthermore, the variants that classified as pathogenic in the European population are expected to be pathogenic in other populations as the African; however, variants that do not occur in the European population are understudied (Dorschner et al., 2013). This may lead to miss some pathogenic variants in the African population; therefore, it is important to update additional African-specific actionable genes.

With the same objective, (Amendola et al., 2015) and (Olfson et al., 2015) both conducted secondary findings with the majority of the identified medically actionable genes found in the European populations, while the lowest found in the African populations. Some studies return secondary findings regarding specific populations, as the study by (Ploug & Holm, 2017), was on the Danish population, they suggested a new policy for reporting incidental findings (IFs) by performing a choice-based conjoint survey. Moreover, (Tang et al., 2017) have done the same project in the east Asian populations among 954 individuals.

Meanwhile, other studies have focused on disease-specific secondary findings, for example, (Tetzlaff et al., 2016) have returned secondary findings in sebaceous carcinoma, other like (Thompson et al., 2018) investigated on secondary findings of developmental delay and intellectual disability in children. There are many research initiatives that have recommended guideline and policies in returning IFs, such as, ACMG (ACMG, 2015) in the United States, EuroGenTest and the European Society of Human Genetics (Matthijs et al., 2016) in Europe, and the Association for Clinical Genetic Science (ACGS) in the United Kingdom (Wallis et al., 2013), yet there are no polices on reporting IFs (Bope et al., 2019) in Africa.

Unfortunately, the absence of major representation of genetic studies on African populations could lead to rule out some novel variants that may consider as a pathogenic actionable genes — taking into consideration that African descent populations have high diversity and genetic variations. Regarding this matter, many researches and enterprises in Africa have increased, such as the resent policy guideline on feeding back findings by the Human Heredity and Health (H3Africa). Moreover, the study report done by (Bope et al., 2019), provided list of available WGS/WES of the African genome data, reviewed in-silico prediction mutation tool, further, they have recommended several points, to benchmark variant calling tools using African populations, develop a reference panel specific to the African genome, in order to improve the clinical outcomes and overall health in the African continent. Regarding previous literature, in this project, we will investigate secondary findings of disease-specific (HIV/AIDS, Malaria, Tuberculosis (TB), and Non-communicable diseases such as Sickle cell disease) in 20 worldwide ethnic groups with focus on the African ethnic groups. We have collected specific gene lists from literature (Sirugo et al., 2008; Dorschner et al., 2013), GWAS Catalog, Actionable Genome Consortium (ACG), and gene-diseases database such DisGeNET as will be discussed in Chapter 4. The full lists of gene-disease pairs are in supplementary information in the Appendix, (**Table** A.1).

Chapter 2

DNA Sequencing Simulation

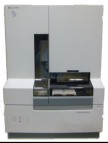
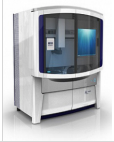
2.1 Introduction

The development of computational algorithms and bioinformatics tools are well-demanded needs, to handle, interpret and to perform analyses on enormous amount of large-scale raw DNA sequence data that yield from Next-Generation sequencing NGS platforms. This with hope to be able to provide answers to current diagnosis and precise treatment and prevention challenges. DNA sequence synthetic data works as a robust appliance to benchmark and to allow new development of bioinformatics algorithms and tools, as they can imitate sequencing error and mutations error, and working as a gold standard.
The benefits of using simulated data is to be able to develop, validate, and pinpoints weakness of bioinformatics tools and assess results (Myers, 1999). In addition, Since, simulated DNA sequence known as synthetic reads; thus, there is no ethical approval or security requirement and the simulation data are produced at a low cost. Moreover, simulation tools allow users to hold control on inserted parameters and expected variants (Holtgrewe, 2010). Furthermore, users can generate as much reads as desired, with predefined parameters of choice for which true values are known to allow validating result agaisnt true/golden values or data. consequently, the genetic and genomics simulated data have become increasingly popular to assess and validate the biological model in order to test a new hypothesis, help to design and extrapolate specific data set (Escalona et al., 2017).

2.1.1 Overview on Different NGS Simulation Tools

Many NGS simulation tools have been developed in the past few years. These tools vary in their features, functionalities and input requirement and their output. Furthermore, a given simulations tool differs to other on sequence parameter features or error profile and rate or it was designed for a sequencing platform, either it represents Illumina, Roche's 454 (454), Thermo Fisher's(SOLiD), Thermo Fisher's IonTorrent, pacific Bioscience (PacBio), Oxford Nanopore sequencing, Sanger sequencing, or either it was a platform-independent such as NEAT-genReads (Stephens et al., 2016). **Table** 2.1 demonstrates different NGS platforms and their features and their error profile and rate, which may be used as an input model for the simulation tools. Moreover, they also differ in several aspects, such as the type of reads, whither it single-end, paired-end or mate-pair, furthermore, the error model, coverage and the presence of genomics variants.

Table 2.1: Main characteristics of NGS technologies and examples of simulation tools representing these platforms (Escalona et al., 2017).

	Sanger	SOLiD	Illumina	Roche's 454	IonTorrent	PacBio	Nanopore
Platforms							
Sequencing principal	chain termination	sequencing by oligonucleotde ligations, detection	sequencing by synthesis	pyrosequencing	Proton detection, semiconductor sequencing	Real-Time sequencing	Single cell DNA template strand sequencing
Run type	SE, PE	SE, PE, MP	SE, PE, MP	SE, PE	SE, PE	SE	SE
Read length	900bp	75bp	300bp	700bp	400bp	14.000bp	9000bp
Error rate	0.001%	0.01-1%	0.003-1%	1.07-1.7%	1.78%	5-10%	10-40%
Error profile	-	nucleotide transition error	substitution error	Indel error	Indel error	high error rate	high error rate
Examples of simulation tools	Mason	FASTQsim	SInC	454sim	BEAR	SimLoRD	NanoSim
Reference	(Holtgrewe, 2010)	(Shcherbina, 2014)	(Pattnaik, Gupta, Rao, & Panda, 2014)	(Lysholm, Andersson, & Persson, 2011)	(Johnson, Trost, Long, Pittet, & Kusalik, 2014)	(Stöcker, Köster, & Rahmann, 2016)	(C. Yang, Chu, Warren, & Birol, 2017)

SE:Single End, PE:Paired End, MP:Mate Pair, bp:base pair

2.1.2 General process of DNA read simulation

Despite the different features of reads simulation tools, they all have numerous features in common with some exceptions. For example, the reference sequence, error profile indicate predefined parameters such as type of variations and error distribution, and last the output is either aligned or unaligned reads in different standard file format, such as FASTA, FASTQ, BAM, VCF and SFF (Escalona et al., 2017). **Figure** 2.1 illustrates the general processes of DNA read simulation.
These error models may differ in their biological features such as GC content, Indels, and substitutions, moreover differ in technological features representing various NGS platforms errors as indicating in (**Table** 2.6). They may differ also in controlling the inserted parameters such as read length, quality score and modelling PCR amplification and artefacts. Additionally, they also differ in using statistical algorithms such as using empirical error probabilities as in Mason (Holtgrewe, 2010), probabilistic model of biased sampling distribution as in FASIM (Hur et al., 2006), using configurable statistical models such as 454sim (Lysholm et al., 2011), or stochastic grammar to model tandem repetitive arrays, large scale duplication as in Celsim (Myers, 1999), and finally, as in Metasim using Yule-Harding model to generate phylogeny tree and Jukes-Cantor formula to estimate probabilities of the change in each base pair (Richter, Ott, Auch, Schmid, & Huson, 2011).

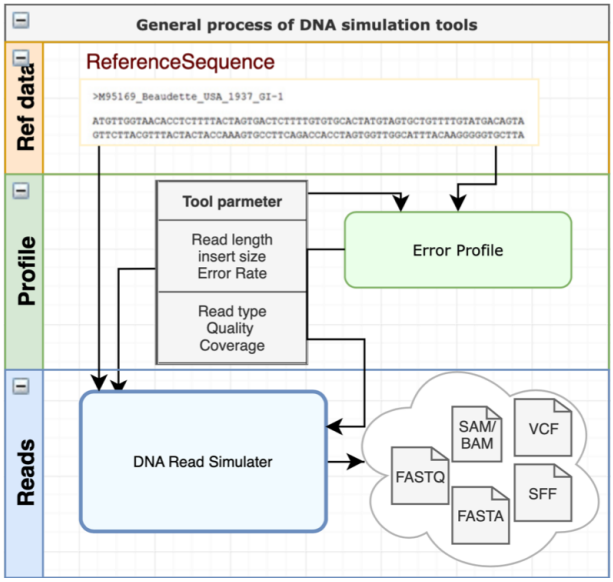


Figure 2.1: Overview of the general Simulation process.

As shown in **Table** 2.1, BEAR (Johnson et al., 2014) can simulate Ion torrent platform even it is a platform-independent, it also used to simulate metagenomic data. Another example of simulation tools that can simulate DNA sequence reads without a reference sequence is XS developed by (Pratas, Pinho, & Rodrigues, 2014). With such many simulations tools and features, one can ask which tools to use when simulating NGS reads. Well, there are two reviews one by (Escalona et al., 2017) and (Alosaimi et al., 2019), these reviews illustrate and define different DNA sequence simulation tools with a decision tree to allow users to make their choices.

In this research project, NEAT-GenRead (Stephens et al., 2016) was chosen; it is a WGS simulation tool. The software allows user to control the mutation model, sequencing error model, and adjusting parameters. It is platform-independent, can simulate SNPs, Indels and any ploidy. Also, it can accept variants as an input, it output Golden VCF and Golden BAM to benchmark other tools. Since, our study aims to choose best variant calling tools that have low FP, FN rate and works best with African populations, NEAT-genReads is the choice as we can mimic the African variation by using real African data as an input in the mutation model, as well as for the European populations.

2.2 Materials and Methods

To simulate DNA sequence samples, we have used NEAT-GenRead (Stephens et al., 2016). **Figure** 2.2 shows NEAT simulation pipeline, NEAT-GenRead is one of the NGS simulating tool written in python and it outputs in three different files: The forward and reverse Fastq and golden Bam and Golden VCF. It can mimic real data by using models learned from specific data sets. We have used two models from NEAT-GenRead which are mutation models (genMutModel.py) and sequence error model (genSeqErrorModel.py), (https://github.com/zstephens/neat-genreads.

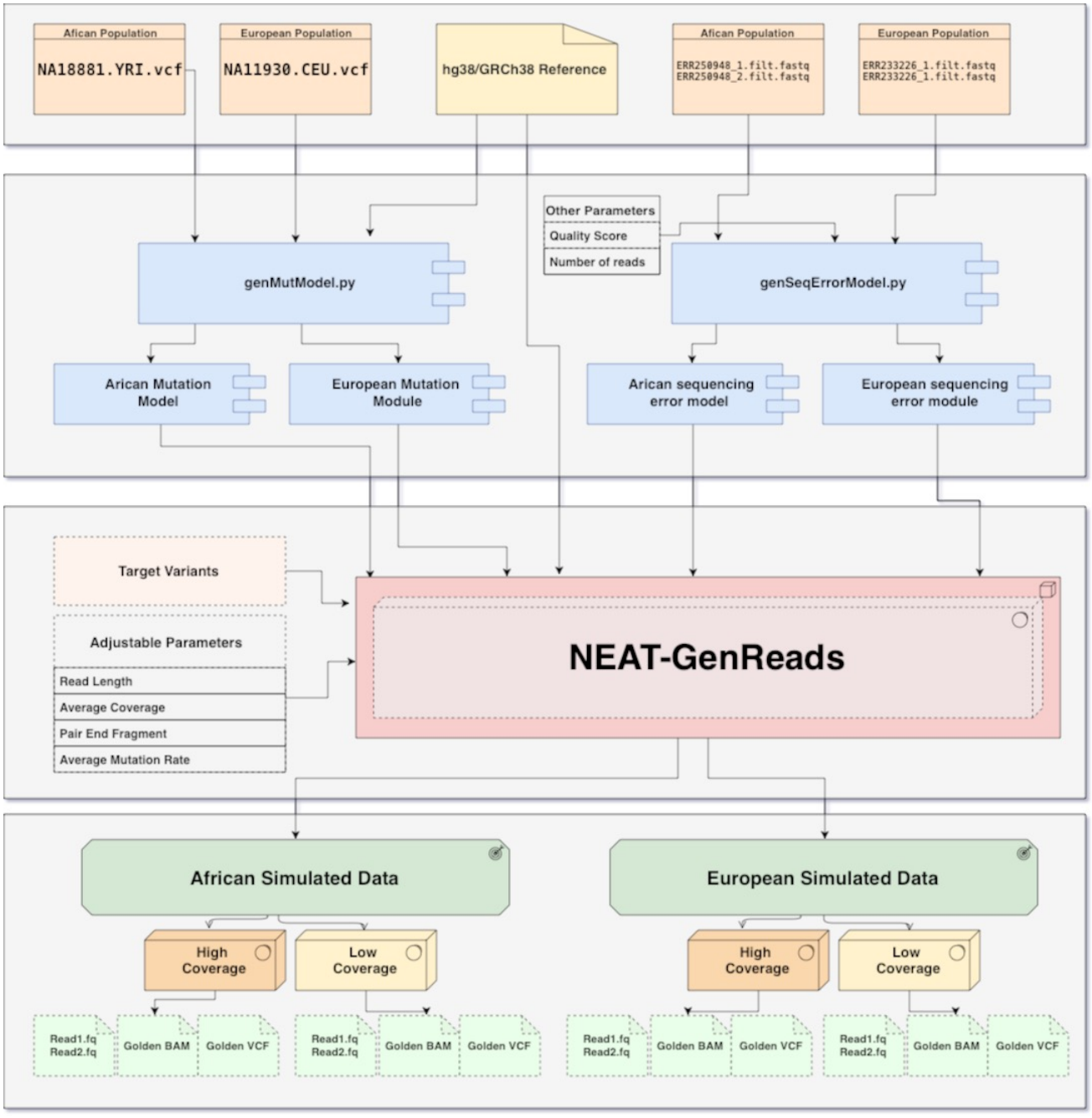


Figure 2.2: Sketch of the method used in NEAT-GenReads.

2.2.1 Data Description

2.2.1.1 Mutation Model

The data we have used is from the 1000 Genome Human Project (Consortium et al., 2012), which is publicly available, and no ethics approval was needed. To generate mutation models, we have used only one random sample (as suggested in NEAT manual) to mimic each population, for the African population we used one random sample (sample ID NA18881) from Yoruba in Ibadan, Nigeria (YRI), and for the European populations, we used a sample ID NA11930 from Utah Residents (CEU) with Northern and Western European Ancestry. Using a population-specific mutation model, we were able to make the result more accurate and confidently representing the population. We have chosen YRI and CEU as they are commonly been used a proxy for African and European populations, respectively.

2.2.1.2 Sequencing Error Model

The forward and reverse Fastq files are the raw files produced from the DNA sequencing platforms. As the same for the mutation model, we used Fastq file specific for each population. Since NEAT is platform-independent, we have chosen to mimic Illumina sequencing Single-pass error rate, and final error rate 0.1, which are mostly from SNPs substitution (Pfeiffer et al., 2018), basically using the same targeted sample, NA1888 for YRI and NA11930 for CEU.

2.2.2 NEAT-GenReads

We performed simulation by using NEAT-GenReads, we used hg38/GRCh38 the latest human reference genome as an input to be used for generating simulated reads. NEAT-GenReads needs target variant. As a result, we have used 50 African samples and 50 European samples as listed in **Table** 2.2. We extracted common variants from each individual by using a custom python script. The mutation rate was set at 0.1, to resemble Illumina sequencing platforms.

Table 2.2: Individuals from 1000 Genome Human Project, used for generating target variants.

African Proxy Population
HG01879, HG01880, HG01882, HG01883, HG01885, HG01886, HG01889, HG02461, HG02462, HG02464, HG02465, HG02561, HG02562, HG02567, HG02922, HG02923, HG02938, HG02941, HG02943, HG02944, HG02946, HG03052, HG03054, HG03055, HG03057, HG03058, HG03060, HG03061, NA18486, NA18487, NA18488, NA18489, NA18498, NA18499, NA18501, NA19023, NA19024, NA19025, NA19026, NA19027, NA19028, NA19030, NA19625, NA19700, NA19701, NA19703, NA19704, NA19707, NA19711, NA19712
European Proxy Population
NA06984, NA06986, NA06989, NA07037, NA07048, NA07051, NA07347, NA11840, NA11843, NA11893, NA11894, NA11918, NA11919, NA11920, NA11931, NA11932, NA11933, NA12045, NA12058, NA12275, NA12282, NA12286, NA12340, NA12341, NA12342, NA12347, NA12348, NA12399, NA12400, NA12413, NA12546, NA12716, NA12748, NA12749, NA12760, NA12775, NA12776, NA12777, NA12778, NA12827, NA12828, NA12829, NA12830, NA12842, NA12843, NA12878, NA12889, NA12890, NA12891, NA12892

Furthermore, we have generated two data-set (African and European) with two different depth coverage. Each population has 50 samples, of which 25 are high coverage, and 25 low coverage.

2.3 Results

One hundred samples successfully simulated, of which 50 representing the African population (25/25 samples with high/low coverage sequence) and 50 representing the European population (25/25 samples with high/low coverage sequence). NEAT-GenReads output forward and reverse FastQ file, golden BAM and golden VCF. We divided it into 4 data sets (AFR high, AFR low, EUR high and EUR low).
The NEAT-GenReads scripts were submitted by using the vertical cloud of the Centre for High-Performance Computing (CHPC) in the Republic of South Africa https://www.chpc.ac.za. The Resulted files are stored at CHPC for further manipulation and analysis. The simulations processes take up around 400 CPUh, almost over two weeks to be done.
As expected and known that African genomics data have more genetics variation than the European population. Hence, the simulation results have met the expectations; the African simulated golden VCF have more SNPs and variation than the European see (**Table** 2.2). Simulated data generated represent whole-genome sequencing (WGS).

2.3.1 Assessing and Examining Simulation Outputs

2.3.1.1 Quality Control Check on the Simulated Forward and Reverse FastQ files

We have checked the quality of the generated FastQ files by using FastQC (Andrews et al., 2012) https://www.bioinformatics.babraham.ac.uk/projects/fastqc/, to ensure that the reads of the simulated raw data are in a good quality prior alignment to a reference genome and to be able to use them for further processes as will be discussed in the next chapter.
We have run FastQC on all the resulted Forward and Reverse FastQ files, then we aggregate the result from FastQC into a single report by using MultiQC (Ewels, Magnusson, Lundin, & Käller, 2016) https://www.github.com/ewels/MultiQC. **Figures** 2.3-2.6 illustrates the report for reads quality control. The summary evaluations of FastQC have resulted into six modules. The first is the sequence counts for each sample (**A**) in **Figures** 2.3- 2.6. Second, comes the mean quality value across each base position in the read, quality values across all bases at all position in the FastQ files which seem to be very good quality calls (green), except for few samples in European High and low, has reasonable score (orange) (**B**) in **Figures** 2.3- 2.6.
Third, per Sequence Quality Scores among all samples are good. **C**) in **Figures** 2.3- 2.6 show the number of reads with average quality scores. These figures illustrates also if a subset of reads has universally poor quality. Per Sequence GC Content (**D**) in **Figures** 2.3- 2.6 show the average GC content of reads, even though, we did not use the GC-content model (was set as the default from the NEAT-GenReads). The mean of CG-content was 40% which has a roughly normal distribution of GC content among all samples.

(**E**) and (**F**) in **Figures** 2.3- 2.6 show the percentage of the calls at each position for whichever an N was called, (**F**) the relative level of duplication found for every sequence, respectively. The per base N content, usually in the simulated samples, the value is always zero as well as our result.
Furthermore, the sequence duplication in the resulted report is good among all samples, the high the level of duplication the more likely to indicate some kind of enrichment bias (e.g. PCR over-amplification).
Accordingly, we have decided to use the Fastq files for further analysis to call the variants by nine different variant calling tools and asses these tools by comparing the resulted VCF files with the Golden VCFs.

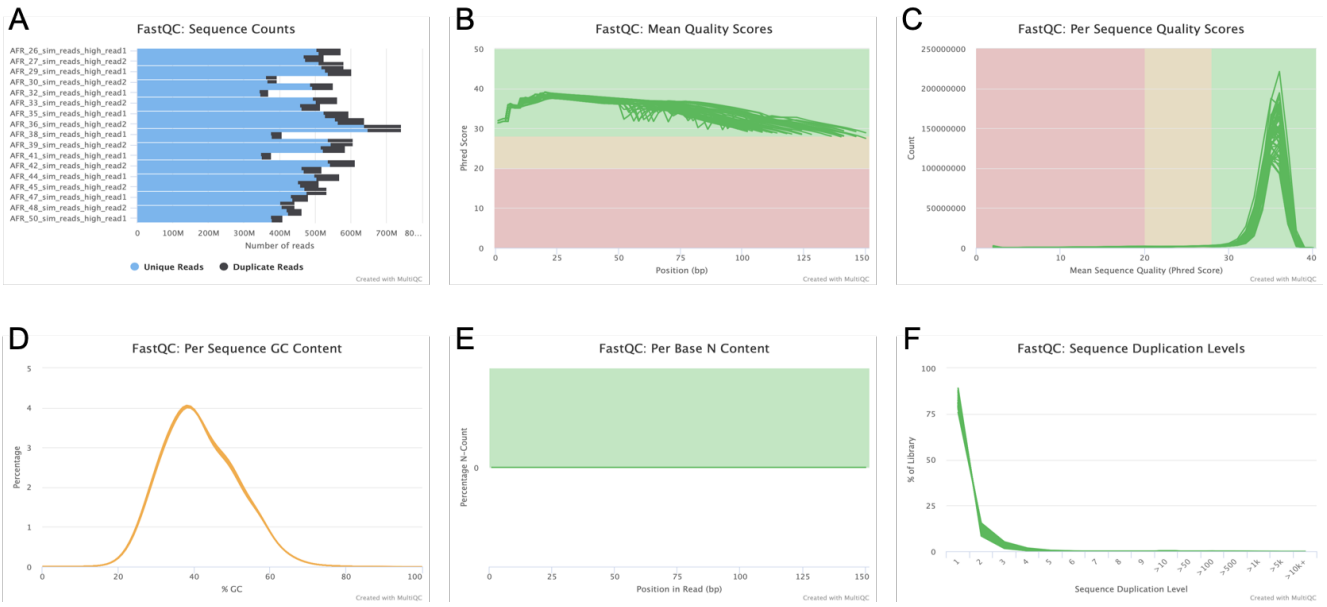


Figure 2.3: Aggregated Report from MultiQC of all the FastQC reports of the simulated African- High coverage samples.

FastQC report shows (A)Sequence Counts. (B)Mean Sequence Quality score. (C)Per Sequence Quality Scores. (D)Per Sequence GC Content. (E)Per Base N Content. (F)Sequence Duplication Levels.

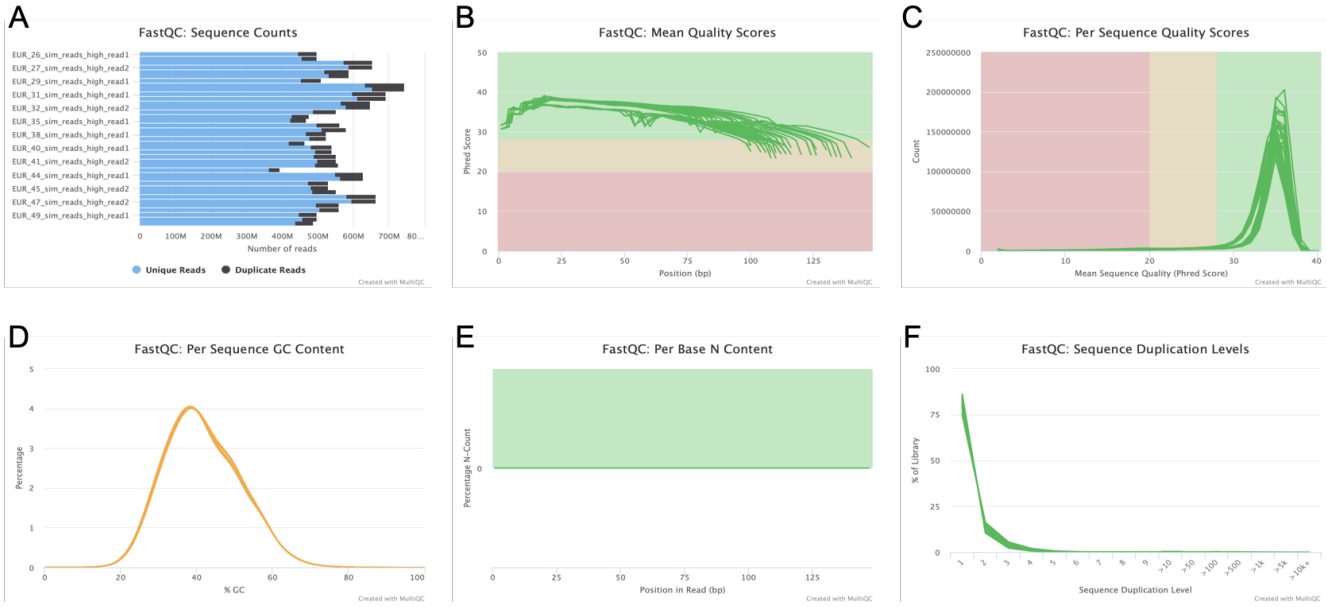


Figure 2.4: Aggregated Report from MultiQC of all the FastQC reports of the simulated European- High coverage samples.

FastQC report shows (A)Sequence Counts. (B)Mean Sequence Quality score. (C)Per Sequence Quality Scores. (D)Per Sequence GC Content. (E)Per Base N Content. (F)Sequence Duplication Levels.

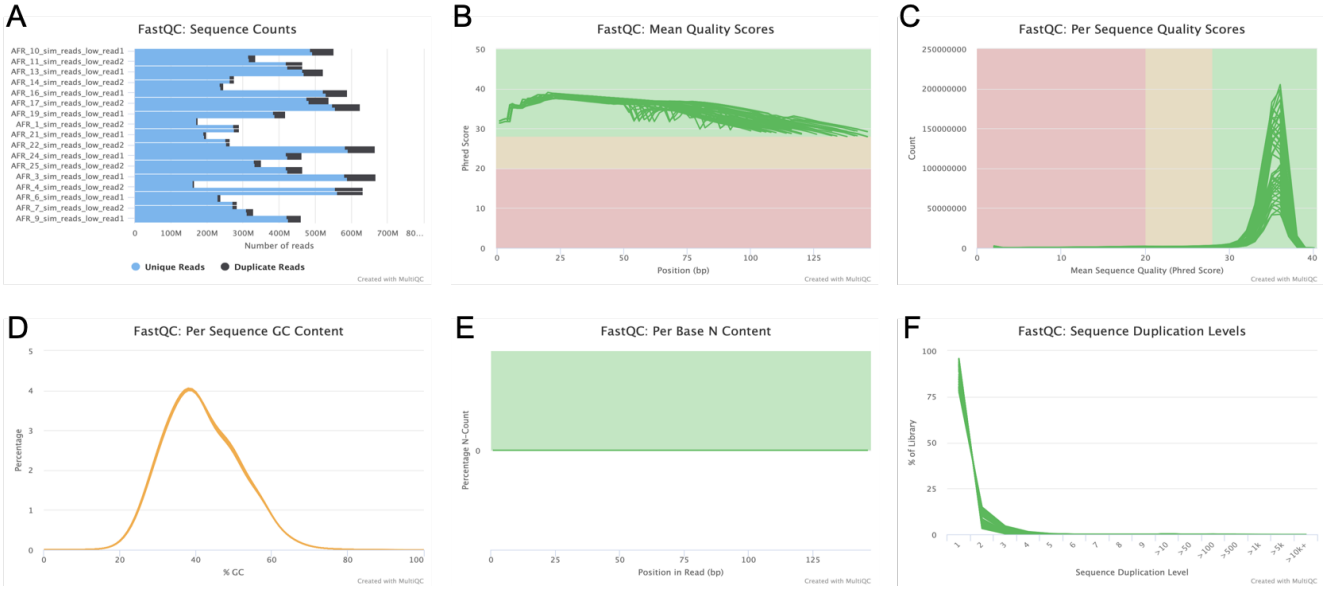


Figure 2.5: Aggregated Report from MultiQC of all the FastQC reports of the simulated African- Low coverage samples.

FastQC report shows (A)Sequence Counts. (B)Mean Sequence Quality score. (C)Per Sequence Quality Scores. (D)Per Sequence GC Content. (E)Per Base N Content. (F)Sequence Duplication Levels.

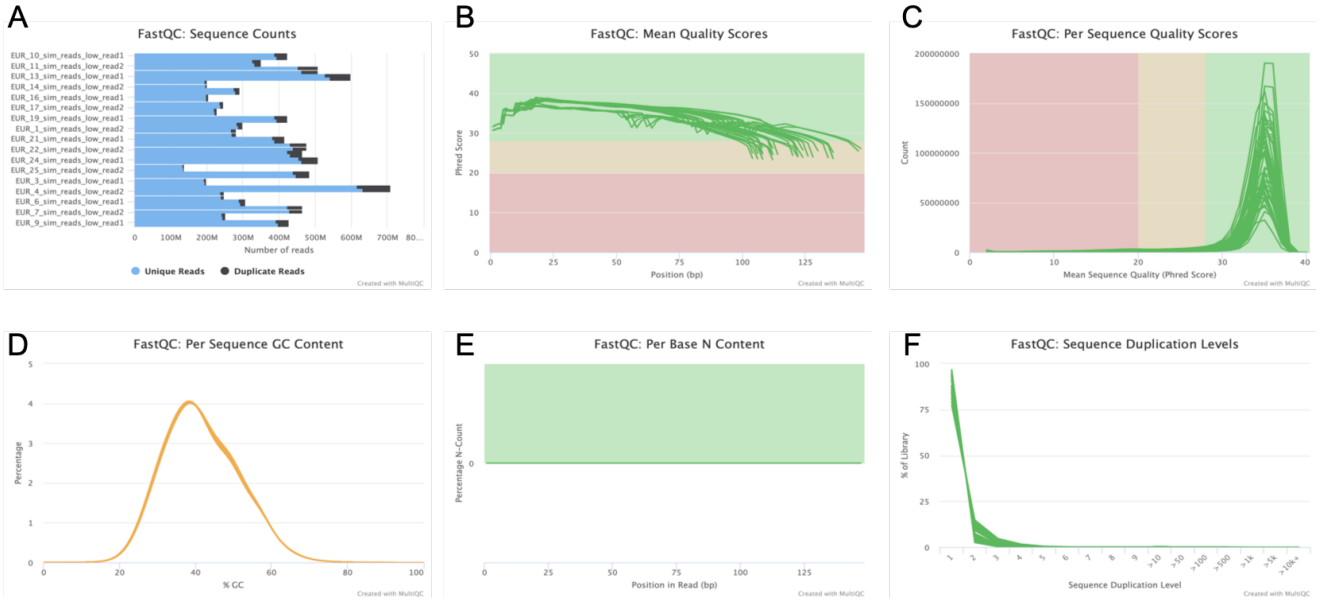


Figure 2.6: Aggregated Report from MultiQC of all the FastQC reports of the simulated European- Low coverage samples.

FastQC report shows (A)Sequence Counts. (B)Mean Sequence Quality score. (C)Per Sequence Quality Scores. (D)Per Sequence GC Content. (E)Per Base N Content. (F)Sequence Duplication Levels.

2.3.1.2 Golden BAM files

The total output was 100 Golden BAM files, 50 African(25 samples high coverage and 25 low coverage) and 50 European (25 samples high coverage and 25 low coverage). The generated Golden BAM files were good. Reads are sorted and coordinate; an example is shown in (**Figure** 2.7). We have examined the Golden BAM files once manually with Samtools falgstat, quickcheck to check if their are qualities are good, the result among all simulated golden BAM was good as indicated in (**Figure** 2.7b).

```
[salosaimi@login1 AFR]$ ${samtool} view -h AFR_11_sim_reads_low_golden.bam
@HD     VN:1.5  SO:coordinate
@SQ     SN:chr1 LN:248956422
@SQ     SN:chr2 LN:242193529
@SQ     SN:chr3 LN:198295559
@SQ     SN:chr4 LN:190214555
@SQ     SN:chr5 LN:181538259
@SQ     SN:chr6 LN:170805979
@SQ     SN:chr7 LN:159345973
@SQ     SN:chr8 LN:145138636
@SQ     SN:chr9 LN:138394717
@SQ     SN:chr10        LN:133797422
@SQ     SN:chr11        LN:135086622
@SQ     SN:chr12        LN:133275309
@SQ     SN:chr13        LN:114364328
@SQ     SN:chr14        LN:107043718
@SQ     SN:chr15        LN:101991189
@SQ     SN:chr16        LN:90338345
@SQ     SN:chr17        LN:83257441
@SQ     SN:chr18        LN:80373285
@SQ     SN:chr19        LN:58617616
@SQ     SN:chr20        LN:64444167
@SQ     SN:chr21        LN:46709983
@SQ     SN:chr22        LN:50818468
@SQ     SN:chrX LN:156040895
@SQ     SN:chrY LN:57227415
```

(a) Golden BAM header.

```
881392057 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
881392057 + 0 mapped (100.00% : N/A)
881392057 + 0 paired in sequencing
441091978 + 0 read1
440300079 + 0 read2
881392057 + 0 properly paired (100.00% : N/A)
881392057 + 0 with itself and mate mapped
0 + 0 singletons (0.00% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

(b) Golden BAM flagstat result.

Figure 2.7: An example of Golden BAM header and flagstat.

However, we could not use BAM files generated fronm NEAT directly for variant calling as their headers, particularly the group read (@RG) tag are the same among all files. @RG refers to a set of reads that were generated from a single run of sequencing platform. Some variant calling tools such as GATK requires unique @RG to be present in the VCF files, otherwise it will terminate with errors the @RG is absent. We have tried to add a new @RG with Picard tool (AddOrReplaceReadGroups), and Picard has failed to correctly replace group read. As a result, we have decided not to use golden BAM as direct input for variant calling. Therefore, we chose to use Fastq files as shown in (**Figure** 1.1).

2.3.1.3 Golden VCF files

100 Golden VCF files were generated, 50 African (25 samples with high coverage and 25 low coverage) and 50 European (25 samples with high coverage and 25 low coverage). Generated Golden VCF will be used to compare and assess the VCFs generated from 9 different state-of-the-art variant calling tools. **Table** 2.3 presents a summary of the variants in the generated VCF files.

Table 2.3: Total variants number present in the golden VCF files gnerated by NEAT-GenReads for African and European Populations.

	Variants numbers present in golden VCF			
Population	**African Populations**		**European Populations**	
Coverage	**High**	**Low**	**High**	**Low**
Number of samples	25	25	25	25
Variant number	1699695258	1719357176	1671991579	1706725018
Total variant in each population	3419052434		3378716597	

2.4 Brief Discussion and Chapter Summary

The DNA sequence simulation was done using NEAT-GenReads. NEAT-GenReads is an excellent DNA sequence simulation tool, user-friendly and easy to use as suggested in (Alosaimi et al., 2019). NEAT-GenReads allows users to adjust the input parameters, setting desired DNA sequence coverage and generating specific mutation model. It implements three models including the mutation, GC-content and sequencing models of which we used two mutation and sequencing models. However, we used default setting for the GC-content model (Stephens et al., 2016).
We generated two simulations data sets each representing a specific population, African and European populations. NEAT-GenReads output three different files, Forward and reverse Fastq, Golden BAM and Golden VCF, thus, the total outputs are 100 sample, 50 samples representing the African population with two different coverage (high and low) to check the effect of sequence coverage on variant calling, as well as for European population (**Figure** 2.2 and **Table** 2.2).

As expected, the variations in the African samples was higher than the European ones. NEAT-GenReads was able to mimic both populations as we allowed NEAT to learn from population-specific data and the generated population specific models. The resulted Fastq files, as stated in the result section, was in good quality for both populations, and we used as for our further analysis. On the other hand, Golden BAM was not used as the group read (@RG tag) was missing, this is one limitation of NEAT. Consequently, we decided to use Fastq files for our further analysis in the next chapter.
Additionally, the generated Golden VCF was good; as a result, it will be used as a baseline to benchmark and evaluate nine different variant calling tools (see Chapter 3).

In summary, in this chapter, we have described the process of DNA sequence simulation tools, and we have illustrated some example for each tool representing different NGS platforms. However, our aim in this chapter was to simulate DNA sequence reads. We choose NEAT-genReads as a simulator, it outcomes three different files format, FASTQ, Golden BAM, and Golden VCF. We have generated two simulated data set each representing the African and European populations, respectively.
We have evaluated the simulation outcome to ensure our result is not truncated or damaged. We have checked the quality of FastQ files by using FastQC tool, checking golden BAM, VCF by using samtools features.

In the next chapter, we will use the simulated data (generated Fastq files) for further investigation and evaluation following the NGS downstream analysis pipeline toward variant discovery from nine different variant calling tools. We will evaluate the obtained resulted variations (VCF file) from each tool against simulated Golden VCF.

Chapter 3

Assessment of Nine Different Variant Calling Tools on African Versus Non-African Populations

3.1 Introduction

Whole-Genome Sequencing (WGS) data have revealed an insight into genome biology and genomics filed. WGS have increased the understanding of diseases and even human history and have enlightened the path toward personalised medicine. Furthermore, using whole-genome sequencing of multiple samples will help improving performance measurements (Highnam et al., 2015). Also, analysing WGS can allow detecting variants that are in a difficult region such as deep introns, which is hard to explicate at DNA level (Caspar et al., 2018). However, WGS may cause uncertain mappability due to the read length, dependency on the sequence coverage and model implement in variant calling tool.

Alongside with the tremendous advancement of Next-Generation Sequencing (NGS) technologies, accompanying with cost reduction in genomics research have yield an enormous amount of DNA sequence raw data, which have challenged the bioinformaticians to develop tools to analyses and process such large-scale data. Furthermore, recent developments in the field of Next-Generation Sequencing (NGS) technologies have led to insight toward disease etiology, diagnosis, and therapy for the world's utmost intractable destructive diseases such as HIV/AIDS, Malaria, Tuberculosis (TB), and Non-communicable diseases such as Sickle cell disease and others (C. Rotimi et al., 2014), these diseases are known to have the highest prevalence rate in Africa. However, despite the development in NSG and the tremendous amount of knowledge in the genetics field, yet studying the genetic background of the African populations still under-represented and have low participation in the genomics and genetics studies (Retshabile et al., 2018). African populations harbour the greatest genetic diversity (Campbell & Tishkoff, 2008), due to geographic and ethnic diversity alongside with long and short migrations, ancient and recent population expansion which have led to complex demographic history (Sirugo et al., 2008) alongside with having highest per capita health burden reported by the (WHO)(Castaño et al., 2011).

It is essential to provide a better understanding of African populations in all the field of genetics and genomic, to learn the ethnic diversity of African populations, as it will help to a better understanding of the genetic basis of phenotypic adaptations and complex diseases and reconstruct human evolution history (Campbell & Tishkoff, 2008). It is also important to develop bioinformatics tools that are able to analyses the complexity of African genetics variations, as well as devolving an African specific reference panel. These concerns are shared with the Human Heredity and Health in Africa (H3Africa) consortium.

Moreover, genomic projects such as the 1000 Human Genome Project, HapMap and The African Genome Variation Project (AGVP) have revealed genetics insights of African populations, for instance,

the high level of genetic variations accompanied with low and more divergent of level linkage disequilibrium (LD). Compare to other populations, African populations show more complex patterns of population substructure and the highest variant site per genome (C. N. Rotimi et al., 2017). Equally important to these variations, it is necessary to investigate and return medically relevant pathogenic variants known as actionable genes as recommended recently by The American College of Medical Genetics and Genomics (ACMG). Given these challenges to pinpoint African populations specific variants and return actionable genes associated with diseases in Africa, many questions have arisen such as "how researchers can improve the detection of these variations ? " Here, we attempt to answer this, therefore, we will focus on the steps of the downstream analysis of NGS, mainly the variants calling step, alongside the mutation prediction and annotations as shown in (**Figure** 1.1).

Variant calling (VC) is one of the important fundamental steps in the downstream analysis of NGS (see **Figure** 1.1), it is important to ensure calling the true positive variants, in addition, to discover true mutations and improving the clinical diagnosis. In the recent few years, many variant calling tools have been developed, they vary in their calling algorithm, and wither they call somatic variant or germline variants as discussed in chapter 1 .
In this chapter, we the most known and open-source variant calling tools, which are (VarScan2, SAMtools, GATK-HaplotypeCaller , SNVer, BCFtools, FreeBayers, Lofreq, PlatyPus and VarDict) see (**Table** 3.1), we excluded other variant calling tools as we follow (Sandmann, De Graaf, et al., 2017) exclusion criteria which is either they are using same tool or they required matched sample or calling Indels. we also excluded somatic variants calling tools.

3.2 Characteristics and Specifications of Variant Calling tools

After the simulation, we used the generated data for reading alignment, and various quality control (QC) steps as Subsequent the simulation process, we align the synthetic Fastq reads to a reference genome, and preformed post-alignment process to ensure the resulted BAM file is good and ready to be called. As discussed in the literature review chapter, many studies have recommended various variant calling tools to use, as demonstrated in (**Table** 1.1). Henceforth, we have chosen nine variant calling tools to use in this downstream analysis to conclude which tools have the most accurate call with a low rate of FP and FN and work best on the African population variations.Here we are introducing the most known and open-source variant calling tools, which are (VarScan2, SAMtools, GATK-HaplotypeCaller, SNVer, BCFtools, FreeBayers, Lofreq, PlatyPus and VarDict), we excluded other variant calling tools as we follow (Sandmann, De Graaf, et al., 2017) exclusion criteria which is either they are using the same tool, or they required matched sample or calling Indels. Here an overview of the tools arranged by releasing date:

1. **VarScan2:** In 2009 koboldt et al., introduce VarScan (Koboldt et al., 2009) to detect SNPs and Indels, as they stated that VarScan is an open-source tool unlike the tools before and compatible with several read aligners. In 2012, the same developers updated the tool to VarScan2, in order to further detect somatic (acquired) mutations and copy number alterations CNAs in cancer by exome sequencing. They used both known variants calling algorithms heuristic and statistical, additionally, they used correlation matrix diagonal segmentation CMDS algorithms to identified CNAs/CNVs (Koboldt et al., 2012).

2. **Samtools:** the most used and popular variant calling tools are: SAMtools (H. Li et al., 2009), which can preform multiple tasks see (**Table** 3.1). The Samtools package consists of two different variant calling tools Samtools and BCFtools `http://github.com/samtools/samtools`.

3. **GATK-HaplotypeCaller:** that have a robust performance and features such as coverage analysis, quality score recalibrations to eliminate FP variants and other reads data manipulations.

It can also detect both germline and somatic variants https://software.broadinstitute.org/gatk/.

4. **SNVer:** In 2011, SNVer developers (Wei et al., 2011) motivated to call the variant in a pooled sequencing, since it was a need back then to both improve the computational cost and accuracy. They also call individuals NGS data and other features. It applies hypothesis testing problem and uses the binomial- binomial model to analyse the pooled or individual NGS data. It allows users to choose FP error rate. When compared to GATK and SNVer, it has the same performance, but since pooled sequenced has advantage of leveraging haplotypes patterns, thus SNVer preformed best according to (Wei et al., 2011) .

5. **BCFtools:** implement SAMtools pileup but with new methods as described in (**Table** 3.1). The difference is that BCFtools introduced multiallelic calling model, while Samtools uses consensus calling model (H. Li et al., 2009; H. Li, 2011) https://github.com/samtools/bcftools.

6. **FreeBayers:** (Garrison & Marth, 2012) developed a haplotyped based variant director and can also detect multiallelic loci with non-uniform copy numbesr.

7. **Lofreq:** Since calling variant in low frequency is a challenging step, Lofreq (Wilm et al., 2012) were developed to detect rare and true variation with frequencies lower than the average error rate.

8. **PlatyPus:** (Rimmer et al., 2016) introduced Platypus with developed assembly algorithms which cope with highly divergent region (**Table** 3.1).

9. **VarDict:** The last variant calling tool will be reviewed is VerDict (Lai et al., 2016), designed to call complex variants and actionable mutations in cancer research. **Table** 3.1 provides these listed variant calling tools use-abilities and limitations. Further, **Table** 3.1 also provides the programming languages, operating system and algorithms used for each tool.

Table 3.1: The Characteristics of the Variant Calling tools used for this study.

Tools Name	Programming Languages	Algorithms Used	Usability	Operation System	Limitations	References
VarScan2	Java, Perl and C	Heuristic and Statistical Algorithm, Bayesian statistic for False Positive filter	Can classify variants on their somatic status, whether they are germline (inherited), somatic or loss of heterozygosity LOH, Apply false positive filters	Platform independent	Uses Fisher's exact test that proven later isn't sufficient to detect less common variants	(Koboldt et al., 2012)
SAMtools	C and Java	Bayesian statistic	Flexible when dealing and manipulating with SAM/BAM, such as sorting and indexing, Call SNPs and short Indels, can perform PCR duplicates removal and provide base-pair information in the pileup format	Linux	Can't handle pooled sequencing data (Wei et al., 2011)	(H. Li et al., 2009)
GATK-Haplotype Caller	Java	Bayesian Algorithm, Posterior Probability	Framework consist of improved stability, CPU, memory efficiency and parallelisation, Can recalibrate quality score, realign and multiple sample SNP genotyping.	Linux		(McKenna et al., 2010)
SNVer	Java and GUI available	Binomial-binomial model	Can detect mutations in pooled data, Users can choose threshold for false-positive rate, which are calculated by statistical algorithms to estimate FDR, It result are unaffected by varying quality values (Wilm et al., 2012)	Linux, Mac OSx and Windows OS	No support for small Indels	(Wei et al., 2011)

Continued from previous page

Table 3.1 – Continued from previous page

Tools Name	Programming Languages	Algorithms Used	Usability	Operation System	Limitations	References
BCFtools	C	Expectation-Maximisation (EM) method	Implement improved equation to call SNPs and thier genotype, Can discover somatic and germline mutations, Estimate site allele frequency, linkage disequilibrium, test Hardy-Weinberg and associations	Linux		(H. Li, 2011)
FreeBayes	C++ and C	Bayesian Algorithm	Capable of modelling multiallelic loci with non-uniforms copy number, Haplotype based variants detector, Applying filters to remove alleles that are unlikely to be true	Linux		(Garrison & Marth, 2012)
Lofreq	Python	Quality aware error correction model	Can call SNVs in very low frequency, very sensitive	Linux	Can't call rare SNVs with non-unique mapping and alignment uncertaintly	(Wilm et al., 2012)
PlatyPus	Python	Bayesian Algorithm	uses local de novo assembly to achieve high sensitivity and specificity for SNPs, Indels and complex variants, Can perform variant calling on raw aligned reads without reprocessing, perform local realignment and probabilistic haplotype estimations	Linux and OSx		(Rimmer et al., 2016)

Continued from previous page

Table 3.1 – Continued from previous page

Tools Name	Programming Languages	Algorithms Used	Usability	Operation System	Limitations	References
VarDict	Perl and R		Call SNV, MNV, Indels , complex and structural variants, Can perform local realignment and allele frequency estimations and-duplication, Reduce false positive calls	Linux		(Lai et al., 2016)

3.3 Materials and Methods

3.3.1 Data Description

We have used the generated simulated data from Chapter 2 also showed in (**Table** 3.2). In fact, two population-specific simulated data sets at two depth coverage (high and low) each set represent African and European population, respectively are considered for variant calling, therefore a total of 4 data sets (AFR high, AFR low, EUR high and EUR low) as described in Chapter 2. These data was generated from NEAT-genReads. We set the sequencing error rate at 0.1 as it represents the illumina sequencing platform error rate. Since most of the NGS downstream analysis tools has been performed using European data (Campbell & Tishkoff, 2008), therefore, we compare African and European populations, to observe if the variant calling tools handle the complex variations presented in the African populations, as well as they, handle the European data.

Table 3.2: Data generated by NEAT-genReads, used to analyse the performance of variant calling tools.

	African Populations		European Populations	
Coverage Depth	High (30- 50x)	Low (20 -10x)	High (30- 50x)	Low (20- 10x)
Sequencer	illumina		illlumina	
Sequencing Error rate	0.1		0.1	
Sample numbers	25	25	25	25
Total	50		50	

3.3.2 Data Generation and Processing

Given the output results from NEAT-genReads in three different forms: (1) forward and reverse FastQ file, (2) golden BAM, and (3) golden VCF, We have used the forward and reverse FastQ files for further analysis, we have performed Quality control on FastQ files as illustrated in Chapter 2. **Figure** 3.1 illustrates the variant calling analysis pipeline that we will in subsequent sections below.

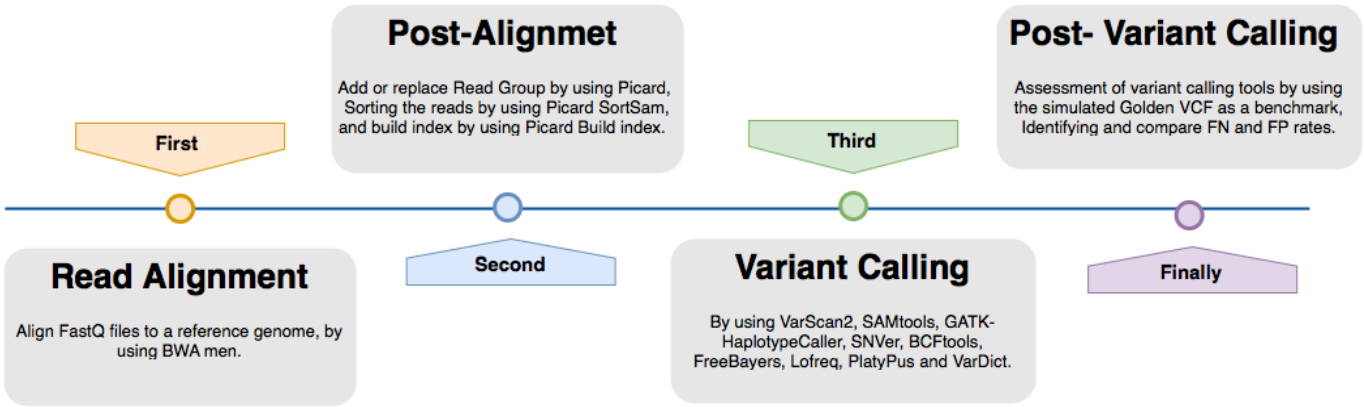


Figure 3.1: Overview of the variant calling analysis pipeline.

Reads Alignment and Mapping to Reference Genome

We align forward and reverse FastQ to the latest human reference genome hg38/GRCh38 by using BWA mem alignment tool developed by (H. Li, 2013), resulted in SAM files.

Post-Alignment Process

First, we replaced the reading group in the SAM file by using Picard (AddOrReplaceReadGroups), as the simulation tool generated one @RG tag for all the generated data, which will not be accepted by GATK and other VC tools, as the reading group must be unique for each sample. Second, we sorted SAMfile and generated BAM files by using Picard (SortSam). Finally, we indexed BAM by Picard (BuildBamIndex). Now the BAM files are ready to be called and used as an input for variant calling tools. Picard (Broad Institute, (Accessed: 2018/02/21; version 2.17.8)) available at http://broadinstitute.github.io/picard/. To not penalise VC tools that can be sensitive to post-alignment quality control and of course since the data used are simulated, we opted to not conduct further post-alignment quality control such as mark duplication, realignment around indels.

3.3.3 Performing Variant Calling

After the post-alignment step, the resulted BAM file was used as an input for each of the nine variant calling tools listed in (**Table** 3.3). All 100 BAM samples were used. We joint call all the samples for SAMtools, BCFtools, FreeBayes, SNVer, GATK, Platypus and VarScan, but, we couldn't apply it on Lofreq and Vardict as they are designed to call per sample.

Table 3.3: List of the variant calling tools used to detect SNPs from simulated WGS data.

Year	Tool	Version	URL	Ref
2009	VarScan2	2.3	http://varscan.sourceforge.net	(Koboldt et al., 2009, 2012)
2009	Samtools	1.6	http://github.com/samtools/samtools	(H. Li et al., 2009)
2010	GATK-HaplotypeCalle	3.3-0-g37228af	https://software.broadinstitute.org/gatk/	(McKenna et al., 2010)
2011	SNVer	0.5.3	http://snver.sourceforge.net	(Wei et al., 2011)
2011	Bcftools	1.2-279-g1bdf8e3	https://github.com/samtools/bcftools	(H. Li et al., 2009; H. Li, 2011)
2012	FreeBayers	1.1.0-46-g8d2b3a0-dirty	https://github.com/ekg/freebayes	(Garrison & Marth, 2012)
2012	Lofreq	2.1.2	https://csb5.github.io/lofreq/	(Wilm et al., 2012)
2016	PlatyPus	0.7.9.1	https://www.well.ox.ac.uk/platypus	(Rimmer et al., 2016)
2016	VarDict	1.70	https://github.com/AstraZeneca-NGS/VarDict	(Lai et al., 2016)

Variant calling parameters

We set the parameters to call from chr1- chr22 only, and if the tool support frequency-based calls such as VarDict and freebayes, we set the rate to 0.01 as the minimum allele frequency. We called the variants simultaneously across all BAM files, and this is known as joint calling and produce one VCF files for each caller, except Vardict, Lofreq per sample-based call was conducted .

Variant calling performance measures and analysis

We compared the VC tools by variant positions using a custom python script to extract variant position from each resulting VCF files, and we computed key measuring performance include True Positive (TP), False Positive (FP) and False Negative(FN). Furthermore, we calculated the performance as

explained in Section 1.9, sensitivity (Recall) in Equation 1.3, precision (PPV) as in Equation 1.4 and F-score as in Equations 1.6. While the Ti/Tv were obtained from using Bcftools-stat. We visualise the intersection of variants simulated versus those in golden VCF by using Intervene tool by (Khan & Mathelier, 2017).

3.4 Results

The variant calling was performed by VarScan, Samtools, GATK-HC, SNVer, Bcftools, FreeBayes, Lofreq, Platypus and Vardict on simulated data (African and European) resulted from NEAT as described in Chapter 2. We divided the simulated data as four sets: (1) African population- High coverage data, (2) African population- Low coverage data, (3) European population- High coverage data (4) European population- Low coverage data as shown in (**Table** 3.2). All resulted VCFs was compared against Golden VCF produced from NEAT.

Table 3.4: Summary of the performance metric regarding eight variant calling tools evaluated from simulated data representing African and European Population.

The samples represent simulated data of different coverages. True Positive (TP), False Positive (FP) and False Negative were used to calculate the performance metric of each variant calling tool.

African Population								
Cov*	**Caller**	**TP**	**FP**	**FN**	**Recall**	**PPV***	**F-score**	**Ti/Tv**
High	VarScan	481,237,109	4,202,307	115,279,037	0.2945	0.991	0.454	1.67
	Samtools	936,138,549	4,801,172	697,888,931	0.572	0.994	0.727	1.73
	GATK-HC	1431,790,559	480,655,083	202,236,921	0.876	0.748	0.807	1.70
	SNVer	1,042,052,763	217,963,522	59,197,4717	0.637	0.827	0.720	1.67
	Bcftools	86,766,719	1,593	154,726,0761	0.053	0.999	0.1008	1.73
	Lofreq	1403,439,343	23,030,134	230,588,137	0.858	0.983	0.917	1.76
	PlatyPus	163,693,637	206,728	1,470,333,843	0.1001	0.998	0.182	2.80
	VarDict	8,462,928	59	1,625,564,552	0.0051	0.999	0.010	1.67
Low	VarScan	46,079,2497	69,094,327	1,258,564,680	0.268	0.869	0.409	1.68
	Samtools	103,3862,412	2,922,554	685,494,765	0.6013	0.997	0.750	1.73
	GATK-HC	1,392,111,583	372,576,395	327,245,594	0.809	0.788	0.799	1.71
	SNVer	1,308,522,590	327,072,554	410,834,587	0.761	0.800	0.780	1.66
	Bcftools	64,417,446	22,509	1,654,939,731	0.037	0.999	0.072	1.73
	Lofreq	1,342,828,800	56,340	376,528,377	0.781	0.999	0.877	1.76
	PlatyPus	99,883,347	73,530,996	1,619,473,830	0.0580	0.575	0.105	2.43
	VarDict	8,100,636	36	1,711,256,541	0.004	0.999	0.009	1.66
European Population								
Cov*	**Caller**	**TP**	**FP**	**FN**	**Recall**	**PPV***	**F-score**	**Ti/Tv**
High	VarScan	395,900,288	60,684,867	1,276,091,292	0.236	0.867	0.371	1.54
	Samtools	977,511,680	2,238,536	694,479,900	0.584	0.997	0.737	1.59
	GATK-HC	1,488,460,421	552,553,957	183,531,159	0.890	0.729	0.8017	1.56
	SNVer	1,017,864,538	207,333,783	654,127,042	0.608	0.830	0.702	1.54
	Bcftools	50,923,622	1,363	1,621,067,958	0.0304	0.999	0.059	1.59
	Lofreq	1,368,140,684	73,175	303,850,896	0.818	0.999	0.900	1.63
	PlatyPus	180,556,949	144,184	1491,434,631	0.107	0.999	0.194	2.39
	VarDict	7,881,251	46	1,664,110,329	0.004	0.999	0.009	1.55
Low	VarScan	484110062	69838149	1222614957	0.283	0.873	0.4282	1.55
	Samtools	325,017,774	8,279	1,381,707,245	0.190	0.999	0.319	1.60
	GATK-HC	691,478,420	111,097,346	1,015,246,599	0.405	0.861	0.551	1.56
	SNVer	1,421,204,610	454,469,916	285,520,409	0.832	0.757	0.793	1.54
	Bcftools	700,731,299	169,5148	1,005,993,720	0.410	0.997	0.581	1.59
	Lofreq	1,344,126,496	47,249	362,598,523	0.787	0.999	0.881	1.63
	PlatyPus	164,588,476	160,790	1,542,136,543	0.096	0.999	0.175	2.23
	VarDict	7,955,439	48	1,698,769,580	0.004	0.999	0.009	1.54

*PPV=Positive Predictive Value, Cov=Coverage

As shown in **Table** 3.4, we were able to calculate TP, FP and other metric performance measures : **The first data set (golden AFR - high)** , the African Golden - high coverage vcf (truth set) contains $1,634,027,480$ SNPs as a total for all the 25 samples, GATK-HC and Lofreq have the highest TP among all result with (sens= 0.87), (sens=0.85) respectively. Since GATK had the highest sensitivity it goes with low PPV (PPV=0.7), all variant calling tools

have good precision, however Vardict achieved the highest with very low FP (PPV=0.999), see (**Table** 3.4). Finally, F-score, which measures the overall performance, indicates that Lofreq had the highest performance among all (F-score=0.91) followed by GATK-HC (F-score=0.80).

The second data set (AFR - low): By the same token GATK-HC had the highest TP among all other tools, the AFR-low Golden vcf contains $1,719,357,177$ SNPs, and Vardict had the lowest FP with high precision (PPV=0.999), and Lofreq had the high F-score (F-score=0.87), see (**Table** 3.4).

The third data set (EUR - high) Golden vcf contains $1,671,991,580$ true SNPs, GATK has the highest FP (sens=0.89) followed by Lofreq (sens=0.81) and high F-score (F-score=0.90), all tools had good PPV (**Table** 3.4).

The fourth data set (EUR - low) , SNVer had the highest TP among all tools (sens=0.82), and Lofreq with high (F-score=0.88), all tools had good PPV (**Table** 3.4). If the Ti/Tv ratio is too low, it is more likely to have false positive. The ratio among all is good, but Platypus had the highest ratio among all in the four data sets, indicating that Platypus tends to infer high rate of false positive variants during the call. Of note, Vardict has a great deal in generating less false positive variant discovery and seems to perform much better in African simulated data sets. In contrast, we weren't able to analyse Freebayes result as the output vcf files have no variant positions.

Regarding sequence depth, the result shows a higher number of calls for high coverage and vice versa, see **Table** 3.4 and **Figure** 3.2. The overall result of each variant calling tools are a bit similar for both European and African populations, GATK-HC, Lofreq, SNVer and Samtools has the lowest FN among all, while Vardict, Bcftools, Lofreq and Samtools has the lowest FP among all, and regarding TPs, the highest result are from GATK-HC, Lofreq, SNVer and Samtools. Concerning, CPU wall-time, almost all variant calling tools took a very long time, we had to use multithreading and GNU Parallel (Tange, 2011), to process data in parallel and to reduce run time. GATK-HC took very long CPU wall-time.

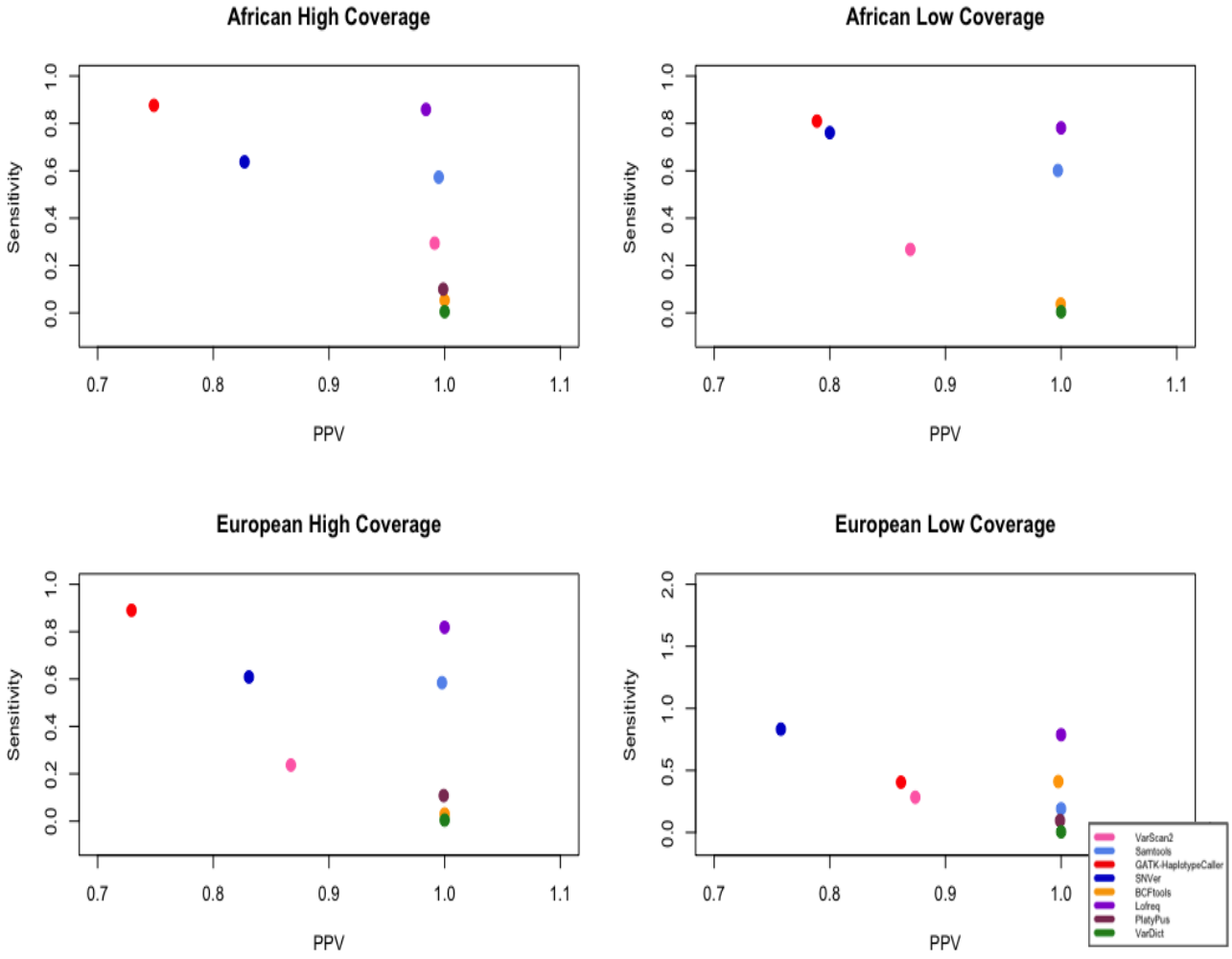


Figure 3.2: Relation between Positive Predictive Value (PPV) and Sensitivity in case of comparing variant calling tools on the African and European Populations regarding different coverage.

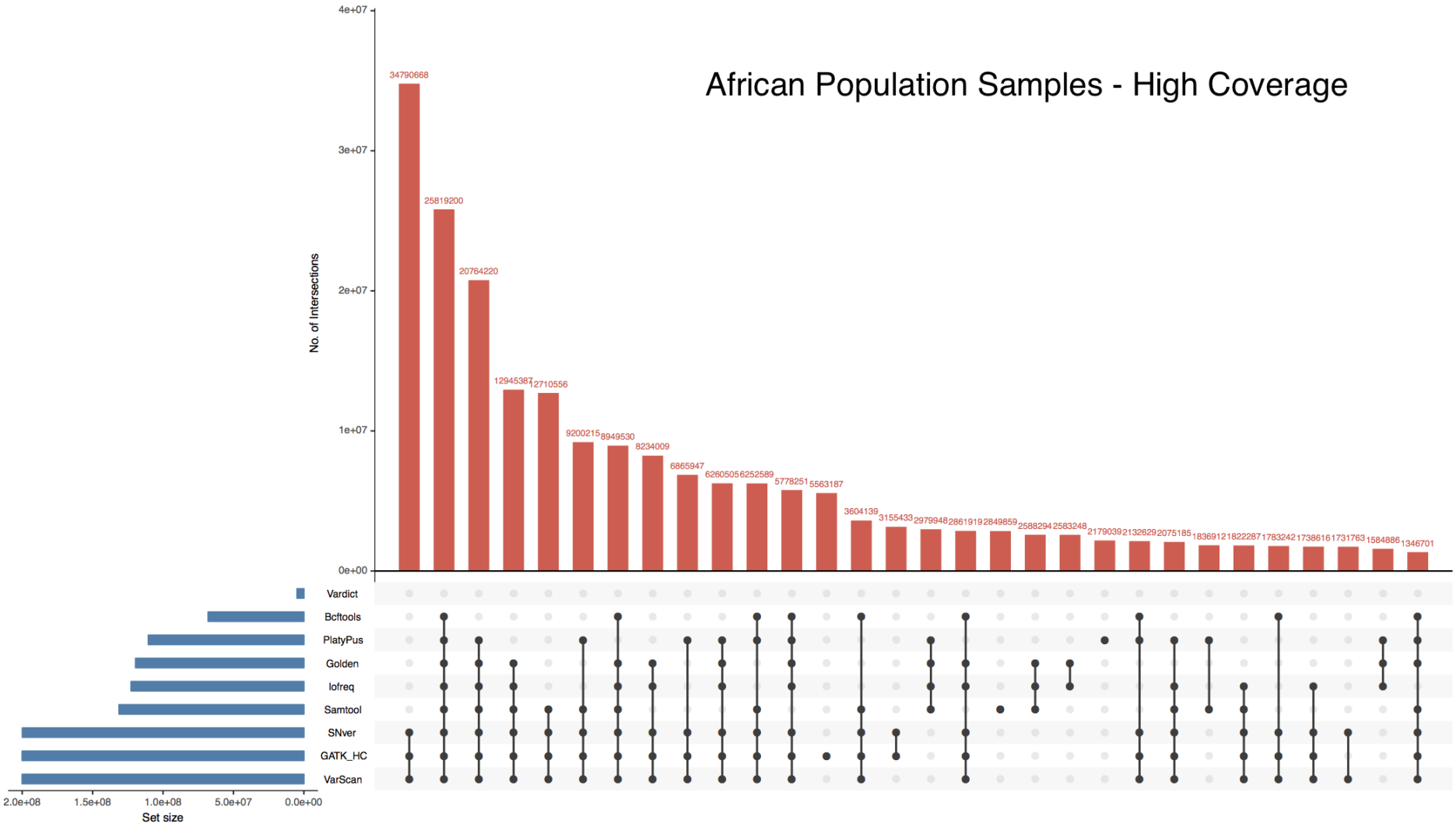


Figure 3.3: An UpSetR diagram visualising the intersections of the variant positions produced by Samtools, BCFtools, SNVer, GATK , Platypus, VarScan, Lofreq and Vardict, on African population- High coverage data.

The red top bar-plot illustrates the size of the intersection, the linked points below display the intersecting sets of variants, while the blue bar-plot on the bottom left shows the set size of each tool.

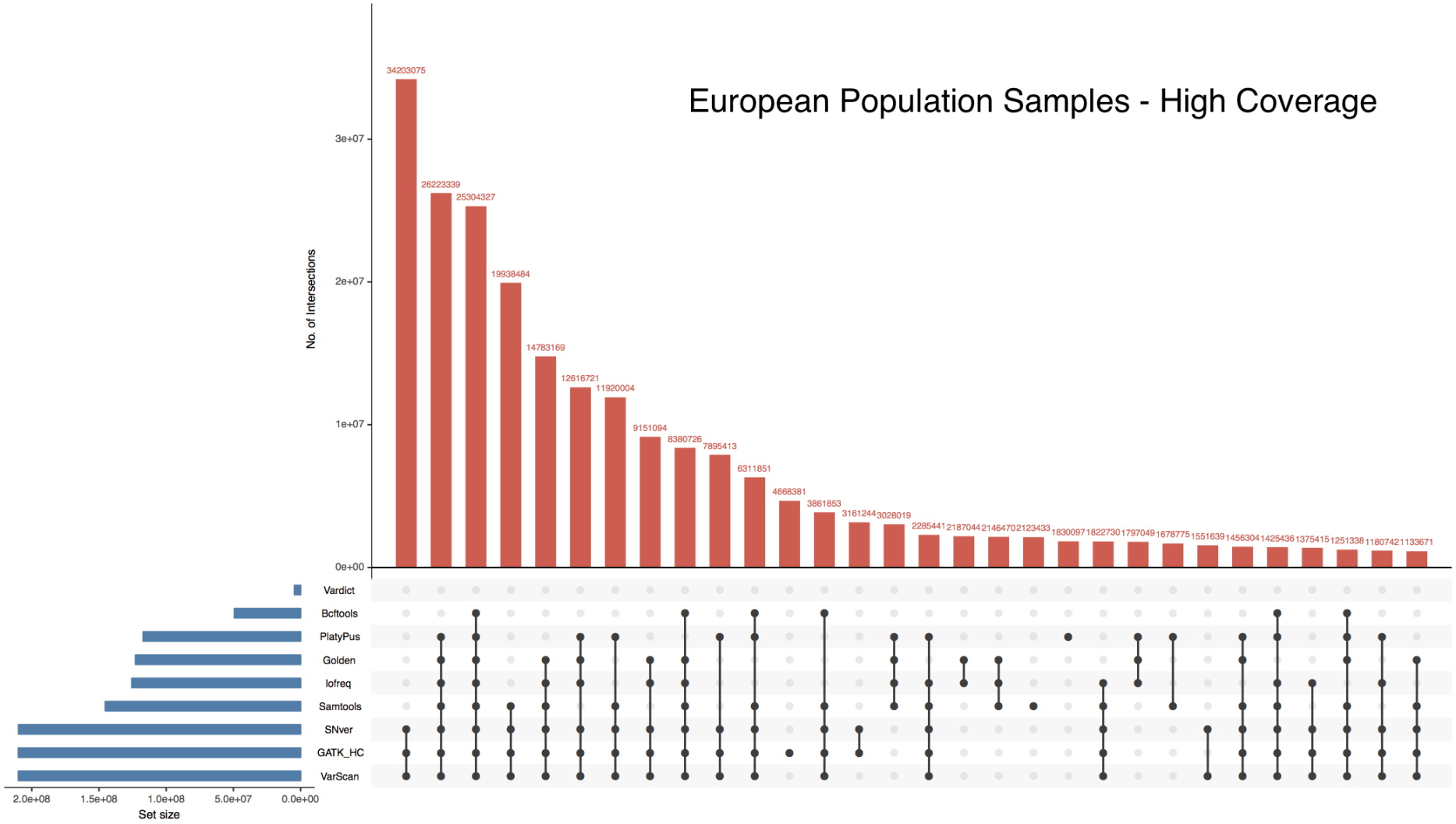


Figure 3.4: An UpSetR diagram visualising the intersections of the variant positions produced by Samtools, BCFtools, SNVer, GATK , Platypus, VarScan, Lofreq and Vardict, on European population- High coverage data.

The red top bar-plot illustrates the size of the intersection, the linked points below display the intersecting sets of variants, while the blue bar-plot on the bottom left shows the set size of each tool.

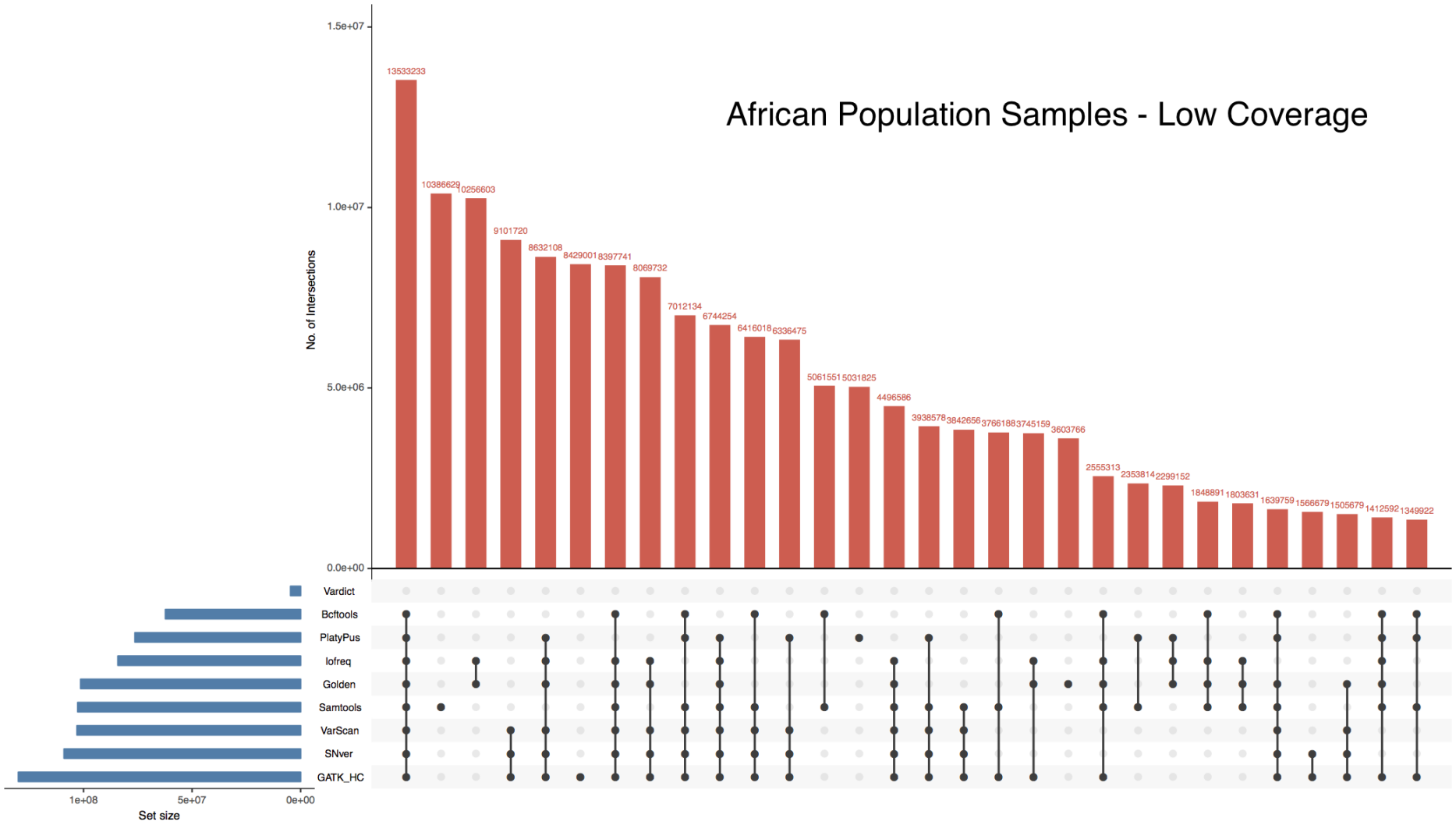


Figure 3.5: An UpSetR diagram visualising the intersections of the variant positions produced by Samtools, BCFtools, SNVer, GATK , Platypus, VarScan, Lofreq and Vardict, on African population- Low coverage data.

The red top bar-plot illustrates the size of the intersection, the linked points below display the intersecting sets of variants, while the blue bar-plot on the bottom left shows the set size of each tool.

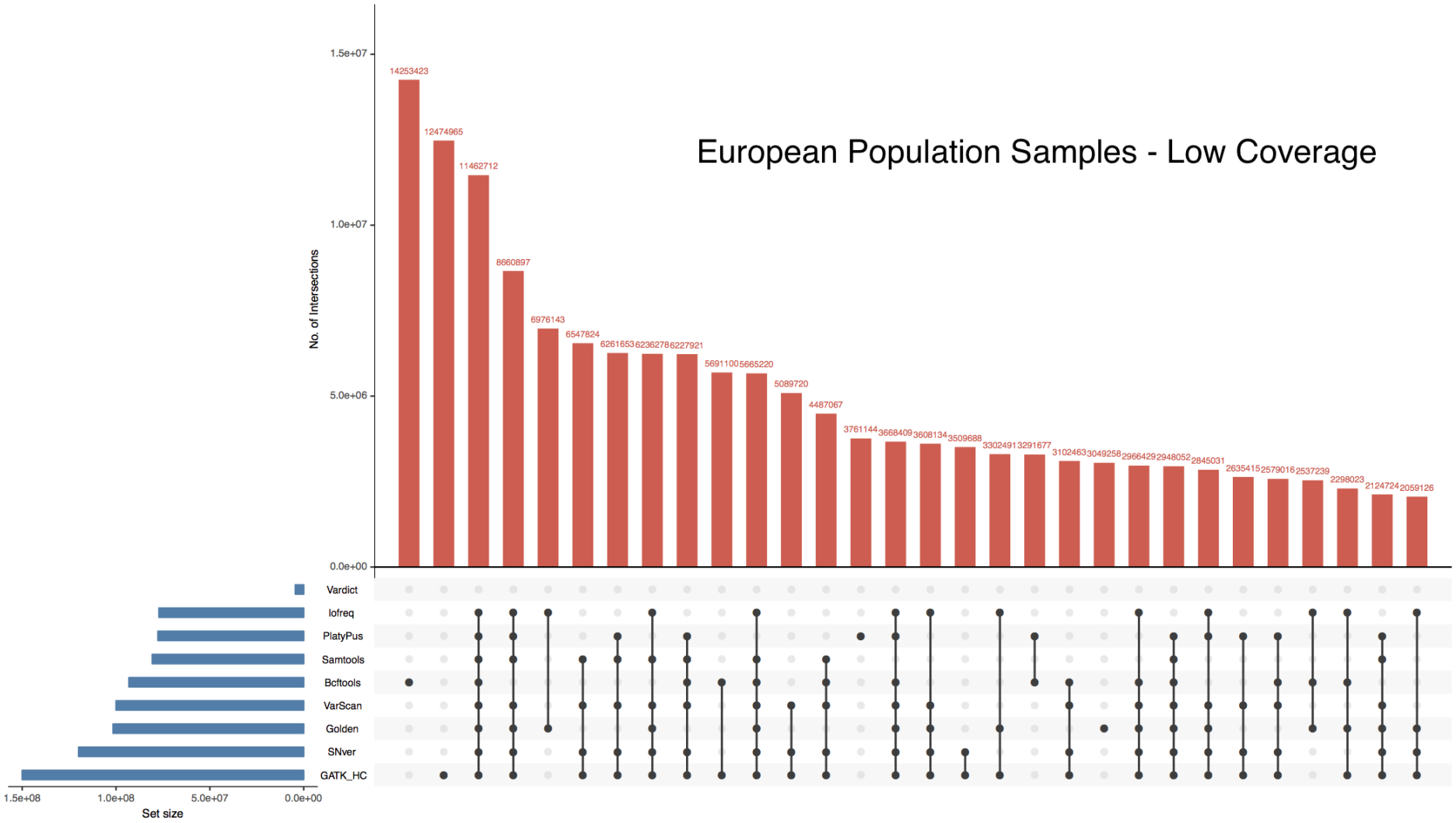


Figure 3.6: An UpSetR diagram visualising the intersections of the variant positions produced by Samtools, BCFtools, SNVer, GATK , Platypus, VarScan, Lofreq and Vardict, on European population- Low coverage data.

The red top bar-plot illustrates the size of the intersection, the linked points below display the intersecting sets of variants, while the blue bar-plot on the bottom left shows the set size of each tool.

3.5 Discussion and Overall Chapter Summary

The tremendous development of Next-Generation sequencing has promoted the evolution of personalised medicine. Such progress in NGS resulted in an enlargement of the downstream analysis tools to handle such data. Such an example, many variant calling tools have been developed, which raise the question of which tool one can choose when dealing with a complex and diverse genome such as the African genome? In this chapter, we were able to answer this question. We compare nine variant calling tools (VarScan2, SAMtools, GATK-HaplotypeCaller, SNVer, BCFtools, FreeBayers, Lofreq, PlatyPus and VarDict) on simulated data representing two population (African and European) at varying coverage (high and low) as a total of 4 data sets. We assessed these tools based on sensitivity and precision and most importantly, the F-score to measure the overall performance, low sensitivity and precision result in low F-score.An increased specificity may result in the loss of true positive data while a prioritised sensitivity will result in increased false positive data. Depending on the desired output for a study, sensitivity or specificity must be favored as there is a trade-off between these two. The total average of FPs for the African population at different coverage ($= 98508519, 31$) are a bit higher than the European populations ($= 91271677, 25$); hence the ability of variant calling tools to detect true variants are higher when dealing with European population which support the previous researches done by honour students at UCT Noëlle van Biljon,2017 and Loratoeng Mpolokeng,2018, that many tools are suitable when dealing with European data than African. Here we discuss the performance of each tool on the four data sets (AFR-high, AFR-low, EUR-High and EUR-Low):

1. **VarScan2:** the result for the first data set was (sens=0.29), (PPV=0.99) and (F-score=0.45).

2. **Samtools:** Performed well with low FPs (PPV=0.99) for all data sets.

3. **Bcftools:** unlike excepted it has the lowest performance among all tools, it may be due to Bcftools-Concat stages when we merged the 22 vcf into one, it could be led to losing some lines resulted in low variant positions.

4. **GATK-HaplotypeCaller:** showed the highest TPs in both coverages for the African population, accompanied by high F-score. From our analysis, when considering sensitivity GATK-HC performed best on both African and European populations.

5. **SNVer:** output two vcf files, one are SNPs call without filtering process, and the other contains Indel calls, the metric results are good with high sensitivity and precision for all four data sets as shown in (**Table** 3.4).

6. **FreeBayers:** detected variants for the (AFR-High $= 109, 317, 033$), (AFR-Low $=159, 320, 395$), (EUR-High $=127, 301, 186$) and (EUR-Low $= 175, 720, 958$) but all without base-pair positions, for this reason we excluded it from comparison.

7. **Lofreq:** showed remarkably good results, F-score for all data sets are the highest among all results alongside with Sensitivity (Recall) and Precision (PPV), especially with the African-high coverage data set. **Figure** 3.2 illustrates the relationship between Positive Predictive Value (PPV) and Sensitivity regarding different coverage, Lofreq in colour purple shows great performance. **Figures** 3.3, 3.4, 3.5 and 3.6 show the very large intersection of variants between Lofreq and Golden vcf (Truth set) only.

8. **PlatyPus:** result are good regarding Precision (PPV=0.9) and (Ti/Tv=$\sim$ 2.4) for all 4 data sets.

9. **VarDict:** From the first run of Vardict, we used a bed file contain all regions of interest. This took very large number of CPUs and exceeded wall time specified by High Performance Computing. The larger the bed file, the longer it took to be processed and done. Therefore, we repeated calling on smaller bed file based on autosomal chromosomes (1 to 22), hence, larger variants were missing. This may affect the result with very high False Negative as it may be missing variants that are not called by Vardict and absent in golden VCF file. Furthermore, Vardict uses post-processed step depending on R and Perl script, R works very slowly as it calculates Fisher's exact test and odds ratio, which took a very long time to process. The metric result of Vardict was good in case of true positive as almost all the variants that are detected are considered to be True positive, the total variants of African-High coverage that are called from Vardict are $8,462,987$, and the true positive are $8,462,928$ (PPV=0.999). VarDict can be suggested for African targeted sequence data.

In summary, higher sequence depths helps variant calling tools to be more confidently call the true variant, this confirms and supports previous findings (Huang et al., 2015). Considering both sensitivity and PPV, Lofreq outperformed all VC tools. It can accurately call such a complex and large variants with high TPs and low FPs even at very low frequency. Our finding support (Huang et al., 2015) result. Our current results suggest that Lofreq can currently be a great option in calling WGS from African populations and VarDict can be considered for targeted sequence, particularly for African data. Moreover, some of the high genetics variations presented in the result may could be artefacts; as a result, it is essential to apply multi variant calling tools to ensure calling true variants. Given huge amount of non-overlapped variants across the results from current VC tools and differing number of variants discovered across these tools, it will be reasonable to consider multi variant calling tools to allow cross validation of variants discovered. In the next chapter, we will use Lofreq to call the variants on WGS of real populations.

Chapter 4

Dissecting Genetic Mutations and Secondary Finding from Twenty World-wide Ethnics Groups

4.1 Introduction

The NGS sequencing analysis contributed to the improvement of patient and clinical care. This development has hoped to bridge the gap between healthcare and genomics. Furthermore, as mentioned earlier, variant calling is an important aspect of genomics studies as polymorphism information can be used to influence the discovery of actionable pathogenic variants and therefore important clinical decisions. Currently, the definition of actionable pathogenic variants varies among scholars.

The Clinical Genome Resource (ClinGen) presents actionability as clinically prescribed interventions to a genetic disorder that is effective for prevention or lowered clinical burden or delay for a clinical disease or improved clinical treatments and outcomes in a previously undiagnosed adult (Hunter et al., 2016). On the other hand, 100,000 Genomes Project protocol presents actionable genes as variants that can significantly prevent (or result in illness or disability that is clinically significant, severely life threatening and clinically actionable) disease morbidity and mortality, if identified before symptoms become apparent. However, in any cases the classification of variants to be clinically actionable or not dependent and can only emerge during the process of seeking ethical approval for the study (Caulfield et al., 2017).

Overall, in current literature and most annotation databases, the classification of pathogenicity differs (Sherry et al., 2001; K. Wang et al., 2010; Z. Wang, Liu, Yang, & Gelernter, 2013; Landrum et al., 2016; McLaren et al., 2016). To illustrate this, study conducted by (Dorschner et al., 2013) leveraged exome data of European and African diaspora to dissect actionable pathogenic variants, however their findings suggested that actionable pathogenic variants were disproportionate between European and African descent with an estimated frequency of approximately 3.4% and 1.2%, respectively. This indicates deficit of identification of pathogenic variant in African population in general. Furthermore, this illustrates a deficit of identification of pathogenic variant in African population in general. A similar study by (Amendola et al., 2015) also confirmed the findings of (Dorschner et al., 2013).

Nevertheless, a common feature to define actionability is to combine many annotation pipelines during filtering and prioritisation mutations, in which casting vote can be applied respectively to allow better prediction of the targeted variant. Furthermore, on top of ethical approval, the ancestral/derived minor allele frequency of the variants, segregation evidence, and the number of the patients affected with the variants and their status as a de novo mutation can highly be considered.

In this chapter, we provide a broad assessment of possible actionability of variants known to be associated to top four burden African diseases and a list of actionable genes from American College of Medical Genetics and Genomics (ACMG) using WGS data of 20 world-wide ethnic groups. In doing so, we aim to

1. apply the best variant calling tool identified in previous chapter 3 on publicly available data, the African Genome Variation and 1000 Genome Project and examine the evolutionary variation of pathogenic mutation based on selected known disease-genes from four big African burden diseases include HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and a set of known actionable genes across 20 world-wide population ethic groups.

2. perform disease-genes population structure from these known disease-genes (HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and a set of known actionable genes) among 20 world-wide ethnic-specific data.

3. examine the heterozygosity ratio, the proportion of ancestral/derived alleles, and the distribution of minor allele frequencies based on these selected disease-genes from HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and a set of known actionable genes across 20 world-wide ethnic-specific data.

4.2 Methods and Material

4.2.1 Data Description and Quality Check

We accessed bam files from 1000 Genomes Project (1KGP) (Consortium et al., 2012) and the African Genome Variation Project (AGVP) (Gurdasani et al., 2015), which has recently characterised the admixture across 18 ethno-linguistic groups from sub-Saharan Africa as shown in (**Table** 4.1). A quality control check was conducted these bam files using samtool. We finally retain 2,504 bam files from 1000 Genomes Project and 2,428 bam files from AGVP, a total of 4,932 samples. Based on the initial sample description including population or country labels, we grouped samples (**Table** 4.1) based on the ethno-linguistic information obtained from (Gudykunst & Schmidt, 1987; Michalopoulos, 2012).

Table 4.1: Data obtained from 1000 Genomes Project (1KGP) (Consortium et al., 2012) and the African Genome Variation Project (AGVP) (Gurdasani et al., 2015) and used for analysis.

Population label	Ethnic group	Population description	Samples ID.
AFR	Afro Asiatic Semitic	Amhara:Ethiopia	22
	African American	Americans of African Ancestry in SW USA (ASW)	60
	African Caribbean	African Caribbeans in Barbado (ACB)	96
	Afro Asiatic	Al-Gharbiyah, NA, Monufia, Kafrel-Sheikh, Mansoura, Alexandria, Dakahlia, Samanoud, Al-Buhayrah, Minya, AlSharqia, El-Mahalla all from Egypt	99
	Afro Asiatic Cushitic	Oromo, Somali from Ethiopia	47
	Afro Asiatic Omotic	Wolayta from Ethiopia	24
	Khoe-San	Khoe-San:Khoesan	84

Table 4.1 – Continued from previous page

Population label	Ethnic group	Population description	Samples ID.
	Niger Congo Bantu	Baganda, Banyarwanda, Barundi, RwandeseUgandan, Banyankole:Uganda Bakiga, Mutanzania, Basoga, other uganda gwas unknown, Mutooro, Batooro, Nyanjiro (Tanzania) from Uganda and Luhya in Webuye, Kenya (LWK)	2158
	Niger Congo Bantu South	Zulu	98
	Niger Congo Volta Niger	Esan in Nigeria (ESN), Yoruba in Ibadan, Nigeria (YRI)	205
	Niger Congo West	Gambian in Western Divisions in the Gambia (GWD), Mende in Sierra Leone (MSL)	198
AMR	Latin American	Puerto Ricans from Puerto Rico (PUR), Colombians from Medellin, Colombia (CLM), Peruvians from Lima, Peru (PEL), Mexican Ancestry from Los Angeles USA (MXL)	347
EAS	East Asian	Southern Han Chinese (CHS), Chinese Dai in Xishuangbanna, China (CDX), Kinh in Ho Chi Minh City, Vietnam (KHV), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT)	504
EUR	European center	British in England and Scotland (GBR)	91
	European North	Finnish in Finland (FIN)	99
	European South	Iberian Population in Spain (IBS), Toscani in Italia (TSI)	214
	European USA	Utah Residents with Northern and Western European Ancestry (CEU)	99
SAS	South Asian	Punjabi from Lahore, Pakistan (PJL), Bengali from Bangladesh (BEB)	180
	UK Indian	Sri Lankan Tamil from the UK (STU), Indian Telugu from the UK (ITU)	204
	USA Indian	Gujarati Indian from Houston, Texas (GIH)	103
		Total	4,932

AFR: African, SAS:South Asian, AMR:Ad Mixed American, EUR:European, EAS:East Asian

4.2.2 Variants Discovery Analysis

As per our result from the evaluation of various variant calling tools (Chapter 3), we adopted Lofreq to conduct joint call across 2,504 samples from 1000 Genomes Project and 2,428 from AGVP for a total of 4,932 samples in 20 world-wide ethnic groups. The best practice specific to Lofreq caller was adopted, the resulting variant sets of all 4,932 samples in VCF file were filtered using the Samtool tool. We added additional filter levels as follows: If 3 SNPs are detected within a window of 10 base-pairs, the site will be flagged as a "SNPCluster" in the FILTER column. If 4 or more alignments having a mapping quality of $MQ = 0$ (which means it maps to different locations equally well) and the number of alignments that mapped ambiguously are more than a tenth of all alignments, it is difficult to decipher artefacts and true differences. These sites will be flagged as "HARD TO VALIDATE". SNPs which are covered by less than 5 reads may be potential artefacts and these sites were flagged as "LowCoverage", SNPs having a SNP quality below 30 are typically artefacts, were flagged as "VeryLowQual", SNPs having a quality score between 30 and 50 are potential artefacts, flagged as "LowQual", SNPs having a

QD score ¡ 1.5 are indicative of false-positive calls and artefacts, flagged as "LowQD" and SNPs covered only by sequences on the same strand are often artefacts, was flagged as "StrandBias". Variants flagged "VeryLowQual", "LowQual", "LowQD" and "StrandBias" were removed in VCF file. The resulting VCF file contained 4,932 samples.

4.2.3 Variant Annotation

The resulting joint call VCF file contained 4,932 samples and samples were split into 20 VCF files per ethnic group as listed in (**Table** 4.1), and we used ANNOVAR (K. Wang et al., 2010) to independently perform gene-based annotation in each final VCF data set to determine whether SNPs cause protein-coding changes and produce a list of the amino acids that are affected. We used ANNOVAR settings, where the population frequency, pathogenicity for each variant was obtained from 1000 Genomes exome21, Exome Aggregation Consortium30 (ExAC), targeted exon datasets and COSMIC31.
Gene functions were obtained from RefGene32 and different functional predictions were obtained from ANNOVAR's library, which contains up to 21 different functional scores including SIFT33,34, LRT35, MutationTaster36, MutationAssessor37,38, FATHMM39, fathmm-MKL39, RadialSVM40, LR40, PROVEAN40, MetaSVM40, MetaLR40, CADD41, GERP++42, DANN29, M-CAP29, Eigen29, GenoCanyon29, Polyphen2 HVAR43, Polyphen2 HDIV43, PhyloP44 and SiPhy44.
We additionally included conservative and segmental duplication sites, dbSNP code and clinical relevance reported in dbSNP45. From each resulting functional annotated data set, we independently filtered for predicted functional status (of which each predicted functional status is of "deleterious"(D), "probably damaging" (D), "disease_ causing_ automatic" (A) or "disease _ causing" (D).46,47,49) from from SIFT (Ng & Henikoff, 2003), LRT (Chun & Fay, 2009), MutationTaster (Schwarz et al., 2010), MutationAssessor (Reva et al., 2011), FATHMM (Shihab et al., 2013), FATHMM-MKL (Shihab et al., 2013), RadialSVM (Liu, 2014), LR (Agresti, 2012), PROVEAN (Choi & Chan, 2015), MetaSVM (Kim, Jhong, Lee, & Koo, 2017), MetaLR (Dong et al., 2014), CADD (Rentzsch et al., 2018), GERP++ (Davydov et al., 2010), DANN (Quang, Chen, & Xie, 2014), M-CAP (Jagadeesh et al., 2016), Eigen (Ionita-Laza, McCallum, Xu, & Buxbaum, 2016), GenoCanyon (Lu et al., 2015), Polyphen2-HVAR (Adzhubei et al., 2010), Polyphen2-HDIV (Adzhubei et al., 2010), PhyloP (Doerks, Copley, Schultz, Ponting, & Bork, 2002), and SiPhy (Garber et al., 2009).
We used a casting vote approach implemented in our custom python script, to retaining only a variant if it had at least 17 predicted functional status "D" or "A"out of 21. Second, the retained variants from each data set were further filtered for rarity, exonic variants, and nonsynonymous mutations and with a high-quality call as described above, yielding a final candidate list of predicted mutant variants in each subject group, including the replication group. We report on the aggregated SiPhy score from all identified mutants SNPs within the gene. Sections below provide details on how SNPs were mapped to genes.

4.2.4 Phased and Haplotypes Inference

To increase the accuracy, the resulting VCF file contained 4,932 samples of 20 ethnic groups, were used to further conduct quality control in removing all structured, indel, multi-allelic variants and those with low minor allele frequency (MAF < 0.05) prior to phasing. We first phased and inferred the haplotypes using Eagle software https://data.broadinstitute.org/alkesgroup/Eagl from the resulting curated data. We further compared sites discordance between these haplotype panels and independently with their original VCF file prior phasing. The only site with phase switch-errors showed discrepancies in MAF and were therefore removed. The whole phased data contained 4,932 samples of 20 ethnic groups were used to conduct downstream analysis below.

4.2.5 Disease- and Actionable Gene-specific Population Structure

We obtained list of genes known as medically actionable from https://www.fredhutch.org/en/news/releases/2014/09/actionable-genome-consortium-world-renowned-cancer-institutions.html, and Actionable Genome Consortium (ACG) and also we collect from GWAS Catalog https://www.ebi

.ac.uk/gwas/, literature and gene-diseases database such DisGeNET http://www.disgenet.org/ list of genes associated with four major African diseases including Malaria, TB, HIV and Sickle cell disease. We obtained 50, 77, 460, 75 and 114 genes known to associate with Tuberculosis, Malaria, Sickle Cell Anemia, HIV and ACG, respectively. We leveraged the dbSNP database to extract SNPs associated with these genes per diseases, as shown in (**Table** 4.2). The obtained SNPs per disease were thus extracted from the whole phased data contained 4,932 samples of these 20 ethnic groups; yield 5 disease-specific phased haplotypes data sets Table 4.2.

Table 4.2: The Number of SNPs after Quality Control (QC) in each group of genes associated with (HIV, TB, SCD, Malaria and actionable genes.)

SNPs	Genes	Diseases
649078	114	ACG
2735797	460	HIV
265427	50	MALARIA
4455648	75	SICKLE
2513341	77	TB

To evaluate the extent of substructure within disease-specific polymorphism across world-wide ethnic groups, we leverage each constructed disease-specific phased haplotypes data set, to perform genetic structure analysis based on Principal Component Analysis (PCA) using smartpca, part of the EIGENSOFT 3.0 package (Patterson, Price, & Reich, 2006). The PCA plot was obtained from using genesishttp://www.bioinf.wits.ac.za/software/genesis.

4.2.6 Proportion of Ancestral/Derived Alleles among Risk conferring Alleles

Each of these four disease-specific phased haplotypes data sets were used to analyse the fraction of derived and ancestral alleles at-risk allele within each ethnic group. Previous work has shown that derived alleles are more often minor alleles ($< 50\%$ allele frequency) and more often associated with risk than ancestral alleles (Gorlova et al., 2012). Therefore, we define risk allele as follow, if gene is being reported to increase the risk of disease (Odd ratio > 1)from either DisGeNET or GWAS Catalog, risk allele were defined as minor allele (for all SNPs associated to the gene) otherwise (Odd ratio < 1) is defined as major allele (for all SNPs associated to the gene).
We downloaded the SNP ancestral alleles from the Ensembl, a 59 comparative 32 species alignment (Paten et al., 2008), and we further checked the SNPs for those present in the dbSNP database. Each of these four disease-specific phased haplotypes data sets was further annotated using the VCFtools 'fillOaa' script (Danecek et al., 2011) with the ancestral allele recorded using the 'AA' INFO tag (McVean et al., 2012).
For each disease-specific data set, we determined the proportion of risk alleles that were ancestral or derived allele. We first computed, for each SNP, the fraction of ancestral allele, which was calculated by dividing the number of times the defined risk allele matched with ancestral allele by the total number of copies of all alternative alleles across all samples (within each ethnic group per disease) for a particular SNP. The fraction of derived allele is equivalent to 1 minus the fraction of ancestral allele. As mentioned earlier, derived alleles are more often minor alleles ($< 50\%$ allele frequency) and more often associated with risk than ancestral alleles, therefore, we investigated the relationship between the fraction of derived allele at-risk allele and ethnic group SNP minor allele frequency. To this end, the alternative (minor) alleles were categories into 6 bins, $(0 - 0.05, > 0.05 - 0.1, > 0.1 - 0.2, > 0.2 - 0.3, > 0.3 - 0.4, > 0.4 - 0.5)$with respect to each data set frequencies and we, independently, computed the fractions of derived alleles in each bin (above).
Furthermore, we computed the fraction of ancestral/derived alleles for all these known disease-specific

genes. To this end, we aggregated (see section 4.2.8 below) the fraction of ancestral/derived alleles at SNP-based level to gene, considering all SNPs located within the gene's downstream or upstream region (Chimusa et al., 2015).

4.2.7 Distribution of Minor Allele Frequency and Gene-specific in SNP Frequencies

To examine the extent to how common variants across these 20 ethnic groups within a specific disease (TB, HIV Sickle Cell Anemia and Malaria) and know actionable genes from ACG, we, therefore, investigated the distribution of the minor allele frequency. To this end, the proportion of minor alleles were categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group with a disease. The minor allele frequency (MAF) per SNP for each category was computed using Plink software (Purcell et al., 2007). Furthermore, the fraction of gene-specific in SNPs frequency for each gene was computed. To this end, the fraction of gene-specific in SNPs frequency was computed, assuming SNPs in upstream and downstream within a gene region are close and possibly in Linkage Disequilibrium (LD). Minor allele frequency per SNP has aggregated a gene level (see section 4.2.8).

4.2.8 Aggregating SNPs Summary Statistics at the Gene level

From each ethnic group, we gene-specific in SNPs allele. In doing so, we aggregated SNP-specific allele frequencies or proportion of ancestral/derived allele from SNPs 40kb downstream and upstream within gene region as per dbSNPs database. Under the null hypothesis, frequency/proportion P_k (k =1,..., L) with a continuous distribution, are uniformly distributed in the interval [0,1].
It follows that a parametric cumulative distribution function F can be chosen and P_k can be transformed into quantile according to $q_k = F^{-1}(P_k)$. The combined frequency/proportion $C^p = \frac{\sum_{k=1}^{L} P_k}{\sqrt{L}}$ is a sum of independent and identically distributed random variables P_k. To account for the independence assumption given correlation among neighboring genomic markers, we implement the Stouffer-Liptak method accounting for spatial correlations among SNPs within a gene or SNPs within a given sub-network. The overall statistic can be obtained by $P = \Phi(C^p)$, in which Φ is the cumulative distribution function of the standard normal distribution.

4.3 Results

4.3.1 Disease- and Actionable Gene-specific Population Structure

HIV variation is observed among Bantu, African-American, Khoisan and Afro-Asiatic, while European are clustering together (**Figure** 4.2). Most African ethnics groups have highest HIV gene-specific frequency (**Figure** 4.2), indicating and confirming that HIV infection has high incidence or prevalence among African ethnic groups compare to other ethnic groups. As for HIV, TB variation on TB-specific genes was observed among Bantu and Khoisan and Afro-Asiatic (**Figure** 4.3), while European are clustering together, except North European (explaining the know high incidence of TB in Central and North Europe). Malaria-specific world-wide ethnic groups genetics structure (**Figure** 4.5) shows that African ethnic groups and African American are still separated to the rest of other ethnic groups. UK/USA Indians and Afro-Asiatic, Latin-American and all Europeans are clustering together based on Malaria-specific genes, justifying low prevalence and/or absence of Malaria in their geographic regions. East/South Asians are clustering apart from African and European descend ethnic groups clusters. While it is known that Malaria has high prevalence among African and Asian populations, the separate cluster between them may indicate differing patterns of linkage and genetics variation at Malaria-specific genes. As expected, and as Malaria and Sickle-Cell are genetics correlated, similar results as for Malaria are observed with Sickle-Cell disease-specific genes (**Figure** 4.4).

Finally, population structure on ACG-specific genes reveals that African ethnic groups, European related ethnic groups, East-Asian, and UK/USA India and South Asian ethnic groups are separated and clustering in three different clusters (**Figure** 4.1). We observed that African-American and Afro-Asian ethics are in the convex these three clusters (**Figure** 4.1), justifying that they are result of admixture these geographic ancestral populations. In addition, Latin-America are close to European and South Asian clusters, as results of admixture, they are mainly in the convex between East-Asian, South-Asian, European and a bit distance to African. This result justifies and indicates that the actionability of these ACG genes may have differing effects on world-wide ethnic groups.

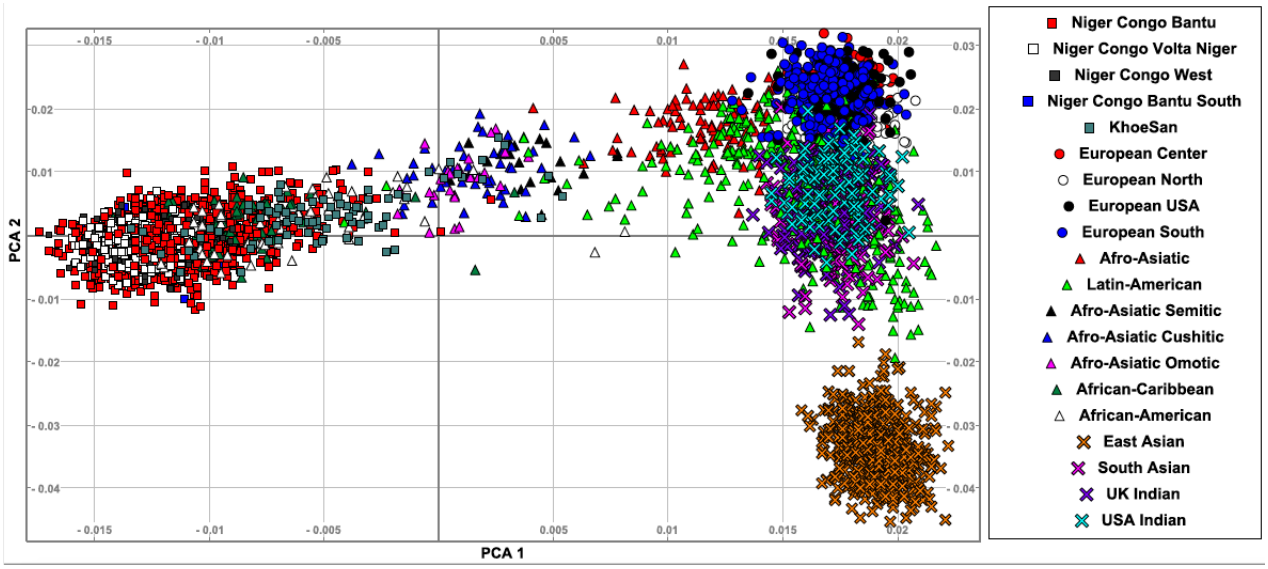


Figure 4.1: Principal Component Analysis (PCA) of the actionable genes, , plot of the first and the second eigenvectors for all populations.

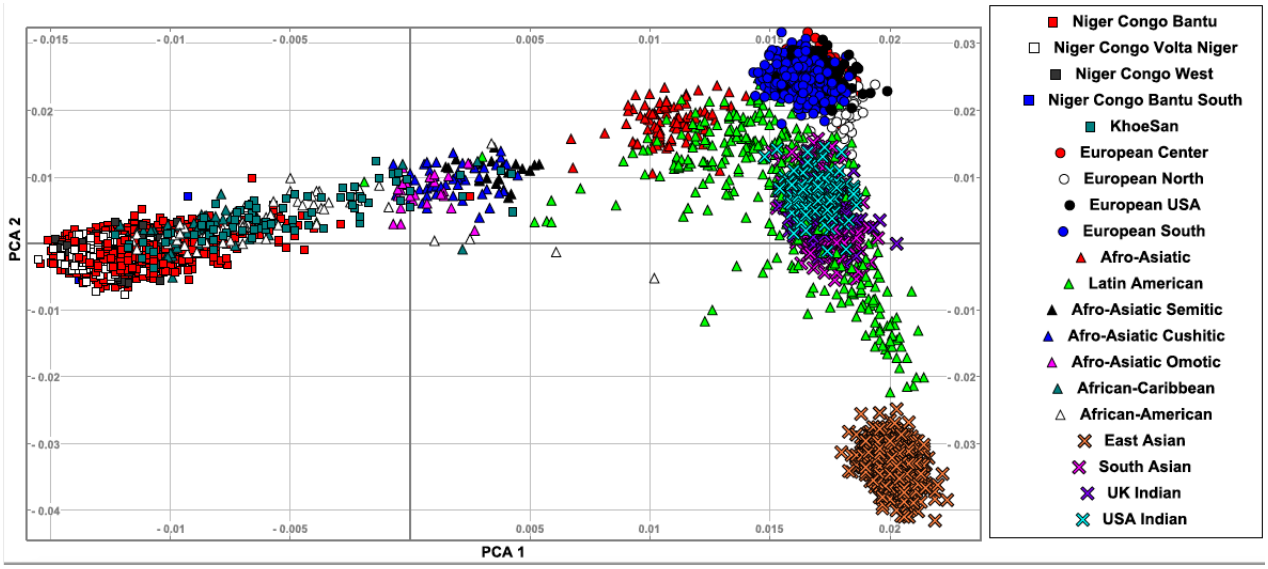


Figure 4.2: Principal Component Analysis (PCA) of genes associated with HIV, plot of the first and the second eigenvectors for all populations.

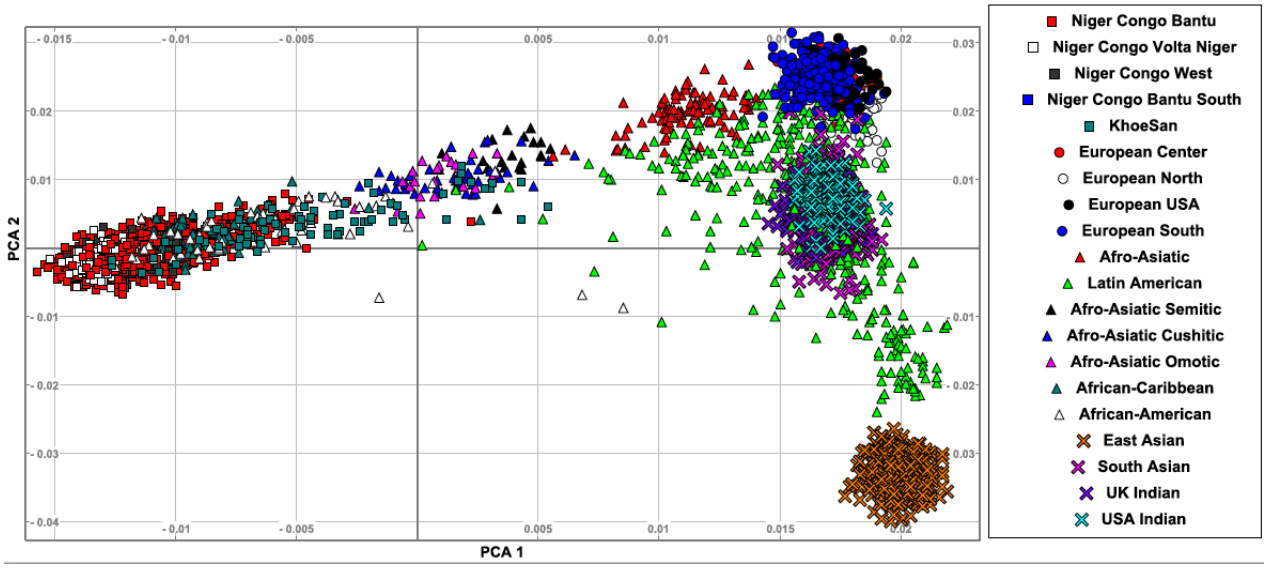


Figure 4.3: Principal Component Analysis (PCA) of genes associated with Tuberculosis, plot of the first and the second eigenvectors for all populations.

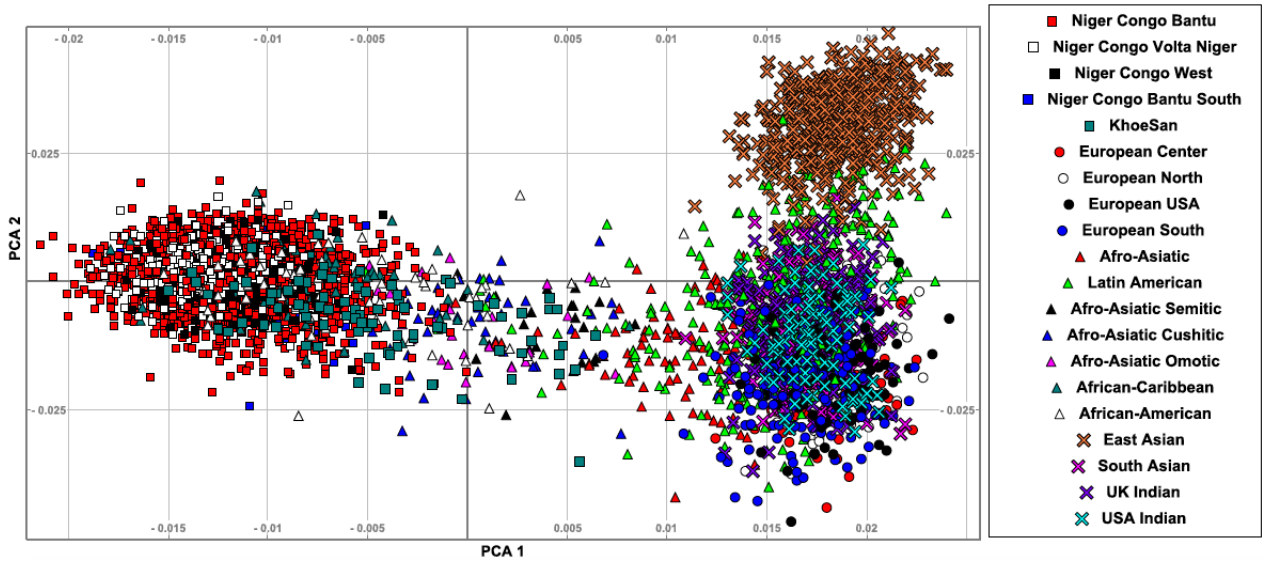


Figure 4.4: Principal Component Analysis (PCA) of genes associated with Sickle Cell Disease, plot of the first and the second eigenvectors for all populations.

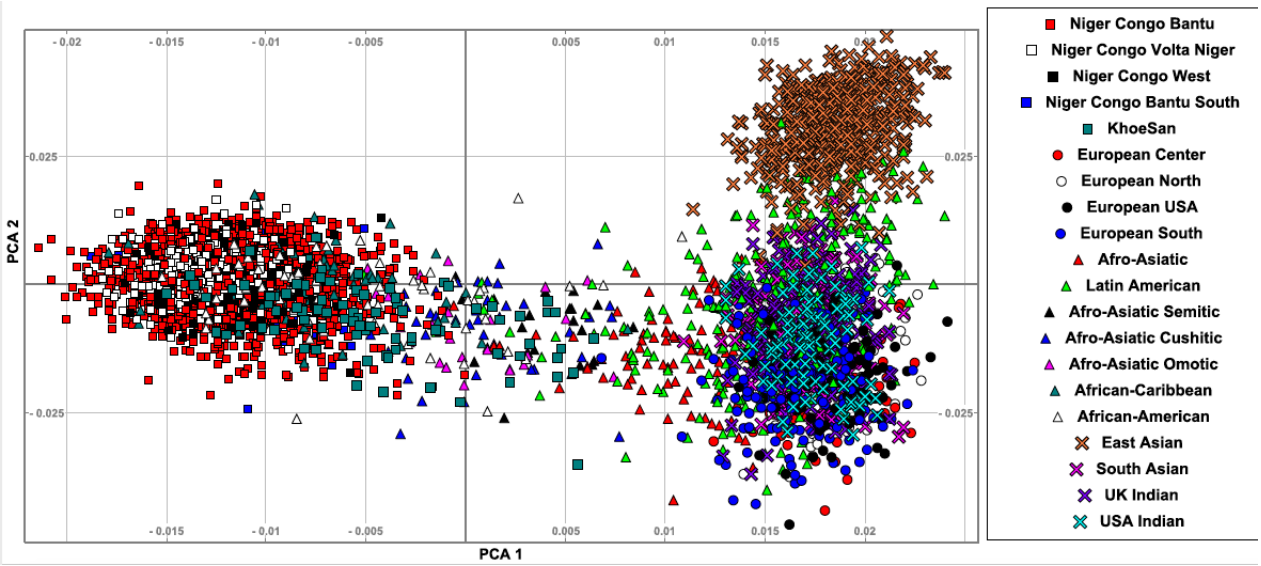


Figure 4.5: Principal Component Analysis (PCA) of genes associated with Sickle Cell Disease, plot of the first and the second eigenvectors for all populations.

4.3.2 Pathogenic Mutation at Polymorphisms within Disease Related Associated Genes

We observed considerable high proportion of pathogenic variants within ACG-specific genes from non-African ethnic groups include Latin America, Afro-Asiatic European related ethnic groups (**Figure ??**), while few genes show high proportion of pathogenic variants in Niger-Bantu, African-American (**Figure ??**). African ethnic groups include Bantu and Latin American and Afro-Asiatic have consistent considerable high proportion of pathogenic variants at these HIV-specific genes (**Figure** 4.7). While Latin American and Afro Asiatic have consistent high proportion of pathogenic variants, we observed that Khoesan group has high proportion of pathogenic variant within TB-specific genes (**Figure** 4.10). Low proportion of pathogenic variants are observed across all Malaria-specific genes in Bantu and Afro-Asiatic and Latin American ethnic groups (**Figure** 4.8), however except for Toll-like receptor 9 (*TLR9*) , *FREM3*, *IL4*, *ICAM-1* and Nitric oxide synthase 1 (neuronal) that Bantu ethnic groups and Latin America have high proportion of pathogenic variants (**Figure** 4.8). Bantu, Afro-Asiatic and Latin America have similar low proportion of pathogenic variants in most of Sickle-Cell Disease-specific genes, except in *MYO7B*, *CPS1*, *COL6A3*, *MTRR*, *SLC22A5*, *ABCC1*, and *RPL3L*.

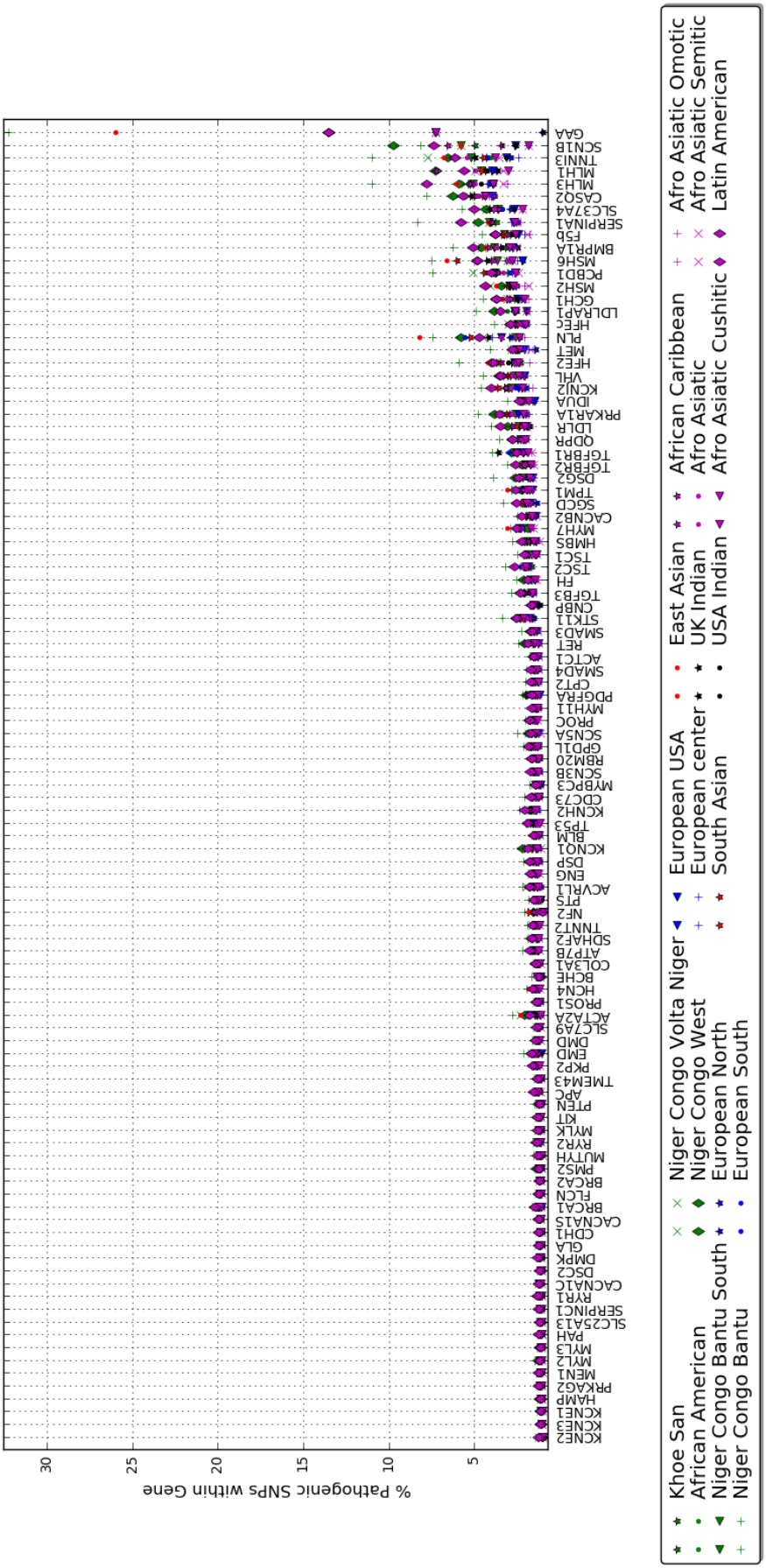


Figure 4.6: The proportion of pathogenic variants within ACG-specific (Actionable Genes) genes among all 20 ethnic groups.

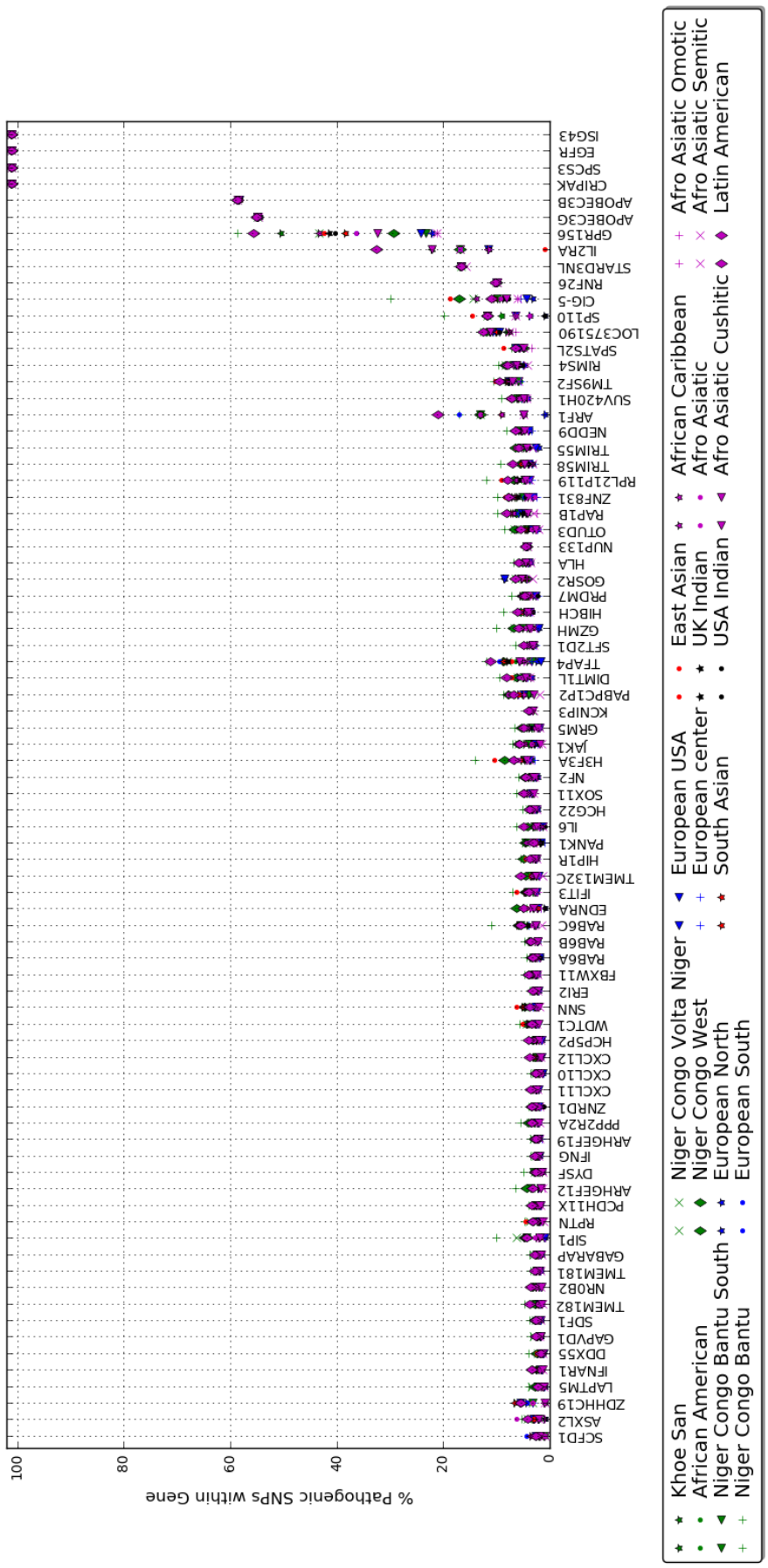


Figure 4.7: The proportion of pathogenic variants within HIV-specific genes among all 20 ethnic groups.

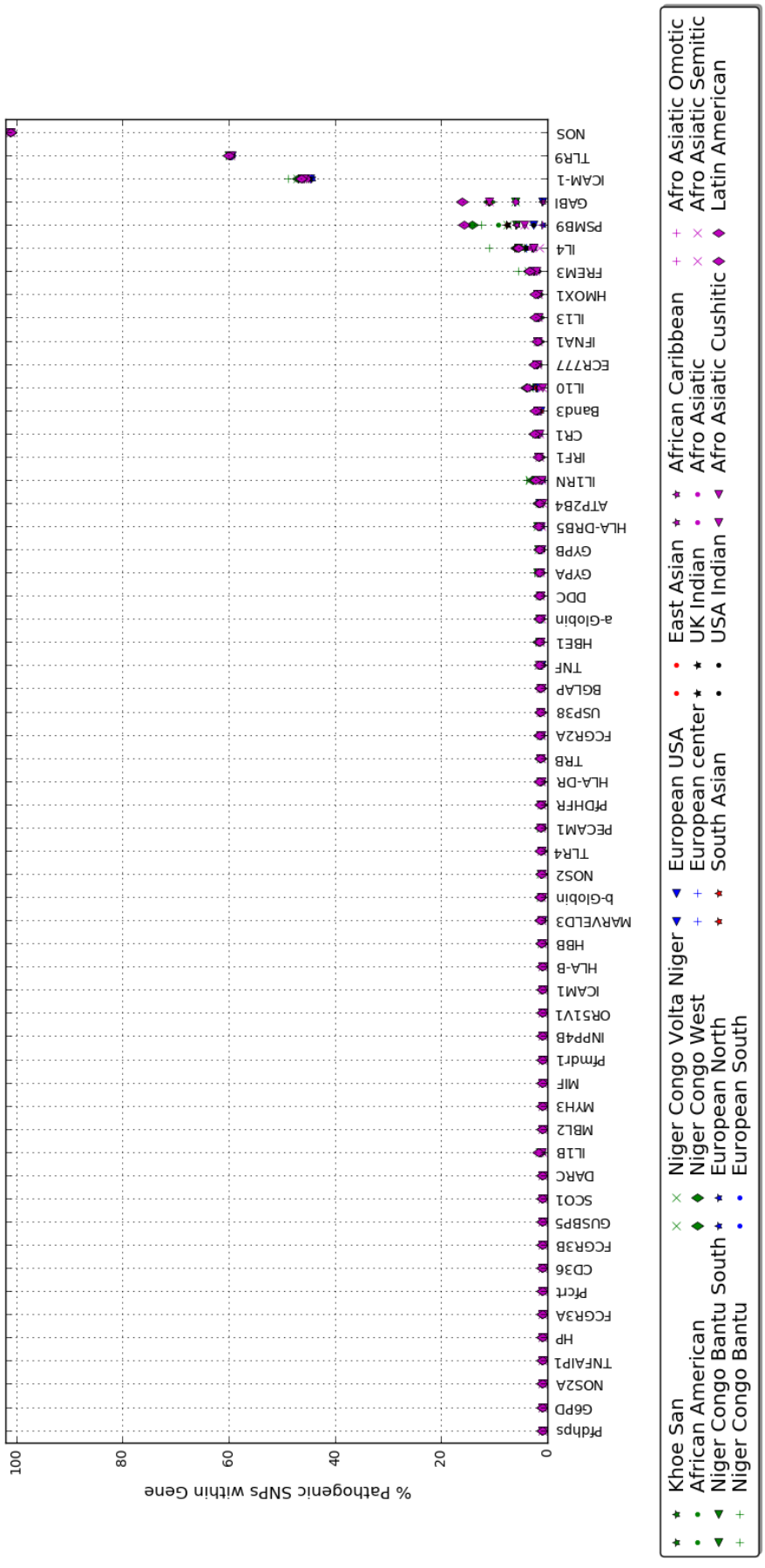


Figure 4.8: The proportion of pathogenic variants within Malaria-specific genes among 20 world-wide ethnic groups.

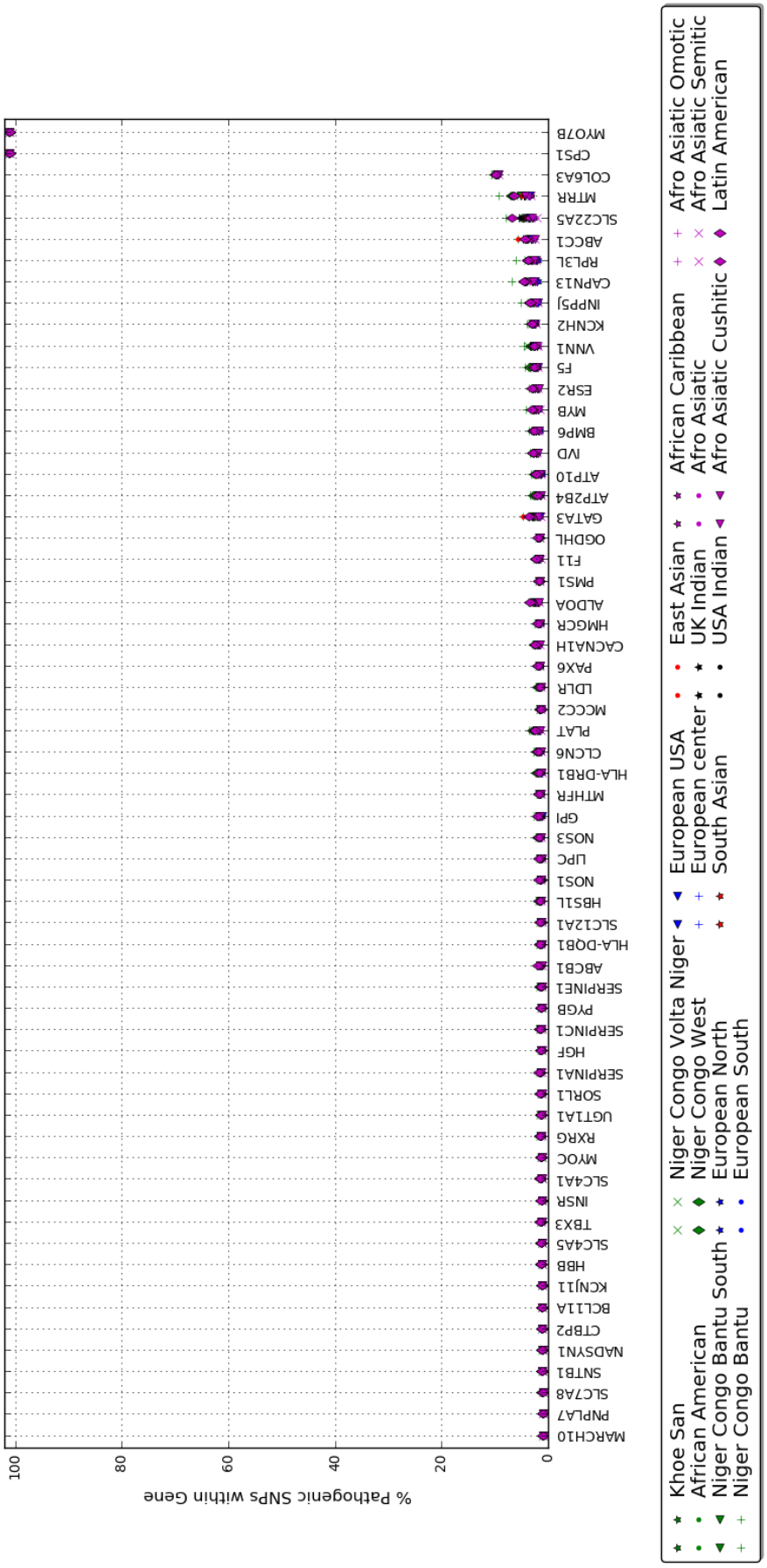


Figure 4.9: The proportion of pathogenic variants within Sickle Cell Disease-specific genes among 20 world-wide ethnic groups.

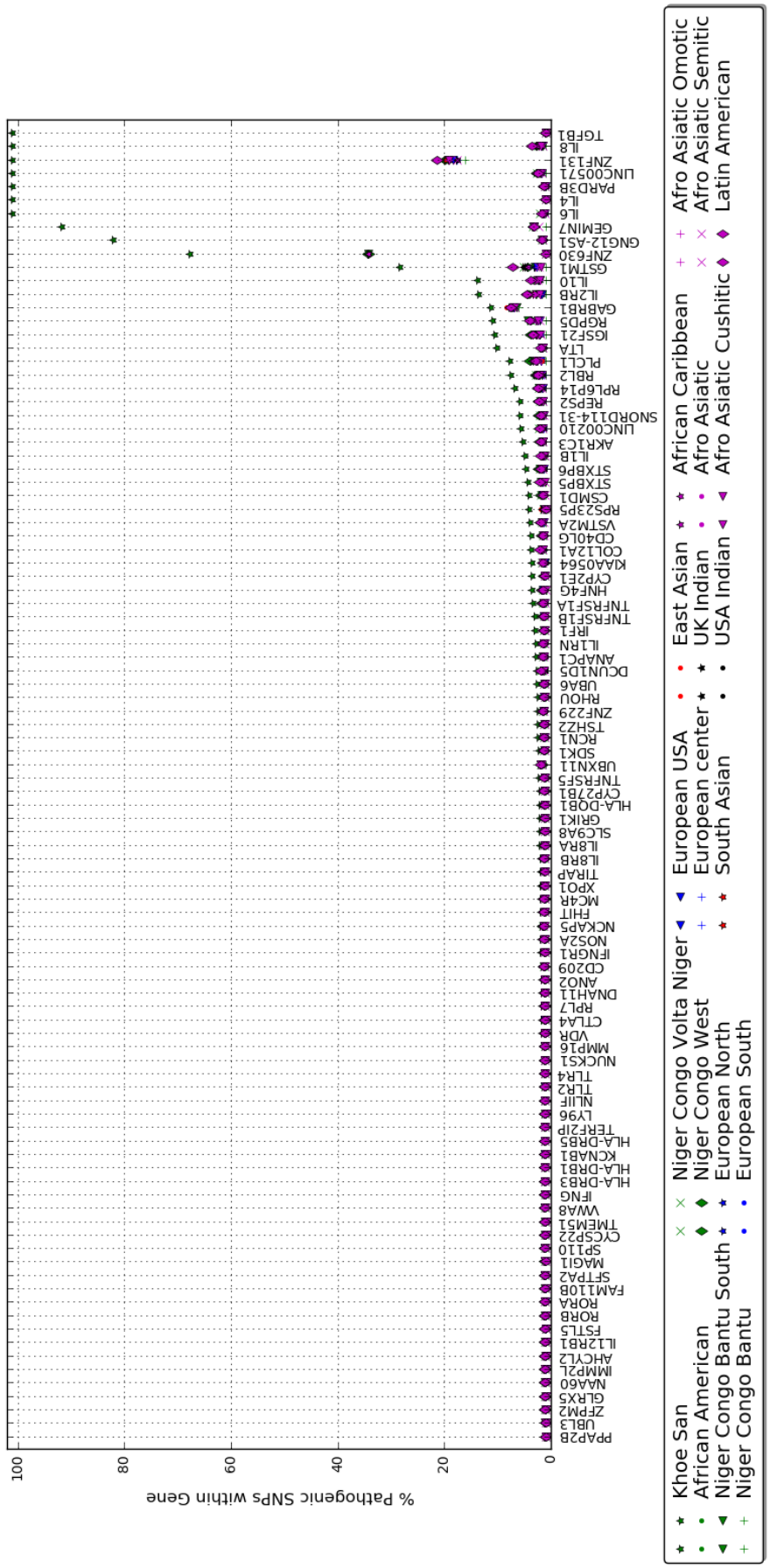


Figure 4.10: The proportion of pathogenic variants within Tuberculosis-specific genes among 20 world-wide ethnic groups.

4.3.3 Distribution of Minor Allele Frequency and Gene-specific in SNPs Frequencies

The distribution of ACG gene-specific in SNPs frequencies in **Figure** 4.11 indicates that all ACG genes have gene-specific in SNP frequencies lower than 0.4% in all ethnic groups. However, the gene-specific in SNP frequencies from most of African ethnic groups are higher compare to those from non-African ethnic groups, supporting potential differing effect and contribution of these actionable genes among world-wide ethnic groups. From **Figure** 4.12, we observe that *BTNL2*,*MOS*, *CDSN*, *USP18*, *MCM8*, *OAS1*, *COG4*, *CCL3L1*, *HLA-G*, *HLA-E*, *STT3A*, *TMED2* and *USP18* have HIV gene-specific in SNPs frequencies ranging between 5% to 15% and that African ethnic groups have the highest. In **Figure** 4.15, 33 genes have TB gene-specific in SNPs frequencies between 5% to 20% of which all African ethnic groups have the most high, suggesting that these genes may harbor common effect and contribution in TB among African ethnic groups. The distribution of Malaria gene-specific in SNPs frequencies **Figure** 4.13 suggests five genes including *GYPB*, *FCGR2A*, *IL13*, and *FREM3* with gene-specific in SNPs frequencies ranging between 4% to 15%, while all Sickle-Cell disease related genes in **Figure** 4.14 have low gene-specific in SNPs frequencies ranging between 0.1% to 0.3% among all 20 ethnic groups, but all African and diaspora ethnic groups have the highest.

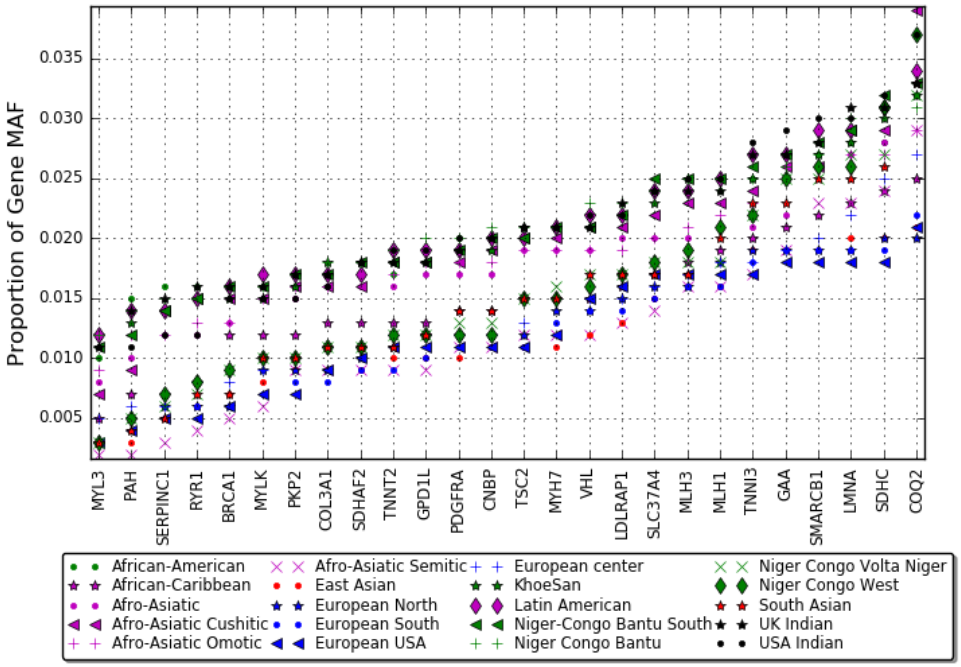


Figure 4.11: The distribution of the minor allele frequency giving a gene level (Actionable Genes) among all ethnic groups.

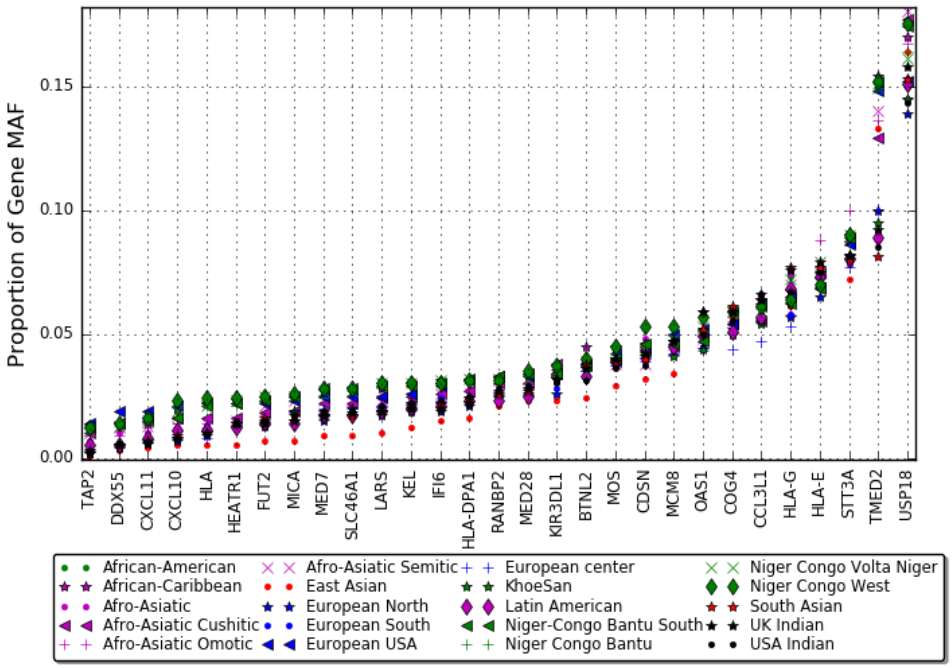


Figure 4.12: The distribution of the minor allele frequency giving a gene level (HIV) among all ethnic groups.

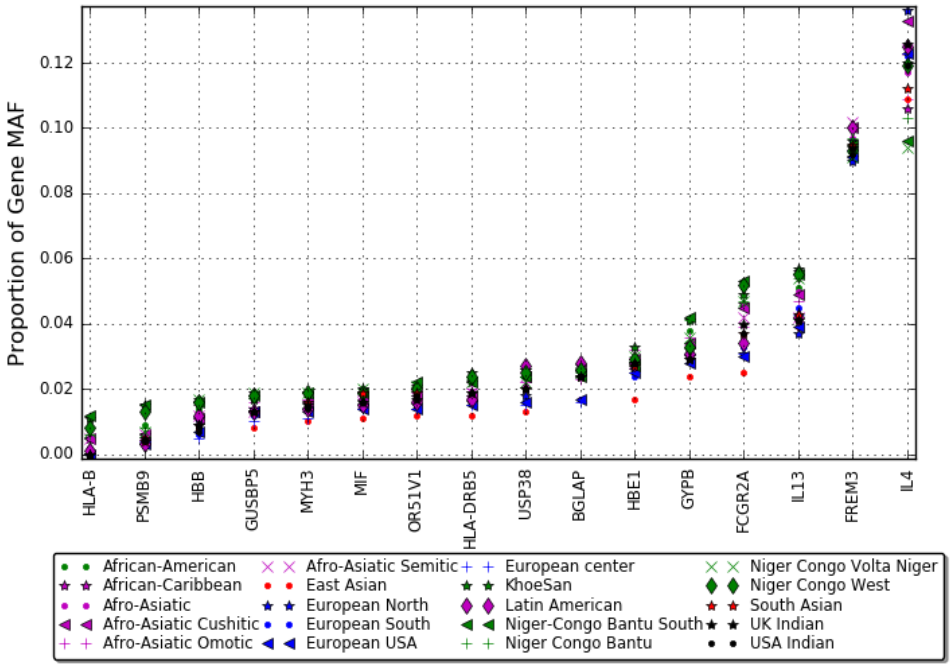


Figure 4.13: The distribution of the minor allele frequency giving a gene level (Malaria) among all ethnic groups.

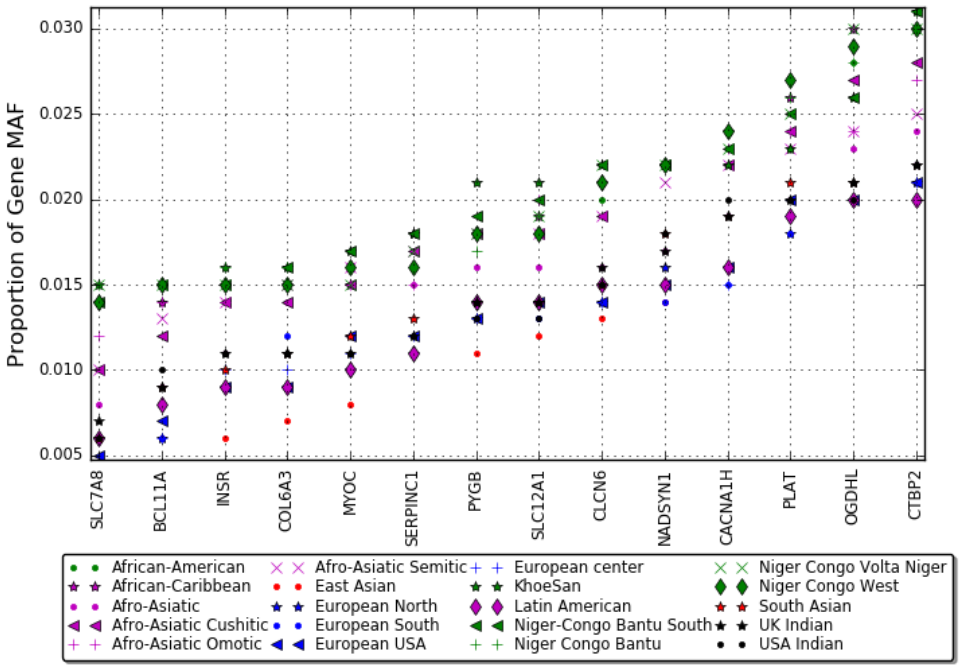


Figure 4.14: The distribution of the minor allele frequency giving a gene level (Sickle Cell Disease) among all ethnic groups.

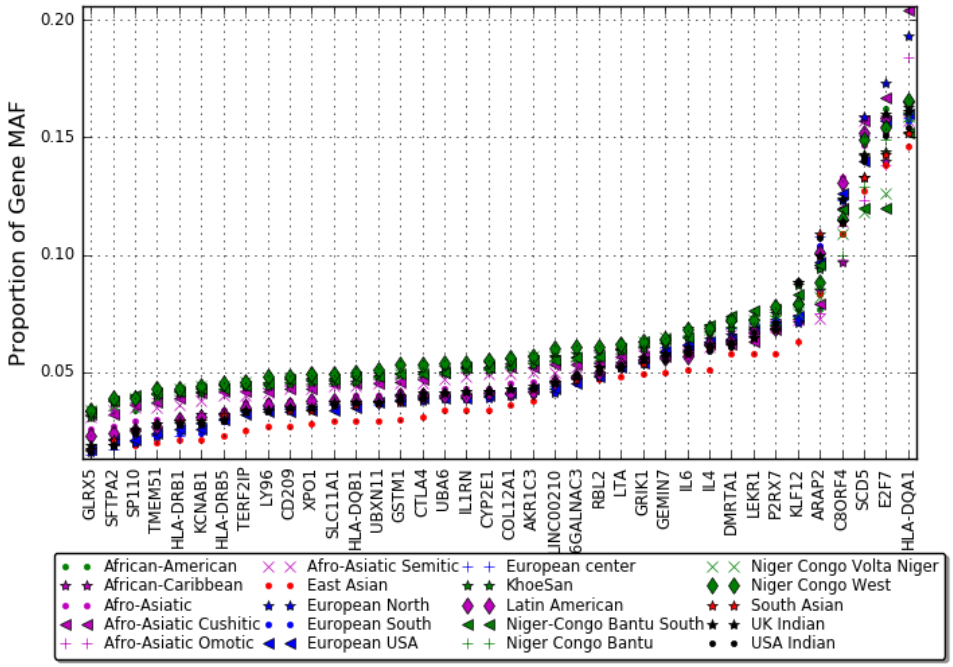


Figure 4.15: The distribution of the minor allele frequency giving a gene level (Tuberculosis) among all ethnic groups.

4.3.4 Gene-specific in Derived Allele Proportion and Relationship between Derived and Ethic-specific Minor Allele

Derived alleles are more often minor alleles ($< 50\%$ allele frequency) and more often associated with risk than ancestral alleles (Gorlova et al., 2012). Our results show high proportion of derived allele at low ethnic-specific minor allele frequency (ranging between 0.0 to 0.1), showing high variation in proportion of derived allele in TB (**Figure** 4.20), HIV (**Figure** 4.17), Malaria (**Figure** 4.18), Sick-Cell Disease (**Figure** 4.19) and set of actionable genes (**Figure** 4.16) across all African ethnics compare to the rest of other ethnic groups, and that most of African ethnics have the highest proportion of derived allele in rang of minor allele frequency bin (0.0-0.1) (**Figure** 4.16), indicating that mutation occurred in rare variants within gene-associated to HIV can play critical role in host HIV variation.

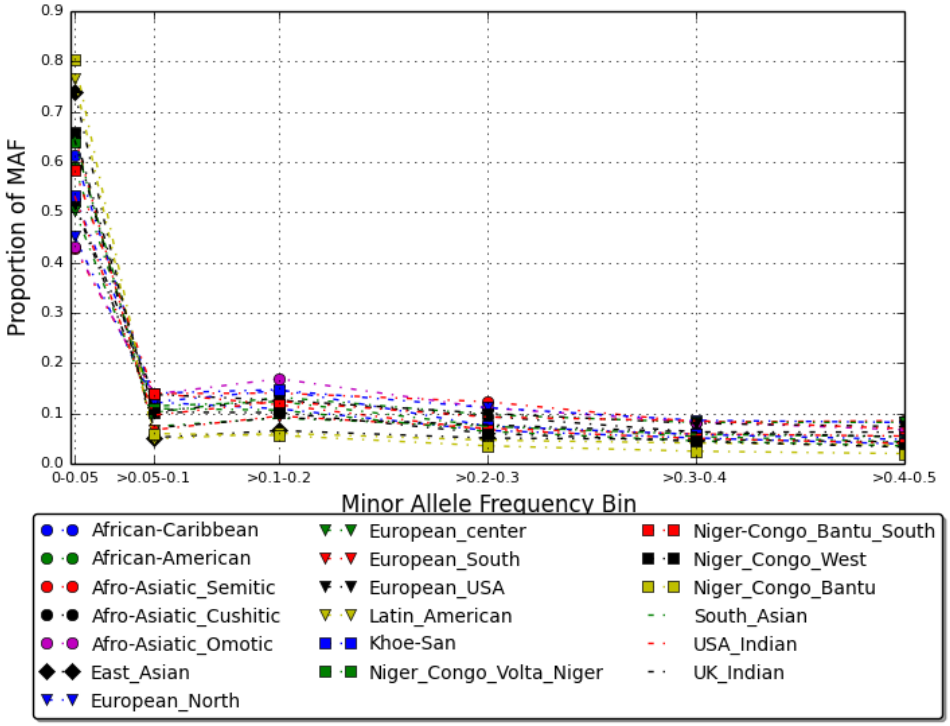


Figure 4.16: The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding Actionable Genes.

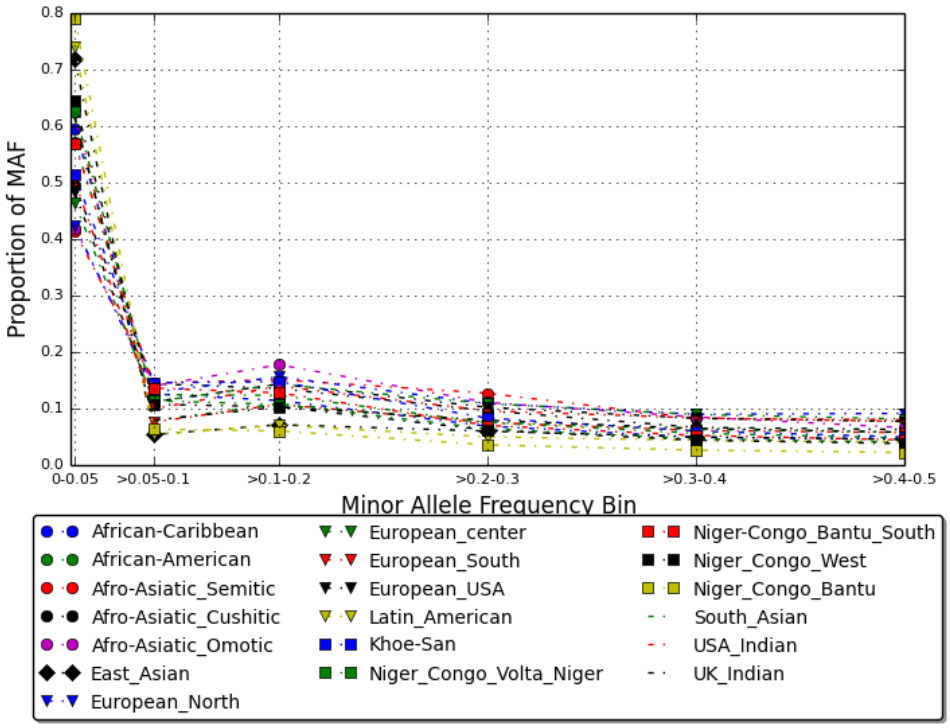


Figure 4.17: The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding HIV genes.

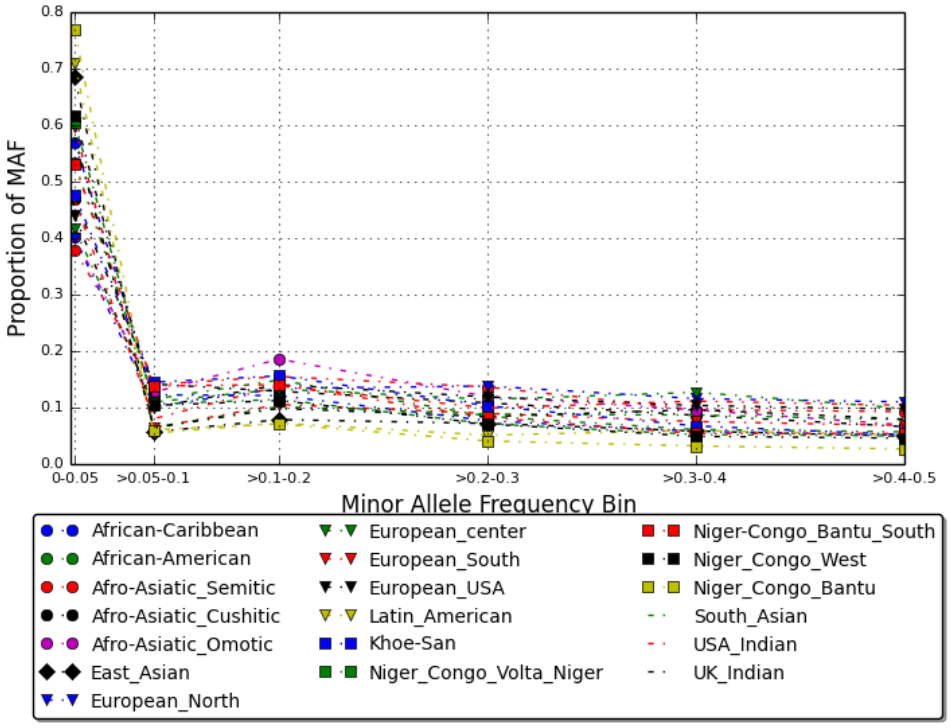


Figure 4.18: The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding Malaria genes.

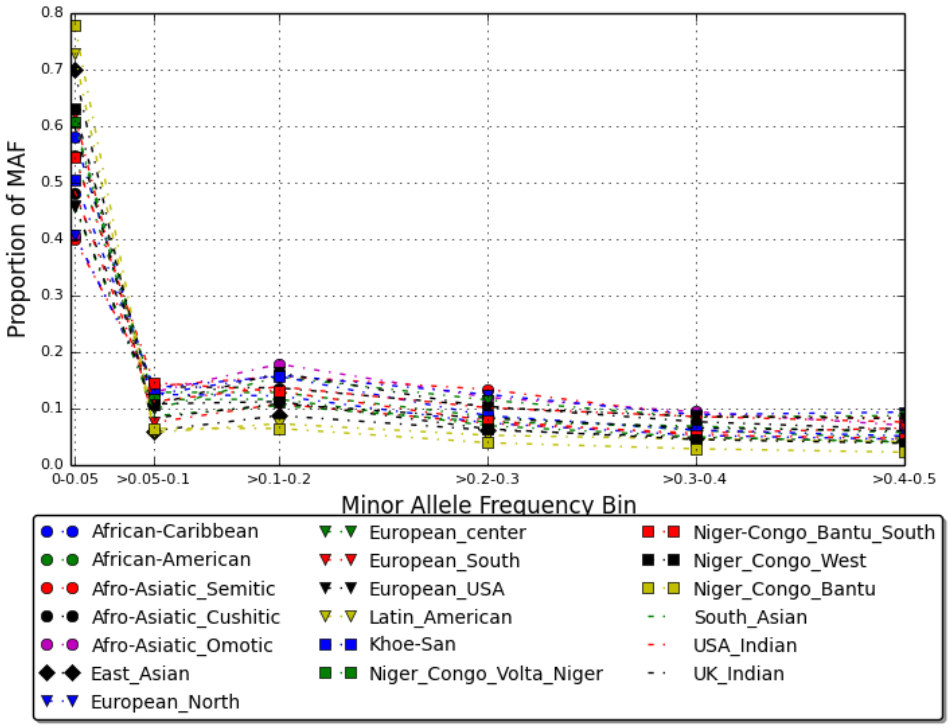


Figure 4.19: The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding Sickle cells diseases genes.

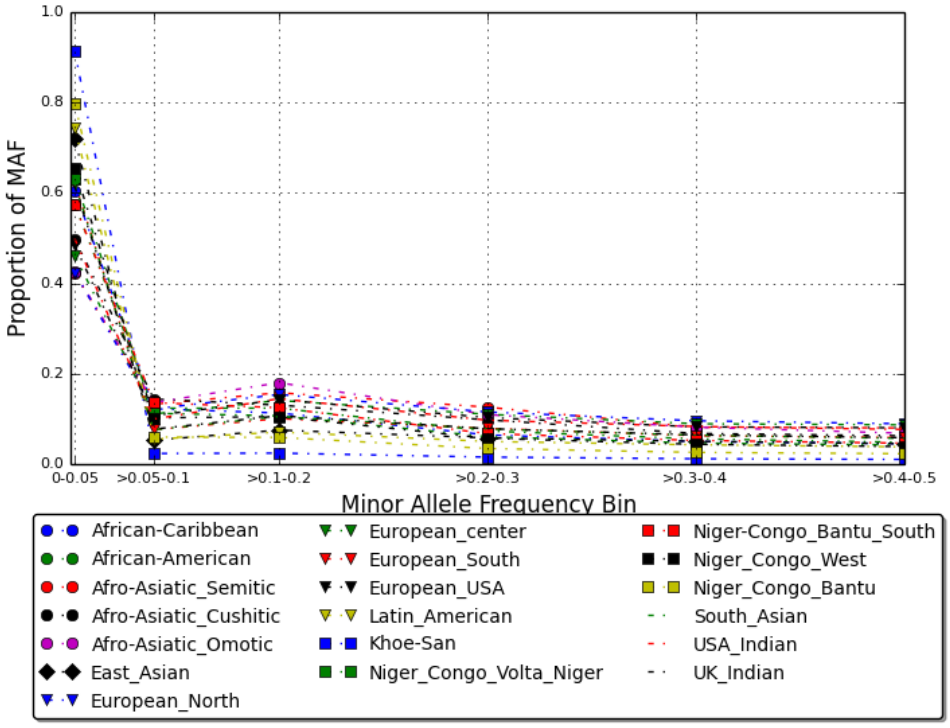


Figure 4.20: The distribution of the minor allele frequency categorised into 6 bins $(0-0.05, >0.05-0.1, >0.1-0.2, >0.2-0.3, >0.3-0.4, >0.4-0.5)$ with respect to each ethnic group regarding Tuberculosis genes.

To obtained gene-specific derived allele, derived allele frequencies were aggregated for all SNPs associated to each of these disease-specific genes (see Materials and Methods section). We observe consistent high ACG gene-specific derive allele in Latin America and most of Afro-Asiatic ethnic groups following most of European related ethnic groups (**Figure** 4.21), while a low ACG gene-specific derive allele are observed in most of African ethnic groups. One can expect actionable gene to have high proportion of derived allele, however this is not the case for most of African ethnic group, indicating that current ACG genes were primarily tailored for non-African.

For all 4 African burden diseases include HIV (**Figure** 4.22), TB (**Figure** 4.25), Malaria (**Figure** 4.23) and Sickle-Cell Diseases (**Figure** 4.24), we observe that Latin America and most of Afro-Asiatic, Bantu and Khoesan ethnic groups have considerable and consistently high proportion of gene-specific derived allele.

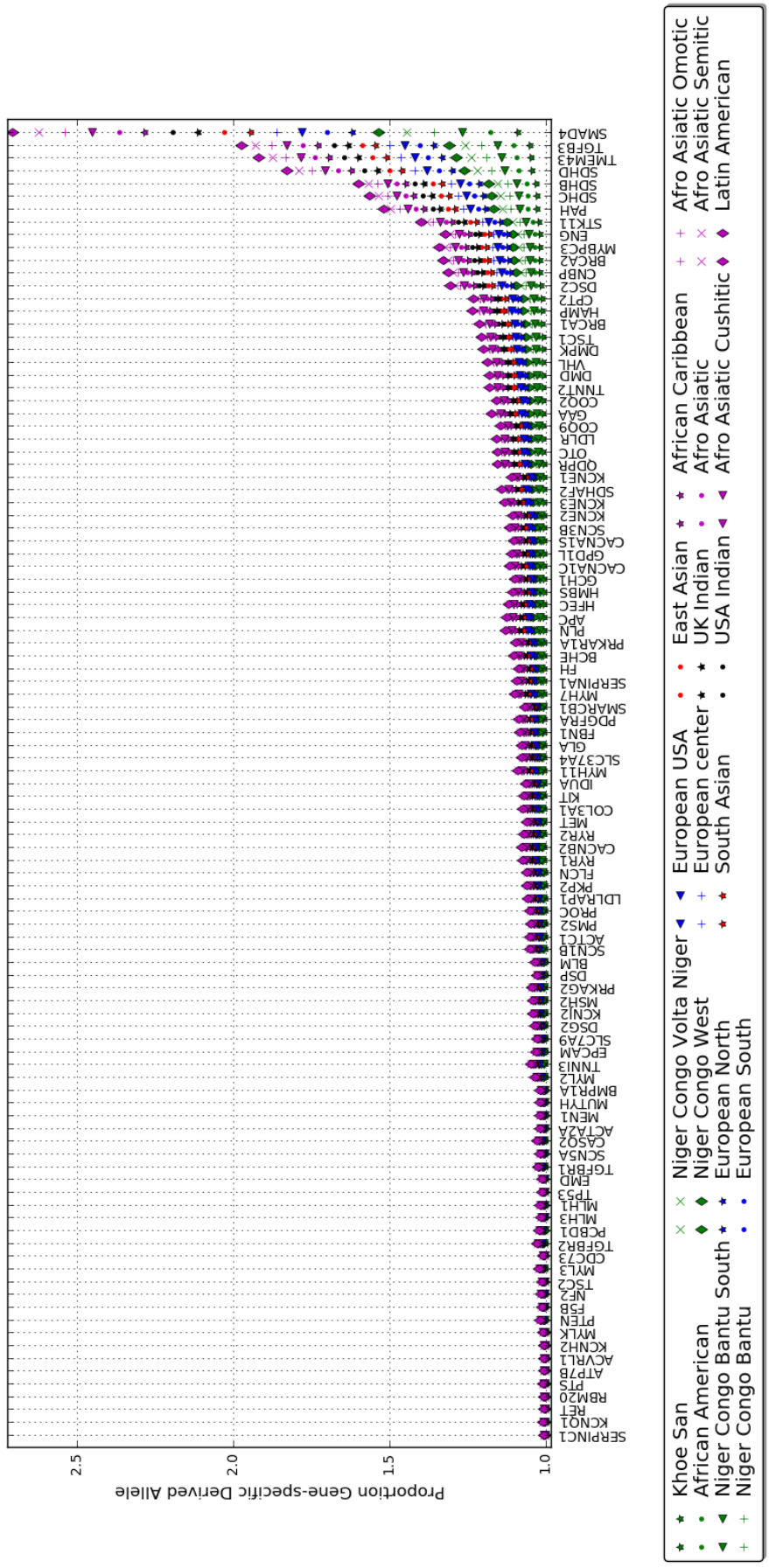


Figure 4.21: ACG Gene-specific proportion of derived allele among 20 world-wide ethnic groups.

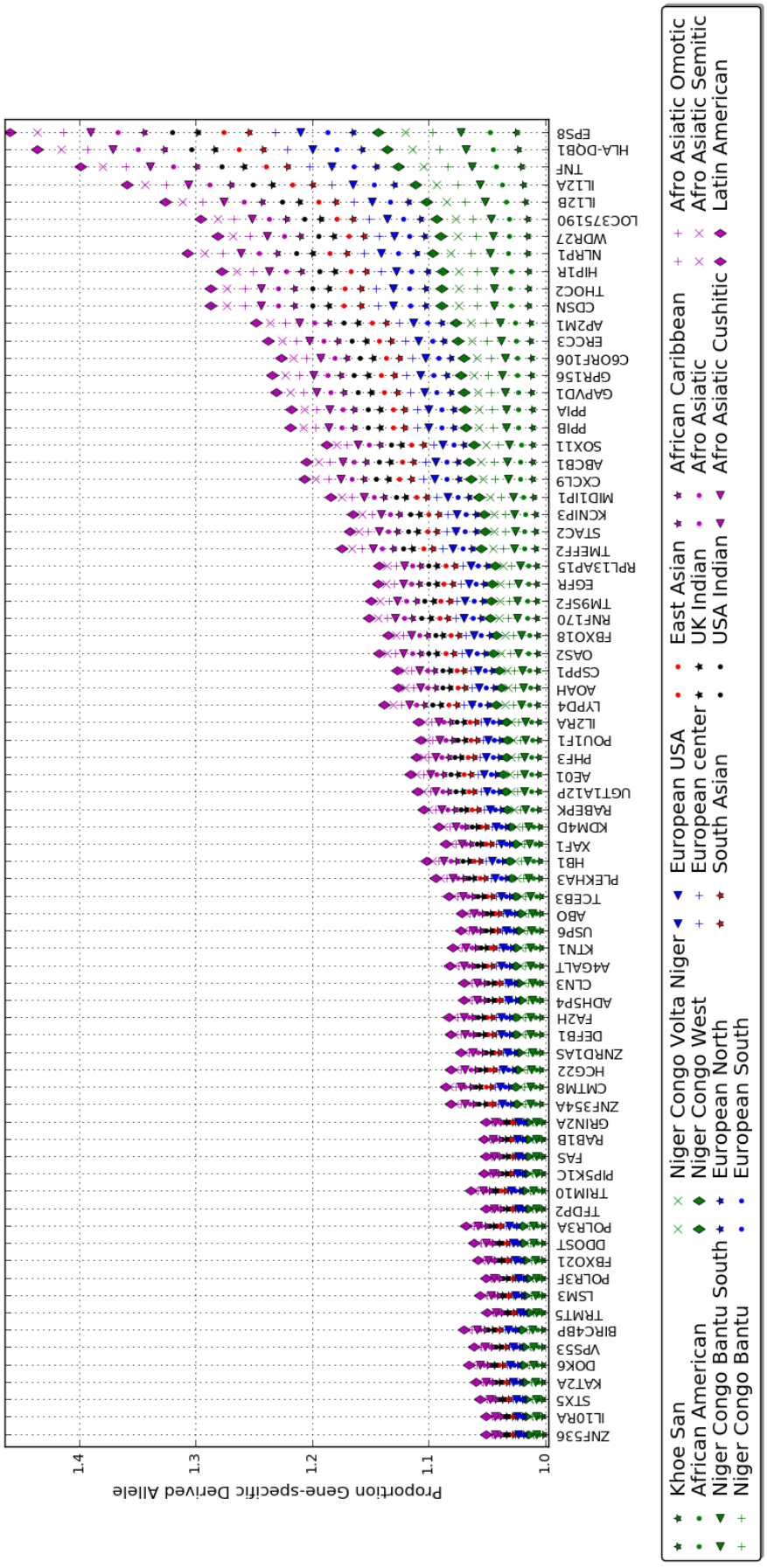


Figure 4.22: HIV Gene-specific proportion of derived allele among 20 world-wide ethnic groups.

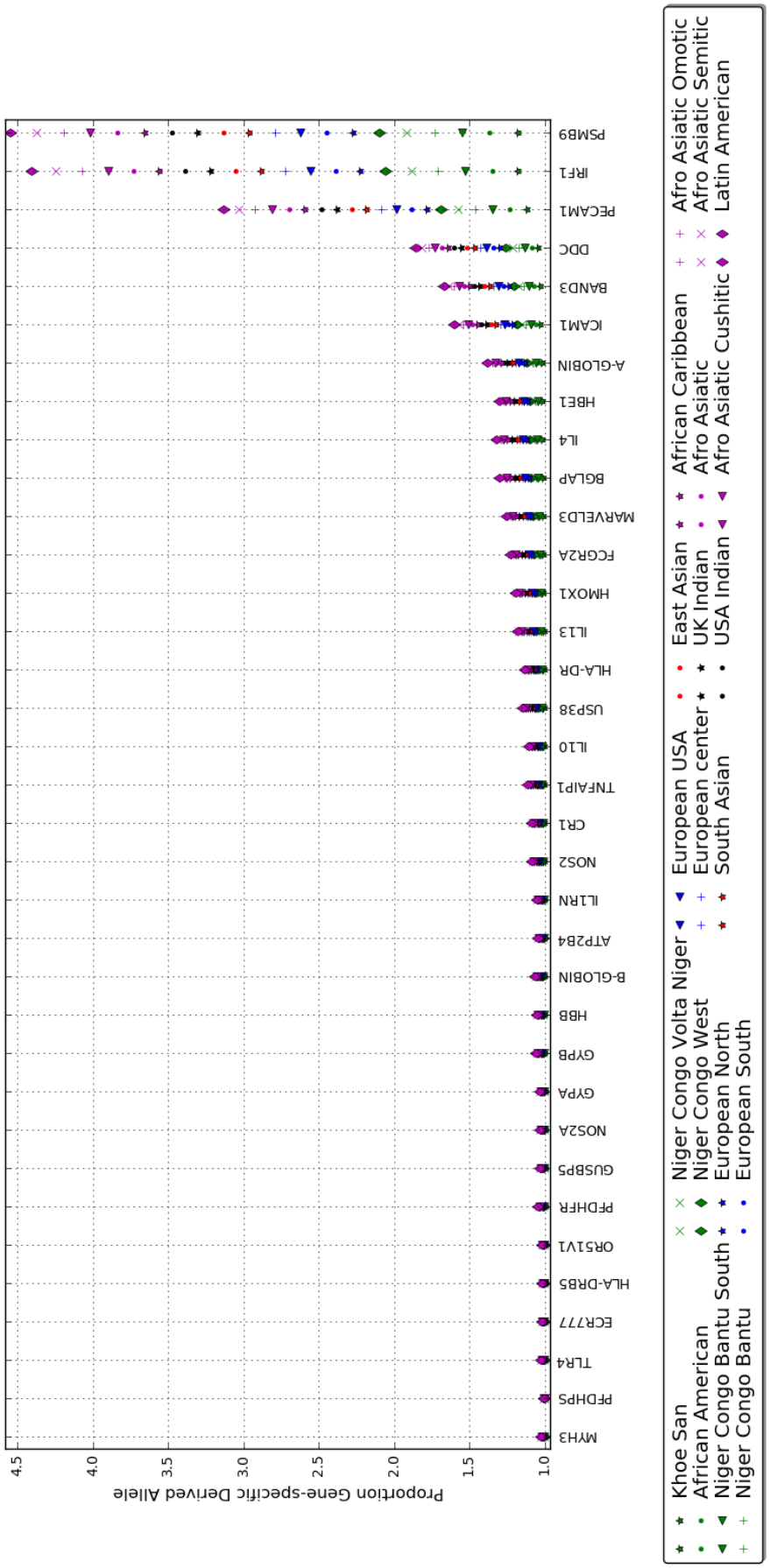


Figure 4.23: Malaria Gene-specific proportion of derived allele among 20 world-wide ethnic groups.

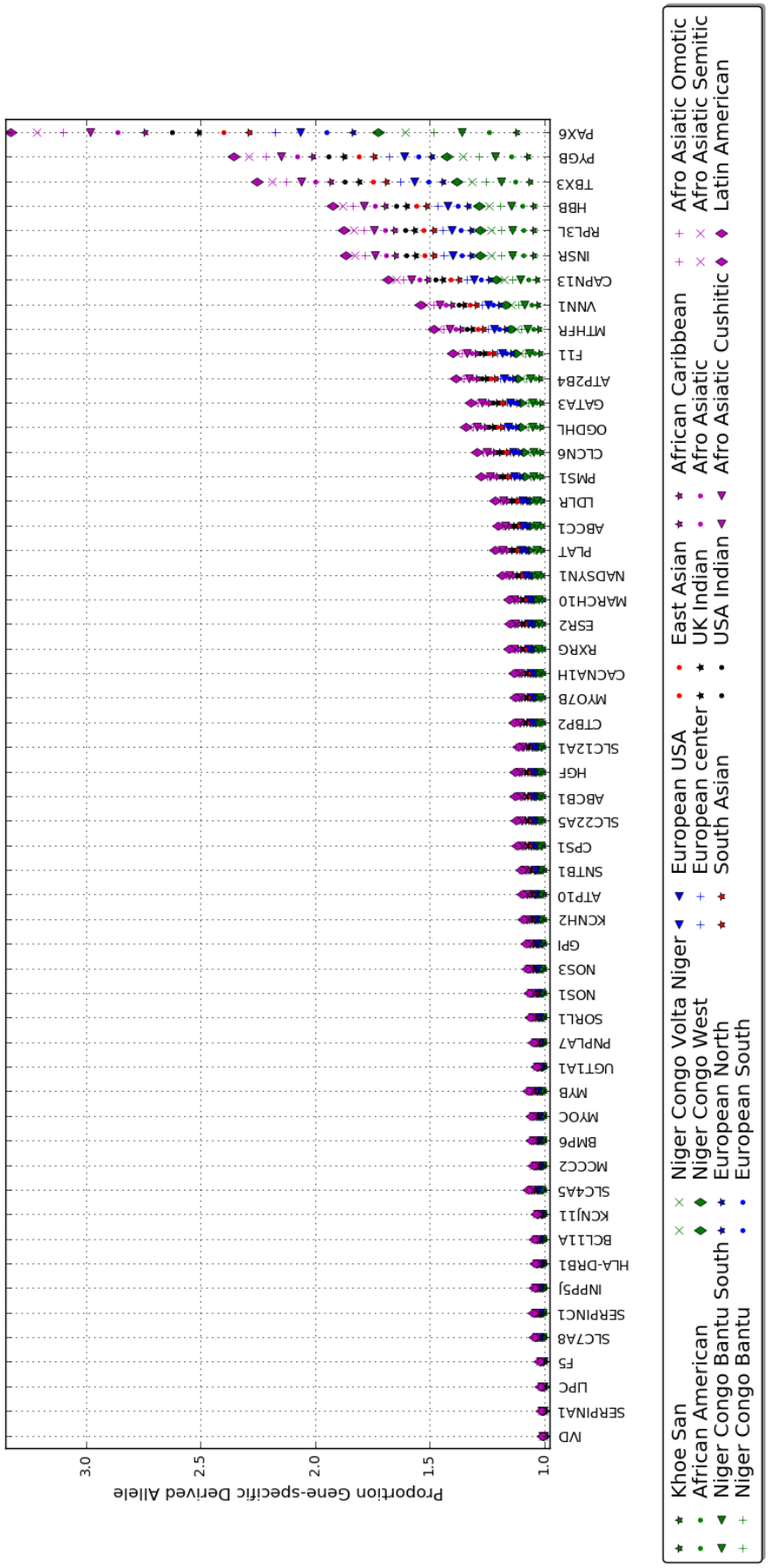


Figure 4.24: Sickle-Cell Disease Gene-specific proportion of derived allele among 20 world-wide ethnic groups.

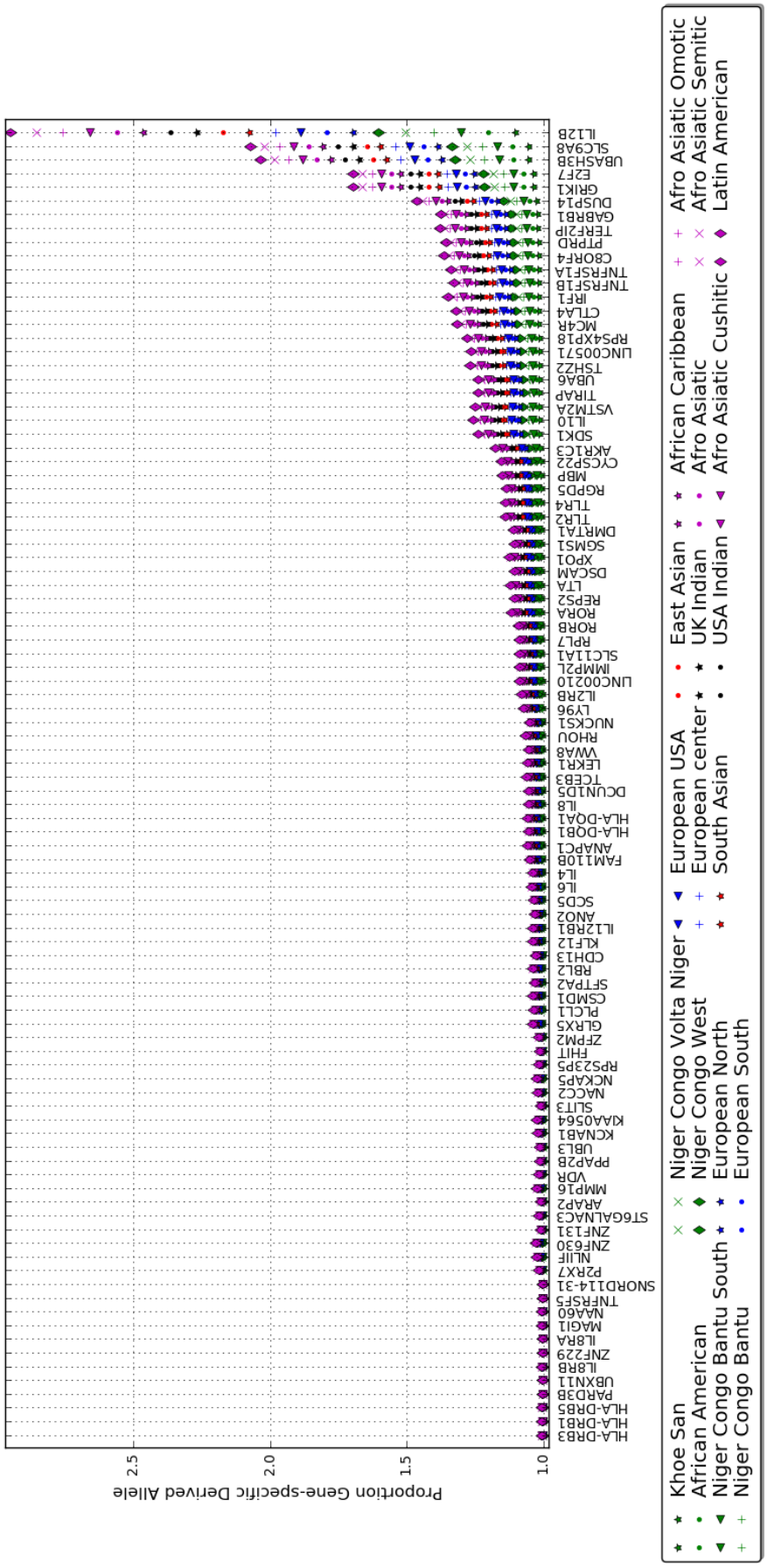


Figure 4.25: TB Gene-specific proportion of derived allele among 20 world-wide ethnic groups.

4.3.5 Genetic Diversity: Observed and Expected Heterozygosity

Gene diversity consists of two elements including the abundance (or evenness) of the alleles and the number of alleles. The abundance (or evenness) of the alleles and the number of alleles would increase the expected heterozygosity. If an ethnic group or population consists of an excess of homozygotes for different alleles this leads to a low observed heterozygosity. In **Figure** 4.26, we observe that African and African diaspora ethnic groups, particularly Bantu and Khoesan ethnic have highest gene diversity in HIV, TB, Malaria, Sickle-Cell disease and ACG associated variants. This result supports highest genetic diversity found in individuals and communities across the African continent, and that the use of personalised medicine will be beneficial both to the continent and worldwide.

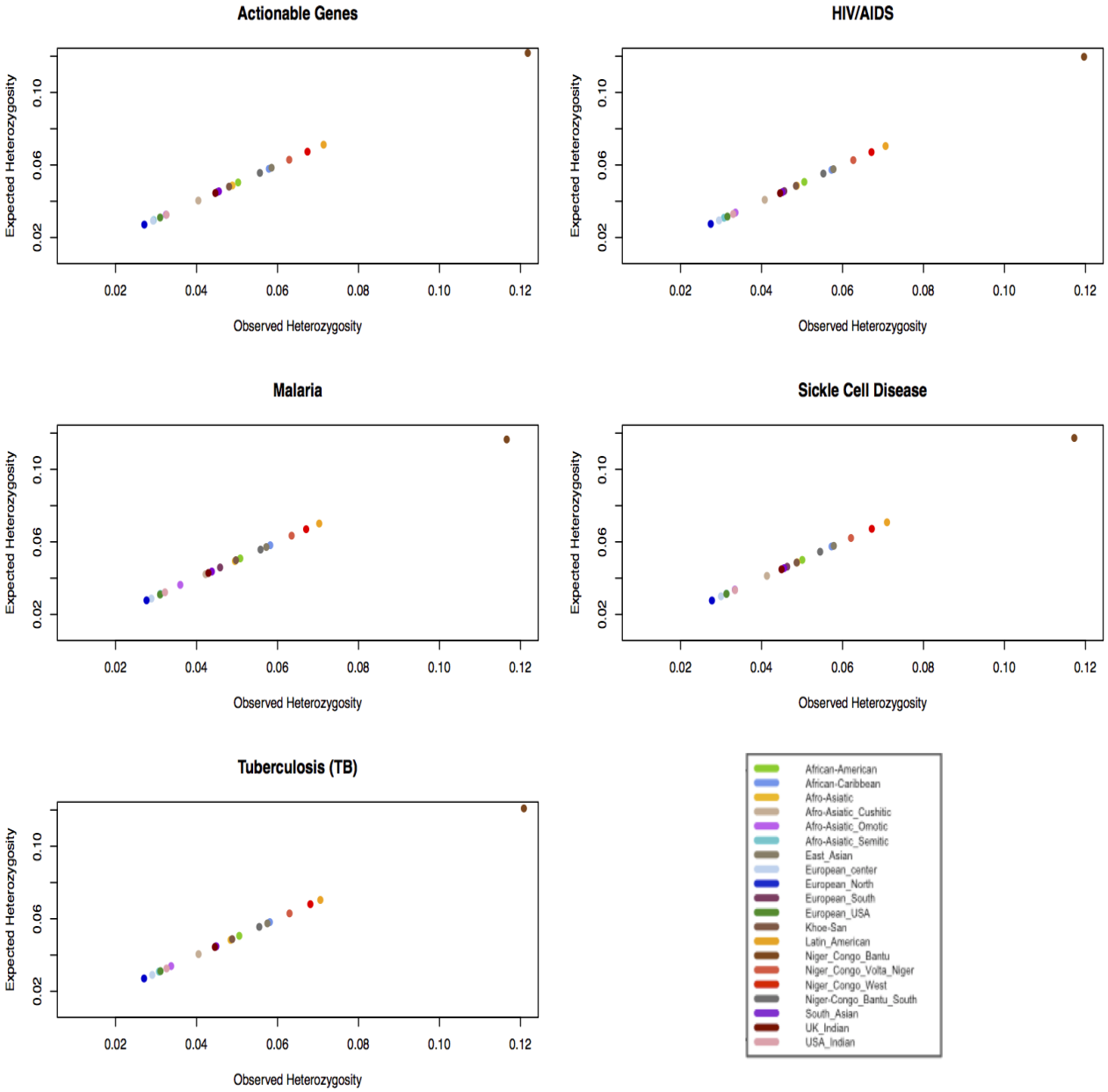


Figure 4.26: Plot Expected heterozygosity as a function of observed heterozygosity per genes of specific diseases within each population.

4.4 Discussion and Chapter Summary

This chapter intended to provide a broad assessment of possible actionability of variants known to be associated to top four burden African diseases and a list of actionable genes from American College of Medical Genetics and Genomics (ACMG) using Whole Genome Sequence data of 20 world-wide ethnic groups from a combined data of the African Genome Variation and 1000 Genome Project. We focused on list of genes related to four of Africa's burden diseases (TB, Malaria, SCD and HIV/AIDS) and most importantly, actionable genes (ACG) proposed by of ACMG. We obtained 77, 50, 75, 460 and 114 genes known to associate with Tuberculosis, Malaria, Sickle Cell disease, HIV and ACG, respectively. We examine the distribution of pathogenic mutation based on these selected known disease-genes across 20 world-wide population ethic groups.

1. **HIV/TB:** Our results indicated that African ethnic groups include Bantu and Latin American and Afro-Asiatic have the highest proportion of pathogenic variants based on 483 HIV-specific genes. From 77 TB-specific genes, we observed that Latin American and Afro Asiatic ethnic groups have the highest proportion of pathogenic variants, important among all African and African diaspora ethnic groups, only Khoesan has high proportion of pathogenic variant within TB-specific genes.

2. **Malaria and Sickle-Cell :** Our result indicate absent of pathogenic variants in most of European related ethnic groups and low proportion of pathogenic variants across all Malaria-specific genes in Bantu and Afro-Asiatic and Latin American ethnic groups, except for Toll-like receptor 9 (*TLR9*) , *FREM3*, *IL4*, *ICAM-1* and Nitric oxide synthase 1 (neuronal) that Bantu ethnic groups and Latin America have high proportion of pathogenic. Furthermore, Bantu, Afro-Asiatic and Latin America have similar low proportion of pathogenic variants in most of Sickle-Cell disease-specific genes, except in *MYO7B*, *CPS1*, *COL6A3*, *MTRR*, *SLC22A5*, *ABCC1*, and *RPL3L*.

3. **ACG :** Our present result showed a considerable high proportion of pathogenic variants within ACG-specific genes from non-African ethnic groups include Latin America, Afro-Asiatic European related ethnic groups compare to most of African related ethnic groups. This result justifies and indicates that the actionability of these ACG genes may have differing effects on world-wide ethnic groups.

We leveraged the dbSNP database to extract SNPs associated with these genes per diseases. The obtained SNPs per disease were thus extracted from the whole phased data contained 4,932 samples of these 20 ethnic groups; yield 5 disease-specific phased haplotypes data sets. From these phased haplotypes data, we conducted disease gene-specific population structure, we examined the distribution and relationship of derived and minor allele frequency and estimate the expected and observed heterozygosity.

Our result suggests significant genetic variation among all non-European ethnic groups, mostly African and African diaspora ethnic groups, while all European ethnic groups are genetically and consistently clustering together based on these diseases or actionable-specific variants. In addition, our result indicates that African and African diaspora ethnic groups, particularly Bantu and Khoesan ethnic have the highest gene diversity in HIV, TB, Malaria, Sickle-Cell disease and ACG associated variants. This supports the highest genetic diversity found in individuals and communities across the African continent, and that the use of personalised medicine will be beneficial both to the continent and worldwide.

1. **HIV/TB :** Most African ethnics groups (Bantu and Khoisan) have highest HIV and TB gene-specific frequency, indicating that HIV infection has high incidence or prevalence among African ethnic groups compare to other ethnic groups. Our result identifies *BTNL2*,*MOS*, *CDSN*, *USP18*, *MCM8*, *OAS1*, *COG4*, *CCL3L1*, *HLA-G*, *HLA-E*, *STT3A*, *TMED2* and *USP18* to have HIV gene-specific in SNPs frequencies ranging between 5% to 15% and that African ethnic groups have the highest. In addition, 33 genes have TB gene-specific in SNPs frequencies between

5% to 20% of which all African ethnic groups have the highest frequencies. This suggests that these genes may harbor common effect and contribution in TB/HIV among African ethnic groups. Furthermore, HIV/TB gene-specific have high proportion of derived allele at low African ethnic-specific minor allele frequency (0.0 to 0.1) and that these proportion derived allele vary among African ethnic groups.

2. **Malaria and Sickle-Cell :** We identify five genes including *GYPB*, *FCGR2A*, *IL13*, and *FREM3* with Malaria gene-specific in SNPs frequencies ranging between 4% to 15%, while all Sickle-Cell disease related genes have low gene-specific in SNPs frequencies ranging between 0.1% to 0.3% among all 20 ethnic groups, but all African and diaspora ethnic groups have the highest in that range.

3. **ACG :** Our result indicates all ACG genes have gene-specific in SNP frequencies low than 0.4% in all ethnic groups. However, the gene-specific in SNP frequencies from most of African ethnic groups are higher compare to those from non-African ethnic groups, supporting potential effect and contribution of these actionable genes among world-wide ethnic groups. High ACG gene-specific derive allele, was observed in Latin America and most of Afro-Asiatic ethnic groups following most of European related ethnic groups, while a low ACG gene-specific derive allele are observed in most of African ethnic groups.

Overall, the results in this chapter suggest that given the highest genetic diversity found in African ethnics and African diaspora related ethnics at these four Africa burden diseases and current actionable gene associated, (1) the use of personalised medicine will be beneficial both to the African continent and worldwide; (2) enabling a recommendation for African-specific actionable list of genes will further improve African and diaspora healthcare.

Chapter 5

General Discussion and Conclusion

Advances in sequencing technologies have facilitated the development of novel statistical genomics approaches with applications ranging from clinical care to pharmaceutical industries. These have led to an unprecedented increase in the computational complexity of downstream data analysis. An obstacle to validating and bench-marking methods for genome analysis is that there are few reference data sets available for which the "ground truth" about the mutational landscape of the sample genome is known and fully validated. Furthermore, accuracy, effectiveness and performance assessments of different analytical methods used to analyse next generation sequence data are important aspects of medical population genetics.

In this project, we provide a broad discussion on DNA sequence simulation tools. Further, we have described the process of DNA sequence simulation tools, and we have illustrated some example for each tool representing different NGS platforms. Since, variant calling (VC) is an important aspect of genomics studies as polymorphism information can be used to influence important clinical decisions, the project has dissected and discussed 9 current state-of-the-art variant calling tools. In doing so, we made use of NEAT-GenReads, to simulate a total of 100 DNA sequence samples of which every 50 samples mimicked the genetics background of the African and European population, respectively at different coverage (high and low), respectively. We have evaluated the simulation outcomes to ensure our result is not truncated or damaged. We have checked the quality of FastQ files by using FastQC tool, checking golden BAM, VCF by using samtools features.

The tremendous development of Next-Generation sequencing has promoted the evolution of personalised medicine. Such progress in NGS resulted in an enlargement of the downstream analysis tools to handle such data. Such an example, many variant calling tools have been developed, which raise the question of which tool one can choose when dealing with a complex and diverse genome such as the African genome? To address this question, we compared nine variant calling tools include VarScan2, SAMtools, GATK-HaplotypeCaller, SNVer, BCFtools, FreeBayers, Lofreq, PlatyPus and VarDict) based on simulated data representing two population (African and European) at varying coverage (high and low) as a total of 4 data sets. These tools were evaluated based on the sensitivity and precision and most importantly, the F-score to measure the overall performance, low sensitivity and precision result in low F-score.

The result of this evaluation suggests that Lofreq can currently be a great option in calling WGS from African populations and VarDict can be considered for targeted sequence, particularly for African data. Given huge amount of non-overlapped variants across the results from current VC tools and differing number of variants discovered across these tools, it will be reasonable to consider multi variant calling tools to allow cross validation of variants discovered. In the next chapter, we will use Lofreq to call the variants on WGS of real populations.

Given the result the above evaluation, we have leveraged Whole Genome Sequence data of 20 world-wide ethnic groups from a combined data of the African Genome Variation and 1000 Genome Project to provide a broad assessment of possible actionability of variants known to be associated to top four burden African diseases (HIV/AIDS, Malaria, Tuberculosis (TB), and Non-communicable diseases: such as Sickle cell disease) and a list of actionable genes from American College of Medical Genetics and Genomics (ACMG).

Our analysis focused on the obtained list of genes related to four of Africa's burden diseases (TB, Malaria, SCD and HIV/AIDS) and most importantly, actionable genes (ACG) proposed by of ACMG. We obtained 77, 50, 75, 460 and 114 genes known to be associated to Tuberculosis, Malaria, Sickle Cell disease, HIV and ACG, respectively. We examine the distribution of pathogenic mutation based on these selected known disease-genes across 20 world-wide population ethic groups.

1. Our results indicated that African ethnic groups include Bantu and Latin American and Afro-Asiatic have the highest proportion of pathogenic variants based on 460 HIV-specific genes. From 77 TB-specific genes, we observed that Latin American and Afro Asiatic ethnic groups have the highest proportion of pathogenic variants, important among all African and African diaspora ethnic groups, only Khoesan has high proportion of pathogenic variant within TB-specific genes.

2. Our result indicate absent of pathogenic variants in most of European related ethnic groups and low proportion of pathogenic variants across all Malaria-specific genes in Bantu and Afro-Asiatic and Latin American ethnic groups, except for Toll-like receptor 9 (*TLR9*) , *FREM3*, *IL4*, *ICAM-1* and Nitric oxide synthase 1 (neuronal) that Bantu ethnic groups and Latin America have high proportion of pathogenic. Furthermore, Bantu, Afro-Asiatic and Latin America have similar low proportion of pathogenic variants in most of Sickle-Cell disease-specific genes, except in *MYO7B*, *CPS1*, *COL6A3*, *MTRR*, *SLC22A5*, *ABCC1*, and *RPL3L*.

3. Our present result illustrated a considerable high proportion of pathogenic variants within ACG-specific genes from non-African ethnic groups include Latin America, Afro-Asiatic European related ethnic groups compare to most of African related ethnic groups. This result justifies and indicates that the actionability of these ACG genes may have differing effects on world-wide ethnic groups.

We leveraged the dbSNP database to extract SNPs associated with these genes per each of these four Africa's burden diseases (TB, Malaria, SCD and HIV/AIDS) and the set of actionable genes (ACG). The obtained SNPs per disease were thus extracted from the whole phased data contained 4,932 samples of these 20 ethnic groups; yield 5 disease-specific phased haplotypes data sets. From these phased haplotypes data sets, we independently conducted disease gene-specific population structure, we examined the distribution and relationship of derived and minor allele frequency and estimate the expected and observed heterozygosity.

Our result illustrated a significant genetic variation among all non-European ethnic groups mostly but all European ethnic groups are genetically and consistently clustering together based on these diseases or actionable-specific variants. In addition, our result indicates that African and African diaspora ethnic groups, particularly Bantu and Khoesan ethnic have the highest gene diversity in HIV, TB, Malaria, Sickle-Cell disease and ACG associated variants. This supports the highest genetic diversity found in individuals and communities across the African continent, and that the use of personalised medicine will be beneficial both to the continent and worldwide. Furthermore, our results suggested the follows,

1. Most African ethnics groups (Bantu and Khoisan) had highest HIV and TB gene-specific frequency, indicating that HIV infection has high incidence or prevalence among African ethnic groups compare to other ethnic groups. Our result identified *BTNL2*,*MOS*, *CDSN*, *USP18*, *MCM8*, *OAS1*, *COG4*, *CCL3L1*, *HLA-G*, *HLA-E*, *STT3A*, *TMED2* and *USP18* to have HIV gene-specific in SNPs frequencies ranging between 5% to 15% and that African ethnic groups had the highest. In addition, 33 genes had TB gene-specific in SNPs frequencies between 5% to 20% of which all African ethnic groups had the highest frequencies. This findings suggests that these genes may harbor common effect and contribution in TB/HIV among African ethnic groups. Furthermore, HIV/TB gene-specific have high proportion of derived allele at low African ethnic-specific minor allele frequency (0.0 to 0.1) and that these proportion derived allele vary among African ethnic groups.

2. We identified five genes including *GYPB*, *FCGR2A*, *IL13*, and *FREM3* with Malaria gene-specific in SNPs frequencies ranging between 4% to 15%, while all Sickle-Cell disease related genes have low gene-specific in SNPs frequencies ranging between 0.1% to 0.3% among all 20 ethnic groups, but all African and diaspora ethnic groups have the highest in that range.

3. Our result illustrated that all ACG genes have gene-specific in SNP frequencies low than 0.4% in all ethnic groups. However, the gene-specific in SNP frequencies from most of African ethnic groups are higher compare to those from non-African ethnic groups, supporting potential effect and contribution of these actionable genes among world-wide ethnic groups. High ACG gene-specific derive allele, was observed in Latin America and most of Afro-Asiatic ethnic groups following most of European related ethnic groups, while a low ACG gene-specific derive allele are observed in most of African ethnic groups.

Overall, given the observed highest genetic diversity found in African ethnics and African diaspora related ethnics at these four Africa burden diseases and current actionable gene associated, our results support (1) the use of personalised medicine as beneficial to both African continent and worldwide; (2) a recommendation for African-specific actionable list of genes to further improve African and diaspora healthcare; (3) future efforts are needed to accurately identify the secondary findings, by improving the downstream analysis such variant calling methods when using African data and developing an African reference panel in order to speed up the translation of genomics into clinical care in Africa.

Appendices

Appendix A

Supplementary Information

Table A.1: Lists of gene-disease pairs of HIV/AIDS, Malaria, Tuberculosis (TB), Sickle cell disease and Actionable genes.

Genes Associated with HIV/AIDS
A4GALT, ATG16L2, CAV2, CTDP1, DNAL1, FASLG, HCG22, IL10, MED14, NUP153, PRKX, SEC14L1, TLR8, TNPO3ABCB1, ATG7, CCDC134, CTLA4, DOK6, FBXO18, HCP5, IL10RA, MED28, NUP155, PSME2, SERPINA1, TLR9, TNS1ABO, ATP6V0A1, CCL11, CUL5, DPCR1, FBXO21, HCP5P2, IL12A, MED4, NUP160, PSORS1C1, SESTD1, TM9SF2, TNXB,ACTR3BP6, BAHD1, CCL2, CX3CR1, DPM1, FBXW11, HCRTR2, IL12B, MED6, NUP85, PSORS1C3, SFT2D1, TMED2, TOMM70AADAM10, BCL9, CCL5, CXCL12, DYRK1A, FCGR2A, HEATR1, IL13, MED7, ODZ4, PURA, SIP1, TMEFF2, TOR2A, ADAM18, BOD1P, CCNT1, CXCR4, DYSF, FGD6, HGS, IL2, MEPE, OTUD3, RAB1B, SLC35F4, TMEM132C, TRAPPC1,ADH5P4, BTNL2, CCR2, CXCR6,ECR777, FHL3, HIBCH, IL2RA, MGAT1, PABPC1P2, RAB28, SLC46A1, TMEM163, C10orf71, CCR5, CYP7B1, EDNRA, FKBP1AP4, HIP1R, IL32, MICA, PANK1, RAB2A, SLCO5A1, TMEM181, TRIM10,AGAP2, C1ORF103, CD209, DAB1, EFEMP1, FLII, HIST1H3B, IL4, MID1IP1, PC, RAB6A, SNN, TMEM182, TRIM55, AGBL5, C21orf96, CD4, DARC, EFHC2, FNTA, IL4R, MKI67, PCDH11X, RAB6B, SOX11, TMTC1, TRIM58, AKT1, C3ORF56, CDSN, DDEF2, EGF, FOXN3, HLA, IL6, MKRN2, PDE7A, RAB6C, SP110, TNF, TRMT5, ALKBH8, C6ORF1, CHORDC1, DDOST, EGFR, FUT2, HLA-A, INTS7, MMADHC, PDIA6, RABEPK, SPAST, ZNF436, TUBAL3 ANKRD30A, C6orf106, chrna3, DDX10, EIF2C3, GABARAP, HLA-B, IQUB, MND1, PHF12, RANBP1, SPATS2L, ZNF512B, UBQLN4, ANKRD43, C6orf48, chrna5, DDX3X, EIF3H, HLA-B, IRF4, MOS, PHF3, RANBP2, SPCS3, ZNF536, UBQLN4P1, ANKRD6, C7orf58, DX53, EPHA5, GAPVD1, HLA-C, ITPKA, MPHOSPH6, PIGH, RAP1B, SPTAN1, ZNF720, ANKRD9, CACNG1, CLN3, DDX55, EPS8, GBAS, HLA-DPA1, JAK1, MR1, PIGK, RAPGEF1, ZNF785, USP26, ANXA1, CADM1, CLNS1A, DEFB1, ERCC3, GCK, HLA-DQA1, JHDM1D, PIGY,RAPGEF2, SSB, ZNF791, USP6, AOAH, CAPN6, CMTM8, DEPDC5, ERI2, GLRX3, KAT2A,NBEA, PIP5K1C, RELA, ST3GAL5, ZNF804A, VANGL2, AP2M1, CARD16, COG2, DHX33, ERP27, GLTSCR1, HLA-E, KBTBD7, NCOR2, PKD1L2, RGP1, STAC2, ZNF804A, VPRBP, APOBEC3G, LNX2,COG3, DIMT1L, ETF1, GNPDA2, HLA-G, KCNIP3, NDUFB7, PLEKHA3, RIMS4, STARD3NL, HLA-B57,VPS53, ARF1, LOC375190, COG4, DMXL1, ETHE1, GOLPH3, HMGXB3, KCNK9, NEDD9, PLOD3, RNF170, STT3A, CXCL11, VWC2L, ARGLU1, LPL, CRIPAK, DNAJB1, EVI5L, GOSR2, HNRNPF, KDM4D, NF2, PM20D1, RNF212, STX5, IFI44, WASF5P, ARHGAP32, LRRC8D, CRTC2, LCP2, EXOSC3, GPC5, HTATSF1, KEL, NGLY1, PNRC1, RNF26, SUV420H1, IFI6, WDR27,ARHGEF12, LSM3, CRTC3, LEFTY1, EXOSC5, GPR156, HUWE1, KIAA1012, NIPSNAP3B, POLR3A, RPL13AP15, TAOK1, IFIT3, WDTC1, ARHGEF19, LY6D, CSPP1, RPL28P3, FA2H, GRIN2A, IDH1, KIF3C, NLRP1, POLR3F, RPL15P15, TAP2, RSAD2, WNK1, ARPC1A, LYPD4, TFE3, RPL32P3, FAM174B, GRM5,IER3, KIR3DL1, NMT1, POU1F1, RPL21P119, TCEB3, CIG-5, WNT1, ASXL2, MAD2L1, THAP3, RPL4P5, FAM200B, GRTP1, IFNAR1, KLHDC2, NOS3, PP2672, RPL21P126, TFAP4, ANKRD22, XKR4, ATG12, MAP4, THOC2, RPTN, FAM5B, H3F3A, IFNG, KLHL1, NOTCH4, PPIB, RPL21P75, TFDP2, CXCL9, YTHDC2, MCM8, MBL2, THRAP3P1, RRAGB, ZNF12, HB1, IFNGR1, KTN1, NR0B2, PPP2R2A, STAT1, CMPK2, CXCL10,ZBTB2, MDN1, MBL2, TIAM2, RSL1D1, ZNF182, LAPTM5, NUP107, PPP3CC, GBP1, XAF1, USP18, ZDHHC19, TIMM8A, RTN2, ZNF354A, LARS, NUP133, PRDM14, ISG43, BIRC4BP, OASL, ZFP90, TLR7, RUSC2, ZNF385D, PRDM7, OAS1, OAS2, SAMD5, PRF1, GZMH, HLA-B, SCFD1, PRKG2, NKG7, SDC1 TRIB1, AE01, GALNT14, CLDND1, UGT1A12P, SPTBN1, NAV2, HLA-DQB1, IGHMBP2

Genes Associated with Sickle Cell Disease
*BCL11A,GPI,SNTB1,MYB,PAX6,SLC4A1,HBS1L,PMS1,VNN1,HBB,SERPINC1,PNPLA7, SORL1,CACNA1H,CTBP2,SERPINA1,CPS1,RPL3L,PLAT,10-Mar,INPP5J,ABCB1,NADSYN1, HBB,SLC12A1,PYGB,HLA-DQB1,LIPC,SLC4A5,HLA,DRB1,KCNJ11,BMP6,MTHFR,ALDOA, COL6A3,NOS1,MYOC,INSR,NOS3,TBX3,MCCC2,NOS3,CAPN13,MTRR,LDLR,NOS3,MYO7B, ATP2B4,RXRG,CLCN6,F11,ESR2,SERPINE1,OGDHL,HMGCR,SLC7A8,KCNH2,ABCC1,ATP10, HGF,F5,IVD,GATA,SLC22A5,HBS1L-MYB,HLA DRB1*1302,HLA-B53*
Genes Associated with Tuberculosis (TB)
IGSF21,SGMS1,SPON1,PPAP2B,RPL7,UBASH3B,NUCKS1,CMKLR1,DCUN1D5,RPL7, UBL3,ANO2,NCKAP5,LINC00571,E2F7,LRP1B,STXBP6,KLF12,ANAPC1,GEMIN7,VWA8, PLCL1,GRIK1,KIAA0564,CTLA4,TCEB3,NPAS3,RGPD5,TMEM51,SNORD114-31,KCNAB1, LINC00210,RORA,MAGI1,GNG12,AS1,RBL2,FSTL5,RHOU,NAA60,UBA6,ST6GALNAC3, SLC9A8,SCD5,UBXN11,TSHZ2,SLIT3,LEKR1,DSCAM,ZNF131,FHIT,IL2RB,SDK1,FSTL5, REPS2,IMMP2L,ARAP2,ZNF630,HNF4G,GABRB1,SLC11A1,C8orf4,RPS23P5,MBL2, ZFPM2,COL12A1,VDR,CSMD1,LOC100508120,DMRTA1,NACC2,KIAA0087,AKR1C3, RORB,VSTM2A,DNAH11,PTPRD,FAM110B
Genes Associated with Malaria
BGLAP,PSMB9,ICAM1,SCO1,SCO1,IFNA1,CD36,TLR4,IL10,CR1,TLR9,IL13, DDC,TNF,IL1B,ECR777,TNFAIP,IL1RN,FCGR2A,TRB,IL4,FCGR3A,ATP2B4,IRF1, FCGR3B,MARVELD3,MBL2,FCGR3B,FREM3,MIF,G6PD,GABI,MYH3,HBB,INPP4B, NOS2,HBE1,USP38,OR51V1,HLA-DRB5,GUSBP5,PECAM1,HMOX1,GYPA,HP,GYPB, DARC
Actionable Genes
ACTA2A,SDHC,SMAD4,TGFBR2,ENG,FLCN,OTC,GAA,LDLRAP1,MSH2, MYH11,CASQ2,CACNB2,DMPK,RYR1,SCN5A,BMPR1A,KCNJ2,MEN1,TPM1,DMD, HMBS,PMS2,PTCH1,SERPINA1,ATP7B,MYLK,SMARCB1,TMEM43,EPCAM,ACTC1, SDHD,HAMP,PAH,MSH6,GCH1,COQ2,CDC73,DSC2,RYR2,MYH7,BRCA1,KCNQ1, MET,TSC1,SDHAF2,KCNE1,PRKAG2,PTEN,SLC25A13,EMD,NF2,STK11,TNNI3, FBN1,BCHE,SERPINC1,HFEc,PCBD1,MUTYH,ACVRL1,COQ9,CDH1,DSG2,SCN1B, GPD1L,BRCA2,KIT,MLH1,TSC2,MYL2,KCNE2,PRKAR1A,RBM20,SLC37A4,SDHB, PDGFRA,TGFB3,TNNT2,FH,GLA,SGCD,HFE2,PTS,MYBPC3,BLM,CPT2,CNBP, DSP,SCN3B,APC,CACNA1C,LDLR,MLH3,VHL,HCN4,KCNE3,PROC,RET,SLC7A9, MYL3,PKP2,TGFBR1,TP53,SMAD3,IDUA,QDPR,F5b,COL3A1,CACNA1S,LMNA, KCNH2,PROS1,PLN

Table A.2: Output files and information of the nine variant calling tools investigated.

Tool	Informations
VarScan2	CHROM POS ID REF ALT QUAL FILTER INFO(ADP;WT;HET;HOM;NC) FORMAT (GT:GQ:SDP:DP:RD:AD:FREQ:PVAL:RBQ:ABQ:RDF:RDR:ADF:ADR)
Samtools	CHROM POS ID REF ALT QUAL FILTER INFO(DP;VDB;SGB;RPB;MQB;MQSB;BQB;MQOF;ICB;HOB;AC;AN;DP4;MQ) FORMAT(PL;GT)
GATK-HaplotypeCaller	CHROM POS ID REF ALT QUAL FILTER INFO (BaseQRankSum;ClippingRankSum; DP;MLEAC;MLEAF;MQ;MQ0;MQRankSum;ReadPosRankSum) FORMAT (GT:DP:GQ:MIN_DP:PL)
SNVer	#CHROM POS ID REF ALT QUAL FILTER INFO(DP;AF;NP;PV) FORMAT(AC:DP)
Bcftools	CHROM POS ID REF ALT QUAL FILTER INFO(DP;VDB;SGB;RPB;MQB;MQSB; BQB;MQOF;ICB;HOB;AC;AN;DP;MQ) FORMAT(GT:PL)
FreeBayers	CHROM POS ID REF ALT QUAL FILTER INFO(AB;ABP;AC;AF;AN;AO;CIGAR;DP; DPB;DPRA;EPP;EPPR;GTI;LEN;MEANALT;MQM;MQMR;NS;NUMALT;ODDS;PAIRED; PAIREDR;PAO;PQA;PQR;PRO;QA;QR;RO;RPL;RPP;RPPR;RPR;RUN;SAF;SAP;SAR; SRF;SRP;SRR;TYPE;technology.illumina) FORMAT(GT:DP:AD:RO:QR:AO:QA:GL)
Lofreq	CHROM POS ID REF ALT QUAL FILTER INFO(DP;AF;SB;DP4)

PlatyPus	CHROM POS ID REF ALT QUAL FILTER INFO(BRF;FR;HP;HapScore;MGOF;MMLQ;MQ ;NF;NR;PP;QD;SC;SbPval;Source;TC;TCF;TCR;TR;WE;WS) FORMAT(GT:GL:GOF:GQ:NR:NV)
VarDict	CHROM POS ID REF ALT QUAL FILTER INFO(SAMPLE;TYPE;VD;BIAS;REFBIAS: ;VARBIAS;PMEAN;PSTD;QUAL;QSTD;SBF;ODDRATIO;MQ;SN;HIAF;ADJAF;SHIFT3;MSI ;MSILEN;NM;HICNT;HICOV;LSEQ;RSEQ;DUPRATE;SPLITREAD;SPANPAIR;DP;AF) FORMAT(GT:DP:VD:AD:AF:RD:ALD)

References

ACMG. (2015). ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. *Genetics in Medicine*, *17*(1), 68–69. Retrieved from https://doi.org/10.1038/gim.2014.151 doi: 10.1038/gim.2014.151

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010, apr). A method and server for predicting damaging missense mutations. , *7*(4), 248–249. doi: 10.1038/nmeth0410-248

Agarwal, A., Guindo, A., Cissoko, Y., Taylor, J. G., Coulibaly, D., Koné, A., ... Diallo, D. (2000, oct). Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood*, *96*(7), 2358 LP – 2363. Retrieved from http://www.bloodjournal.org/content/96/7/2358.abstract

Agresti, A. (2012). *Categorical Data Analysis* (3rd Editio ed.). New York: Wiley-Interscience. Retrieved from https://www.wiley.com/en-us/Categorical+Data+Analysis{%}2C+3rd+Edition-p-9780470463635

Alioto, T. S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M. D., Hovig, E., ... Gut, I. G. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*, *6*. doi: 10.1038/ncomms10001

Alosaimi, S., Bandiang, A., van Biljon, N., Awany, D., Thami, P. K., Tchamga, M. S. S., ... Chimusa, E. R. (2019, 12). A broad survey of DNA sequence data simulation tools. *Briefings in Functional Genomics*. Retrieved from https://doi.org/10.1093/bfgp/elz033 (elz033) doi: 10.1093/bfgp/elz033

Amendola, L. M., Dorschner, M. O., Robertson, P. D., Salama, J. S., Hart, R., Shirts, B. H., ... Jarvik, G. P. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Research*, *25*(3), 305-315. Retrieved from http://genome.cshlp.org/content/25/3/305.abstract doi: 10.1101/gr.183483.114

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2012, January). *FastQC*. Babraham Institute. Babraham, UK.

Bao, R., Huang, L., Andrade, J., Tan, W., Kibbe, W. a., Jiang, H., & Feng, G. (2014). Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Libertas Academica*, *13*, 67–82. doi: 10.4137/CIN.S13779.Received

Bauer, D., & Bauer, D. (2011). Variant calling comparison CASAVA1.8 and GATK. *Nature Precedings*, *1*. doi: 10.1038/npre.2011.6107.1

Berg, J. S., Amendola, L. M., Eng, C., Van Allen, E., Gray, S. W., Wagle, N., ... Jarvik, G. P. (2013, nov). Processes and preliminary outputs for identification of actionable genes as incidental findings in genomic sequence data in the Clinical Sequencing Exploratory Research Consortium. *Genetics in medicine : official journal of the American College of Medical Genetics*, *15*(11), 860–867. doi: 10.1038/gim.2013.133

Bope, C. D., Chimusa, E. R., Nembaware, V., Mazandu, G. K., Vries, J. D., Wonkam, A., & Marti, M. A. (2019). Dissecting in silico Mutation Prediction of Variants in African Genomes : Challenges and Perspectives. , *10*(June), 1–9. doi: 10.3389/fgene.2019.00601

Botstein, D., & Risch, N. (2003, mar). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, *33*, 228. Retrieved from http://dx.doi.org/10.1038/ng1090http://10.0.4.14/ng1090

Broad Institute. ((Accessed: 2018/02/21; version 2.17.8)). *Picard tools*. http://broadinstitute.github.io/picard/.

Campbell, M. C., & Tishkoff, S. A. (2008). African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual*

Review of Genomics and Human Genetics, *9*(1), 403–433. doi: 10.1146/annurev.genom.9.081307.164258

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. *BMC genomics*, *14 Suppl 3*(Suppl 3), S3. doi: 10.1186/1471-2164-14-S3-S3

Caspar, S. M., Dubacher, N., Kopps, A. M., Meienberg, J., Henggeler, C., & Matyas, G. (2018). Clinical sequencing: From raw data to diagnosis with lifetime value. *Clinical Genetics*, *93*(3), 508–519. doi: 10.1111/cge.13190

Castaño, A., Maurer, M. S., Inoshima, I., Inoshima, N., Wilke, G., Powers, M., ... Hanchard, N. A. (2011). HHS Public Access. *Genome biology*, *20*(1), 1310–1314. doi: 10.1186/1471-2148-11-16

Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., ... McCarroll, J. (2017). The 100,000 Genomes Project Protocol. *Genomics England*(November), 1–112. Retrieved from https://www.genomicsengland.co.uk/wp-content/uploads/2017/03/GenomicEnglandProtocol{_}151117-v4-Wales.pdf doi: 10.6084/M9.FIGSHARE.4530893.V2

Cheng, A. Y., Teo, Y. Y., & Ong, R. T. H. (2014). Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*, *30*(12), 1707–1713. doi: 10.1093/bioinformatics/btu067

Chimusa, E. R., Meintjies, A., Tchanga, M., Mulder, N., Seioghe, C., Soodyall, H., & Ramesar, R. (2015). A Genomic Portrait of Haplotype Diversity and Signatures of Selection in Indigenous Southern African Populations. *PLoS Genetics*, *11*(3), 1–28. doi: 10.1371/journal.pgen.1005052

Chimusa, E. R., Zaitlen, N., Daya, M., Möller, M., Helden, P. D., Nicola, J. M., ... Hoal, E. G. (2014). Genome-wide association study of ancestry-specific TB risk in the South African coloured population. *Human Molecular Genetics*, *23*(3), 796–809. doi: 10.1093/hmg/ddt462

Choi, Y., & Chan, A. P. (2015, aug). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics (Oxford, England)*, *31*(16), 2745–2747. doi: 10.1093/bioinformatics/btv195

Choudhury, A., Aron, S., Sengupta, D., & Hazelhurst, S. (2018). African genetic diversity provides novel insights into evolutionary history and local adaptations. , *27*(May), 209–218. doi: 10.1093/hmg/ddy161

Chun, S., & Fay, J. C. (2009, sep). Identification of deleterious mutations within three human genomes. *Genome research*, *19*(9), 1553–1561. doi: 10.1101/gr.092619.109

Consortium, T. . G. P., McVean, G. A., Altshuler (Co-Chair), D. M., Durbin (Co-Chair), R. M., Abecasis, G. R., Bentley, D. R., ... McVean, G. A. (2012, oct). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*, 56. Retrieved from https://doi.org/10.1038/nature11632http://10.0.4.14/nature11632https://www.nature.com/articles/nature11632{#}supplementary-information

Cornish, A., & Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International*, *2015*, 1–11. doi: 10.1155/2015/456479

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Durbin, R. (2011, aug). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. doi: 10.1093/bioinformatics/btr330

David, S., Aguiar, P., Antunes, L., Dias, A., David, S., Aguiar, P., ... Morais, A. (2018). Variants in the non-coding region of the TLR2 gene associated with infectious subphenotypes in pediatric sickle cell anemia To cite this version : HAL Id : pasteur-02003876. *Immunogenetics*, 70: 37. doi: https://doi.org/10.1007/s00251-017-1013-7

Davila, S., Hibberd, M. L., Hari Dass, R., Wong, H. E. E., Sahiratmadja, E., Bonnard, C., ... Seielstad, M. (2008, oct). Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. *PLoS genetics*, *4*(10), e1000218. doi: 10.1371/journal.pgen.1000218

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010, dec). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, *6*(12), e1001025. doi: 10.1371/journal.pcbi.1001025

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–501. doi: 10.1038/ng.806

Diallo, D., & Tchernia, G. (2002). Sickle cell disease in Africa. *Current Opinion in Hematology*, *9*(2), 111–116.

Doerks, T., Copley, R. R., Schultz, J., Ponting, C. P., & Bork, P. (2002, jan). *Systematic identification of novel protein domain families associated with nuclear functions.* (Vol. 12) (No. 1). doi: 10.1101/gr.203201

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2014, 12). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics*, *24*(8), 2125-2137. Retrieved from https://doi.org/10.1093/hmg/ddu733 doi: 10.1093/hmg/ddu733

Dorschner, M. O., Amendola, L. M., Turner, E. H., Robertson, P. D., Shirts, B. H., Gallego, C. J., ... Jarvik, G. P. (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. *American Journal of Human Genetics*, *93*(4), 631–640. Retrieved from http://dx.doi.org/10.1016/j.ajhg.2013.08.006 doi: 10.1016/j.ajhg.2013.08.006

Durtschi, J., Margraf, R. L., Coonrod, E. M., Mallempati, K. C., & Voelkerding, K. V. (2013). VarBin, a novel method for classifying true and false positive variants in NGS data. *BMC bioinformatics*, *14 Suppl 1*(Suppl 13), S2. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3849648{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1186/1471-2105-14-S13-S2

Dye, C., Scheele, S., ¿Dolin, P., Pathania, V., Raviglione, M. C., for the WHO Global Surveillance, & Project, M. (1999, 08). Global Burden of Tuberculosis: Estimated Incidence, Prevalence, and Mortality by Country. *JAMA*, *282*(7), 677-686. Retrieved from https://doi.org/10.1001/jama.282.7.677 doi: 10.1001/jama.282.7.677

Escalona, M., Rocha, S., & Posada, D. (2017). Europe PMC Funders Group Europe PMC Funders Author Manuscripts A comparison of tools for the simulation of genomic next-generation sequencing data. , *17*(8), 459–469. doi: 10.1038/nrg.2016.57.A

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016, 06). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047-3048. Retrieved from https://doi.org/10.1093/bioinformatics/btw354 doi: 10.1093/bioinformatics/btw354

Fang, H., Wu, Y., Yang, H., Yoon, M., Jiménez-Barrón, L. T., Mittelman, D., ... Lyon, G. J. (2017). Whole genome sequencing of one complex pedigree illustrates challenges with genomic medicine. *BMC Medical Genomics*, *10*(1), 10. Retrieved from http://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-017-0246-5 doi: 10.1186/s12920-017-0246-5

Farrer, R. A., Henk, D. A., MacLean, D., Studholme, D. J., & Fisher, M. C. (2013). Using false discovery rates to benchmark SNP-callers in next-generation sequencing projects. *Scientific reports*, *3*, 1512. doi: 10.1038/srep01512

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009, jun). Identi-

fying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics (Oxford, England)*, *25*(12), i54–62. doi: 10.1093/bioinformatics/btp190

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*, 9. Retrieved from http://arxiv.org/abs/1207.3907 doi: arXiv:1207.3907[q-bio.GN]

Gorlova, O. Y., Ying, J., Amos, C. I., Spitz, M. R., Peng, B., & Gorlov, I. P. (2012, apr). Derived SNP alleles are used more frequently than ancestral alleles as risk-associated variants in common human diseases. *Journal of bioinformatics and computational biology*, *10*(2), 1241008. doi: 10.1142/S0219720012410089

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., ... Biesecker, L. G. (2013, jul). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics*, *15*(7), 565–574. doi: 10.1038/gim.2013.73

Gudykunst, W. B., & Schmidt, K. L. (1987). Language and Ethnic Identity: An Overview and Prologue. *Journal of Language and Social Psychology*, *6*(3-4), 157–170. Retrieved from https://doi.org/10.1177/0261927X8763001 doi: 10.1177/0261927X8763001

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., ... Sandhu, M. S. (2015). The African Genome Variation Project shapes medical genetics in Africa. *Nature*, *517*(7534), 327–332. Retrieved from https://doi.org/10.1038/nature13997 doi: 10.1038/nature13997

H3 Africa Working Group on Ethics. (2017). Ethics and Governance Framework for Best Practice in Genomic Research and Biobanking in Africa. (February). Retrieved from http://rfi.cohred.org/wp-content/uploads/2018/04/Final-Framework-for-African-genomics-and-biobanking{_}SC-.pdfhttps://www.sun.ac.za/english/faculty/healthsciences/rdsd/Documents/FinalFrameworkforAfricangenomicsandbiobanking{_}SC{_}February2017II.pdf

Highnam, G., Wang, J. J., Kusler, D., Zook, J., Vijayan, V., Leibovich, N., & Mittelman, D. (2015). An analytical framework for optimizing variant discovery from personal genomes. *Nature Communications*, *6*, 1–6. Retrieved from http://dx.doi.org/10.1038/ncomms7275 doi: 10.1038/ncomms7275

Hintzsche, J. D., Robinson, W. A., & Tan, A. C. (2016). A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. *International Journal of Genomics*, *2016*, 1–16. doi: 10.1155/2016/7983236

Holtgrewe, M. (2010). Mason – A Read Simulator for Second Generation Sequencing Data. *Life Sciences*(October), 18. Retrieved from http://publications.mi.fu-berlin.de/962/{%}5Cnhttp://svn.seqan.de/seqan/trunk/core/apps/mason/README

Huang, H. W., Mullikin, J. C., & Hansen, N. F. (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics*, *16*(1), 235. Retrieved from http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0624-y doi: 10.1186/s12859-015-0624-y

Hunter, J. E., Irving, S. A., Biesecker, L. G., Buchanan, A., Jensen, B., Lee, K., ... Goddard, K. A. B. (2016, dec). A standardized, evidence-based protocol to assess clinical actionability of genetic disorders associated with genomic variation. *Genetics in medicine : official journal of the American College of Medical Genetics*, *18*(12), 1258–1268. doi: 10.1038/gim.2016.40

Hur, C. G., Kim, S., Kim, C. H., Yoon, S. H., In, Y. H., Kim, C., & Cho, H. G. (2006). FASIM: Fragments assembly simulation using biased-sampling model and assembly simulation for microbial genome shotgun sequencing. *Journal of Microbiology and Biotechnology*, *16*(5), 683–688.

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling

pipelines using gold standard personal exome variants. *Scientific Reports*, *5*(December), 1–8. Retrieved from http://dx.doi.org/10.1038/srep17875 doi: 10.1038/srep17875

Ionita-Laza, I., McCallum, K., Xu, B., & Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, *48*(2), 214–220. Retrieved from https://doi.org/10.1038/ng.3477 doi: 10.1038/ng.3477

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., ... Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature Genetics*, *48*(12), 1581–1586. Retrieved from https://doi.org/10.1038/ng.3703 doi: 10.1038/ng.3703

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., ... Kwiatkowski, D. P. (2009, jun). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature genetics*, *41*(6), 657–665. doi: 10.1038/ng.388

Johnson, S., Trost, B., Long, J. R., Pittet, V., & Kusalik, A. (2014). A better sequence-read simulator program for metagenomics. *BMC Bioinformatics*, *15*(9), S14. Retrieved from http://www.biomedcentral.com/1471-2105/15/S9/S14 doi: 10.1186/1471-2105-15-S9-S14

Joubert, B. R., Lange, E. M., Franceschini, N., Mwapasa, V., North, K. E., Meshnick, S. R., & Immunology, t. N. C. f. H. V. (2010). A whole genome association study of mother-to-child transmission of HIV in Malawi. *Genome Medicine*, *2*(3), 17. Retrieved from https://doi.org/10.1186/gm138 doi: 10.1186/gm138

Khan, A., & Mathelier, A. (2017). Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics*, *18*(1), 287. Retrieved from https://doi.org/10.1186/s12859-017-1708-7 doi: 10.1186/s12859-017-1708-7

Kim, S., Jhong, J.-H., Lee, J., & Koo, J.-Y. (2017). Meta-analytic support vector machine for integrating multiple omics data. *BioData mining*, *10*, 2. doi: 10.1186/s13040-017-0126-8

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., ... Ding, L. (2009). VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, *25*(17), 2283–2285. doi: 10.1093/bioinformatics/btp373

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., Mclellan, M. D., Lin, L., ... Wilson, R. K. (2012). VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. doi: 10.1101/gr.129684.111

Kumaran, M., Subramanian, U., & Devarajan, B. (2019). Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. *BMC Bioinformatics*, *20*(1), 1–11. doi: 10.1186/s12859-019-2928-9

Kwiatkowski, D. P. (2005). Doi:10.1086/432519. , 1–22. Retrieved from papers3://publication/uuid/22668F87-86F6-4048-93C6-52275B30A283

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., Mcewen, R., ... Dry, J. R. (2016). VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Research*, *44*(11), 1–11. doi: 10.1093/nar/gkw227

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., ... Maglott, D. R. (2016, jan). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, *44*(D1), D862–8. doi: 10.1093/nar/gkv1222

Laurie, S., Fernandez-Callejo, M., Marco-Sola, S., Trotta, J. R., Camps, J., Chacón, A., ... Beltran, S. (2016). From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. *Human Mutation*, *37*(12), 1263–1271. doi: 10.1002/humu.23114

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*,

27(21), 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H. (2013, mar). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv e-prints*, arXiv:1303.3997.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, *11*(5), 473–483. doi: 10.1093/bib/bbq015

Li, Z., Wang, Y., & Wang, F. (2018). A study on fast calling variants from next-generation sequencing data using decision tree. *BMC Bioinformatics*, *19*(1), 1–14. doi: 10.1186/s12859-018-2147-9

Liu, X. (2014). *dbNSFP v2.5.*

Liu, X., Han, S., Wang, Z., Gelernter, J., & Yang, B. Z. (2013). Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, *8*(9), 1–11. doi: 10.1371/journal.pone.0075619

Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K.-H., & Zhao, H. (2015, may). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific reports*, *5*, 10576. doi: 10.1038/srep10576

Lysholm, F., Andersson, B., & Persson, B. (2011). An efficient simulator of 454 data using configurable statistical models. *BMC Research Notes*, *4*(1), 449. Retrieved from http://www.biomedcentral.com/1756-0500/4/449 doi: 10.1186/1756-0500-4-449

Macharia, A. W., Mochamah, G., Uyoga, S., Ndila, C. M., Nyutu, G., Makale, J., ... Williams, T. N. (2018). The clinical epidemiology of sickle cell anemia In Africa. (November 2017), 363–370. doi: 10.1002/ajh.24986

Makani, J., Williams, T. N., & Marsh, K. (2007). Sickle cell disease in Africa : burden and research priorities. , *101*(1), 3–14. doi: 10.1179/136485907X154638

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, *24*(3), 133–141. Retrieved from http://www.sciencedirect.com/science/article/pii/S0168952508000231 doi: https://doi.org/10.1016/j.tig.2007.12.007

Martin, A. R., Teferra, S., Möller, M., Hoal, E. G., & Daly, M. J. (2018). The critical needs and challenges for genetic architecture studies in Africa. *Current Opinion in Genetics & Development*, *53*, 113–120. Retrieved from http://www.sciencedirect.com/science/article/pii/S0959437X18300558 doi: https://doi.org/10.1016/j.gde.2018.08.005

Martin, E. R., Kinnamon, D. D., Schmidt, M. A., Powell, E. H., Zuchner, S., & Morris, R. W. (2010). SeqEM: An adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, *26*(22), 2803–2810. doi: 10.1093/bioinformatics/btq526

Matthijs, G., Souche, E., Alders, M., Corveleyn, A., Eck, S., Feenstra, I., ... Bauer, P. (2016). Guidelines for diagnostic next-generation sequencing. *European Journal of Human Genetics*, *24*(1), 2–5. Retrieved from https://doi.org/10.1038/ejhg.2015.226 doi: 10.1038/ejhg.2015.226

Mboowa, G. (2019). Role of genomics literacy in reducing the burden of common genetic diseases in Africa. (February), 1–8. doi: 10.1002/mgg3.776

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010, sep). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297–1303. doi: 10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., ... Cunningham, F. (2016, jun). The Ensembl Variant Effect Predictor. *Genome biology*, *17*(1), 122. doi: 10.1186/s13059-016-0974-4

McVean, G. A., Altshuler (Co-Chair), D. M., Durbin (Co-Chair), R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., ... McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. Retrieved from http://www.nature.com/doifinder/10.1038/nature11632 doi: 10.1038/nature11632

Metzker, M. L. (2010, jan). Sequencing technologies [mdash] the next generation. *Nat Rev Genet*, *11*(1), 31–46. Retrieved from http://dx.doi.org/10.1038/nrg2626

Michalopoulos, S. (2012, jun). The Origins of Ethnolinguistic Diversity. *American Economic Review*, *102*(4), 1508–1539. Retrieved from http://www.aeaweb.org/articles?id=10.1257/aer.102.4.1508 doi: 10.1257/aer.102.4.1508

Mielczarek, M., & Szyda, J. (2016). *Review of alignment and SNP calling algorithms for next-generation sequencing data.* doi: 10.1007/s13353-015-0292-7

Möller, M., & Hoal, E. G. (2010). Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis*, *90*(2), 71–83. Retrieved from http://www.sciencedirect.com/science/article/pii/S1472979210000193 doi: https://doi.org/10.1016/j.tube.2010.02.002

Mpye, K. L., Matimba, A., Dzobo, K., Chirikure, S., Wonkam, A., & Dandara, C. (2017). Disease burden and the role of pharmacogenomics in African populations. *Global health, epidemiology and genomics*, *2*, e1. doi: 10.1017/gheg.2016.21

Myers, G. (1999). A dataset generator for whole genome shotgun sequencing. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 202–10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10786303 doi: 10.1007/s13398-014-0173-7.2

Ng, P. C., & Henikoff, S. (2003, jul). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, *31*(13), 3812–3814. doi: 10.1093/nar/gkg509

Olfson, E., Cottrell, C. E., Davidson, N. O., Gurnett, C. A., Heusel, J. W., Stitziel, N. O., ... Bierut, L. J. (2015). Identification of Medically Actionable Secondary Findings in the 1000 Genomes. *PLOS ONE*, *10*(9), 1–18. Retrieved from https://doi.org/10.1371/journal.pone.0135193 doi: 10.1371/journal.pone.0135193

Oni, T., Youngblood, E., Boulle, A., Mcgrath, N., Wilkinson, R. J., & Levitt, N. S. (2015). Patterns of HIV , TB , and non-communicable disease multi-morbidity in peri-urban South Africa- a cross sectional study. , 1–8. doi: 10.1186/s12879-015-0750-1

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., ... Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. *Genome Medicine*, *5*(3). doi: 10.1186/gm432

Orkin, S. H., & Bauer, D. E. (2019). Emerging Genetic Therapy for Sickle Cell Disease.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., ... Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, *15*(2), 256–278. doi: 10.1093/bib/bbs086

Pantano, B. Y. L. (2016). Blue Collar Bioinformatics. , 1–11.

Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., & Birney, E. (2008, nov). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome research*, *18*(11), 1829–1843. doi: 10.1101/gr.076521.108

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, *2*(12), e190. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17194218{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1713260 doi: 10.1371/journal.pgen.0020190

Pattnaik, S., Gupta, S., Rao, A. A., & Panda, B. (2014). SInC: An accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data. *BMC Bioinformatics*, *15*(1), 1–9. Retrieved from BMCBioinformatics doi: 10.1186/1471-2105-15-40

Peer, N. (2015). The converging burdens of infectious and non-communicable diseases in rural-to-urban migrant Sub-Saharan African populations : a focus on HIV / AIDS , tuberculosis and cardio-metabolic diseases. *Tropical Diseases, Travel Medicine and Vaccines*, 1–8. Retrieved from http://dx.doi.org/10.1186/s40794-015-0007-4 doi: 10.1186/s40794-015-0007-4

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., & Mayer, G. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Scientific Reports*, *8*(1), 10950. Retrieved from https://doi.org/10.1038/s41598-018-29325-6 doi: 10.1038/s41598-018-29325-6

Picton, A. C. P., Paximadis, M., & Tiemessen, C. T. (2010). Genetic variation within the gene encoding the HIV-1 CCR5 coreceptor in two South African populations. *Infection, Genetics and Evolution*, *10*(4), 487–494. Retrieved from http://www.sciencedirect.com/science/article/pii/S1567134810000468 doi: https://doi.org/10.1016/j.meegid.2010.02.012

Piel, F. B., Steinberg, M. H., & Rees, D. C. (2017). Sickle Cell Disease. *New England Journal of Medicine*, *376*(16), 1561–1573. Retrieved from https://doi.org/10.1056/NEJMra1510865 doi: 10.1056/NEJMra1510865

Pipek, O., Ribli, D., Molnár, J., Póti, Á., Krzystanek, M., Bodor, A., ... Szüts, D. (2017). Fast and accurate mutation detection in whole genome sequences of multiple isogenic samples with IsoMut. *BMC Bioinformatics*, *18*(1), 73. Retrieved from http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1492-4 doi: 10.1186/s12859-017-1492-4

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, *8*(1), 14. Retrieved from http://humgenomics.biomedcentral.com/articles/10.1186/1479-7364-8-14 doi: 10.1186/1479-7364-8-14

Ploug, T., & Holm, S. (2017). Clinical genome sequencing and population preferences for information about 'incidental' findings-From medically actionable genes (MAGs) to patient actionable genes (PAGs). *PLoS ONE*, *12*(7), 1–13. doi: 10.1371/journal.pone.0179935

Popejoy, A. B., & Fullerton, S. M. (2016, oct). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. doi: 10.1038/538161a

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., der Auwera, G. A. V., ... Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178. Retrieved from https://www.biorxiv.org/content/early/2017/11/14/201178.1 doi: 10.1101/201178

Pratas, D., Pinho, A. J., & Rodrigues, J. M. O. S. (2014). XS : a FASTQ read simulator Open Access XS : a FASTQ read simulator. *BMCResearch Notes*. doi: 10.1186/1756-0500-7-40

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007, sep). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, *81*(3), 559–575. doi: 10.1086/519795

Quang, D., Chen, Y., & Xie, X. (2014, 10). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, *31*(5), 761-763. Retrieved from https://doi.org/10.1093/bioinformatics/btu703 doi: 10.1093/bioinformatics/btu703

Rabbani, B., Tekin, M., & Mahdieh, N. (2013, nov). The promise of whole-exome sequencing in medical genetics. *Journal Of Human Genetics*, *59*, 5.

Rees, D. C., Williams, T. N., & Gladwin, M. T. (2010). *Sickle-cell disease* (Vol. 376) (No. 9757). [London] ;: Elsevier Science. doi: 10.1016/S0140-6736(10)61029-X

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2018). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. Retrieved from https://doi.org/10.1093/nar/gky1016 doi: 10.1093/nar/gky1016

Retshabile, G., Mlotshwa, B. C., Williams, L., Mwesigwa, S., Mboowa, G., Huang, Z., ... Hanchard, N. A. (2018). Whole-Exome Sequencing Reveals Uncaptured Variation and Distinct Ancestry in the Southern African Population of Botswana. *American Journal of Human Genetics*, *102*(5), 731–743. Retrieved from https://doi.org/10.1016/j.ajhg.2018.03.010 doi: 10.1016/j.ajhg.2018.03.010

Reva, B., Antipin, Y., & Sander, C. (2011, sep). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, *39*(17), e118. doi: 10.1093/nar/gkr407

Richter, D. C., Ott, F., Auch, A. F., Schmid, R., & Huson, D. H. (2011). MetaSim: A Sequencing Simulator for Genomics and Metagenomics. *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, *3*(10), 417–421. doi: 10.1002/9781118010518.ch48

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., & Twigg, S. R. F. (2016). Europe PMC Funders Group approaches for calling variants in clinical sequencing applications. , *46*(8), 912–918. doi: 10.1038/ng.3036.Integrating

Roberts, N. D., Kortschak, R. D., Parker, W. T., Schreiber, A. W., Branford, S., Scott, H. S., ... Adelson, D. L. (2013). A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*, *29*(18), 2223–2230. doi: 10.1093/bioinformatics/btt375

Rotimi, C., Abayomi, A., Abimiku A, Adabayeri, V., Adebamowo, C., Adebiyi, E., ... Joubert, F. (2014). Enabling the genomic revolution in Africa. *Science*, *344*(6190), 1346–1348. doi: 10.1126/science.1251546.Enabling

Rotimi, C. N., Bentley, A. R., Doumatey, A. P., Chen, G., Shriner, D., & Adeyemo, A. (2017). The genomic landscape of African populations in health and disease. , *26*(June), 225–236. doi: 10.1093/hmg/ddx253

Said Mohammed, K., Kibinge, N., Prins, P., Agoti, C. N., Cotten, M., Nokes, D., ... Githinji, G. (2018). Evaluating the performance of tools used to call minority variants from whole genome short-read data. *Wellcome Open Research*, *3*(0), 21. doi: 10.12688/wellcomeopenres.13538.2

Salie, M., Daya, M., Lucas, L. A., Warren, R. M., van der Spuy, G. D., van Helden, P. D., ... Möller, M. (2015). Association of toll-like receptors with susceptibility to tuberculosis suggests sex-specific effects of TLR8 polymorphisms. *Infection, Genetics and Evolution*, *34*, 221–229. Retrieved from http://www.sciencedirect.com/science/article/pii/S156713481500266X doi: https://doi.org/10.1016/j.meegid.2015.07.004

Sandmann, S., De Graaf, A. O., Karimi, M., Van Der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., & Dugas, M. (2017). Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports*, *7*, 1–12. Retrieved from http://dx.doi.org/10.1038/srep43169 doi: 10.1038/srep43169

Sandmann, S., de Graaf, A. O., Karimi, M., van der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., & Dugas, M. (2017). Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports*, *7*, 43169. Retrieved from http://www.nature.com/articles/srep43169 doi: 10.1038/srep43169

Schurz, H., Daya, M., Möller, M., Hoal, E. G., & Salie, M. (2015). TLR1, 2, 4, 6 and 9 Variants Associated with Tuberculosis Susceptibility: A Systematic Review and Meta-Analysis. *PLOS ONE*, *10*(10), 1–24. Retrieved from https://doi.org/10.1371/journal.pone

.0139711 doi: 10.1371/journal.pone.0139711
Schurz, H., Kinnear, C. J., Gignoux, C., Wojcik, G., van Helden, P. D., Tromp, G., ... Möller, M. (2018). A Sex-Stratified Genome-Wide Association Study of Tuberculosis Using a Multi-Ethnic Genotyping Array. *Frontiers in genetics*, *9*, 678. doi: 10.3389/fgene.2018.00678
Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods*, *7*(8), 575–576. Retrieved from https://doi.org/10.1038/nmeth0810-575 doi: 10.1038/nmeth0810-575
Shcherbina, A. (2014). FASTQSim: Platform-independent data characterization and in silico read generation for NGS datasets. *BMC Research Notes*, *7*(1), 1–12. doi: 10.1186/1756-0500-7-533
Shen, T., de Stadt, S. H., Yeat, N. C., & Lin, J. C.-H. (2015). Clinical applications of next generation sequencing in cancer: from panels, to exomes, to genomes. *Frontiers in Genetics*, *6*, 215. Retrieved from https://www.frontiersin.org/article/10.3389/fgene.2015.00215 doi: 10.3389/fgene.2015.00215
Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001, jan). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, *29*(1), 308–311. doi: 10.1093/nar/29.1.308
Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M., & Gaunt, T. R. (2013, jun). Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics (Oxford, England)*, *29*(12), 1504–1510. doi: 10.1093/bioinformatics/btt182
Sirugo, G., Hennig, B. J., Adeyemo, A. A., Matimba, A., Newport, M. J., Ibrahim, M. E., ... Williams, S. M. (2008). *Erratum: Genetic studies of African populations: An overview on disease susceptibility and response to vaccines and therapeutics (Human Genetic (2008) vol. 123 (557-598) 10.1007/s00439-008-0511-y)* (Vol. 124) (No. 2). doi: 10.1007/s00439-008-0534-4
Søborg, C., Bengaard, A., Range, N., Malenganisho, W., Friis, H., Magnussen, P., ... Garred, P. (2007). Influence of candidate susceptibility genes on tuberculosis in a high endemic region. , *44*, 2213–2220. doi: 10.1016/j.molimm.2006.11.002
Spencer, D. H., Tyagi, M., Vallania, F., Bredemeyer, A. J., Pfeifer, J. D., Mitra, R. D., & Duncavage, E. J. (2014). Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data. *Journal of Molecular Diagnostics*, *16*(1), 75–88. Retrieved from http://dx.doi.org/10.1016/j.jmoldx.2013.09.003 doi: 10.1016/j.jmoldx.2013.09.003
Spinella, J. F., Mehanna, P., Vidal, R., Saillour, V., Cassart, P., Richer, C., ... Sinnett, D. (2016). SNooPer: A machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics*, *17*(1), 1–11. Retrieved from http://dx.doi.org/10.1186/s12864-016-3281-2 doi: 10.1186/s12864-016-3281-2
Stead, L. F., Sutton, K. M., Taylor, G. R., Quirke, P., & Rabbitts, P. (2013). Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: Applications in tumor subclone resolution. *Human Mutation*, *34*(10), 1432–1438. doi: 10.1002/humu.22365
Stephens, Z. D., Hudson, M. E., Mainzer, L. S., Taschuk, M., Weber, M. R., & Iyer, R. K. (2016). Simulating next-generation sequencing datasets from empirical mutation and sequencing models. *PLoS ONE*, *11*(11), 1–18. doi: 10.1371/journal.pone.0167047
Stöcker, B. K., Köster, J., & Rahmann, S. (2016). SimLoRD: Simulation of Long Read Data. *Bioinformatics*, *32*(17), 2704–2706. doi: 10.1093/bioinformatics/btw286
Talwalkar, A., Liptrap, J., Newcomb, J., Hartl, C., Terhorst, J., Curtis, K., ... Patterson, D. (2014). SMaSH: A benchmarking toolkit for human genome variant calling. *Bioinformatics*, *30*(19), 2787–2795. doi: 10.1093/bioinformatics/btu345
Tan, A., Abecasis, G. R., & Kang, H. M. (2015). Unified representation of genetic variants.

Bioinformatics, *31*(13), 2202–2204. doi: 10.1093/bioinformatics/btv112

Tang, C. S.-m., Dattani, S., So, M.-t., Cherny, S. S., Tam, P. K. H., Sham, P. C., & Garcia-Barcelo, M.-M. (2017). Actionable secondary findings from whole-genome sequencing of 954 East Asians. *Human Genetics*(0123456789). Retrieved from http://link.springer.com/10.1007/s00439-017-1852-1 doi: 10.1007/s00439-017-1852-1

Tange, O. (2011, Feb). Gnu parallel - the command-line power tool. *;login: The USENIX Magazine*, *36*(1), 42-47. Retrieved from http://www.gnu.org/s/parallel

Tetzlaff, M. T., Singh, R. R., Seviour, E. G., Curry, J. L., Hudgens, C. W., Bell, D., ... Esmaeli, B. (2016). Next-generation sequencing identifies high frequency of mutations in potentially clinically actionable genes in sebaceous carcinoma. *The Journal of Pathology*, *240*(1), 84–95. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/path.4759 doi: 10.1002/path.4759

Thompson, M. L., Finnila, C. R., Bowling, K. M., Brothers, K. B., Neu, M. B., Amaral, M. D., ... Cooper, G. M. (2018). Genomic sequencing identifies secondary findings in a cohort of parent study participants. *Genetics in Medicine*, *20*(12), 1635–1643. Retrieved from https://doi.org/10.1038/gim.2018.53 doi: 10.1038/gim.2018.53

Thye, T., Owusu-dabo, E., Vannberg, F. O., Crevel, R. V., Sahiratmadja, E., Balabanova, Y., ... Muntau, B. (2012). Europe PMC Funders Group Common variants at 11p13 are associated with susceptibility to tuberculosis. , *44*(3), 257–259. doi: 10.1038/ng.1080.Common

Vannberg, F. O., Chapman, S. J., & Hill, A. V. S. (2011). Human genetic susceptibility to intracellular pathogens. *Immunological Reviews*, *240*(1), 105–116. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-065X.2010.00996.x doi: 10.1111/j.1600-065X.2010.00996.x

Wallis, Y., Payne, S., McAnulty, C., Bodmer, D., Sistermans, E., Robertson, K., ... Devereau, A. (2013). Practice guidelines for the evaluation of pathogenicity and the reporting of sequence variants in clinical molecular genetics.

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), 1–7. doi: 10.1093/nar/gkq603

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., ... Zhao, Z. (2013). Detecting somatic point mutations in cancer genome sequencing data: A comparison of mutation callers. *Genome Medicine*, *5*(10), 1–8. doi: 10.1186/gm495

Wang, Z., Liu, X., Yang, B.-Z., & Gelernter, J. (2013). The Role and Challenges of Exome Sequencing in Studies of Human Diseases. *Frontiers in Genetics*, *4*, 160. Retrieved from https://www.frontiersin.org/article/10.3389/fgene.2013.00160 doi: 10.3389/fgene.2013.00160

Warden, C. D., Adamson, A. W., Neuhausen, S. L., & Wu, X. (2014). Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ*, *2*, e600. doi: 10.7717/peerj.600

Wei, Z., Wang, W., Hu, P., Lyon, G. J., & Hakonarson, H. (2011). SNVer: A statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research*, *39*(19), 1–13. doi: 10.1093/nar/gkr599

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., ... Nagarajan, N. (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research*, *40*(22), 11189–11201. doi: 10.1093/nar/gks918

Wonkam, A., Bitoungui, V. J. N., Vorster, A. A., Ramesar, R., Cooper, R. S., Tayo, B., ... Ngogang, J. (2014). Association of Variants at BCL11A and HBS1L-MYB with Hemoglobin F and Hospitalization Rates among Sickle Cell Patients in Cameroon. , *9*(3). doi: 10.1371/journal.pone.0092506

Wonkam, A., Mnika, K., Josiane, V., Bitoungui, N., Chemegni, C., Chimusa, E. R., ... Africa, S. (2019). HHS Public Access. , *180*(1), 134–146. doi: 10.1111/bjh.15011.Clinical

Wood, R., Liang, H., Wu, H., Middelkoop, K., Oni, T., Rangaka, M. X., ... Lawn, S. D. (2010). Changing prevalence of tuberculosis infection with increasing age in high-burden townships in South Africa. , *14*(September 2009), 406–412.

Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, *16*, 15–24. Retrieved from https://doi.org/10.1016/j.csbj.2018.01.003 doi: 10.1016/j.csbj.2018.01.003

Xu, H., DiCarlo, J., Satya, R. V., Peng, Q., & Wang, Y. (2014). Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics*, *15*(1), 1–10. doi: 10.1186/1471-2164-15-244

Yang, C., Chu, J., Warren, R. L., & Birol, I. (2017). NanoSim: Nanopore sequence read simulator based on statistical characterization. *GigaScience*, *6*(4), 1–6. doi: 10.1093/gigascience/gix010

Yang, H., & Wang, K. (2015, oct). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature protocols*, *10*(10), 1556–1566. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4718734/ doi: 10.1038/nprot.2015.105

Yi, M., Zhao, Y., Jia, L., He, M., Kebebew, E., & Stephens, R. M. (2014). Performance comparison of SNP detection tools with illumina exome sequencing data - An assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Research*, *42*(12), 1–14. doi: 10.1093/nar/gku392

Yu, X., & Sun, S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics*, *14*(1). doi: 10.1186/1471-2105-14-274

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, *32*(3), 246–251. doi: 10.1038/nbt.2835