Genetic diversity and population structure within Botswana: association with HIV-1 infection

Prisca Kerapetse Thami
THMPRI004

Thesis Presented for the Degree of
DOCTOR OF PHILOSOPHY
in the Division of Human Genetics, Department of Pathology,
Faculty of Health Sciences
UNIVERSITY OF CAPE TOWN

January 2021

Supervisor: Associate Prof. Emile R. Chimusa
Co-supervisors: Dr Simani Gaseitsiwe, Dr Vlad Novitsky, Dr Melvin Leteane

Declaration

I, Prisca Kerapetse Thami, hereby declare that the work on which this dissertation is based is my original research 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:

Date: 12 January 2021

Publications

Publications out of this thesis

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

1. **Prisca K. Thami** and Emile R. Chimusa. 2019. Population Structure and Implications on the Genetic Architecture of HIV-1 Phenotypes Within Southern Africa. *Front. Genet*. 10, 905. doi:10.3389/fgene.2019.00905.
2. **Prisca K. Thami**, Wonderful T. Choga, Delesa Damena Mulisa, Collet Dandara, Andrey K. Shevchenko, Melvin M. Leteane, Vlad Novitsky, Stephen J. O'Brien, Myron Essex, Simani Gaseitsiwe and Emile R. Chimusa. 2020. Whole Genome Sequencing-based Characterization of Human Genome Variation and Mutation Burden in Botswana. *bioRxiv*. https://doi.org/10.1101/2020.12.15.422821
3. **Prisca K. Thami**, Wonderful T. Choga, Delesa Damena Mulisa, Collet Dandara, Andrey K. Shevchenko, Melvin M. Leteane, Vlad Novitsky, Stephen J. O'Brien, Myron Essex, Simani Gaseitsiwe and Emile R. Chimusa. 2020. Whole Genome Rare-Variant Association Study of HIV-1 progression in a Southern African population. *medRxiv*. https://doi.org/10.1101/2020.12.16.20248307

Signature: Date: 12 January 2021

Student Name: Prisca Kerapetse Thami Student Number: THMPRI004

Publications as a collaborator

Additional publications that I collaborated in during the course of my PhD, although these are not part of my thesis:

1. Alosaimi S, Bandiang A, van Biljon N, Awany D, **Thami PK**, Tchamga MSS, Kiran A, Messaoud O, Hassan RIM, Mugo J, Ahmed A, Bope CD, Allali I, Mazandu GK, Mulder NJ and Chimusa ER. 2020. A broad survey of DNA sequence data simulation tools. *Briefings in Functional Genomics*, 19(1), pp.49-59.

2. Chimusa ER, Defo J, **Thami PK**, Awany D, Mulisa DD, Allali I, Ghazal H, Moussa A and Mazandu GK. 2018. Dating admixture events is unsolved problem in multi-way admixed populations. *Briefings in bioinformatics*, doi: 10.1093/bib/bby112.
3. Alosaimi S, van Biljon N, Awany D, **Thami PK**, Allali I, Mazandu GK, Mulder NJ, Chimusa ER. 2020. Simulation of African and non-African low and high coverage whole genome sequence data to assess variant calling approaches. *Briefings in bioinformatics*. BIB-20-0573.R2. Accepted for publication.
4. Shevchenko AK, Zhernakova DV, Malov SV, Cherkasov N, Komissarov A, Tamazian G, Antonik A, Kolchanova SM, Kliver S, Turenko A, Kirk GD, Vlahov D, **Thami PK**, Gaseitsiwe S, Novitsky V, Essex M, O'Brien SJ. 2020. Genome-wide Association Study Reveals Novel Genetic Variants Associated with HIV-1C Infection in a Botswana Study Population. Under review at *PLOS PATHOGENS*
5. Maseng MJ, Tawe L, **Thami P**, Seatla KK, Moyo S, Martinelli A, Kasvosve I, Novitsky V, Essex M, Russo G, Gaseitsiwe S, Paganotti GM. 2020. Association between CYP2B6 fast metabolism and development of efavirenz and nevirapine resistance in HIV-1 patients on antiretroviral therapy from Botswana. Under review at *Pharmacogenomics and Personalized Medicine*

Acknowledgements

I thank Christ Jesus our Lord who has enabled me to fulfil this entrustment [1 Timothy 1: 12].

My utmost gratitude goes to Associate Professor Emile R. Chimusa, my supervisor and technical advisor, for pushing me beyond my limits, for outpouring a deeper knowledge and exceptional skill of bioinformatics that I needed to successfully complete my PhD, for his patience and great mentorship that have shaped me into a better academic researcher. I thank my co-supervisors Dr Simani Gaseitsiwe, Dr Vlad Novitsky and Dr Melvin M. Leteane; my thesis advisory committee, Professor Collet Dandara, Professor Scott Hazelhurst and Professor Nicki Tiffin for their invaluable guidance.

A special thank you to Dr Simani Gaseitsiwe, Dr Vlad Novitsky, Professor Stephen J. O'Brien and Emeritus Professor Myron "Max" E. Essex for facilitating the data acquisition of the whole genome sequences. My deepest appreciation to the participants, investigators and key personnel of the "Host Genetics of HIV-1C Infection, Progression, and Treatment in Africa/GWAS on Determinants of HIV-1C Infection" study at Botswana Harvard AIDS Institute Partnership.

Thanks to my colleagues in the UCT Human Genetics Division, especially Dr Delesa Damena Mulisa, Denis Awany and Shatha Alosaimi for the team spirit and inspiration. To Professor Nicola J. Mulder, Jacquiline Mugo, Mamana Mbiyavanga, Suresh Maslamoney and Gerrit Botha, thank you for the data transfer and analysis support.

All analyses were performed through the computing clusters at the Centre for High Performance Computing (CHPC), Cape Town, South Africa. This work was supported through the sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006 with funding from the Wellcome Trust [grant # 107752/Z/15/Z] and the UK government. Thank you Dr Denis Chopera for believing in me and my supervisor Associate Professor Emile R. Chimusa, and for vehemently supporting my transfer to the University of Cape Town.

Finally, I am deeply grateful for the love, tangible support and encouragement from my lovely husband Thami and beautiful daughter Charis Brielle, my parents Thandi Lesole, Mphoeng Joromea and Pitsoemang Joromea. To my entire family, you have been my source of inspiration.

Abstract

Southern Africa is disproportionately affected by HIV-1, with Botswana being among the most affected countries. The interindividual heterogeneity in susceptibility or resistance to HIV-1 and progression upon infection is attributable to, among other factors, host genetic variation. Characterisation of human genetic variations can contribute towards understanding the genetic aetiology of HIV-1 and foster development of novel preventive and treatment strategies against HIV-1. Despite the high burden of HIV-1 in Botswana, the population of Botswana is significantly underrepresentation in genomics studies of HIV-1. Furthermore, the bulk of previous genomics studies evaluated common human genetic variations, however, there is increasing evidence of the influence of rare variants in the outcome of diseases which may be uncovered by comprehensive complete and deep genome sequencing. This research aimed to characterise human genomic variations of Batswana in order to elucidate mutation burden, assess population structure and evaluate the role of these genomic variations in susceptibility to HIV-1 and progression through bioinformatics analyses.

Whole genome sequences (WGS) of 265 HIV-1 positive and 125 were HIV-1 negative unrelated individuals from Botswana were computationally analysed. The sequences were mapped to the human reference genome GRCh38. Population joint variant calling was performed using Genome Analysis Tool Kit (GATK) and BCFTools. Variant characterisation was achieved by annotating the variants with a suite of databases in ANNOVAR. The genomic architecture of Botswana was assessed through principal component analysis and structure analysis and F_{ST}. Cumulative effects of rare variant sets on susceptibility to HIV-1 and progression (CD4+ T-cell decline) were determined with optimized Sequence Kernel Association Test (SKAT-O). Functional analysis of the prioritized variants was performed through gene-set enrichment using databases in GeneMANIA and Enrichr.

Variant characterization revealed 24 damaging variants with the most damaging variants being *ACTRT2* rs3795263, *HOXD12* rs200302685, *ABCB5* rs111647033, *ATP8B4* rs77004004 and *ABCC12* rs113496237. There was admixture of Khoe-San, Niger-Congo and European ancestries observed in the population of Botswana, however, there was no evidence of overall substructure among the HIV-1 positive/negative individuals of Botswana, indicating similar genetic exposure among HIV-1 samples. No variant set was significantly associated with susceptibility to HIV-1, while sets of novel rare-variants within the *ANKRD39* (8.48 x 10^{-}

^{8}), *LOC105378523* (7.45 x 10^{-7}) and *GTF3C3* (1.36 x 10^{-6}) genes were significantly associated with HIV-1 progression. Functional analysis revealed that the variants affected several pathways including chemokine signalling, glycolysis, glycosylation, HIV-1 and host receptor glycoprotein biosynthesis, intracellular transport of molecules and transcription pathways. These findings highlight the significance of whole genome sequencing in pinpointing rare variants of clinical relevance. This PhD thesis unravelled novel genes and novel rare variants that are putatively linked to HIV-1 progression. The thesis contributes towards a deeper understanding of the host genetics HIV-1 and offers promise of population specific interventions against HIV-1.

Table of Contents

List of Figures

List of Tables

Abbreviations

1KGP	1000 Genomes Project
AA	Amino-acid
ABCB5	ATP binding cassette subfamily B member 5
ABCC12	ATP Binding Cassette Subfamily C Member 12
ACTRT2	Actin Related Protein T2
AFR	African (or African/American)
AGVP	African Genome Variation Project
AIDS	Acquired immune deficiency syndrome
AKR1	Aldo-keto reductase
ALT	Alternative allele
ANKRD39	Ankyrin Repeat Domain 39
ANOVA	Analysis of variance
APOBEC3G	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G
ARG	AIDS restriction gene
ART	Antiretroviral therapy
ATP8B4	ATPase phospholipid transporting
BAM	Sequence Alignment/Map binary format
BWA-MEM	Burrows-Wheeler Aligner (memory efficient)
C6orf48	Small nucleolar RNA host gene 32
cART	Combination ART
AZT	Azidothymidine
CCL5	Chemokine (C-C motif) ligand 5
CCNG1	Cyclin G1
CCR2	C-C chemokine receptor type 2
CCR5	C-C chemokine receptor type 5
CD4	Cluster of differentiation 4
cDNA	Complementary DNA
CDS	Coding sequence
CHR	Chromosome
cM	Centimorgan
CNV	Copy number variation
CV	Cross Validation
CXCR6	C-X-C motif chemokine receptor 6
dbSNP	Single Nucleotide Polymorphism Database
DDX40	DEAH-Box Helicase 40
DNA	Deoxyribonucleic acid
EIF3K	Eukaryotic Translation Initiation Factor 3 subunit K
EMMAX	Efficient Mixed-Model Association eXpedited
ENO	Enolase
EU	European
Euro-CHAVI	Center for HIV/AIDS Vaccine Immunology
ExAC	Exome Aggregation Consortium
FUT10	Fucosyltransferase 10
FS	Frameshift

F_{ST}	Wright's fixation index
GATK	Genome Analysis Toolkit
GCTA	Genome-wide Complex Trait Analysis
gnomAD	Genome Aggregation Database
GTF3C3	General Transcription Factor IIIC Subunit 3
GWAS	Genome-wide association studies
HAART	Highly active ART
H3Africa	Human Heredity and Health in Africa Consortium
HGP	Human Genome Project
HIST1H4A	Histone Cluster 1 H4 Family Member A
HIST1H4B	Histone Cluster 1 H4 Family Member B
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HCG8	HLA complex group 8
HCG2	HLA Complex Group 22
HCP5	HLA Complex P5
HIF-1	Hypoxia-inducible factor 1
HOXD12	Homeobox D12
hPNPase^{old-35}	Human polynucleotide phosphorylase
HRDC	Health Research Development Committee
HREC	Human Research Ethics Committee
HWE	Hardy-Weinberg Equilibrium
IBD	Identity-by-descent,
IGSF21	Immunoglobin Superfamily Member 21
IL	Interleukin
Indel	Insertion/Deletion variants
IRB	Institutional Review Board
IRIS	Immune reconstitution inflammatory syndrome
Kb	Kilobases
KIR	Natural killer cell immunoglobin-like receptors
KLF3	Kruppel like factor 3
LCT	Lactase
LD	Linkage disequilibrium
LOC105378523	RNA Gene
LP	Lactase persistance
LOF	Loss-of-function
MAF	Minimum allele frequency
Mb	Megabases
MeSH	Medical subject headings
MHC	Major histocompatibility complex
MICB	MHC class I polypeptide-related sequence B
mRNA	Messenger RNA
MTCT	Mother to child transmission
mtDNA	Mitochondrial DNA
MTX3	Metaxin
$NADP^+$	Nicotinamide adenine dinucleotide phosphate

NCBI	National Center for Biotechnology Information
NCBP2	Nuclear Cap Binding Protein Subunit 2
ncRNAs	non-coding RNA
NGS	Next generation sequencing
NOTCH4	Notch receptor 4
NRG1	Neuregulin 1
nsSNV	non-synonymous SNV
OOA	Out of Africa
OR	Odds ratio
PARD3B	Par-3 family cell polarity regulator beta
PCA	Principal Components Analysis
PDCD	Pyruvate dehydrogenase complex deficiency
PDH	Pyruvate dehydrogenase
PROX1	Prospero homeobox 1
PSORS1C1	Psoriasis Susceptibility 1 Candidate 1
RANTES	Regulated upon activation, normal T cell expressed and secreted
REF	Reference allele
RH	Relative hazards
RICH2	Rho GTPase activating protein 44
RNA	Ribonucleic acid
RNF39	Ring Finger Protein 39
ROH	Runs of homozygosity
RXRG	Retinoid X receptor gamma
SAC	The “Coloureds” of South Africa
SAHGP	South African Human Genome project
SCFD1	Sec1 Family Domain Containing 1
SDC2	Syndecan 2
SE	Standard error of the mean
SKAT	Sequence Kernel Association Test
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOX5	Sex determining region Y -box 5
sSNV	synonymous SNV
TB	Tuberculosis
TBC1D1	TBC1 domain family member 1
TCA	Tricarboxylic acid
TGFBRAP1	Transforming growth factor beta receptor associated protein 1
TI/TV	Transition-transversion
TNF	Tumor necrosis factor
TNXB	Tenascin XB
TRIM-5	Tripartite Motif Containing 5
TRIM10	Tripartite motif containing 10
TSG101	Tumor Susceptibility 101
UNAIDS	The Joint United Nations Programme on HIV/AIDS
UTR	Untranslated region
VCF	Variant call format
WES	Whole exome sequencing

WGS	Whole genome sequencing
YPEL2	Yippee Like 2
Y-STR	Y-chromosome short tandem repeat
ZDHHC19	Zinc Finger DHHC-Type Palmitoyltransferase 19
ZNRD1	Zinc ribbon domain-containing 1

Chapter 1. General Introduction

1.1 Introduction

In the pursuit of understanding the human physio-biology, the Human Genome Project (HGP) released the first human genome in February 2001 (production period 1999-2003). The project used Sanger sequencing; mainly short-gun sequencing was performed through shearing of the genome into segments which are then sequenced randomly (International Human Genome Sequencing Consortium, 2001). The short-gun method was cumbersome in that it produced redundancy and cloning bias during the sequencing of large genomes (International Human Genome Sequencing Consortium, 2001). The finished human genome sequence (also known as hg17) that covered about 99% of the human genome was published by National Center for Biotechnology Information (NCBI) (International Human Genome Sequencing Consortium, 2004). The HGP was performed on slab-gels and capillary systems with a throughput of 96 parallel sequence reads at a time. This project unravelled 3 billion base pairs (2,851,330,913 nucleotides) of the human genome, about 25,000 genes (Pavlopoulos et al., 2013) and just over 1 million single nucleotide variants (SNVs) (International Human Genome Sequencing Consortium, 2004).

The then state-of-the-art sequencing technique (first generation sequencing) was replaced by next generation sequencing (NGS) techniques which could sequence millions to billions of sequence reads at a time at a cheaper price (Mardis, 2008; Metzker, 2010). NGS methods have been applied to re-sequence the human genome to enhance the understanding of how genomic variations affect health (Metzker, 2010). After the completion of the first HGP, another project, International HapMap Project, was established to understand variability within the human populations (Gonzalez-Garay, 2014). The aim of the HapMap project was to determine common variation patterns in the human genome by characterising sequence variants, variant frequencies, correlations between the variants, in populations of African, Asian and European ancestry (The International HapMap Consortium, 2003). Since the HapMap variations had a minimum allele frequency (MAF) of 5% in the population, there arose a need to document variants of a much lower frequency. The 1000 Genomes Project

sought to resequence approximately 2,500 human genomes and identify rare variants of population-specific frequency even less 1% to provide a more comprehensive catalogue of human genome variations (Buchanan et al., 2012; The 1000 Genomes Project Consortium, 2012). The 1000 Genomes Project (1KGP) was accomplished through implementation of NGS methods (The 1000 Genomes Project Consortium, 2010).

Genomic variations occur at different frequencies within different populations, which infers diversity. Genetic diversity of sub-Saharan Africa has been documented in projects such as the HapMap, the 1KGP, the African Genome Variation Project (AGVP) by Gurdasani *et. al.*, 2014 and the South African Human Genome project (SAHGP) by Choudhury *et. al.* 2017, to name a few. Populations that have been studied in these projects are from Nigeria, Kenya, Ethiopia, Ghana, Gambia, Namibia and South Africa (Choudhury et al., 2017; Gurdasani et al., 2015; Schuster et al., 2010; The 1000 Genomes Project Consortium, 2012; Zhang and Dolan, 2010). Africa has the highest population genetic diversity (Campbell and Tishkoff, 2008; Reed and Tishkoff, 2006), however, a large part of Southern Africa had been missed from previous population genetics studies. It is less likely that the findings of previous studies can be translated in the Southern African region due to possible genetic diversity as a result of geographic distance (genetic drift) (Handley et al., 2007; May et al., 2013).

Advances in the HGP and other human genotyping projects have driven the scientific community into an era of genome-wide association studies (GWAS) in which variations across the genomes of many individuals are investigated for potential associations with diseases (Hirschhorn and Daly, 2005). Two main approaches of mapping common disease associated genes are candidate gene studies and linkage mapping. There are two types of linkage analysis: parametric or model-based method in which the genetic model of inheritance is explicitly specified and non-parametric or model-free method in which few or no assumptions about the genetic model are made (Strauch et al., 2000). Parametric linkage analysis has traditionally been used in Mendelian traits while non-parametric linkage analysis has been useful in pedigree studies of complex traits (Kruglyak et al., 1996). Linkage requires that the variants be in the same chromosome within 10–20 cM. This means linkage is less powered to identify common disease variants as the multiple alleles that may not necessarily be linked can influence an outcome of disease (Hirschhorn and Daly, 2005; Risch and Merikangas,

1996).

A practical alternative to linkage analysis has been candidate gene association studies in which genes are selected either by being located in a region of linkage or by evidence of association with a disease (Hirschhorn and Daly, 2005). Likewise, candidate gene approaches have not been successful in identifying genes that underlie complex diseases as the approach targets specific genes that are known to cause a disease (Tabor et al., 2002). A superior approach that is not biased as it does not rely on previous knowledge of gene-disease association is the genome-wide association (GWA) approach. In this approach the genome is surveyed for all possible causal variants (Hirschhorn and Daly, 2005) with assumption based on common diseases and common variants. This means only common causal variants can be identified using this method.

Despite successes in antiretroviral drugs administration and prevention of new HIV infections, Sub-Saharan Africa remains the most affected region. Over half of the 38 million global cases of HIV live in sub-Saharan Africa (UNAIDS, 2020). Due to the high disease burden and high genetic diversity of African populations, it is very compelling to study the influence of genomic variations on disease. Botswana is one of the most affected countries in Southern Africa with a general population HIV-1 prevalence of 20.7% [18.2 - 22.1] (UNAIDS, 2019) in adults. The aetiology of HIV-1 is an interplay of several factors such as socio-behavioural factors, environmental factors, viral genetics and host genetics (Fellay et al., 2010; McLaren et al., 2015; Thami and Chimusa, 2019; Tough and McLaren, 2019). It is well known that increasing GWAS in diverse populations may further contribute in understanding the mechanism underpinning genetic diseases and enhance evidence of GWAS results for their clinical utility.

Variations in the human genome have been implicated in HIV-1 susceptibility, acquisition, progression and treatment response (Ballana and Este, 2013; International Human Genome Sequencing Consortium, 2001). Previous GWA studies of association with HIV used microarray analysis of SNVs for the identification of common variants (Dalmasso et al., 2008; Fellay et al., 2007; Joubert et al., 2010; Limou and Zagury, 2013; Pelak et al., 2010). The first GWA study on HIV performed by the Euro-CHAVI consortium identified three SNVs associated with HIV control and disease progression in European samples. This study used Illumina

HumanHap550 BeadChip microarray with 555,352 SNVs. The study revealed an association between human leukocyte antigen-C (*HLA*-C; rs9264942), *HLA* complex P5 (*HCP5*; rs2395029) and zinc ribbon domain-containing 1 (*ZNRD1*; rs9261174) with HIV (Fellay et al., 2007). Another study in the French PRIMO cohort additionally identified rs11725412, a SNV located between *TBC1* domain family member 1 (*TBC1D1*) and (Kruppel like factor 3) *KLF3* genes (Dalmasso et al., 2008). In addition to the rs2395029 (*HCP5*) variant, the study also reported rs3093662 (tumor necrosis factor, *TNF*), rs13199524 and rs12198173 (tenascin XB, *TNXB*) SNVs which were had higher frequencies in long term HIV controllers (Dalmasso et al., 2008). *HLA-B**5703 was also suggested to be associated with regulating HIV-1 viremia in African Americans (Pelak *et al.*, 2010).

In a GWAS that was performed to investigate an association between SNVs and transmission of HIV from mother to child in Malawi, the rs8069770 variant was associated with a lower risk of HIV MTCT (Joubert et al., 2010). Among other SNPs, *HLA* SNVs such *as HLA-B**57, C-*X-C motif chemokine receptor 6* (*CXCR6*) rs2234358, an intronic SNV in *PARD3B* (par-3 family cell polarity regulator beta) and *HLA-C* rs9264942 have been implicated in viral load control and disease progression (Fellay et al., 2007; Limou and Zagury, 2013; Pelak et al., 2010). *CCR5*-Δ32, *HCP5* rs2395029 and rs6996198 have been associated with resistance to HIV-1 (Fellay et al., 2007; Limou and Zagury, 2013; Xie et al., 2017).

A recent study in Botswana discovered two variants which had not been associated with HIV before (Xie et al., 2017). The variants were rs2535307 in chromosome 6, located near the *HLA* Complex Group 22 (*HCG22*) gene and kgp22385164 in chromosome 5 near the Cyclin G1 (*CCNG1*) gene. *HCG22* is a mucin-like, immune regulatory gene, its expression is stimulated by interleukin-1 (IL-1) (Jeong et al., 2015; Xie et al., 2017); while the *CCNG1* gene codes for a cyclin that regulates cell growth (Xie et al., 2017). The variants were significantly associated with HIV disease progression (CD4 and viral load dynamics) (Xie et al., 2017). As seen in the aforementioned studies, most HIV susceptibility genes are part of the HLA pathway.

Imputation is a commonly used method for fine-mapping the causal variant following GWAS (Chundru et al., 2019; Mitt et al., 2017; Servin and Stephens, 2007). However, imputation has limitations such as that 1) very rare variants cannot be imputed or only with very poor

accuracy (Mitt et al., 2017; Servin and Stephens, 2007) and 2) although the TOPMed imputation panel (Kowalski et al., 2019) includes individuals of African American ancestry there are no imputation panels that are available that would be appropriate to use for the Botswana population. As such, a comprehensive and unbiased approach such as whole genome sequencing (WGS) that can reveal and capture most variants including rare variants, is needed (An and Winkler, 2010; Cirulli et al., 2010). There is scant data on human genetic diversity let alone association with HIV-1 in Botswana population. Previous studies on genetic diversity of Botswana were either focusing on a particular ethnic group, focusing on single candidate gene investigation (Pickrell et al., 2012; Tau et al., 2015) or targeted sequencing (Tau et al., 2017). This data may not fully capture the genetic diversity or be a representative of the general population structure of Botswana. More recently an assessment of population structure within Botswana using whole exome sequencing revealed that there was little evidence of population substructure (Retshabile et al., 2018). A study by Xie *et. al.* (2017) that reported on variations associated with HIV was also performed on microarray with common/known variants (Xie et al., 2017).

1.2 Challenges and Opportunities

Host genetics play a major role in the clinical course of HIV-1. Previous studies in this field were performed using methods ascertained in populations of European ancestry. This poses a challenge because these methods are not fully applicable in African populations due to underlying genetic diversity and population structure. The African genome is the most diverse, with very different linkage patterns (shorter haplotype blocks) compared to other populations. We have fully discussed the complexity of African genomes in relation to HIV-1 in a peer reviewed paper (Thami and Chimusa, 2019). Due to the diverse genetic architecture, we need a high-resolution method that can pinpoint population specific variants that may be of clinical relevance to HIV-1. Whole genome sequencing (WGS) offers such an opportunity, where rare and novel variants can be uncovered in an understudied African population. Our study will be the first to use WGS in Botswana to answer the following objectives.

1.3 Aims of the project

1.3.1 Hypothesis/Research Question

The hypothesis of this study is that additional novel and some rare-variants in the human genome are in the Botswana population, and some variations may predispose individuals to acquiring HIV infection and progressing to disease.

1.3.2 Research Objectives

In this study, we intend to mine variants from whole genome sequences using bioinformatics tools to reveal a comprehensive catalogue of genetic variations in the population of Botswana and to investigate associations between the identified variations with HIV acquisition and progression.

Aim 1: To analyze whole genome sequences of Botswana nationals (Batswana) and investigate genomic variations and the genetic architecture of a Botswana population.

Aim 2: To leverage the whole genome sequences of HIV-1 positive and negative participants to investigate the association between the genomic variations and HIV-1 infection (susceptibility to HIV-1 acquisition and progression).

1.3.3 Specific Objectives

1. To analyze raw human whole genome sequences of 390 individuals from Botswana using bioinformatics tools.
2. To characterize common and rare variations within the human whole genome sequences of Botswana.
3. To describe the genetic architecture and investigate population structure within Botswana.
4. To elucidate the mutation burden in the complete human genomes from Botswana.
5. To identify HIV-1 associated genetic loci in the human whole genome sequences by aggregating the effects of rare-variants.

1.4 Significance of the project

Botswana has limited data on the genetic structure and genetic diversity of the general population. The country carries one of the highest HIV prevalence in the world. It would be interesting to study genetic variations within Botswana population which would also shed light on association of host genetics with HIV infection. Analysing whole genome sequences from Botswana would also give a comprehensive catalogue of genetic variations within the population and give insights of population homogeneity or stratification. The study will contribute insights that may lead to the discovery of novel variants unique to Botswana population and possibly "lead to the discovery of new host genetic associations with HIV that could lead to the development of new antiretroviral therapy or vaccine" (Limou and Zagury, 2013). This study will be adapting and optimizing a human whole genome sequence analysis method in context with HIV-1 that can be referred to for subsequent NGS projects. Though the cost of sequencing has significantly reduced since the introduction of NGS methods, GWA studies still require thousands of samples to be sequenced for a study to be of significant power. This means GWAS on Whole Genome Sequences are expensive to perform. The aggregation of the effects of HIV-1 associated variants can aid in identifying relevant genetic loci in a small sample size.

1.5 Outline of the thesis

This thesis began in the current chapter (**Chapter 1**) with a general background highlighting the evolution of human genomics research that was catalysed by the Human Genome Project. Chapter 1 highlights the importance of genomic variation in population diversity and the implication of this in diseases particularly HIV. The chapter ends with a mention of challenges and how we planned to mitigate these.

Chapter 2 is a comprehensive account of the genetic structure of the human populations of Southern Africa presented as a published systematic review (Thami and Chimusa, 2019). In the same breath, the implications of the genetic landscape of Southern African populations on HIV-1 are dissected. The chapter concludes by suggesting methods that can better unravel novel/rare variants associated with HIV-1 in Southern African Populations.

Chapter 3 presents state-of-the-art methods of deciphering human genetic variations and subsequently the findings of characterizing genetic variations and predicting mutation burden

in a Southern African population of Botswana.

Chapter 4 is an account of the assessment of population structure, genetic relatedness and admixture in Botswana using genetic variants identified in Chapter 3.

Chapter 5 illustrates the role of rare-variants in susceptibility to HIV-1 acquisition and progression in the population of Botswana.

Then **Chapter 6** is the general discussion of the main research findings and conclusion of the thesis.

1.6 Contribution to the field

We identified 2.8 million novel single nucleotide variants from the whole genome sequences of a cohort of 390 individuals from Botswana. The number of novel variants is more than what was reported by a previous whole exome study in Botswana, indicating that we have identified additional novel variants in the Botswana population that is not captured in public databases. Our study is the first to characterize a comprehensive catalogue (whole genome) of genomic variations and evaluate the burden of human genomic mutations in Botswana. To this effect we identified single nucleotide variants which could potentially disrupt the function of 24 genes, the most deleterious (damaging) variants being *ACTRT2* rs3795263, *HOXD12* rs200302685, *ABCB5* rs111647033, *ATP8B4* rs77004004 and *ABCC12* rs113496237. Eighteen of the most damaging variants had higher minor allele frequencies (MAFs) than those of Europeans (our data compared to the Genome Aggregation Database – gnomAD data) which accentuates that African populations have the highest genetic diversity. The disparity in MAFs drives population structure and differential risk of disease in different ethnicities. Our study is also the first to evaluate the role of rare-variants in susceptibility to HIV-1 and progression to disease in Botswana at a larger scale. We observed novel associations of rare-variants in or near the *ANKRD39*, *LOC105378523* and *GTF3C3* genes with HIV-1 progression. Through *in silico* functional analysis of the human genomic variations prioritized by both mutation burden analysis and genetic association tests, we observed that among other effects, the variants potentially affect HIV-1 glycoprotein synthesis and host receptor composition which are crucial for controlling susceptibility to HIV-1 acquisition and progression. These findings have a great potential to illuminate the budding field of immunometabolism in which biological processes such as glycolysis and glycosylation are being investigated to develop preventive, diagnostic and treatment strategies against HIV-1. Our findings not only uncovered HIV-1 related pathways affected by the discovered genomic variations, but also suggested an interplay of HIV-1 and cancer that both have a potential to reprogram the host cell metabolism. This can inform host genetics strategies of prevention, management and treatment against the 2 comorbidities in an African setting.

Chapter 2. Population Structure and Implications on the Genetic Architecture of HIV-1 Phenotypes within Southern Africa

Original publication

Prisca K. Thami and Emile R. Chimusa. (2019). Population Structure and Implications on the Genetic Architecture of HIV-1 Phenotypes Within Southern Africa. *Front. Genet.* 10, 905. doi:10.3389/fgene.2019.00905.

Nature of publication: Systematic review

Journal: Frontiers in Genetics

Pubmed link: https://pubmed.ncbi.nlm.nih.gov/31611910/

Journal link: https://doi.org/10.3389/fgene.2019.00905

Candidate's contribution: Conceived the structure, conducted the literature searches, drafted and edited the manuscript.

Co-author contribution: ERC supervised the research and edited the manuscript.

Synopsis of paper 1: This paper gives a comprehensive background on the genetics of the people of Southern Africa, and the implications of this genetic landscape on susceptibility to HIV-1 and progression. We also dissected the genome-wide association studies (GWAS) of susceptibility to HIV-1 and progression within the past 12 years. We discovered that the populations of Southern Africa are genetically structured and have complex scenarios of admixture. Of the three GWAS of common-variants from Southern Africa, one from Botswana identified novel statistically significant single nucleotide polymorphisms (SNPs) that were associated with susceptibility to HIV-1 acquisition and progression within the *HCG22* and *CCNG1* genes. The identification of novel associations of HIV-1 implies that more associations could be unravelled in the human population of Botswana using a more robust method. This paper gives a solid background on population diversity and the clinical relevance of genomic variations which are significant components of this thesis.

frontiers in Genetics

REVIEW
published: 27 September 2019
doi: 10.3389/fgene.2019.00905

Population Structure and Implications on the Genetic Architecture of HIV-1 Phenotypes Within Southern Africa

Prisca K. Thami[1,2] *and Emile R. Chimusa*[1]*

[1] Division of Human Genetics, Department of Pathology, University of Cape Town, Cape Town, South Africa,
[2] Research Laboratory, Botswana Harvard AIDS Institute Partnership, Gaborone, Botswana

The interesting history of Southern Africa has put the region in the spotlight for population medical genetics. Major events including the Bantu expansion and European colonialism have imprinted unique genetic signatures within autochthonous populations of Southern Africa, this resulting in differential allele frequencies across the region. This genetic structure has potential implications on susceptibility and resistance to infectious diseases such as human immunodeficiency virus (HIV) infection. Southern Africa is the region affected worst by HIV. Here, we discuss advances made in genome-wide association studies (GWAS) of HIV-1 in the past 12 years and dissect population diversity within Southern Africa. Our findings accentuate that a plethora of factors such as migration, language and culture, admixture, and natural selection have profiled the genetics of the people of Southern Africa. Genetic structure has been observed among the Khoe-San, among Bantu speakers, and between the Khoe-San, Coloureds, and Bantu speakers. Moreover, Southern African populations have complex admixture scenarios. Few GWAS of HIV-1 have been conducted in Southern Africa, with only one of these identifying two novel variants (*HCG22*rs2535307 and *CCNG1*kgp22385164) significantly associated with HIV-1 acquisition and progression. High genetic diversity, multi-wave genetic mixture and low linkage disequilibrium of Southern African populations constitute a challenge in identifying genetic variants with modest risk or protective effect against HIV-1. We therefore posit that it is compelling to assess genome-wide contribution of ancestry to HIV-1 infection. We further suggest robust methods that can pin-point population-specific variants that may contribute to the control of HIV-1 in Southern Africa.

Keywords: population structure, diversity, genome-wide association studies (GWAS), host genetics, Southern Africa, HIV-1

OPEN ACCESS

Edited by:
William Scott Bush,
Case Western Reserve University,
United States

Reviewed by:
Fabio Marroni,
University of Udine, Italy
Mark Z. Kos,
University of Texas Rio Grande Valley
Edinburg, United States

***Correspondence:**
Emile R. Chimusa
emile.chimusa@uct.ac.za

Specialty section:
This article was submitted to Applied Genetic Epidemiology, a section of the journal Frontiers in Genetics

Received: *16 May 2019*
Accepted: *26 August 2019*
Published: *27 September 2019*

Citation:
Thami PK and Chimusa ER (2019) Population Structure and Implications on the Genetic Architecture of HIV-1 Phenotypes Within Southern Africa. Front. Genet. 10:905. doi: 10.3389/fgene.2019.00905

INTRODUCTION

Southern Africa extends across a 2.7 million km^2 land in the southernmost part of Africa and is the home to about 66 million people (Worldometers, 2019). The region comprises 10 mainland countries: Angola, Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa, eSwatini, Zambia, and Zimbabwe (Marks, 2014). The region carries the highest burden of human immunodeficiency virus/acquired immune deficiency syndrome (HIV/AIDS). Around 37 million people live with HIV globally,

and over half of these live in Southern and eastern Africa (UNAIDS, 2017). Despite high exposure to HIV, some people stay uninfected (Fowke et al., 1996), while those infected exhibit heterogeneous clinical outcomes of HIV infection and to an extent also differential antiretroviral drug metabolism (Carr et al., 2017). This heterogeneity is partly due to underlying diversity in host genetics (Telenti and Goldstein, 2006).

Genetic diversity within and between human populations can be inferred from the observed distributions of genetic allele frequencies in populations, a concept also known as population (genetic) structure (Wright, 1950; Underhill and Kivisild, 2007). The classic measure of genetic population structure is Wright's fixation index, F_{ST}, a type of ANOVA statistic (Tishkoff and Williams, 2002; Jakobsson et al., 2013). The value ranges from zero to one, with F_{ST} scores close to one indicating a high degree of population divergence (i.e., most genetic variation between populations), and F_{ST} scores close to zero indicating strong population similarity and gene flow (i.e., most genetic variation within populations) (Tishkoff and Williams, 2002). Geographic isolation, migration, population bottleneck, admixture, language and culture, and natural selection have played significant roles in contributing to variations in allele frequencies and the genetic landscape of Southern Africa (Campbell and Tishkoff, 2008; Campbell and Tishkoff, 2010; Botigue et al., 2013; Breton et al., 2014; Gurdasani et al., 2014; Li et al., 2014; Macholdt et al., 2014; Chimusa et al., 2015) as illustrated in **Figure 1**.

Genome-wide association studies (GWAS) are approaches of scanning the genome to identify *in silico* variants that confer susceptibility or resistance to a particular disease through statistical association models (Hirschhorn and Daly, 2005; Hutcheson et al., 2008). Unlike candidate gene-based methods that require *a priori* knowledge of suspected genes, GWAS have the potential to discover novel genomic loci (Telenti and Goldstein, 2006). Although GWAS are powerful techniques, the variants discovered through these methods (Purcell et al., 2007; Kang et al., 2010; Yang et al., 2011; Wen et al., 2018) have not accounted for all of the variability in viral load (Fellay et al., 2007; Fellay et al., 2009; Pereyra et al., 2010). The overall heritability of set point viral load in populations of European ancestry measured through GWAS was estimated to be 24.6%. Common variants contributed largely to this estimate of heritability (McLaren et al., 2015; Tough and McLaren, 2019). Like other complex traits, this highlights the importance of solving the missing heritability of HIV-1 infection phenotypes which might be uncovered by discovering factors such as rare variants, structural variants, and gene-gene and gene-environment interactions responsible for inter-host variability of viral load (Verma and Ritchie, 2018).

Confounders such as population structure can affect GWAS results. These have to be controlled to avoid spurious results (Hirschhorn and Daly, 2005; Price et al., 2006; Tishkoff et al., 2009; McLaren and Carrington, 2015). Moreover, characterizing genetic structure is crucial for reconstruction of human population history (Tishkoff et al., 2009). In general, African populations have the highest genetic variation and lower linkage disequilibrium (LD) among loci (Campbell and Tishkoff, 2008;

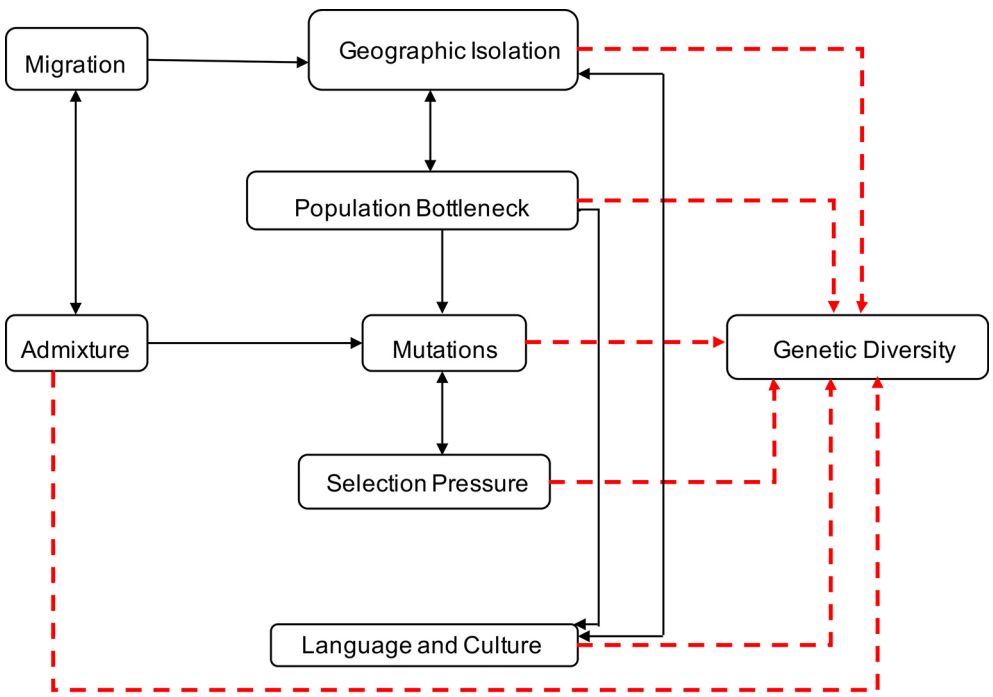


FIGURE 1 | Factors contributing to human genetic diversity within Southern Africa. Although it is complex to represent the interplay between geographic isolation, migration, population bottleneck, admixture, language, culture, and natural selection on the genetic structure of Southern Africa, this figure provides a basic illustration of some of the factors that shaped the genetics of the people of Southern Africa. Geographic distribution and migration into Southern Africa have influenced population admixture and geographic isolation, yielding to population bottleneck, mutation, and selection, that shaped the genetics, culture, and language diversity in Southern Africa.

The 1000 Genomes Project Consortium, 2010; Choudhury et al., 2018); therefore, not all tag-single nucleotide polymorphism (SNPs) selected from other populations can be used as proxies in African populations. Risk alleles can be structured in populations due to multiple demographic factors and genetic ancestry contributions (Botigue et al., 2013; Gurdasani et al., 2014; Chimusa et al., 2015; Skoglund et al., 2017). The people of Southern Africa are culturally, linguistically, and genetically diverse; the region has been underrepresented in previous genetic diversity studies (Awany et al., 2018; Choudhury et al., 2017; Sirugo et al., 2019).

Most GWAS were performed in non-African populations (Awany et al., 2018; Sirugo et al., 2019) in which HIV-1B is the prevalent subtype. It is possible that the genetics underlying the control of HIV-1 in Southern African is different from these other populations. Considering these genetic differences between African and other populations, and due to the enormous burden of HIV within Southern Africa, it is imperative to dissect human genetic diversity and investigate the role of genetic landscape on HIV acquisition and progression within the region. Deducing a comprehensive architecture of HIV host genetics in Southern Africa will assist in the development of population-specific interventions against HIV. Hence, this review aims to present a comprehensive discussion of the advances made in the GWAS of HIV-1 and document common variants within Southern Africa associated with HIV-1 infection.

We used PubMed search engine to retrieve HIV-1 GWAS studies which have been published in the past 12 years (2007–2019); species was restricted to the human species. The specific search terms were the following:((“genome”[MeSH Terms] OR “genome”[All Fields]) AND wide[All Fields] AND (“association”[MeSH Terms] OR “association”[All Fields]) AND (“hiv-1”[MeSH Terms] OR “hiv-1”[All Fields])) AND (“2007/01/01”[PDat]: “2019/04/30”[PDat] AND “humans”[MeSH Terms]). Ninety-eight items were retrieved; articles relevant to Southern Africa were used in the review. Cited studies which were not in the search results were directly searched for. To review population structure and admixture in Southern Africa, a relaxed search of the terms (population structure and Southern Africa; human genetic diversity and Southern Africa; admixture and Southern Africa) was performed in PubMed, and relevant articles were selected for this review. SNP annotations were confirmed on dbSNP (Sherry et al., 2001). A map of migration routes (refer to the *Migration Into Southern Africa* section) was created using maps package in R and edited using MacOS Preview software. We conclude with a discussion of research areas where further work on GWAS of HIV-1 is needed.

MIGRATION INTO SOUTHERN AFRICA

Demic diffusion has been shown to have a paramount role in genetic drift and gene flow, the two most important events that can trigger population structure. The major expansions of humans date as far back as 100,000 years ago (the Out of Africa (OOA) model) (Campbell and Tishkoff, 2008). One of the major human expansions that occurred in relatively recent years is the Bantu expansion that happened 3,000–5000 years ago. The Bantu dispersed from western Africa, in the region known as the proto-Bantu homeland in eastern Nigeria and western Cameroon (Cavalli-Sforza et al., 1993; Soodyall, 2006; de Filippo et al., 2010; Li et al., 2014).

Migration routes of the Bantu populations have been well described (Li et al., 2014). According to archeological, linguistics, and genetics evidence, two main theories of the split of Bantu speakers from their western African homeland have been postulated: a) migration of the eastern Bantu speakers directly from western Africa to eastern Africa, and thereafter the ancestors of Nguni and Sotho-Tswana speakers moved between 1,200 and 1,000 years into Southern Africa through Great Zimbabwe (Soodyall, 2006; Li et al., 2014) and b) migration of western Bantu speakers through southern Cameroon directly to Southern Africa. A hypothesis supported by Bayesian tree methods stated that the eastern and western Bantu speakers might have split later after passing through the central African rainforest (Li et al., 2014).

In more recent history, Southern Africa has seen an influx of migrants from other continents such as Europe and Asia. By 1750 the Dutch and Portuguese had colonized Southern Africa on the southwest and east coasts, respectively (Hall, 1993). These colonial powers brought slaves and exiles from Indonesia, Malaysia, east coast of Africa, Madagascar and India, to provide labor in wine and wheat farms in Cape Town, the then Cape of Good Hope (Hall, 1993; de Wit et al., 2010). Intermarriages and liaison between these populations resulted in progeny who are now termed “Coloureds” (de Wit et al., 2010). The major migration routes into Southern Africa are depicted in **Figure 2**.

LANGUAGE AND CULTURAL DIVERSITY IN SOUTHERN AFRICA

The inhabitants of Southern Africa could be distinctly classified according to ways of subsistence and food production: hunters and gatherers (and fishers) and the agro-pastoralists (Phillipson, 1977). It has been suggested that the distribution of languages in Southern Africa is an effect of demic diffusion rather than language diffusion alone (Huffman, 1970; Li et al., 2014). Southern Africa was previously occupied by hunter-gatherers who had their own unique culture. Food production was a major drive in human expansions as evidenced in the spread of farming from the cradle in western Africa (Neumann et al., 2012; Russell et al., 2014). Farming was able to support higher population densities (population explosion) than hunting and gathering, which gave farmers an edge over the indigenous inhabitants. As such the culture and languages of hunter-gatherers have been replaced by that of early farmers (Diamond and Bellwood, 2003).

In general, African populations can be classified into four main ethnolinguistic groups: Niger-Kordofanian, Afroasiatic, Nilo-Saharan, and Khoisan (Tishkoff et al., 2009). Over 200 million sub-Saharan inhabitants speak a collective of related languages known as Bantu languages. Bantu speakers are a subgroup of the Niger-Kordofanian ethnolinguistic group (Heine and Nurse, 2000; Reed and Tishkoff, 2006; Campbell and Tishkoff, 2010; Li et al., 2014; Veeramah and Hammer, 2014). Majority of Southern African inhabitants are Bantu speakers and Khoisan-speakers. Over 50 Bantu languages are spoken in Southern Africa

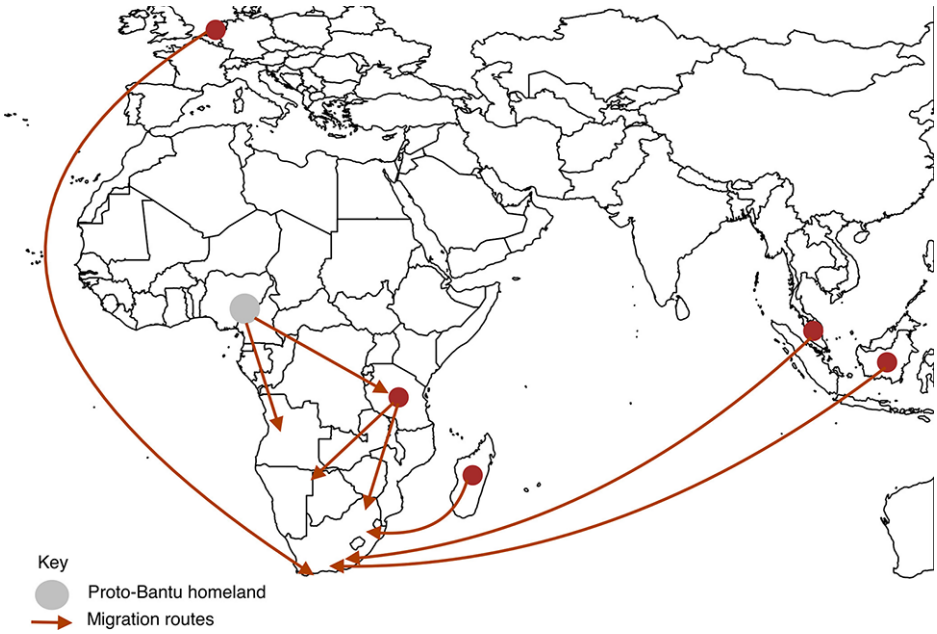


FIGURE 2 | Migration routes into Southern Africa. We present here the main routes of pre-historic migration of the ancestors of the Southern African inhabitants. The circles represent places of origin, the gray circle being the proto-Bantu land and dark red circles being other places of origin. The arrows represent movement of people from places of origin into Southern Africa.

Eberhard et al. (2019), these include but are not limited to Tswana, Nama, Xhosa, Zulu, Ndebele, Swati, Sotho, Shona, Bemba, and Makua (Heine and Nurse, 2000; Soodyall, 2006).

Khoisan language families include Khoe, !Xun (Ju), Ta'a (including !Xo), !Wi, and Southwestern or Cape /Xam. The Khoe and the San intermixed resulting in linguistic affinities and diffusion of culture between the two groups; this led to them collectively being referred to as the Khoe-San (Schlebusch, 2010; Schlebusch et al., 2016). The present-day Khoe-San inhabit the central Kalahari Desert and Okavango swamps of Botswana, much of Namibia, Angola, Zimbabwe, Cape Province, and some parts of Lesotho. It is believed that the Khoe-San occupied most of Southern Africa before the arrival of farmers who displaced the Khoe-San to the desert (Barnard, 1992; Li et al., 2014). The Khoe-San are the indigenous inhabitants of Southern Africa. Analysis of genome-wide data and the lactase (LCT) region has shown that the Khoe originated from the assimilation of the San into eastern African Bantu-speaking populations (Schlebusch et al., 2012; Breton et al., 2014; Sadr, 2015; Schlebusch et al., 2016).

Using autosomal SNPs, the Pygmy (a Central Africa population), the Hadza, and Sandawe clustered together with Southern African Khoe-Sans. This suggested that the Pygmy, Southern African Khoe-San, Hadza, and Sandawe populations are a remnant of a proto-Khoe-San-Pygmy population (Tishkoff et al., 2009). Although there have been postulations of the San also originating in East Africa (Soodyall, 2006; Campbell and Tishkoff, 2008; Gonder et al., 2007; Barbieri et al., 2013a; Morris et al., 2014), the origin of the San in Southern Africa cannot be ruled out. In fact, Nurse et al. (1985) suggests that by the 17th century the San lived in most parts of Southern Africa. Before the arrival of Bantu speakers, the Southern San inhabited the foothills of the Drakensberg mountains, the northern, eastern, and western Cape provinces of South Africa, and the southmost parts of Botswana and Namibia [Supplementary information (Schlebusch et al., 2016)]. The rock paintings in Tsodilo Hills in northwestern Botswana show that the Northern San had been living around Tsodilo Hills by the first millennium AD before the arrival of the Khoe and the Bantu speakers.

The Southern African Khoe-San underwent a population bottleneck due to barriers such culture, language, and geographic isolation (i.e., the Kalahari Desert, which could not allow their interaction and contact with other populations). The Khoe-San also underwent genocide at the hands of Bantu speakers and European settlers. This stemmed from conflicts over land which the colonists had invaded while the Khoe-San also hunted the colonists' cattle, as they used to hunt antelopes and elands (Adhikari, 2010; Schlebusch et al., 2016). The population decline was also due to the Khoe-San succumbing to diseases such as smallpox which were introduced by the settlers (Chimusa et al., 2015). It is speculated that smallpox might have been introduced into the Cape through passengers from India and then spread to European settlers who might have passed it to the Khoe-San. Since the Khoe-San had less resistance to the disease, this drastically reduced their population numbers. The acquisition of Khoe-San women as wives by the European settlers due to the small number of European women also diminished the gene pool of the Khoe-San, this further exacerbating the bottleneck of the Khoe-San population (Nurse et al., 1985).

Bantu speakers were agro-pastoralists, while the Khoe were pastoralists and the San were hunter-gatherers (Breton et al.,

2014). There has been some cultural (Breton et al., 2014) and language diffusion within Southern Africa. Through interaction, the San acquired the pastoralism lifestyle from the Bantu and Khoe populations (Breton et al., 2014). Although the click consonants have been connected to the Khoe-San, these have been borrowed into some of the Bantu groups such as the Zulu and Xhosa populations (Barbieri et al., 2013a; Marks et al., 2015). Interactions and trade among Southern African populations has resulted in cultural diffusion which has driven intermixing of the different populations; this has in turn shaped the genetics of Southern African populations.

POPULATION GENETIC DIVERSITY IN SOUTHERN AFRICA

The gene pool of Southern Africa is a mosaic of alleles from multiple ancestries (Montinaro et al., 2017). Past historical events such as geographical isolation, colonialism, and the Bantu expansion have helped to shape the genetic landscape of Southern Africa (Choudhury et al., 2017). When determining variation between the Khoe and San, using two genetic marker systems, mitochondrial DNA (mtDNA) and Y-chromosome, Soodyall and colleagues could not unambiguously distinguish the Khoe from the San (Soodyall et al., 2008). According to the authors, the Khoe and the San share a common gene pool, which suggests that they branched off from a common ancestry.

Conversely, genome-wide studies analyzing millions of SNPs revealed that the Khoe and the San are a genetically diverse group (Schlebusch et al., 2012). Schlebusch et al. (2012) analyzed 2.3 million SNPs of autosomal chromosomes from 220 individuals representing 11 populations from Southern Africa. These populations were Ju/'hoansi, !Xun, /Gui and //Gana, Karretjie, ǂKhomani, Nama, Khwe, "Coloured" (Colesberg), "Coloured" (Wellington), Herero, and Bantu speakers (South Africa). In their report, there was a distinct stratification in the Khoe-San group; there was a separation between the Ju speakers (!Xun and Ju/'hoansi) who are the San, and the Nama representing Khoe speakers. This is consistent with the findings of a study by Uren et al. (2016) who found fine-scale structure between the ǂKhomani and Nama of South Africa.

The Khoe-San are genetically distinct from the Bantu populations (**Table 1**). In the Schlebusch et al. (2012) study, a variant associated with "fast-twitching" muscles and elite

TABLE 1 | Genome-wide population diversity studies of Southern African populations.

Population	Country	Genotyping method	Marker	Major findings	Reference
Khoe-San and Bantu speakers	Namibia, South Africa	Microarray (genome-wide)	Autosomal SNPs and CNVs	Khoe-San and Bantu speakers are genetically different.	(Jakobsson et al., 2008; Li et al., 2008)
Bantu speakers, SAC and Khoe-San	South Africa	Panel sequencing	Autosomal SNPs, indels, and microsatellites	Khoe-San, SAC and Bantu speakers are genetically divergent. SAC show the highest level of intercontinental admixture.	(Tishkoff et al., 2009)
Khoe-San and Bantu speakers	Botswana, Namibia, South Africa	Microarray (genome-wide)	Autosomal SNPs	Structure observed among the Khoisan, and between the Khoisan, SAC, and Bantu speakers.	(Pickrell et al., 2012; Schlebusch et al., 2012)
Khoe-San	South Africa	Microarray (genome-wide)	Autosomal SNPs	Precolonial Eurasian admixture observed in Khoe-San populations.	(Pickrell et al., 2014)
Khoe-San	Angola, Botswana, South Africa	Microarray (genome-wide)	Autosomal SNPs, mtDNA, and Y-chromosome	Khoe-San show high genetic differentiation and have sex-biased Bantu speakers and European admixture.	(Henn et al., 2011; Montinaro et al., 2017; Uren et al., 2016; Choudhury et al., 2017; Oliveira et al., 2019)
Bantu speakers and SAC	South Africa	Microarray (genome-wide) and WGS	Autosomal SNPs	Multiway admixture including Khoe-San ancestry observed in the SAC population. Differential Khoe-San admixture detected in Bantu speakers.	(de Wit et al., 2010; Chimusa et al., 2013a, Gurdasani et al., 2014; Chimusa et al., 2015; Uren et al., 2016; Choudhury et al., 2017; Chimusa et al., 2018)
Bantu speakers	South Africa	Microarray (genome-wide)	Autosomal SNPs	Weak clustering of southeastern Bantu speakers possibly due to admixture with Khoe-San.	(May et al., 2013)
Khoe-San and Bantu speakers	South Africa, Namibia, Malawi	Microarray (genome-wide)	Autosomal SNPs	Admixture in Malawian population supports the late split of the Eastern Bantu speakers. Eurasian admixture detected in South African Khoe-San, dating back to European colonial period settlement in Southern Africa.	(Busby et al., 2016)
Bantu speakers and Khoe-San	Botswana	Microarray (whole exome) and WES	Autosomal SNPs	Within population structure of Botswana was not observed. Genetic differentiation was observed between Sotho, Zulu (of South Africa), and the Botswana population.	(Retshabile et al., 2018)

Indels, insertions/deletions; SAC, The "Coloured" of South Africa; SNP, single nucleotide polymorphism; CNV, copy number variation; mtDNA, mitochondrial DNA; WES, whole exome sequencing; WGS, whole genome sequencing. Details on specific groups within Khoe-San and Bantu speakers can be obtained from original publications.

athletic performance had greater frequencies (>90%) in the Khoe-San groups than in other Southern African populations. Moreover, a strong signal of selection was found around the human leukocyte antigen (HLA) complex in several genes that are known to protect against infectious diseases in the ‡Khomani and Karretjie, this effect probably owing to early and extensive contact with European colonists and novel infectious diseases such as smallpox (Schlebusch et al., 2012). Likewise, Tau and colleagues also reported significant genetic differences between the Khoe-San and Bantu-speaking groups through analysis of short tandem repeats (STRs) (Tau et al., 2017).

In another study, SNPs associated with lactase persistence (LP) from eight Khoe-San and seven Bantu-speaking groups from Botswana, Namibia, and Zambia were characterized (Macholdt et al., 2014). LP is a distinctive trait of pastoralism. LP alleles were first identified in eastern Africa (Tishkoff et al., 2007) where the alleles have faced a strong positive selection. In the study conducted by Macholdt et al. (2014), -14010*C was the most frequent SNP with a collective higher prevalence in pastoralists (20.2%) than in foragers (6.7%) or agriculturalists (1.3%). Within the pastoralists group, -14010*C was observed at the highest frequency of 36% in the Nama population. This further shows the implication of migration and gene flow from the east of Africa into Southern Africa through Khoe speakers.

Y-STR analysis of Southern African populations revealed regional variation between the populations of Botswana, Zimbabwe, Namibia, and South Africa (Tau et al., 2015). Whole genome sequences of Southern African populations also revealed genetic variation among the Bantu of South Africa and among Khoe-San groups. Principal component analysis and structure analysis revealed significant genetic differentiation ($p < 10^{-6}$) between the Xhosa and Sotho groups (Choudhury et al., 2017). F_{ST} analysis revealed genomic regions with high divergence between the two groups (Average $F_{ST} \geq 0.3$) (Choudhury et al., 2017; Supplementary Table 11). There have been reports of a weak clustering of Bantu-speaking populations that might be underpinned by admixture (May et al., 2013; Retshabile et al., 2018). Southern (and eastern) African populations have greater genetic diversity than other populations closer to the western African origin. This might be due to admixture that arose from the mixing of eastern and Southern African Bantu populations with indigenous populations of the regions (Schlebusch et al., 2012; Chan et al., 2015). Admixture between genetically isolated and differentiated populations can result in a mosaic of population-specific chromosomal blocks (Daya et al., 2014; Sanderson et al., 2015; Chimusa et al., 2018).

PATTERNS OF ADMIXTURE IN SOUTHERN AFRICA

Bantu groups reached Southern Africa 1,200–2,000 years ago (Ramsay et al., 1996; Soodyall, 2006; Barbieri et al., 2014) and reached the southeastern Cape 1,300 years ago (Breton et al., 2014). Arriving in Southern Africa, Bantu groups did not replace indigenous populations entirely; instead, there have been assimilations of the indigenous populations into the Bantu groups through intermixing and intermarriages especially between migrant men and Khoe-San women (Nurse et al., 1985; Diamond and Bellwood, 2003; Barbieri et al., 2014; Li et al., 2014).

L0d and L0k mtDNA haplogroups have been previously shown to be localized in the Khoe-San (Behar et al., 2008; Barbieri et al., 2013b; Schlebusch et al., 2013). Barbieri et al. (2014) examined mtDNA haplogroups L0d and L0k to measure admixture between the Khoe-San and Bantu speakers. The presence of these mtDNA lineages in Bantu-speaking populations may suggest very ancient admixture with the Khoe-San (Barbieri et al., 2013b). The authors showed that these haplogroups were shared with the Khoe-San, with a group such as the Kgalagadi of southern Botswana harboring 53% frequency of the haplogroups (Barbieri et al., 2014). A considerable amount of admixture was also observed between the Bantu speakers of South Africa (Zulu and Xhosa populations) through elucidation of Y-chromosome haplogroups (Naidoo et al., 2010). Consistent with this finding, Gurdasani et al. (2014) found hunter-gatherers admixture across African populations, the greatest level of admixture being in Zulu and Xhosa populations.

Y-chromosome and mtDNA analyses of Southern African populations suggest patterns of sex-biased migration and admixture (Underhill and Kivisild, 2007). The Bantu expansion was possibly a male-dominated event which led to Khoe-San females being assimilated into Bantu-speaking populations (Bajic et al., 2018). Henn et al. found that approximately 35% of paternal lineages in the Khoe-San group were either of Bantu or European origin (Henn et al., 2011). This sex-biased interaction between migrant men and Khoe-San females has led to asymmetrical patterns of admixture in the region (**Table 1**). A common pattern is the flow of Y-chromosome from migrant men into Khoe-San populations or the flow of mtDNA from Khoe-San women into migrant populations (Choudhury et al., 2017; Bajic et al., 2018; Oliveira et al., 2019).

Precolonial admixture of indigenous Southern African populations with populations of west Eurasian ancestry has also been documented (Pickrell et al., 2014). Due to the age of this admixture and suggestive linguistic, archeological, and genetic evidences of migration into Southern Africa from eastern Africa, Pickrell et al. (2014) concluded that the precolonial west Eurasian ancestry in Southern Africa is due to the indirect gene flow through eastern Africa. Signatures of admixture in Khoe-San groups have also been observed. In their study, Pickrell and colleagues detected ~6% non-Khoe-San ancestry in the Ju'|hoan North (Pickrell et al., 2012). Moreover, a genetic component common in Southeastern Bantu speakers was observed in most Khoe-San groups from Botswana, Lesotho, and South Africa (Montinaro et al., 2017).

The "Coloured" of South Africa (SAC) are another popular, well-studied population in Southern Africa (**Table 1**). The SAC show the highest level of intercontinental admixture, with high levels of Southern African Khoisan, Niger-Khodofanian, Indian, and European ancestries and low levels of Asian and Cushitic ancestries (Tishkoff et al., 2009). Consistent to this insight, a multi-faceted admixture scenario (five-way admixture) has been documented in a study that sought to develop a method (PROXYANC) that could detect the best proxy ancestry in admixture in the SAC population (Chimusa et al., 2013a). The

results of this study suggested the admixture to be a combination of Europeans (16%), Xhosa (33%), Gujarati (12%), Chinese (7%), and ‡Khomani (31%) ancestries.

Recent direct admixture involving Eurasian groups has been detected using GLOBETROTTER, an R program, (Hellenthal et al., 2014) in the ‡Khomani and Karretjie of South Africa, dating back to five generations (220 years ago) which aligns with the period of European settlers' arrival in Southern Africa (Busby et al., 2016). Congruent with this, in a recent study Chimusa et al. (2018) dated admixture in the SAC population to approximately 4 ± 1 generations ago, which likely reflects the wave of admixture that occurred during the era of colonialism in Southern Africa.

Admixture mapping is increasingly becoming an indispensable tool for predicting diseases in admixed populations. Admixture mapping is a method of measuring the contribution of ancestry to a phenotype (Smith et al., 2004; Chimusa et al., 2015; Chi et al., 2019). In Southern Africa, a GWAS of tuberculosis (TB) in the SAC population revealed a link between San ancestry and susceptibility to *Mycobacterium tuberculosis* (Chimusa et al., 2013b; Daya et al., 2014).

HISTORY OF HIV IN SOUTHERN AFRICA

AIDS was first recognized as a new disease in 1981 when multiple young homosexual men in Los Angeles and New York presented with opportunistic infections and a rare form of Kaposi's Sarcoma (Greene, 2007; Sharp and Hahn, 2011). HIV was later confirmed as the causative agent of AIDS (Gallo et al., 1984). There are two types of HIV: HIV-1 and HIV-2. HIV-1 is classified into three groups; Main group (M), Outlier group (O), and non-M-non-O group (N). Due to high mutation and recombination rates of HIV-1, the pandemic group, M, is further divided into genetic subtypes A–D, F–H, J–K, and intersubtype recombinants, as well as circulating recombinant forms (CRFs) and unique recombinant forms (URFs). Over half of global HIV infections are caused by HIV-1. HIV-1 group M has a worldwide distribution; subtype A is more prevalent in eastern Europe, subtype B in North America, South America, and western Europe, and subtype C in Southern Africa and India. HIV-2, which is more prevalent in western Africa, is divided into subtypes A–H and is less prevalent and less pathogenic than HIV-1 (Hemelaar et al., 2011; Eberle and Gurtler, 2012; Wilkinson et al., 2015).

Southern Africa has the highest HIV prevalence globally with countries such as eSwatini leading with 27.4% [**Figure 3**, (Taha, 2011; Vermund et al., 2015)]. The predominant subtype in the region is HIV-1C (Wilkinson et al., 2015). Using coalescence and molecular clock methods, HIV-1C was traced to the late 1930s in the Democratic Republic of Congo (Wilkinson et al., 2015). This strain is believed to have later spread to Southern and eastern Africa (Hemelaar et al., 2011; Faria et al., 2014; Wilkinson et al., 2015). Although the first samples of HIV-1 were sampled in South Africa in 1985, the origin of HIV-1C epidemic in Southern Africa was placed around 1960 [95% highest posterior density (HPD) 1956–1964] (Wilkinson et al., 2015).

The first anti-HIV drug, azidothymidine (AZT) also known as zidovudine, was approved in 1987. This was an anti-cancer drug that was also found to inhibit reverse transcriptase (Greene,

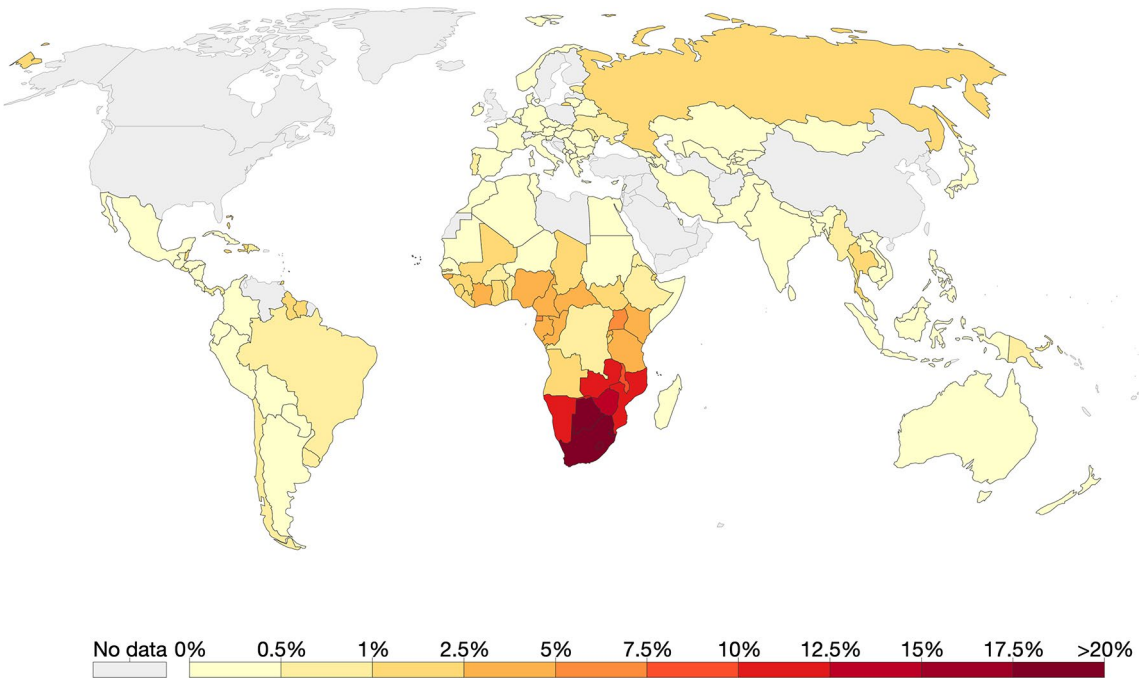


FIGURE 3 | The global prevalence of HIV as of 2017. Of all the countries that had data, Southern Africa had the highest HIV prevalence. Source: Our World in Data; data provided by UNAIDS.

2007). AZT and subsequent monotherapies were short-lived due to limitations including low genetic barrier to resistance and toxicity. This led to the introduction of combination therapy of multiple antiretrovirals [cART or highly active antiretroviral therapy (HAART)] with better tolerability and potency (Arts and Hazuda, 2012; Vella et al., 2012). In the early 1990s, there was high morbidity and mortality as a result of HIV in most Southern African countries (Kagaayi and Serwadda, 2016). The region has since seen a steep decline of 30% in HIV-related mortality between 2010 and 2017 due to rapid scale-up of antiretroviral therapy (UNAIDS, 2018). However, these drugs only suppress HIV to undetectable levels; they do not eliminate HIV from cellular reservoirs (McLaren and Carrington, 2015). While current therapeutic measures are geared towards elimination of HIV in cellular reservoirs and the use of broadly neutralizing antibodies against HIV, effective vaccination against HIV remains a hurdle due to ineffective immune responses against HIV and high variability of HIV sequence (Halper-Stromberg and Nussenzweig, 2016; Siliciano and Siliciano, 2016; Hsu and O'Connell, 2017).

Southern African countries concomitantly bear the brunt of both HIV and TB (Corbett et al., 2003; John et al., 2007; Abdool Karim et al., 2009; Dokubo et al., 2014; Tafuma et al., 2014). In 2017 an estimate of 10 million people developed active TB, and 72% of these were HIV-infected people living in Africa (WHO, 2018). HIV is the greatest predictor of development of active TB; susceptibility to active TB development increases 20-fold in people infected with HIV (Pawlowski et al., 2012). HIV/TB coinfection is the leading cause of mortality in Southern Africa even in people receiving antitubercular and antiretroviral therapy (Bisson et al., 2015; Campa et al., 2017). In a HIV/TB coinfection scenario, TB has a negative impact on immune response to HIV, thus exacerbating progression to AIDS (Bruchfeld et al., 2015). Another complication associated with HIV infection is the manifestation of adverse restoration of immune responses, known as immune reconstitution inflammatory syndrome (IRIS), which occurs during the initial months of antiretroviral treatment (Lawn et al., 2005). Although IRIS unmasks a spectrum of pre-existing opportunistic infections, it is commonly associated with mycobacterial infections (Lawn et al., 2005; Meintjes et al., 2008). TB-IRIS is a major health challenge in the developing world (Bruchfeld et al., 2015; Walker et al., 2018).

PRE-GWAS OF HIV-1: CANDIDATE AIDS RESTRICTION GENES

One of the striking characteristics of HIV infection is the highly heterogeneous viral set point across infected individuals. Set point is the viral load of an individual that fluctuates around a steady point during the asymptomatic phase of HIV infection (Fraser et al., 2007). The difference in viral set point between individuals can be as high as 1,000-fold (Mellors et al., 1996; Fellay et al., 2007; Fraser et al., 2007). Some of this variability is underpinned by host genetic factors. Following the discovery that HIV infected cells by binding to CD4 cells and CCR5, the first AIDS restriction gene (ARG) to be discovered was *CCR5*-Δ32 (Dean et al., 1996). Subsequently, a number of SNPs associated with HIV-1 were identified by candidate gene analysis in other ARGs (O'Brien and Nelson, 2004; Hutcheson et al., 2008; Winkler, 2008). These ARGs include those encoding HLA class I antigen presentation molecules (Carrington et al., 1999; Gao et al., 2001; Carrington and O'Brien, 2003; Welzel et al., 2007); chemokines and chemokine receptor genes involved in HIV-1 cell entry such as CCR2, CXCR6, and CCL5 (RANTES) (Smith et al., 1997; An et al., 2002; Duggal et al., 2003); natural killer cell immunoglobin-like receptors (KIR) genes (Martin et al., 2002); cytokines (Shin et al., 2000); and the intrinsic viral restriction factors TRIM5 (Javanbakht et al., 2006) and APOBEC3G (An et al., 2004). A detailed discussion of these SNPs and harboring genes can be found in these reviews (O'Brien and Nelson, 2004; Hutcheson et al., 2008).

Candidate gene methods have been pivotal in revealing genes that may be used as targets for drug development (van Manen et al., 2011). For instance, *CCR5* variants have enabled the development of HIV-1 entry inhibitor drugs (Troyer et al., 2011). Maraviroc, which is a CCR5 antagonist, is currently the only chemokine receptor in clinical use (Henrich and Kuritzkes, 2013). Through the identification of *CCR5*-Δ32, to date two people have maintained complete HIV-1 remission after a *CCR5*-Δ32 stem transplant (Hutter et al., 2009; UNAIDS, 2019). The ARGs identified by candidate gene analysis explained about 10% of the variability in HIV-1 infection (O'Brien and Nelson, 2004; Hutcheson et al., 2008). HapMap annotation of about 3 million human SNPs has facilitated the development of high-density arrays with 500,000 to 1,000,000 variants that could be utilized in the screening of variants associated with any disease (The International HapMap Consortium, 2003; Hutcheson et al., 2008; Winkler, 2008); this set the stage for HIV-1 GWAS.

A GLOBAL PERSPECTIVE OF GWAS OF HIV-1

GWAS are useful tools for identifying genetic contributions towards predisposition to infection, progression to disease, and differential drug metabolism (Adebamowo et al., 2017). GWAS offer an unbiased opportunity to screen the entire genome and pinpoint variants implicated in a given phenotype of interest. Here and in the subsequent section, we present an account of GWAS of HIV-1 as illustrated in **Table 2**.

The first HIV GWAS report unraveled genome-wide contributions of host genetic markers in the control of HIV infection in Europeans and Caucasians (Fellay et al., 2007). Consistent with previous candidate gene studies, in the Fellay et al. (2007) study, the major genetic determinants of HIV set point were in *HLA B* and *C* loci. These findings were replicated in subsequent GWAS of HIV-1 in European/Caucasian populations, with additional loci being identified (Dalmasso et al., 2008; Fellay et al., 2009; Le Clerc et al., 2009; Limou et al., 2009; Herbeck et al., 2010; Pereyra et al., 2010; Le Clerc et al., 2011). In most studies, the HIV-1 phenotypes of interest were viral set point and progression to disease (Fellay et al., 2007; Dalmasso et al., 2008; Fellay et al., 2009; Le Clerc et al., 2009;

TABLE 2 | GWAS significant genes associated with HIV-1 acquisition, viral load set point, and progression.

Known gene	Description	Potential protein function	Effect	Population	Reference
HCP5	HLA complex P5	Related in sequence to human endogenous retroviruses and possibly interacts directly with HIV.	Minor alleles of SNPs rs2395029*, rs2255221, and rs2523608** were associated with low viral loads and delayed progression	Europeans, African Americans, and Africans	(Fellay et al., 2007; Dalmasso et al., 2008; Fellay et al., 2009; Limou et al., 2009; Pereyra et al., 2010; Le Clerc et al., 2011)
HLA-B	MHC class I B	Plays a critical role in the immune system; peptide presentation *via* endoplasmic reticulum (ER) pathway.	SNPs found in the gene were associated with low viral loads and delayed progression.	Europeans, Chinese, African Americans, and Africans	(Fellay et al., 2009; Herbeck et al., 2010; Pelak et al., 2010; Pereyra et al., 2010)
HLA-C	MHC class I C	Plays a critical role in the immune system; peptide presentation *via* endoplasmic reticulum (ER) pathway.	Minor allele of rs9264942 was associated with low viral loads and delayed progression.	Europeans	(Fellay et al., 2007; Fellay et al., 2009; Pereyra et al., 2010)
ZNRD1	Zinc ribbon domain containing 1	Plays a role in regulation of cell proliferation.	SNPs identified in this gene were associated with low viral loads and delayed progression.	Europeans	(Fellay et al., 2007, Fellay et al., 2009)
TNXB	Tenascin XB	Plays a role in matrix maturation during wound healing.	SNPs identified in this gene were associated with low viral loads.	Europeans	(Dalmasso et al., 2008)
TNF	Tumor necrosis factor	Cytokine. Involved in the regulation of biological processes including cell proliferation, differentiation, apoptosis, lipid metabolism, and coagulation.	Minor allele of SNP rs3093662 was associated with low viral loads.	Europeans	(Dalmasso et al., 2008)
SDC2	Syndecan 2	Mediates cell binding, cell signaling, and cytoskeletal organization. Syndecan receptors are required for internalization of the HIV-1 tat protein.	Minor allele of SNP rs2575735 was associated with reduction in HIV reservoir.	Europeans	(Dalmasso et al., 2008)
DDX40YPEL2	DEAH-Box Helicase 40 Yippee Like 2	NR	Minor allele of intergenic rs6503919 was associated with low viral reservoir.	Europeans	(Dalmasso et al., 2008)
TRIM10	Tripartite motif containing 10	Plays a role in terminal differentiation of erythroid cells.	SNP rs9468692 was associated with low viral loads.	Europeans	(Fellay et al., 2009)
NOTCH4	Notch receptor 4	Regulates interactions between physically adjacent cells.	Minor allele of SNP rs8192591 was associated with low viral loads and delayed progression.	Europeans	(Fellay et al., 2009; Le Clerc et al., 2011)
RNF39	Ring finger protein 39	Plays a role in early phase of synaptic plasticity.	SNPs found in the gene were associated with low viral loads and delayed progression.	Europeans	(Fellay et al., 2009; Limou et al., 2009)
C6orf48	Small nucleolar RNA host gene 32	NR	Minor allele of SNP rs9368699 was associated with low viral loads and delayed progression.	Europeans	(Limou et al., 2009; Le Clerc et al., 2011)
PSORS1C1	Psoriasis susceptibility 1 candidate 1	Confers susceptibility to psoriasis and systemic sclerosis.	SNPs found in the gene were associated with low viral loads.	Europeans	(Limou et al., 2009)
MICB	MHC class I polypeptide-related sequence B	Involved in immune response. Activates natural killer cells, CD8 alphabeta T cells, and gammadelta T cells.	SNPs found in the gene were associated with low viral loads.	Europeans	(Limou et al., 2009; Herbeck et al., 2010)
SOX5	Sex determining region Y -box 5	Regulation of embryonic development and determination of the cell fate.	Minor allele of SNP rs1522232 was associated with delayed progression.	Europeans	(Le Clerc et al., 2009)
RXRG	Retinoid X receptor gamma	Mediates the antiproliferative effects of retinoic acid.	Minor allele of SNP rs10800098 was associated with rapid progression.	Europeans	(Le Clerc et al., 2009)
TGFBRAP1	Transforming growth factor beta receptor associated protein 1	Acts as a chaperone in signaling downstream of TGF-beta.	Minor allele of SNP rs1020064 was associated with delayed progression.	Europeans	(Le Clerc et al., 2009)
PROX1	Prospero homeobox 1	Plays a role in development.	A haplotype of rs17762192, rs17762150, and rs1367951 was associated with delayed progression.	Europeans	(Herbeck et al., 2010)

(Continued)

TABLE 2 | Continued

Known gene	Description	Potential protein function	Effect	Population	Reference
MICA	MHC class I polypeptide-related sequence A	A ligand for the NKG2-D type II receptor that acts as a stress-induced antigen that is broadly recognized by intestinal epithelial gamma delta T cells.	Minor allele of SNP rs4418214 was associated with low viral loads.	Europeans	(Pereyra et al., 2010)
PSORS1C3	Psoriasis susceptibility 1 candidate 3	NR	Minor allele of SNP rs3131018 was associated with low viral loads.	Europeans	(Pereyra et al., 2010)
AL671883.2 DHFRP2	Pseudogenes	NR	Minor allele of intergenic SNP rs2523590 was associated with low viral loads.	Europeans, African Americans, Hispanics	(Pereyra et al., 2010)
HCG22	HLA complex group 22	NR	Minor allele of SNP rs9262632 was associated with low viral loads in African Americans. Minor allele of SNP rs2535307 was associated with rapid progression and increased susceptibility.	African Americans and Southern Africans	(Pereyra et al., 2010; Xie et al., 2017)
RICH2	Rho GTPase activating protein 44	Involved in GTPase activation activity and phospholipid binding.	Major allele of SNP rs2072255 was associated with rapid progression.	Europeans	(Le Clerc et al., 2011)
PARD3B	Par-3 family cell polarity regulator beta	May play a role in asymmetrical cell division and cell polarization processes.	SNP rs11884476 was associated with delayed progression.	Europeans	(Troyer et al., 2011)
CCR5	C-C motif chemokine receptor 5	Acts as a co-receptor for macrophage-tropic viruses to enter host cells.	CCR5-Δ32 was associated with resistance to HIV-1 infection.	Europeans	(McLaren et al., 2013)
CCNG1	Cyclin G1	May play a role in cell proliferation and regulation	SNP kgp22385164 was associated with rapid progression.	Southern Africans	(Xie et al., 2017)

*Potential function of genes according to Pubmed and GeneCards. MHC, major histocompatibility complex; SNP, single nucleotide polymorphism; *, SNP tags HLA-B57*01 in Europeans; **, SNP tags HLA-B57*03 in Africans; NR, no records.*

Limou et al., 2009; Herbeck et al., 2010; Pereyra et al., 2010; Le Clerc et al., 2011; Troyer et al., 2011; van Manen et al., 2011; Bartha et al., 2013).

The effects of the following SNPs were consistently replicated in European samples: *HCP5* rs2395029, *HLA-C rs9264942, ZNRD1 rs9261174, NOTCH4 rs8192591, and C6orf48 rs9368699.* The most replicated SNP was *HCP5* rs2395029 (which tags *HLA-B**5701). The minor allele G of *HCP5* rs2395029 led to a reduction in viral load and delayed progression (Fellay et al., 2007; Dalmasso et al., 2008; Fellay et al., 2009; Limou et al., 2009; Pereyra et al., 2010; Le Clerc et al., 2011; Bartha et al., 2013). All these SNPs are located within the HLA region in chromosome 6. The HLA region is attributed by a long pattern of LD which makes it challenging to pinpoint the causal variant within this region. It has been observed that the minor allele of *HCP5* rs2395029 is often observed with the controlling C allele *HLA-C*rs9264942, both leading to a protective effect (Fellay et al., 2007; Fellay et al., 2009). At gene level, besides *HCP5*, *HLA-B*, and *HLA-C*, the following were also overrepresented: *TNXB*, *ZNRD1*, *RNF39*, and *MICB*. The role of these genes was delayed progression [*ZNRD1*, *RNF39* (Fellay et al., 2007; Fellay et al., 2009; Limou et al., 2009)] and reduction in viral load [*TNXB* and *MICB* (Dalmasso et al., 2008; Limou et al., 2009; Herbeck et al., 2010)]. Collectively, discovered SNPs within the European populations explained not more than 20% of the variability in viral load. More details on the identified SNPs can be found in the original publications and **Table 2**.

One of the major caveats of GWAS is the stringent threshold (p-value = 5.0×10^{-8}) which accounts for multiple testing of about one million SNPs. As a result of this, SNPs with small effect sizes are not detected by GWAS and thousands of samples are required to effectively perform GWAS. In a study of progression in 404 Europeans, no SNP reached genome-wide significance (van Manen et al., 2011). The first GWAS of HIV-1 acquisition in European populations detected 11 SNPs which were genome-wide significant. Despite the large sample size, over 6,000 participants in each case/control category, these effects were disqualified due to frailty bias. Nonetheless, the study revealed a genome-wide association significance between imputed *CCR5*-Δ32 and HIV-1 acquisition [**Supplementary Table 1** (McLaren et al., 2013)]. The outcomes of these studies also show that the defining and selection of GWAS phenotypes should be handled with care (van Manen et al., 2011; McLaren et al., 2013).

GWAS of HIV-1 have also been performed in other populations, African Americans, Hispanics, Chinese, and Africans (Pelak et al., 2010; Pereyra et al., 2010; Petrovski et al., 2011; Wei et al., 2015; Xie et al., 2017). Pereyra et al. identified an intergenic SNP rs2523590 in Hispanics, Europeans, and Americans. The minor allele of this SNP was associated with

low viral loads (Pereyra et al., 2010). Two GWAS of African Americans identified the SNP rs2523608 which tags the *HLA-B**5703 allele in people of African descent. Having the minor allele of the SNP leads to a reduction in viral load (Pelak et al., 2010; Pereyra et al., 2010). When considered alone, *HLA-B**5703 had the strongest signal and accounted for about 10% of the variation in viral load set point. Participants who had the *HLA-B**5703 allele (absent in Europeans) showed improved viral control of the same magnitude with that observed in Europeans carrying the *HLA-B**5701 allele (Pelak et al., 2010). The *HLA-B**5701 allele was present in African Americans because of admixture at 0.3% and had little contribution to HIV-1 control. The implications of different alleles of *HLA-B**57 in two populations of different ancestries emphasizes the importance of investigating rare variants that might affect HIV infection distinctly in different populations.

The first and only HIV-1 GWAS in the Chinese population was reported in 2015 (Wei et al., 2015). Several novel SNPs that correlated with HIV-1 viral load set point were identified, albeit not genome-wide significant. The top signals such as rs2442719 were observed in the HLA region. The authors highlighted that they observed no significant association for the highest peaks identified in Caucasians and African Americans (*HCP5* rs2395029 and *HLA-B* rs2523608 respectively), this indicating ethnic specificity of the variant associations [**Supplementary Table 1** (Wei et al., 2015)].

The *CCR5*-Δ32 variant that is associated with resistance to HIV infection in people of European ancestry is rare in African populations (Fowke et al., 1996; Joubert et al., 2010; Lingappa et al., 2011; Petrovski et al., 2011). Since *CCR5* variants were the only confirmed genetic markers that could influence HIV-1 acquisition in people of non-African ancestry, few studies were conducted to discover host genetics underlying HIV-1 acquisition in African populations (Kenya, Uganda and Tanzania, South Africa and Botswana (Lingappa et al., 2011), and Malawi (Petrovski et al., 2011). Lingappa et al. additionally investigated host genetics of viral control. These studies failed to detect signals of association between genetic variants and susceptibility to HIV-1 nor viral control. The authors postulated several arguments to explain the failure to detect signals: (1) susceptibility to HIV-1 might be due to non-genetic factors such as mode of transmission and viral sequence variability, (2) common or rare variants in an African populations might have not been represented in the chip used for genotyping, or (3) these weak signals may be due to the authors' insufficient sample size (Lingappa et al., 2011; Petrovski et al., 2011).

The most recent HIV-1 GWAS was also conducted in an African (Botswana) population (Xie et al., 2017). This is the first GWAS to reveal significant variants associated with HIV-1 acquisition and progression in an African population. Moreover, two SNPs which had never been reported to associate with HIV infection nor progression were identified. These were SNPs rs2535307 and kgp22385164 located near *HCG22* (immune regulatory gene within chromosome 6) and *CCNG1* (encodes a cyclin that controls cell cycle; within chromosome 5) genes, respectively. The minor alleles of both SNPs led to lower *CD4* T-cell counts and high viral loads. *HCG22* was associated with progression and acquisition of HIV1-C, while *CCNG1* was associated with progression only in the Botswana cohort [**Table 2** (Xie et al., 2017)]. A validation test was done in data from three independent American cohorts, and no significant association was observed around the two genes (*HCG22*and *CCNG1*). SNPs that were found to associate with HIV infection and progression in previous studies did not appear in the top 100,000 p-values in this study (Xie et al., 2017).

Most of the SNPs that were identified in GWAS of HIV-1 are located within genes that play a role in immune response (**Table 2** and **Figure 4**). We present in **Figure 4**, a pathway interaction network of host candidate genes from GWAS of HIV-1 (**Table 2**) using GeneMANIA software (Warde-Farley et al., 2010). Out of the 28 genes in **Table 2**, GeneMANIA was able to predict the functions of 23 genes. Other genes that GeneMANIA deemed biologically and/or transcriptionally similar to the HIV-1 candidate genes from GWAS are also added to the network. According to databases accessed by GeneMANIA, the interactive genes are co-expressed with identified genes from GWAS of HIV-1 (**Table 2**). This suggests that more robust GWAS methods may pinpoint novel HIV-associated variants within genes from this broad gene-based network.

GWAS OF HIV-1 IN SOUTHERN AFRICA

Only three GWAS of adult African populations have been published to date (Lingappa et al., 2011; Petrovski et al., 2011; Xie et al., 2017). These studies investigated common variants in Southern African populations of Malawi, Botswana, and South Africa. No variant reached genome-wide significance in two of these studies (Lingappa et al., 2011; Petrovski et al., 2011), while only two SNPs that had never been associated with HIV-1 acquisition and progression before were reported in the Botswana cohort (Xie et al., 2017). This lack of significant GWAS results is likely due to (1) lack of power from both sample size and statistical approaches used in both studies, (2) that the variants on microarrays could not tag the causal variants due to different LD patterns and different haplotype block structures between populations, or (3) that populations of Southern Africa use different mechanisms to control HIV-1C as a result of genetic diversity compared to previously studied populations.

CONCLUSIONS AND PERSPECTIVES

Southern Africa has a unique genetic architecture with evident structure between the populations within the region. The haplotype data of Khoe-San and Southern Bantu ethnolinguistic groups are currently absent from existing haplotype reference panels such as 1000 Genomes Project. Studies of population genetics and ancient DNA in Southern Africa are likely to provide new opportunities to discover novel disease susceptibility loci and refine gene-disease association signals. Southern Africans harbor ancient genetic diversity, as well as historical admixture, which

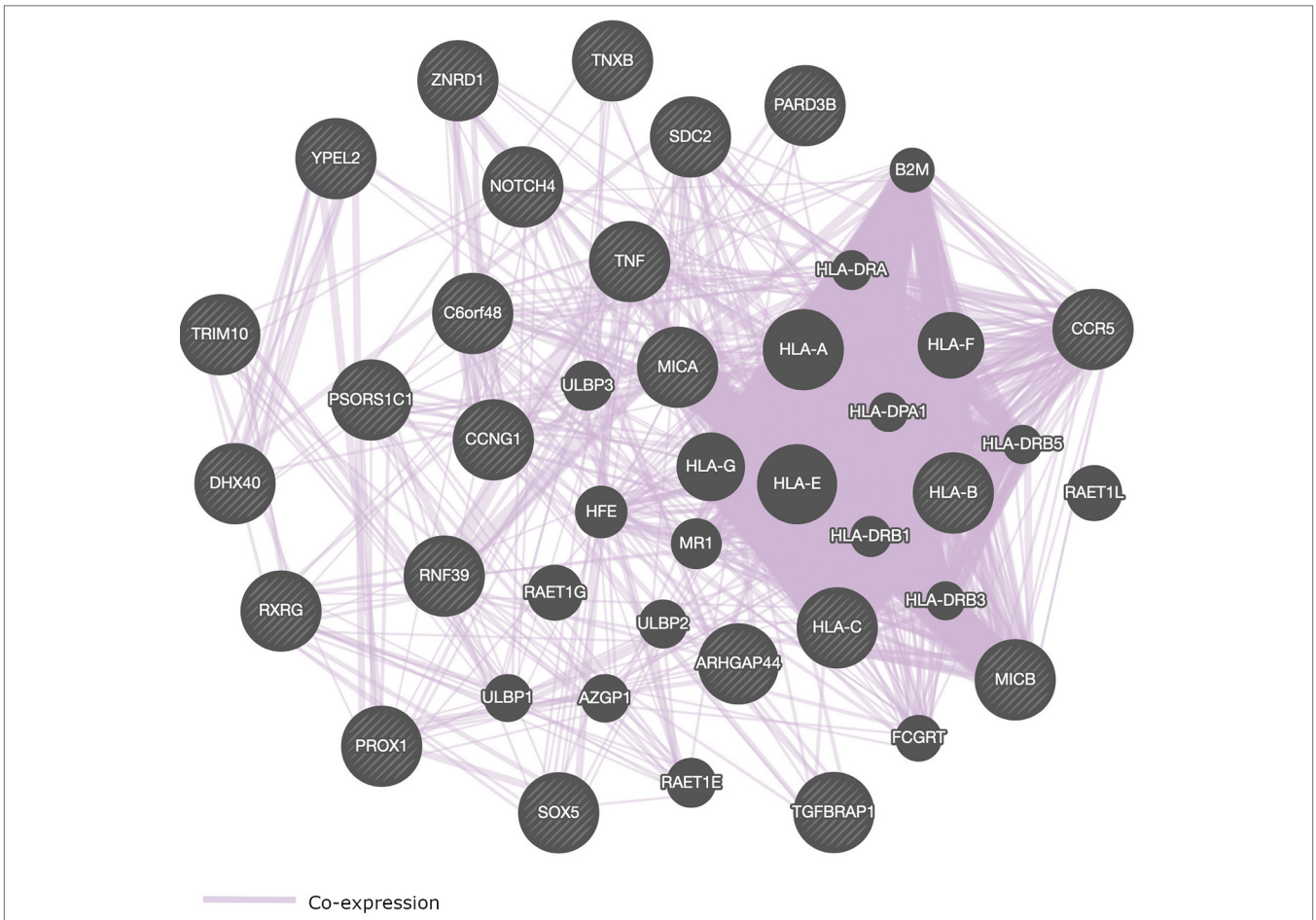


FIGURE 4 | Pathway interaction network of genome-wide significant genes that control HIV-1 phenotypes. SNPs within or close to these genes are associated with the control of HIV-1. Striped circles indicate genes which have been identified from GWAS of HIV-1. Circles without stripes indicate genes putatively involved in the same pathway with genes implicated in HIV-1 phenotypes *via* GWAS. Purple lines represent gene co-expressions. The following genes could not be recognized by GeneMANIA software: *HCP5*, *PSORS1C3*, *AL671883.2*, *DHFRP2*, and *HCG22*.

can lead to complexities in (a) the design of studies assessing the genetic determinants of diseases and human variation, and (b) approaches for reconstructing DNA segments (sequence alignment) and variant discovery.

This review provided a broad discussion of genetic diversity in Southern Africa and its implication in the genetic architecture of HIV-1 phenotypes. In addition, we posited the advances made in HIV-1 GWAS. Few GWAS of HIV have been conducted using adult populations from Southern Africa. This is not enough considering that the region has the highest social and economic burden of HIV. Despite the insufficient sample sizes, HIV-1 GWA studies have reaffirmed the role of HLA in controlling HIV infection. *HLA-B*, *HLA-C*, *ZNRD1*, *RFN39*, and *HCP5* genes continue to be the most influential players in the outcome of HIV infection. However, a lot of implicated genes in GWAS reached significance by virtue of LD between the putative genes and the discovered variants. It is worthy to note that current findings from HIV-1 GWAS have not yet had a major impact on therapeutic optimization.

Critically, there has been a considerable amount of failure to replicate HIV-1 GWAS results across studies with some of the reasons being small sample sizes, different LD patterns between populations (causal SNPs probably not in the same haplotype block as tag-SNPs), and different criteria of phenotypic stratification. Nonetheless, many studies revealed novel variants which were not reported in previous studies. Two novel SNPs associated with HIV progression and acquisition were discovered in a Southern African population. This might imply that different networks of genes control HIV-1 in different regions. Given this, we hypothesize that functional categories of the genome may contribute disproportionately to the predisposition and resistance to HIV-1.

High genetic diversity, multi-wave genetic mixture, and low LD in the case of Southern African populations constitute a challenge in identifying genetic variants with modest risk or protective effect against HIV-1. Current GWAS of HIV-1 in Southern Africa were performed mostly in homogeneous populations; Malawi GWAS had a total of 1,532

Chichewa-speaking individuals, while the Botswana GWAS had a total of 556 individuals of undisclosed ethnicity. The Lingappa et al. study had 798 in total including 191participants of undisclosed ethnicity from Southern Africa. These sample sizes are not well powered to detect genome-wide significant SNPs of intermediate and small effect. Furthermore, considering the high genetic diversity and diverse patterns of admixture within Southern Africa, the populations used in the two GWAS of HIV-1 do not represent the diversity in the region. Thus, the findings of these studies cannot be effectively applied in other populations within the region. Hence, we need an inclusion of other populations in GWAS of HIV-1 in Southern Africa. More studies of admixture mapping should also be done to investigate the genetic role of different ancestries in HIV-1 phenotypes.

Moreover, factors such as host microbiome, viral genomics, epigenetics, social factors, and environmental factors can further contribute to HIV-1 acquisition, transmission, and progression. However, these components are not accounted for in current developed GWAS approaches that are mostly tailored for populations with long range of LD and haplotypes of European populations. These current GWAS approaches may limit the power of detecting possible associated variants and the ability to predict HIV-1 predisposition/resistance in the African context, in particular, Southern Africa. Therefore, failure to leverage these characteristics in modeling their joint contributions will be an obstacle to fully understand HIV-1 phenotype variability for an efficient and effective personalized therapy.

To date, GWAS of HIV-1 used Eurocentric microarrays for genotyping. Like other complex traits, HIV-1 phenotypes are polygenic; these traits are influenced by many rare variants of small effect sizes that microarray-based GWAS may fail to detect. However, other unbiased methods such as whole-genome or whole-exome sequences have the power to identify not only rare genetic variants but can also uncover more novel variants. Additionally, GWAS of Southern Africa may use SNP arrays that have a better coverage and representation of African genotypes and LD structures such as the recently developed Human Heredity and Health in Africa Consortium (H3Africa) SNP array. Other cost-effective approaches would be to develop or optimize targeted sequencing approaches such as the recently developed single primer enrichment technology. The advantage of target enrichment approaches is that they offer scalability and avoid the ascertainment bias that is common with microarray genotyping.

Exploring information from GWAS summary statistics provides a new paradigm to GWAS and might enable a complete characterization of genetic susceptibility/resistance to a disease. To account for the missing heritability of HIV-1, we suggest to further (a) examine whole genome or whole exome sequences of Southern African populations to uncover population-specific variants, (b) investigate functional categories of the genome and cell type-specific elements to estimate their enrichment and polygenic contribution to heritability of HIV-1, (c) perform a meta-analysis of previous HIV-1 GWAS results of African populations, (d) predict HIV-1 polygenic risk score, and (e) use robust methods for aggregation of GWAS signals to augment the identification of implicated genes and pathways or sub-networks. Importantly, it is worthy to note that HIV-1 polygenic risk scores based on European GWAS results may likely be poor predictors in Southern African populations because of differences in haplotype structure, patterns of LD (i.e., LD-tagging), and population-specific variation.

Changes in the host gene expression (or regulation) are a crucial stage in biological mechanisms underlying resistance and predisposition to HIV-1, and yet our current understanding of gene regulation is still limited. Without deep understanding of host gene regulation and the paucity in available tools for examining regulation, the transition from GWAS results to biological insights (biological mechanisms, genes involved, and the direction of causality) will remain a challenge.

Studies of different populations (including sub-Saharan Africa) have revealed that women with diverse vaginal microbiota have an increased risk of acquiring and transmitting HIV. However, these findings were not correlated with host genetics. Most studies investigating HIV and the gut microbiome were focused in non-African populations. It is known that the gut microbiome of populations differ geographically; therefore, it will be challenging to translate data from developed countries to populations in developing countries. Given the influences of (1) host genetics on the microbiome (gut, oral, genital, rectal, and lung), (2) environmental factors on microbiome profile, (3) HIV-1 infection on the microbiome, and (4) microbiome on host phenotypes, it is apparent that integrating the microbiome in host GWAS will reveal important insights and launch the first steps towards establishing causality in HIV-1 GWAS. The integrative analysis will be critical to comprehend the role and mechanisms of host genetics, the microbiome, and environment in the manifestation of HIV-1 infections.

AUTHOR CONTRIBUTIONS

PT and EC conceived, structured, and wrote the content of the manuscript.

FUNDING

This work was supported through the sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant # 107752/Z/15/Z] and the UK government. The views expressed in this publication are those of the authors and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, or the UK government. The authors would also like to thank the National Research

Foundation of South Africa for funding (NRF) [grant # RA171111285157/119056].

ACKNOWLEDGEMENT

We would like to thank Simani Gaseitsiwe, PhD; Vlad Novitsky, PhD; Melvin Leteane, PhD for their support and guidance; Denis Awany for the useful discussions and The Centre for High-Performance Computing (CHPC, www.chpc.ac.za) for computing resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00905/full#supplementary-material

REFERENCES

Abdool Karim, S. S., Churchyard, G. J., Karim, Q. A., and Lawn, S. D. (2009). HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. *Lancet (London, England)* 374, 921–933. doi: 10.1016/S0140-6736(09)60916-8

Adebamowo, S. N., Tekola-Ayele, F., Adeyemo, A. A., and Rotimi, C. N. (2017). Genomics of cardiometabolic disorders in sub-Saharan Africa. *Public Health Genomics* 20, 9–26. doi: 10.1159/000468535

Adhikari, M. (2010). A total extinction confidently hoped for: the destruction of Cape San society under Dutch colonial rule, 1700–1795. *J. Genocide Res.* 12, 19–44. doi: 10.1080/14623528.2010.508274

An, P., Bleiber, G., Duggal, P., Nelson, G., May, M., Mangeat, B., et al. (2004). APOBEC3G genetic variants and their influence on the progression to AIDS. *J. Virol.* 78, 11070–11076. doi: 10.1128/JVI.78.20.11070-11076.2004

An, P., Nelson, G. W., Wang, L., Donfield, S., Goedert, J. J., Phair, J., et al. (2002). Modulating influence on HIV/AIDS by interacting RANTES gene variants. *Proc. Natl. Acad. Sci. U. S. A.* 99, 10002–10007. doi: 10.1073/pnas.142313799

Arts, E. J., and Hazuda, D. J. (2012). HIV-1 antiretroviral drug therapy. *Cold Spring Harb. Perspect. Med.* 2, a007161. doi: 10.1101/cshperspect.a007161

Awany, D., Allali, I., Dalvie, S., Hemmings, S., Mwaikono, K. S., Thomford, N. E., et al. (2018). Host and microbiome genome-wide association studies: current state and challenges. *Front. Genet.* 9, 637. doi: 10.3389/fgene.2018.00637

Bajic, V., Barbieri, C., Hubner, A., Guldemann, T., Naumann, C., Gerlach, L., et al. (2018). Genetic structure and sex-biased gene flow in the history of Southern African populations. *Am. J. Phys. Anthropol.* 167, 656–671. doi: 10.1002/ajpa.23694

Barbieri, C., Butthof, A., Bostoen, K., and Pakendorf, B. (2013a). Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. *Eur. J. Hum. Genet.* 21, 430–436. doi: 10.1038/ejhg.2012.192

Barbieri, C., Vicente, M., Oliveira, S., Bostoen, K., Rocha, J., Stoneking, M., et al. (2014). Migration and interaction in a contact zone: mtDNA variation among Bantu-speakers in Southern Africa. *PLoS One* 9, e99117. doi: 10.1371/journal.pone.0099117

Barbieri, C., Vicente, M., Rocha, J., Mpoloka, S. W., Stoneking, M., and Pakendorf, B. (2013b). Ancient substructure in early mtDNA lineages of Southern Africa. *Am. J. Hum. Genet.* 92, 285–292. doi: 10.1016/j.ajhg.2012.12.010

Barnard, A. (1992). *Hunters and herders of Southern Africa: a comparative ethnography of the Khoisan peoples*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9781139166508

Bartha, I., Carlson, J. M., Brumme, C. J., McLaren, P. J., Brumme, Z. L., John, M., et al. (2013). A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* 2, e01123. doi: 10.7554/eLife.01123

Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., et al. (2008). The dawn of human matrilineal diversity. *Am. J. Hum. Genet.* 82, 1130–1140. doi: 10.1016/j.ajhg.2008.04.002

Bisson, G. P., Zetola, N., and Collman, R. G. (2015). Persistent high mortality in advanced HIV/TB despite appropriate antiretroviral and antitubercular therapy: an emerging challenge. *Curr. HIV/AIDS Rep.* 12, 107–116. doi: 10.1007/s11904-015-0256-x

Botigue, L. R., Henn, B. M., Gravel, S., Maples, B. K., Gignoux, C. R., Corona, E., et al. (2013). Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. *Proc. Natl. Acad. Sci.* 110, 11791–11796. doi: 10.1073/pnas.1306223110

Breton, G., Schlebusch, C. M., Lombard, M., Sjodin, P., Soodyall, H., and Jakobsson, M. (2014). Lactase persistence alleles reveal partial East African ancestry of Southern African Khoe pastoralists. *Curr. Biol.* 24, 852–858. doi: 10.1016/j.cub.2014.02.041

Bruchfeld, J., Correia-Neves, M., and Källenius, G. (2015). Tuberculosis and HIV coinfection. *Cold Spring Harb. Perspect. Med.* 5, a017871. doi: 10.1101/cshperspect.a017871

Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., et al. (2016). Admixture into and within sub-Saharan Africa. *Elife* 5, e15266. doi: 10.7554/eLife.15266

Campa, A., Baum, M. K., Bussmann, H., Martinez, S. S., Farahani, M., van Widenfelt, E., et al. (2017). The effect of micronutrient supplementation on active TB incidence early in HIV infection in Botswana. *Nutr. Diet.* Suppl. 2017, 37–45. doi: 10.2147/NDS.S123545

Campbell, M. C., and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.* 9, 403–433. doi: 10.1146/annurev.genom.9.081307.164258

Campbell, M. C., and Tishkoff, S. A. (2010). The evolution of human genetic and phenotypic variation in Africa. *Curr. Biol.* 20, R166–R173. doi: 10.1016/j.cub.2009.11.050

Carr, D. F., Bourgeois, S., Chaponda, M., Takeshita, L. Y., Morris, A. P., Cornejo Castro, E. M., et al. (2017). Genome-wide association study of nevirapine hypersensitivity in a sub-Saharan African HIV-infected population. *J. Antimicrob. Chemother.* 72, 1152–1162. doi: 10.1093/jac/dkw545

Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., et al. (1999). HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* 283, 1748–1752. doi: 10.1126/science.283.5408.1748

Carrington, M., and O'Brien, S. J. (2003). The influence of HLA genotype on AIDS. *Annu. Rev. Med.* 54, 535–551. doi: 10.1146/annurev.med.54.101601.152346

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1993). Demic expansions and human evolution. *Science (80-.)* 259, 639–646. doi: 10.1126/science.8430313

Chan, E. K. F., Hardie, R.-A., Petersen, D. C., Beeson, K., Bornman, R. M. S., Smith, A. B., et al. (2015). Revised timeline and distribution of the earliest diverged human maternal lineages in Southern Africa. *PLoS One* 10, e0121223. doi: 10.1371/journal.pone.0121223

Chi, C., Shao, X., Rhead, B., Gonzales, E., Smith, J. B., Xiang, A. H., et al. (2019). Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. *PLoS Genet.* 15, 1–27. doi: 10.1371/journal.pgen.1007808

Chimusa, E. R., Daya, M., Moller, M., Ramesar, R., Henn, B. M., van Helden, P. D., et al. (2013a). Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PLoS One* 8, e73971. doi: 10.1371/journal.pone.0073971

Chimusa, E. R., Defo, J., Thami, P. K., Awany, D., Mulisa, D. D., Allali, I., et al. (2018). Dating admixture events is unsolved problem in multi-way admixed populations. *Brief. Bioinform.* doi: 10.1093/bib/bby112

Chimusa, E. R., Meintjies, A., Tchanga, M., Mulder, N., Seoighe, C., Soodyall, H., et al. (2015). A genomic portrait of haplotype diversity and signatures of selection in indigenous Southern African populations. *PLoS Genet.* 11, e1005052. doi: 10.1371/journal.pgen.1005052

Chimusa, E. R., Zaitlen, N., Daya, M., Moller, M., van Helden, P. D., Mulder, N. J., et al. (2013b). Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Hum. Mol. Genet.* 23, 796–809. doi: 10.1093/hmg/ddt462

Choudhury, A., Aron, S., Sengupta, D., Hazelhurst, S., and Ramsay, M. (2018). African genetic diversity provides novel insights into evolutionary history and local adaptations. *Hum. Mol. Genet.* 27, R209–R218. doi: 10.1093/hmg/ddy161

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., et al. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* 8, 1–12. doi: 10.1038/s41467-017-00663-9

Corbett, E. L., Watt, C. J., Walker, N., Maher, D., Williams, B. G., Raviglione, M. C., et al. (2003). The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch. Intern. Med.* 163, 1009–1021. doi: 10.1001/archinte.163.9.1009

Dalmasso, C., Carpentier, W., Meyer, L., Rouzioux, C., Goujard, C., Chaix, M.-L., et al. (2008). Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study. *PLoS One* 3, e3907. doi: 10.1371/journal.pone.0003907

Daya, M., van der Merwe, L., Gignoux, C. R., van Helden, P. D., Moller, M., and Hoal, E. G. (2014). Using multi-way admixture mapping to elucidate TB susceptibility in the South African Coloured population. *BMC Genomics* 15, 1021. doi: 10.1186/1471-2164-15-1021

de Filippo, C., Heyn, P., Barham, L., Stoneking, M., and Pakendorf, B. (2010). Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. *Am. J. Phys. Anthropol.* 141, 382–394. doi: 10.1002/ajpa.21155

de Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Moller, M., van Helden, P. D., et al. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Hum. Genet.* 128, 145–153. doi: 10.1007/s00439-010-0836-1

Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Smith, M. W., Allikmets, R., et al. (1996). Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE. *Science* 273, 1856–1862. doi: 10.1126/science.273.5283.1856

Diamond, J., and Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science (80-.)* 300, 597–603. doi: 10.1126/science.1078208

Dokubo, E. K., Baddeley, A., Pathmanathan, I., Coggin, W., Firth, J., Getahun, H., et al. (2014). Provision of antiretroviral therapy for HIV-positive TB patients—19 countries, sub-Saharan Africa, 2009–2013. *MMWR. Morb. Mortal. Wkly. Rep.* 63, 1104–1107.

Duggal, P., An, P., Beaty, T. H., Strathdee, S. A., Farzadegan, H., Markham, R. B., et al. (2003). Genetic influence of CXCR6 chemokine receptor alleles on PCP-mediated AIDS progression among African Americans. *Genes Immun.* 4, 245–250. doi: 10.1038/sj.gene.6363950

Eberhard, D. M., Simons, G. F., and Fennig, C. D. Ethnologue: Languages of the World. Ethnologue. Available at: https://www.ethnologue.com/region/saf [Accessed February 10, 2019].

Eberle, J., and Gurtler, L. (2012). HIV types, groups, subtypes and recombinant forms: errors in replication, selection pressure and quasispecies. *Intervirology* 55, 79–83. doi: 10.1159/000331993

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., et al. (2014). HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346, 56–61. doi: 10.1126/science.1256739

Fellay, J., Ge, D., Shianna, K. V., Colombo, S., Ledergerber, B., Cirulli, E. T., et al. (2009). Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.* 5, e1000791. doi: 10.1371/journal.pgen.1000791

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science (80-.)* 317, 944–947. doi: 10.1126/science.1143767

Fowke, K. R., Nagelkerke, N. J., Kimani, J., Simonsen, J. N., Anzala, A. O., Bwayo, J. J., et al. (1996). Resistance to HIV-1 infection among persistently seronegative prostitutes in Nairobi, Kenya. *Lancet (London, England)* 348, 1347–1351. doi: 10.1016/S0140-6736(95)12269-2

Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F., and Hanage, W. P. (2007). Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 104, 17441–17446. doi: 10.1073/pnas.0708559104

Gallo, R. C., Salahuddin, S. Z., Popovic, M., Shearer, G. M., Kaplan, M., Haynes, B. F., et al. (1984). Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* 224, 500–503. doi: 10.1126/science.6200936

Gao, X., Nelson, G. W., Karacki, P., Martin, M. P., Phair, J., Kaslow, R., et al. (2001). Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N. Engl. J. Med.* 344, 1668–1675. doi: 10.1056/NEJM200105313442203

GeneCards. (2019). GeneCards—human gene database. https://www.genecards.org/ [Accessed 26, July 2019].

Gonder, M. K., Mortensen, H. M., Reed, F. A., de Sousa, A., and Tishkoff, S. A. (2007). Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* 24, 757–768. doi: 10.1093/molbev/msl209

Greene, W. C. (2007). A history of AIDS: looking back to see ahead. *Eur. J. Immunol.* 37 , S94–102. doi: 10.1002/eji.200737441

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al. (2014). The African Genome Variation Project shapes medical genetics in Africa. *Nature* 517, 327–332. doi: 10.1038/nature13997

Hall, M. (1993). The archaeology of colonial settlement in Southern Africa. *Annu. Rev. Anthropol.* 22, 177–200. doi: 10.1146/annurev.an.22.100193.001141

Halper-Stromberg, A., and Nussenzweig, M. C. (2016). Towards HIV-1 remission: potential roles for broadly neutralizing antibodies. *J. Clin. Invest.* 126, 415–423. doi: 10.1172/JCI80561

Heine, B., and Nurse, D. (2000). *African languages: An introduction*. 1st ed. Cambridge: Cambridge University Press.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., et al. (2014). A genetic atlas of human admixture history. *Science (80-.)* 343, 747–751. doi: 10.1126/science.1243518

Hemelaar, J., Gouws, E., Ghys, P. D., and Osmanov, S. (2011). Global trends in molecular epidemiology of HIV-1 during 2000–2007. *AIDS* 25, 679–689. doi: 10.1097/QAD.0b013e328342ff93

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J. M., Kidd, J. M., et al. (2011). Hunter-gatherer genomic diversity suggests a Southern African origin for modern humans. *Proc. Natl. Acad. Sci. U. S. A.* 108, 5154–5162. doi: 10.1073/pnas.1017511108

Henrich, T. J., and Kuritzkes, D. R. (2013). HIV-1 entry inhibitors: recent development and clinical use. *Curr. Opin. Virol.* 3, 51–57. doi: 10.1016/j.coviro.2012.12.002

Herbeck, J. T., Gottlieb, G. S., Winkler, C. A., Nelson, G. W., An, P., Maust, B. S., et al. (2010). Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. *J. Infect. Dis.* 201, 618–626. doi: 10.1086/649842

Hirschhorn, J. N., and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108. doi: 10.1038/nrg1521

Hsu, D. C., and O'Connell, R. J. (2017). Progress in HIV vaccine development. *Hum. Vaccin. Immunother.* 13, 1018–1030. doi: 10.1080/21645515.2016.1276138

Huffman, T. N. (1970). The early Iron Age and the spread of the Bantu. *S. Afr. Archaeol. Bull.* 25, 3–21. doi: 10.2307/3888762

Hutcheson, H. B., Lautenberger, J. A., Nelson, G. W., Pontius, J. U., Kessing, B. D., Winkler, C. A., et al. (2008). Detecting AIDS restriction genes: from candidate genes to genome-wide association discovery. *Vaccine* 26, 2951–2965. doi: 10.1016/j.vaccine.2007.12.054

Hutter, G., Nowak, D., Mossner, M., Ganepola, S., Mussig, A., Allers, K., et al. (2009). Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N. Engl. J. Med.* 360, 692–698. doi: 10.1056/NEJMoa0802905

Jakobsson, M., Edge, M. D., and Rosenberg, N. A. (2013). The relationship between F(ST) and the frequency of the most frequent allele. *Genetics* 193, 515–528. doi: 10.1534/genetics.112.144758

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H.-C., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451, 998–1003. doi: 10.1038/nature06742

Javanbakht, H., An, P., Gold, B., Petersen, D. C., O'Huigin, C., Nelson, G. W., et al. (2006). Effects of human TRIM5alpha polymorphisms on antiretroviral function and susceptibility to human immunodeficiency virus infection. *Virology* 354, 15–27. doi: 10.1016/j.virol.2006.06.031

John, M.-A., Menezes, C. N., Chita, G., Sanne, I., and Grobusch, M. P. (2007). High tuberculosis and HIV coinfection rate, Johannesburg. *Emerg. Infect. Dis.* 13, 795–796. doi: 10.3201/eid1305.060908

Joubert, B. R., Lange, E. M., Franceschini, N., Mwapasa, V., North, K. E., and Meshnick, S. R. (2010). A whole genome association study of mother-to-child transmission of HIV in Malawi. *Genome Med.* 2, 17. doi: 10.1186/gm138

Kagaayi, J., and Serwadda, D. (2016). The history of the HIV/AIDS epidemic in Africa. *Curr. HIV/AIDS Rep.* 13, 187–193. doi: 10.1007/s11904-016-0318-8

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Lawn, S. D., Bekker, L.-G., and Miller, R. F. (2005). Immune reconstitution disease associated with mycobacterial infections in HIV-infected individuals receiving antiretrovirals. *Lancet. Infect. Dis.* 5, 361–373. doi: 10.1016/S1473-3099(05)70140-7

Le Clerc, S., Coulonges, C., Delaneau, O., Van Manen, D., Herbeck, J. T., Limou, S., et al. (2011). Screening low-frequency SNPS from genome-wide association study reveals a new risk allele for progression to AIDS. *J. Acquir. Immune Defic. Syndr.* 56, 279–284. doi: 10.1097/QAI.0b013e318204982b

Le Clerc, S., Limou, S., Coulonges, C., Carpentier, W., Dina, C., Taing, L., et al. (2009). Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J. Infect. Dis.* 200, 1194–1201. doi: 10.1086/605892

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science (80-.)* 319, 1100–1104. doi: 10.1126/science.1153717

Li, S., Schlebusch, C., and Jakobsson, M. (2014). Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proceedings. Biol. Sci.* 281. doi: 10.1098/rspb.2014.1448

Limou, S., Le Clerc, S., Coulonges, C., Carpentier, W., Dina, C., Delaneau, O., et al. (2009). Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J. Infect. Dis.* 199, 419–426. doi: 10.1086/596067

Lingappa, J. R., Petrovski, S., Kahle, E., Fellay, J., Shianna, K., McElrath, M. J., et al. (2011). Genomewide association study for determinants of HIV-1 acquisition and viral set point in HIV-1 serodiscordant couples with quantified virus exposure. *PLoS One* 6, e28632. doi: 10.1371/journal.pone.0028632

Macholdt, E., Lede, V., Barbieri, C., Mpoloka, S. W., Chen, H., Slatkin, M., et al. (2014). Tracing pastoralist migrations to Southern Africa with lactase persistence alleles. *Curr. Biol.* 24, 875–879. doi: 10.1016/j.cub.2014.03.027

Marks, S. E. (2014). Southern Africa. https://www.britannica.com/place/Southern-Africa [Accessed 20 January 2019].

Marks, S. J., Montinaro, F., Levy, H., Brisighelli, F., Ferri, G., Bertoncini, S., et al. (2015). Static and moving frontiers: the genetic landscape of Southern African Bantu-speaking populations. *Mol. Biol. Evol.* 32, 29–43. doi: 10.1093/molbev/msu263

Martin, M. P., Gao, X., Lee, J.-H., Nelson, G. W., Detels, R., Goedert, J. J., et al. (2002). Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat. Genet.* 31, 429–434. doi: 10.1038/ng934

May, A., Hazelhurst, S., Li, Y., Norris, S. A., Govind, N., Tikly, M., et al. (2013). Genetic diversity in black South Africans from Soweto. 14, 644. doi: 10.1186/1471-2164-14-644

McLaren, P. J., and Carrington, M. (2015). The impact of host genetic variation on infection with HIV-1. *Nat. Immunol.* 16, 577–583. doi: 10.1038/ni.3147

McLaren, P. J., Coulonges, C., Bartha, I., Lenz, T. L., Deutsch, A. J., Bashirova, A., et al. (2015). Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proc. Natl. Acad. Sci.* 112, 14658–14663. doi: 10.1073/pnas.1514867112

McLaren, P. J., Coulonges, C., Ripke, S., van den Berg, L., Buchbinder, S., Carrington, M., et al. (2013). Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS Pathog.* 9, e1003515. doi: 10.1371/journal.ppat.1003515

Meintjes, G., Lawn, S. D., Scano, F., Maartens, G., French, M. A., Worodria, W., et al. (2008). Tuberculosis-associated immune reconstitution inflammatory syndrome: case definitions for use in resource-limited settings. *Lancet. Infect. Dis.* 8, 516–523. doi: 10.1016/S1473-3099(08)70184-1

Mellors, J. W., Rinaldo, C. R. J., Gupta, P., White, R. M., Todd, J. A., and Kingsley, L. A. (1996). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 272, 1167–1170. doi: 10.1126/science.272.5265.1167

Montinaro, F., Busby, G. B. J., Gonzalez-Santos, M., Oosthuitzen, O., Oosthuitzen, E., Anagnostou, P., et al. (2017). Complex ancient genetic structure and cultural transitions in Southern African populations. *Genetics* 205, 303–316. doi:10.1534/genetics.116.189209.

Morris, A. G., Heinze, A., Chan, E. K. F., Smith, A. B., and Hayes, V. M. (2014). First ancient mitochondrial human genome from a prepastoralist Southern African. *Genome Biol. Evol.* 6, 2647–2653. doi: 10.1093/gbe/evu202

Naidoo, T., Schlebusch, C. M., Makkan, H., Patel, P., Mahabeer, R., Erasmus, J. C., et al. (2010). Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. *Investig. Genet.* 1, 6. doi: 10.1186/2041-2223-1-6

Neumann, K., Bostoen, K., Höhn, A., Kahlheber, S., Ngomanda, A., and Tchiengué, B. (2012). First farmers in the central African rainforest: a view from southern Cameroon. *Quat. Int.* 249, 53–62. doi: 10.1016/j.quaint.2011.03.024

Nurse, G. T., Weiner, J. S., Jenkins, T., et al., (1985). *The peoples of Southern Africa and their affinities*. New York: Oxford University Press.

O'Brien, S. J., and Nelson, G. W. (2004). Human genes that limit AIDS. *Nat. Genet.* 36, 565–574. doi: 10.1038/ng1369

Oliveira, S., Hubner, A., Fehn, A.-M., Aco, T., Lages, F., Pakendorf, B., et al. (2019). The role of matrilineality in shaping patterns of Y chromosome and mtDNA sequence variation in southwestern Angola. *Eur. J. Hum. Genet.* 27, 475–483. doi: 10.1038/s41431-018-0304-2

Pawlowski, A., Jansson, M., Sköld, M., Rottenberg, M. E., and Källenius, G. (2012). Tuberculosis and HIV co-infection. *PLoS Pathog.* 8, e1002464. doi: 10.1371/journal.ppat.1002464

Pelak, K., Goldstein, D. B., Walley, N. M., Fellay, J., Ge, D., Shianna, K. V., et al. (2010). Host determinants of HIV-1 control in African Americans. *J. Infect. Dis.* 201, 1141–1149. doi: 10.1086/651382

Pereyra, F., Jia, X., McLaren, P. J., Telenti, A., de Bakker, P. I. W., Walker, B. D., et al. (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 330, 1551–1557. doi: 10.1126/science.1195271

Petrovski, S., Fellay, J., Shianna, K. V., Carpenetti, N., Kumwenda, J., Kamanga, G., et al. (2011). Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population. *AIDS* 25, 513–518. doi: 10.1097/QAD.0b013e328343817b

Phillipson, D. W. (1977). *The later prehistory of Eastern and Southern Africa*. London: Heinemann. Available at: https://books.google.co.za/books?id=pRC1AAAAIAAJ.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., et al. (2012). The genetic prehistory of Southern Africa. *Nat. Commun.* 3, 1143. doi: 10.1038/ncomms2140

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., et al. (2014). Ancient west Eurasian ancestry in Southern and Eastern Africa. *Proc. Natl. Acad. Sci. U. S. A.* 111, 2632–2637. doi: 10.1073/pnas.1313787111

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Ramsay, J., Morton, B., and Mgadla, T., (1996). *Building a nation: a history of Botswana from 1800 to 1910*. Gaborone: Longman Botswana.

Reed, F. A., and Tishkoff, S. A. (2006). African human diversity, origins and migrations. *Curr. Opin. Genet. Dev.* 16, 597–605. doi: 10.1016/j.gde.2006.10.008

Retshabile, G., Mlotshwa, B. C., Williams, L., Mwesigwa, S., Mboowa, G., Huang, Z., et al. (2018). Whole-exome sequencing reveals uncaptured variation and distinct ancestry in the Southern African population of Botswana. *Am. J. Hum. Genet.* 102, 731–743. doi: 10.1016/j.ajhg.2018.03.010

Roser, M., and Ritchie, H. (2019). HIV / AIDS. Our World In Data. Available at: https://ourworldindata.org/hiv-aids [Accessed April 16, 2019].

Russell, T., Silva, F., and Steele, J. (2014). Modelling the spread of farming in the Bantu-speaking regions of Africa: an archaeology-based phylogeography. *PLoS One* 9, e87854. doi: 10.1371/journal.pone.0087854

Sadr, K. (2015). Livestock first reached Southern Africa in two separate events. *PLoS One* 10, e0134215. doi: 10.1371/journal.pone.0134215

Sanderson, J., Sudoyo, H., Karafet, T. M., Hammer, M. F., and Cox, M. P. (2015). Reconstructing past admixture processes from local genomic ancestry using wavelet transformation. *Genetics* 200, 469–481.

Schlebusch, C. (2010). Issues raised by use of ethnic-group names in genome study. *Nature* 464, 487. doi: 10.1038/464487a

Schlebusch, C. M., Lombard, M., and Soodyall, H. (2013). MtDNA control region variation affirms diversity and deep sub-structure in populations from Southern Africa. *BMC Evol. Biol.* 13, 56. doi: 10.1186/1471-2148-13-56

Schlebusch, C. M., Prins, F., Lombard, M., Jakobsson, M., and Soodyall, H. (2016). The disappearing San of southeastern Africa and their genetic affinities. *Hum. Genet.* 135, 1365–1373. doi: 10.1007/s00439-016-1729-8

Schlebusch, C. M., Skoglund, P., Sjodin, P., Gattepaille, L. M., Hernandez, D., Jay, F., et al. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* 338, 374–379. doi: 10.1126/science.1227721

Sharp, P. M., and Hahn, B. H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harb. Perspect. Med.* 1, a006841. doi: 10.1101/cshperspect.a006841

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shin, H. D., Winkler, C., Stephens, J. C., Bream, J., Young, H., Goedert, J. J., et al. (2000). Genetic restriction of HIV-1 pathogenesis to AIDS by promoter alleles of IL10. *Proc. Natl. Acad. Sci. U. S. A.* 97, 14467–14472. doi: 10.1073/pnas.97.26.14467

Siliciano, J. D., and Siliciano, R. F. (2016). Recent developments in the effort to cure HIV infection: going beyond N = 1. *J. Clin. Invest.* 126, 409–414. doi: 10.1172/JCI86047

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell* 177. 26–31 doi: 10.1016/j.cell.2019.02.048

Skoglund, P., Thompson, J. C., Prendergast, M. E., Mittnik, A., Sirak, K., Hajdinjak, M., et al. (2017). Reconstructing prehistoric African population structure. *Cell* 171, 59–71.e21. doi: 10.1016/j.cell.2017.08.049

Smith, M. W., Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Lomb, D. A., et al. (1997). Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort. *Science* 277, 959–965. doi: 10.1126/science.277.5328.959

Smith, M. W., Patterson, N., Lautenberger, J. A., Truelove, A. L., McDonald, G. J., Waliszewska, A., et al. (2004). A high-density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* 74, 1001–1013. doi: 10.1086/420856

Soodyall, H. (2006). *The prehistory of Africa: tracing the lineage of modern man*. Soodyall Jeppestown, H, editor. (South Africa: Jonathan Ball Publisher).

Soodyall, H., Makkan, H., Haycock, P., and Naidoo, T. (2008). The genetic prehistory of the Khoe and San. *S. Afr. Humanity* 20, 37–48.

Tafuma, T. A., Burnett, R. J., and Huis In 't Veld, D. (2014). National guidelines not always followed when diagnosing smear-negative pulmonary tuberculosis in patients with HIV in Botswana. *PLoS One* 9, e88654. doi: 10.1371/journal.pone.0088654

Taha, T. E. (2011). Mother-to-child transmission of HIV-1 in sub-Saharan Africa: past, present and future challenges. *Life Sci.* 88, 917–921. doi: 10.1016/j.lfs.2010.09.031

Tau, T., Davison, S., and D'Amato, M. E. (2015). Polymorphisms at 17 Y-STR loci in Botswana populations. *Forensic Sci. Int. Genet.* 17, 47–52. doi: 10.1016/j.fsigen.2015.03.001

Tau, T., Wally, A., Fanie, T. P., Ngono, G. L., Mpoloka, S. W., Davison, S., et al. (2017). Genetic variation and population structure of Botswana populations as identified with AmpFLSTR identifiler short tandem repeat (STR) loci. *Sci. Rep.* 7, 6768. doi: 10.1038/s41598-017-06365-y

Telenti, A., and Goldstein, D. B. (2006). Genomics meets HIV-1. *Nat. Rev. Microbiol.* 4, 865–873. doi: 10.1038/nrmicro1532

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

The International HapMap Consortium (2003). The International HapMap Project. *Nature* 426, 789–796. doi: 10.1038/nature02168

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., et al. (2009). The genetic structure and history of Africans and African Americans. *Science* 324, 1035–1044. doi: 10.1126/science.1172257

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39, 31–40. doi: 10.1038/ng1946

Tishkoff, S. A., and Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* 3, 611–621. doi: 10.1038/nrg865

Tough, R. H., and McLaren, P. J. (2019). Interaction of the host and viral genome and their influence on HIV disease. *Front. Genet.* 9, 720. doi: 10.3389/fgene.2018.00720

Troyer, J. L., Nelson, G. W., Lautenberger, J. A., Chinn, L., McIntosh, C., Johnson, R. C., et al. (2011). Genome-wide association study implicates PARD3B-based AIDS restriction. *J. Infect. Dis.* 203, 1491–1502. doi: 10.1093/infdis/jir046

UNAIDS. (2017). UNAIDS data 2017. *Jt. United Nations Program. HIV/AIDS*, 1–248. doi: 978-92-9173-945-5

UNAIDS. (2018). UNAIDS data 2018. *Jt. United Nations Program. HIV/AIDS*, 1–370.

UNAIDS (2019). UNAIDS is greatly encouraged by news of a possible cure of an HIV-positive man | UNAIDS. Available at: http://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2019/march/20190305_PS_cure [Accessed May 10, 2019].

Underhill, P. A., and Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* 41, 539–564. doi: 10.1146/annurev.genet.41.110306.130407

Uren, C., Kim, M., Martin, A. R., Bobo, D., Gignoux, C. R., Van Helden, P. D., et al. (2016). Fine-scale human population structure in Southern Africa reflects ecogeographic boundaries. *Genetics* 204, 303–314. doi: 10.1534/genetics.116.187369

van Manen, D., Delaneau, O., Kootstra, N. A., Boeser-Nunnink, B. D., Limou, S., Bol, S. M., et al. (2011). Genome-wide association scan in HIV-1-infected individuals identifying variants influencing disease course. *PLoS One* 6, e22208. doi: 10.1371/journal.pone.0022208

Veeramah, K. R., and Hammer, M. F. (2014). The impact of whole-genome sequencing on the reconstruction of human population history. *Nat. Publ. Gr.* 15, 149–162. doi: 10.1038/nrg3625

Vella, S., Schwartlander, B., Sow, S. P., Eholie, S. P., and Murphy, R. L. (2012). The history of antiretroviral therapy and of its implementation in resource-limited areas of the world. *AIDS* 26, 1231–1241. doi: 10.1097/QAD.0b013e32835521a3

Verma, S. S., and Ritchie, M. D. (2018). Another round of "Clue" to uncover the mystery of complex traits. *Genes (Basel).* 9, 61. doi: 10.3390/genes9020061

Vermund, S. H., Sheldon, E. K., and Sidat, M. (2015). Southern Africa: the highest priority region for HIV prevention and care interventions. *Curr. HIV/AIDS Rep.* 12, 191–195. doi: 10.1007/s11904-015-0270-z

Walker, N. F., Stek, C., Wasserman, S., Wilkinson, R. J., and Meintjes, G. (2018). The tuberculosis-associated immune reconstitution inflammatory syndrome: recent advances in clinical and pathogenesis research. *Curr. Opin. HIV AIDS* 13, 512–521. doi: 10.1097/COH.0000000000000502

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220. doi: 10.1093/nar/gkq537

Wei, Z., Liu, Y., Xu, H., Tang, K., Wu, H., Lu, L., et al. (2015). Genome-wide association studies of HIV-1 host control in ethnically diverse Chinese populations. *Sci. Rep.* 5, 10879. doi: 10.1038/srep10879

Welzel, T. M., Gao, X., Pfeiffer, R. M., Martin, M. P., O'Brien, S. J., Goedert, J. J., et al. (2007). HLA-B Bw4 alleles and HIV-1 transmission in heterosexual couples. *AIDS* 21, 225–229. doi: 10.1097/QAD.0b013e3280123840

Wen, Y.-J., Zhang, H., Ni, Y.-L., Huang, B., Zhang, J., Feng, J.-Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform.* 19, 700–712. doi: 10.1093/bib/bbw145

WHO. (2018). Global tuberculosis report.

Wilkinson, E., Engelbrecht, S., and de Oliveira, T. (2015). History and origin of the HIV-1 subtype C epidemic in South Africa and the greater Southern African region. *Sci. Rep.* 5, 16897. doi: 10.1038/srep16897

Winkler, C. A. (2008). Identifying host targets for drug development with knowledge from genome-wide studies: lessons from HIV-AIDS. *Cell Host Microbe* 3, 203–205. doi: 10.1016/j.chom.2008.04.001

Wright, S. (1950). Genetical structure of populations. *Nature* 166, 247–249. doi: 10.1038/166247a0

Xie, W., Agniel, D., Shevchenko, A., Malov, S. V., Svitin, A., Cherkasov, N., et al. (2017). Genome-wide analyses reveal gene influence on HIV disease progression and HIV-1C acquisition in Southern Africa. *AIDS Res. Hum. Retroviruses* 33, 597–609. doi: 10.1089/aid.2016.0017

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Worldometers. (2019). Southern Africa population. Available at http://www.worldometers.info/world-population/southern-africa-population/ [Accessed 14 January 2019].

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter 3. Whole Genome Sequencing based Characterization of HIV-1 Mutation Burden in Southern Africa

Chapter 3 and chapter 4 have been published in the following original paper:

Prisca K. Thami, Wonderful T. Choga, Delesa Damena Mulisa, Collet Dandara, Andrey K. Shevchenko, Melvin M. Leteane, Vlad Novitsky, Stephen J. O'Brien, Myron Essex, Simani Gaseitsiwe and Emile R. Chimusa. 2020. Whole Genome Sequencing-based Characterization of Human Genome Variation and Mutation Burden in Botswana. *bioRxiv*. https://doi.org/10.1101/2020.12.15.422821

Nature of publication: Original research

Journal: BioRxiv

Journal link: https://doi.org/10.1101/2020.12.15.422821

Candidate's contribution: Conceived the structure, conducted all bioinformatics analyses, drafted the manuscript, incorporated comments from the primary supervisor, submitted the manuscript to the journal.

Co-author contribution: ERC supervised the research and edited the manuscript. WTC, DDM, CD, AKS, MML, VN, SJO, ME, SG edited the manuscript.

Synopsis of paper 2: This paper addresses the following objectives of the thesis: to computationally analyse human whole genome sequences, delineate the human genetic landscape and elucidate the mutational burden in the complete human genomes of Botswana. We identified 27.7 million genomic variations from the complete genomes of Batswana (the people of Botswana), of which 2.8 million were novel. We observed a 3-way admixture of the Khoe-San, Niger-Congo and European ancestries in the population of Botswana and an overall largely homogenous sample of the Botswana population. Furthermore, we identified 24 potentially damaging variants, the most damaging variants being *ACTRT2* rs3795263, *HOXD12* rs200302685, *ABCB5* rs111647033, *ATP8B4* rs77004004 and *ABCC12* rs113496237, that putatively affected the transport of molecules including anti-HIV drugs (among other functions). Our findings highlight the potential of whole genome

sequencing in contributing to a deeper understanding of human diversity and the clinical relevance of genomic variations in African populations.

3.1 Introduction

Variants (or variations) are differences between the human reference genome (for instance GRCh38) and a genome of interest (Eichler, 2019). Genomic variations (variants) include single nucleotide variants (SNVs), short insertions and deletions (indels) of less than 50 bases and structural variations (Eichler, 2019; Steward et al., 2017; The 1000 Genomes Project Consortium, 2010, 2012; The 1000 Genomes Project Consortium et al., 2015). Most genomic variations are structural variations (Conrad et al., 2010). Structural variations are variants of a greater than 1kb that include large indels, copy number variations (CNVs), inversions, retrotransposon elements and duplications (Eichler, 2019; Mills et al., 2011; Steward et al., 2017; Sudmant et al., 2015). At the core of variant characterization is annotation of the discovered variants. Variant annotation involves interpreting the variants by determining their types, predicting the genomic location and functional elements of the discovered variants.

Variants can be classified according to genomic regions such as coding region (exons), downstream or upstream of genes, intergenic, untranslated regions (5'-UTR and 3'-UTR) of exons, introns, splice sites and non-coding regions (Antonarakis and Cooper, 2019). The coding sequence (CDS) is a stretch of exons interspersed by often 10- or 100-times larger introns. On each side of the CDS is a UTR. Many SNVs associated with diseases are found within non-coding regions (ENCODE Project Consortium, 2012). According to the central dogma of molecular biology, genetic material differentiates from DNA to RNA to protein (through transcription and translation respectively). However, about 4-9% of the human genome gets transcribed into RNA transcripts that do not code for proteins (Derrien et al., 2012; Ponting et al., 2009), known as non-coding RNAs. In the recent past years there has been a surge of the discovery of non-coding RNAs (ncRNAs), many of which have not been fully characterized and are of unclear functions (Tsai et al., 2017).

Variant annotation also deals with determining the consequences of variants. The main types of variant effects are: synonymous (silent), nonsynonymous (missense) and nonsense effects.

Synonymous variants cause a change in the codon trimer but due to the degeneracy of the genetic code do not change the amino acid. However, synonymous variants can cause amino acid change if the variants are located within splice sites (Sauna and Kimchi-Sarfaty, 2011). Nonsynonymous variants lead to a change in the amino acid. Nonsense variants change the affected codon into a stop codon which leads to a truncated gene product or change a stop codon into a non-stop codon which may cause elongation of the protein. In a scenario where a stop codon is introduced, this variant is termed a stop-gain variant. When a stop codon is lost due to alteration of the codon, the variant is termed a stop-loss variant. Although nonsynonymous variants can alter protein function, often leading to disease, the variant can also be compensated by surrounding variants in the human genome to obtain optimal functionality. A type of variants known as loss-of-function (LOF) variants lead to a severe disruption of genes. LOF variants can be any form of variants, single nucleotide substitutions (nonsense SNVs), splice site indels that can shift the CDS reading frame (frameshift variants) or large deletions that can delete an entire genetic element (MacArthur and Tyler-Smith, 2010).

Another form of annotation is adding minor allele frequency (MAF) labels of other populations to the genomic data of interest. This requires that the variant be common to both reference populations and the population under study. Commonly used reference datasets for this aim include The 1000 Genomes Project (1KGP), Hapmap and Exome Aggregation Consortium (ExAC) which has since been replaced by the Genome Aggregation Database (gnomAD). If the interest is to analyse African genomes, gnomAD is a more suitable database as it contains 21,042 genomes of African/African-American descent (Karczewski et al., 2020). Annotating the study variants with the reference populations MAF allows for population comparisons of MAF which fosters the identification of population-specific variants. Discrepancies in MAF across populations have implications in health. Prioritizing variants in medical genetics mainly entails distinguishing background benign variants from pathogenic variants that can lead to disease phenotypes (Conrad et al., 2010; Torkamani et al., 2011). According to comparative genomics, if a variant occurs in a gene that is conserved among species, this variant is likely to be pathogenic. In this regard a number of conservation methods can be used to identify deleterious mutations (Kircher et al., 2014).

Variant characterization of human whole exome sequences (WES) has been pivotal in diagnosis and clinical management of genetic diseases. However, WES examines protein-coding elements which constitute only 2% of the human genome. With WGS, a comprehensive catalogue of the human genomic variants that include non-coding regions (such as ncRNAs) not covered by WES can be obtained (Wells et al., 2019). Therefore, the aim of this chapter is to characterize genomic variations and elucidate mutation burden within a population of Botswana using whole genome sequencing.

3.2 Materials and Methods

3.2.1 Ethical approval

This study is part of a bigger protocol titled "Host Genetics of HIV-1 Subtype C Infection, Progression and Treatment in Africa/GWAS on determinants of HIV-1 Subtype C Infection" conducted by Botswana Harvard AIDS Institute Partnership. Ethics approval was obtained according to The Declaration of Helsinki. All participants consented to participate in the study. Institutional Review Board (IRB) approval was obtained for these samples from Botswana Ministry of Health and Wellness - Health Research Development Committee (HRDC) & Harvard School of Public Health IRB (reference number: HPDME 13/18/1) and the University of Cape Town - Human Research Ethics Committee (HREC reference number: 316/2019).

3.2.2 Patients and controls

This is a retrospective study that used samples from previous studies conducted at Botswana Harvard AIDS Institute Partnership between 2001 and 2007. Of the 390 participants, 265 were HIV-1 positive and 125 were HIV-1 negative. The participants were recruited from four locations within the southern region of Botswana (Mochudi, Molepolole, Lobatse and Gaborone) (**Figure 5**). The HIV-1 positive participants were previously part of the Mashi study (Shapiro et al., 2006; Thior et al., 2006), while HIV-1 negative participants were previously part of the Tshedimoso study (Novitsky et al., 2008).

Figure 5. Whole genome sequencing sampling sites in Botswana.

GB: Gaborone, LB: Lobatse, MC: Mochudi, ML: Molepolole. The map was produced with Maps package in R (code by Richard A. Becker et al., 2018).

3.2.3 DNA and Genomic characterisation

DNA was extracted from buffy coat using Qiagen DNA isolation kit following manufacturer's instructions. DNA concentration was quantified using Nanodrop® 1000 (Thermo Scientific, USA). Whole genome sequences of Botswana nationals were generated using paired end libraries on Illumina HiSeq 2000 sequencer at BGI (Cambridge, MA, US).

3.2.4 Variant Calling and Downstream Data Description

Quality assessment was performed on paired-end WGS (minimum of 30X depth) in FASTQ format (Cock et al., 2010) using FastQC (Van Der Auwera et al., 2014). Low-quality sequence bases and adapters were trimmed using Trimmomatic with default parameters (Bolger et al., 2014). The sequencing reads were aligned to the GRCh38 human reference genome using Burrows-Wheeler Aligner (BWA-MEM) (Li et al., 2008; Li and Durbin, 2009) and post-alignment quality control including adding of read groups, marking duplicates, fix mating and recalibration of base quality scores was performed using Picard tools, SAMtools (Li, 2011) and Genome Analysis Toolkit (McKenna et al., 2010). Four samples (HIV-1 positive females) were excluded due to poor quality of sequences, the remaining dataset had 390 individuals. We have run FastQC on all final BAM files prior the variant calling, then we aggregated the results from FastQC into a single report by using MultiQC (Ewels et al., 2016).

Variant calling is a process of determining nucleotide differences between the reference sequence and the sequence of a sample. In population genetics, it is best to perform the identification of variants from different individuals simultaneously – a process known as population joint calling (Nielsen et al., 2011; Pfeifer, 2017). We performed variant calling using two different population joint calling methods to leverage the quality and accuracy of our results: GATK's HaplotypeCaller (DePristo et al., 2011; McKenna et al., 2010) and BCFtools (Li, 2011). The variant call format (VCF) dataset was filtered using VCFTOOLS (Danecek et al., 2011), GATK's Variant Quality Score Recalibration and BCFtools. The specific filtering parameters employed for both call-sets have been detailed below. Downstream analyses were performed with GATK call-set and BCFtools call-set used as a validation set.

3.2.4.1 Variant Calling parameters applied to the sequence data

All the sequences passed quality control. Over 90% of the samples had a high sequence quality (at least 30 Phred score). This means that for these sequences a 99.9% accuracy in base calling was achieved. The acceptable Phred score for sequencing downstream analyses (population structure and genetic association) is at least 20. Although the sequence quality of the study samples was good, for variant calling bases with a Phred score of at least 30 were considered to ensure accurate identification of variants.

For **GATK** variant calling and filtration, we used GATK version 4.1.4.1. We performed variant calling using GATK's HaplotypeCaller to identify potential variants in each sample, then performed population joint genotyping to ensure high accuracy of calling. The variant calling followed GATK's best practices the following parameters:

```
gatk --java-options " -Xmx8g " HaplotypeCaller \
-R hg38.fasta \
-I SAMPLE.bam \
--dbsnp dbsnp151.vcf.gz \
--emit-ref-confidence GVCF \
-stand-call-conf 30 \
-O SAMPLE.g.vcf
```

The genotype VCF files were combined into a cohort file then population joint-calling was performed using the following parameters:

```
gatk --java-options " -Xmx100g" GenotypeGVCFs \
-R hg38.fasta \
--dbsnp dbsnp151.vcf.gz \
-V combine.vcf.gz \
-stand-call-conf 30.0 \
-A Coverage -A FisherStrand -A StrandOddsRatio -A MappingQualityRankSumTest \
-A QualByDepth -A RMSMappingQuality -A ReadPosRankSumTest \
--allow-old-rms-mapping-quality-annotation-data \
-O gatk.cohort.vcf.gz
```

For **BCFTOOLS** variant calling we used multiallelic calling model with the following parameters:

```
bcftools mpileup -Q30 -Ou -f hg38.fasta \
SAMPLE1.bam \
SAMPLE2.bam \
SAMPLE390.bam | bcftools call -mv -Oz -o bcf.cohort.vcf.gz
```

Raw variants called from GATK were filtered on minimum depth (DP) of 10, minimum genotype quality (GQ) of 20 and genotype call rate of 90% using VCFTOOLS. Following GATK guidelines, we filtered variants that had excess heterozygosity (> 54.69) and further filtered the data using machine learning framework implemented in GATK's Variant Quality Score Recalibration (VQSR) with the above-mentioned annotation features and the following training data sets:

INDELs

Mills_and_1000G_gold_standard.indels.hg38.vcf (prior=12.0)

dbsnp151.vcf (prior=2.0)

SNVs

hapmap_3.3.hg38.vcf (prior=15.0)

1000G_omni2.5.hg38.vcf (prior=12.0)

1000G_phase1.snps.high_confidence.hg38.vcf (prior=10.0)

dbsnp151.vcf (prior=7.0)

The training models were applied to the data using GATK's ApplyVQSR with truth sensitivity level of 99.9%.

We used the following filters on the BCFtools call set: and the following filters: depth (DP) > 10; mapping quality (MQ) > 30; variant quality (QUAL) > 20; <10% missing genotypes; <10% heterozygosity; and filtering SNP within 3bp around a gap (--SnpGap 3).

3.2.5 Variants Annotation and Mutation Prioritization

Functional annotation in the Botswana HIV-1 positive/negative VCF file to determine whether the variants putatively cause protein coding changes was performed using ANNOVAR (Wang et al., 2010), with minor validations done through snpEFF version 4.3T (Cingolani et al., 2012). We used ANNOVAR "2016Dec18" setting, where the population frequency, pathogenicity for each variant was obtained from 1000 Genomes exome (The 1000 Genomes Project Consortium et al., 2015), Exome Aggregation Consortium (Karczewski et al., 2017) (ExAC), targeted exon datasets and COSMIC (Forbes et al., 2015). Gene functions were obtained from RefGene (O'Leary et al., 2016) and different functional predictions were obtained from ANNOVAR's library. A total of 14 predictions that included 7 functional prediction scores (SIFT (Ng and Henikoff, 2003; Sim et al., 2012), LRT (Chun and Fay, 2009), MutationTaster (Schwarz et al., 2014), MutationAssessor (Reva et al., 2011), FATHMM (Shihab et al., 2013, 2014), Polyphen2 HVAR (Adzhubei et al., 2010), Polyphen2 HDIV (Adzhubei et al., 2010)), 3 ensemble scores (RadialSVM (Kircher et al., 2014), LR (Kircher et al., 2014), CADD (Kircher et al., 2014; Rentzsch et al., 2019)), and 4 conservation scores (GERP++ (Cooper et al., 2005), PhyloP-placental (Garber et al., 2009), PhyloP-vertebrate (Garber et al., 2009) and SiPhy (Garber et al., 2009)). 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)) from

SIFT, LRT, MutationTaster, MutationAssessor, FATHMM, RadialSVM, LR, CADD, GERP++, Polyphen2 HVAR, Polyphen2 HDIV, PhyloP-placental, PhyloP-vertebrate and SiPhy.

We prioritized the variants by retaining a variant only if it had at least 10 predicted functional status "D" or "A" out of 14 (Wonkam et al., 2020). We classified the top variants as those that were assigned "D" by FATHMM (Dong et al., 2015; Shihab et al., 2013, 2014), a disease-specific weighting scheme, which uses a Hidden Markov Models prediction algorithm capable of discriminating between disease-causing mutations and neutral polymorphisms. FATHMM has been found to have the most discriminative power among other individual *in silico* mutation prediction tools (Dong et al., 2015). We identified additional deleterious variants within the prioritized genes with snpEFF loss-of-function (LOF) module (Cingolani et al., 2012).

3.2.6 Distribution of pathogenic SNVs in known HIV-1 specific host genes

Following the functional annotation of the discovered variants, we evaluated the share of pathogenic SNVs between HIV-1 positive and HIV-1 negative individuals from Botswana. We further classified the SNVs as pathogenic or population specific if their MAFs were lower than 5%. The proportion of pathogenic SNVs within a gene was defined as the count of observed pathogenic variants over the total number of variants in the given gene (Wonkam et al., 2020). We obtained a list of 730 HIV associated genes from GWAS Catalog (www.ebi.ac.uk/gwas/), literature and gene-diseases database such DisGeNET (disgenet.org). We leveraged the dbSNP151 database (https://www.ncbi.nlm.nih.gov/snp/ (Sherry et al., 2001)) to extract SNVs associated with these genes in the Botswana data set (**Table A1**).

3.2.7 Pathways enrichment analysis and gene-gene interactions

The GeneMANIA (Warde-Farley et al., 2010) tool was used to analyse how the genes harbouring the identified variants interact in a biological network. This allowed us to obtain an enrichment of related genes within the obtained sub-network with potential biological pathways, processes, and molecular functions. Enrichment analysis was performed using Enrichr package (Chen et al., 2013; Kuleshov et al., 2016) in R (R Core Team, 2019).

3.3 Results

3.3.2 Characterization of variants and variants effect

We identified 27.7 million variants from 390 individuals of Botswana. Of these variations, we found 25.1 million SNVs and 2.6 million indels (**Table A2**); 13.4% of these variations were novel, i.e. not found in dbSNP151, 1KGP, AGVP and gnomAD (Karczewski et al., 2020) (**Figure 6a**). The average transition-transversion (TI/TV) ratio was 2.1. The novel variants were classified into genomic region and functional classes. Of the 2,789,599 novel variants, intergenic variants were observed at the highest frequency (1,461,193), followed by intronic (1,066,166) and ncRNA (178,178) variants (**Figure 6b** and **Table A3**). A majority of the novel variants were singletons, rare (MAF <= 0.01) and low frequency variants (MAF >0.01-0.05) (**Table A4**, **Figure 6a** and **c**). Nonsynonymous SNVs, stop gain and stop loss variants formed 65.6% of the exonic variants (**Figure 6d** and **Table A3**).

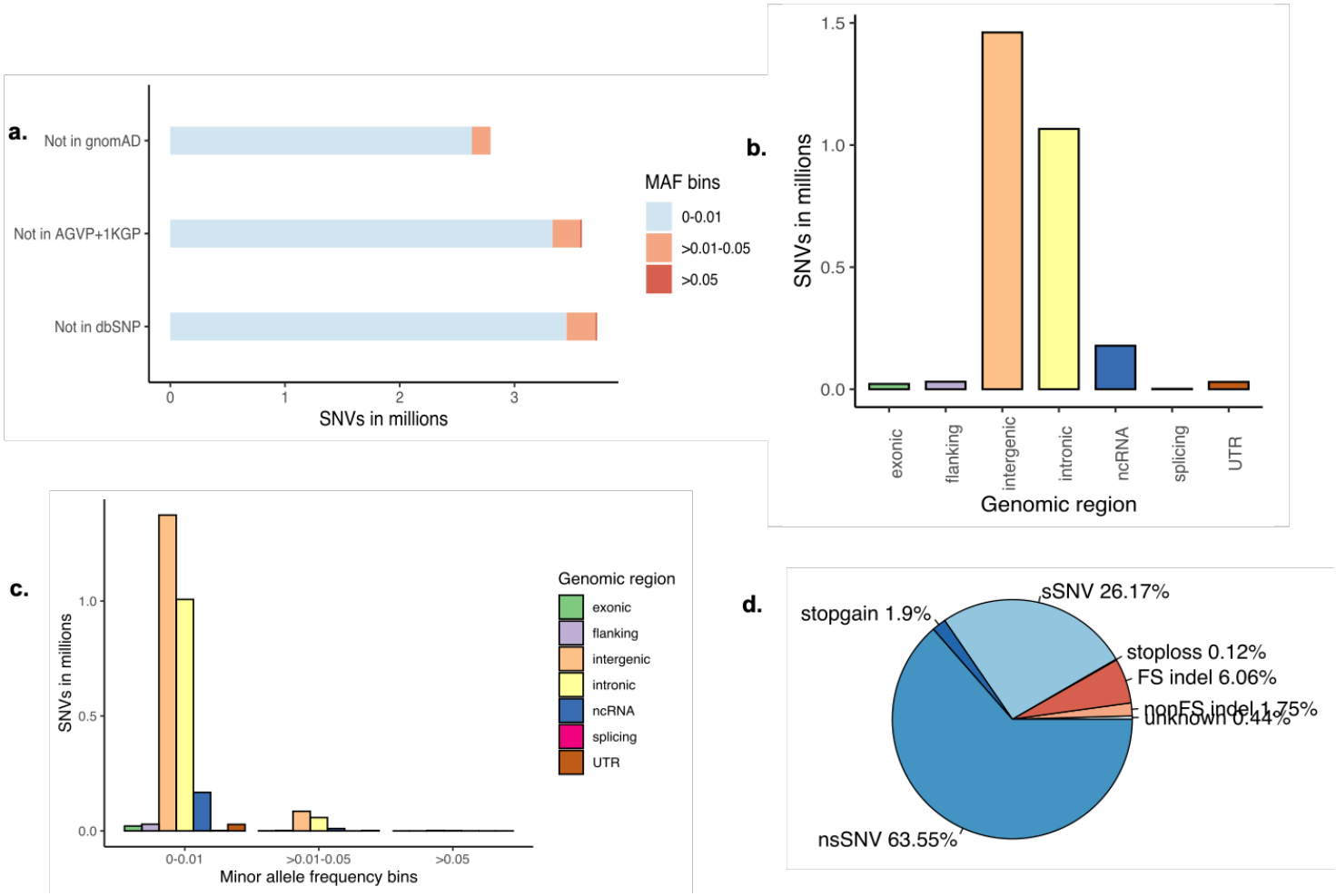


Figure 6. The distribution of novel variants in the Botswana population genomes.
a. Novel variants, absent from dbSNP151, the African Genome Variation Project (AGVP), the 1000 Genomes Project (1KGP) and gnomAD. **b.** Genome-wide distribution of novel variant effects by functional elements. **c.** Distribution of novel functional elements across MAF bins d. Distribution of exonic variants by functional elements. FS, frameshift; sSNV (synonymous SNVs); nsSNV (non-synonymous SNVs).

3.3.3 Variant Prioritization and prediction of mutation burden

Potentially pathogenic SNVs were identified by selecting those that had at least 10 predictions of deleteriousness (**Table A5**). We also observed that 8 of the genes identified by at least 10 predictions in ANNOVAR harboured additional loss-of-function (LOF) variants according to snpEFF (**Table A5**). A trimmed list of five SNVs that were further classified as "damaging" by FATHHM is hereby presented. The most deleterious mutations were found within the *ACTRT2*, *HOXD12*, *ABCB5*, *ATP8B4* and *ABCC12* genes (**Table 3**).

Table 3. The most deleterious nonsynonymous single nucleotide variants.

CHR	ID	cDNA change	AA change	Gene	Botswana	1KGP	gnomAD$_{AFR}$
1	rs3795263	exon1:c.G739A	p.G247R	*ACTRT2*	A=0.0013	A=0.12	A=0.044
2	rs200302685	exon2:c.G790C	p.E264Q	*HOXD12*	C=0.032	.	C=0.00038
7	rs111647033	exon13:c.G1319C	p.R440P	*ABCB5*	C=0.026	C=0.0004	C=0.0022
15	rs77004004	exon13:c.C1112A	p.P371H	*ATP8B4*	T=0.019	T=0.0038	T=0.013
16	rs113496237	exon5:c.G796C	p.G266R	*ABCC12*	G=0.013	.	G=0.000071

CHR: chromosome, AA change: amino acid change, 1KGP: The 1000 Genomes Project MAF, gnomAD$_{AFR}$: MAF of an African population from the gnomAD database.

3.3.4 Distribution of pathogenic SNVs in known HIV-1 specific host genes

Discrepancies in pathogenic SNV proportions were observed between HIV-1 positive (HIV-1 cases) and HIV-1 negative (HIV-1 controls) in the Sec1 Family Domain Containing 1 (*SCFD1*), Histone Cluster 1 H4 Family Member B (*HIST1H4B*), Histone Cluster 1 H4 Family Member A (*HIST1H4A*), Immunoglobin Superfamily Member 21 (*IGSF21*), Nuclear Cap Binding Protein Subunit 2 (*NCBP2*) and Zinc Finger DHHC-Type Palmitoyltransferase 19 (*ZDHHC19*) genes. Lower proportions were observed for *SCFD1*, *HIST1H4B*, *HIST1H4A* and *ZDHHC19* genes, and higher proportions were observed for *IGSF21* and *NCBP2* genes in HIV-1 cases (**Figure 7**).

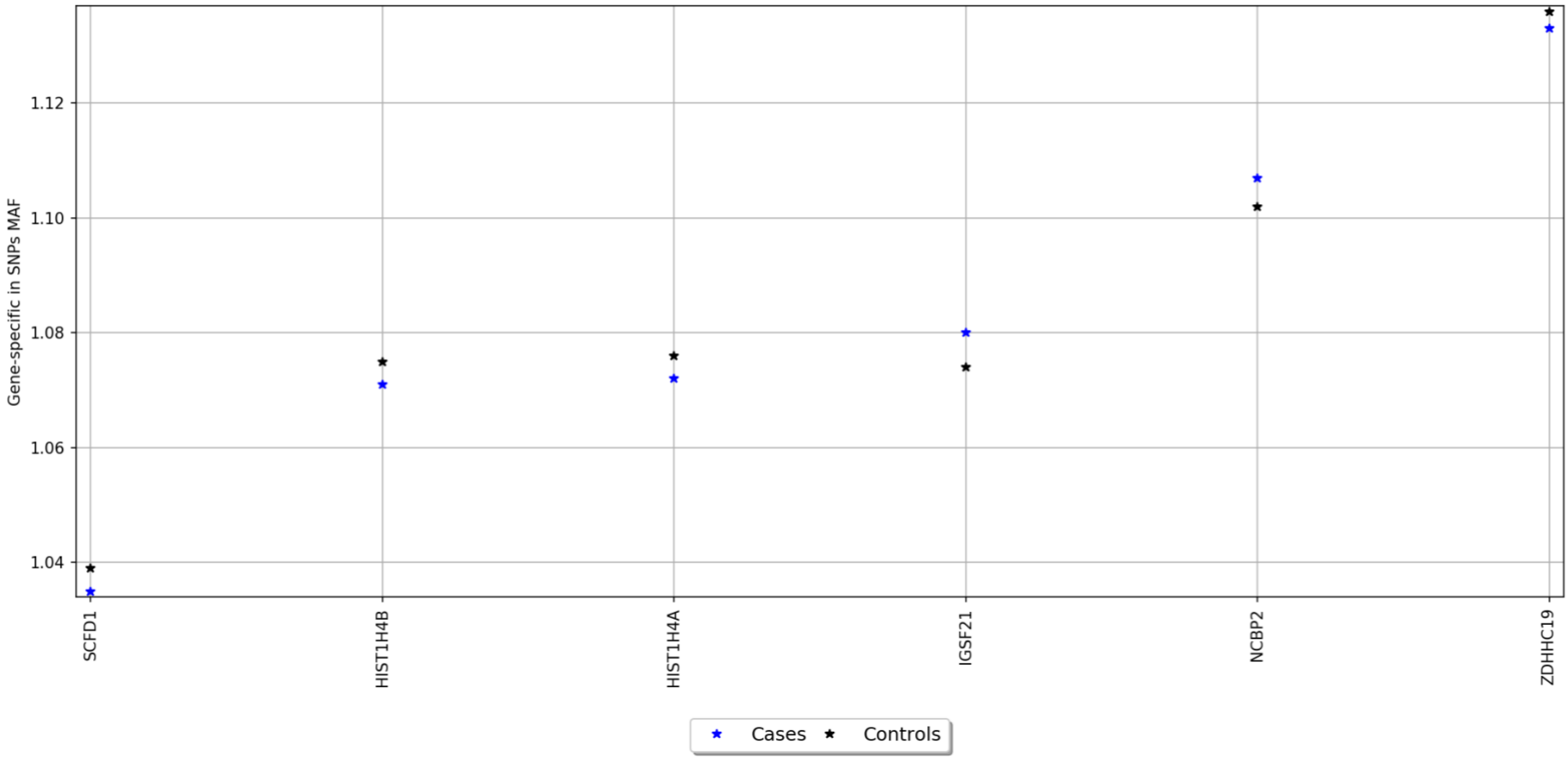


Figure 7. Distribution of pathogenic SNVs in known HIV-1 specific host genes.

3.3.5 Pathways enrichment analysis and gene-gene interactions

The 24 genes harbouring the potentially pathogenic variants were subjected to enrichment analysis using GeneMANIA (Warde-Farley et al., 2010) and Enrichr (Kuleshov et al., 2016) bioinformatics tools to identify biological processes and pathways putatively affected (**Figure 8**, **Table 4**). To successfully enrich for biological processes and pathways, the identified genes were used to "fish" 20 more related genes that are predicted to physically interact, co-express and co-localize with the identified genes (**Figure 8**).

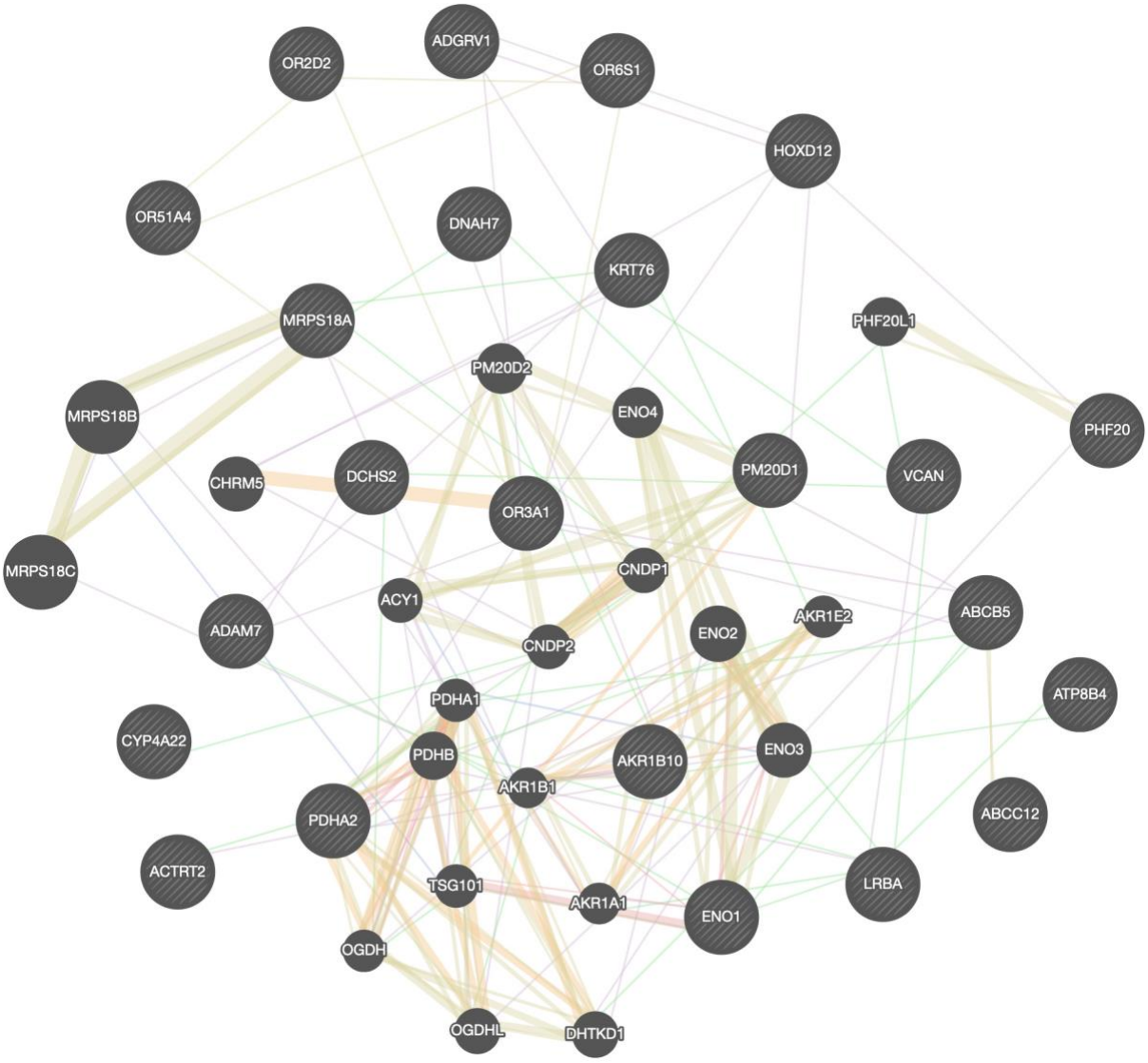


Figure 8. Gene-gene interaction network of genes harbouring the most deleterious variants. The different colours of branches represent how the genes are related; pink: physical interactions, purple: co-expression, orange: predicted, navy blue: co-localization, blue: Pathway, green: Genetic interactions, yellow: shared protein domains. Black and stripped nodes: genes provided as input into GeneMANIA. Black only nodes: genes predicted by GeneMANIA to interact with the input list. Connecting lines represent interactions.

The products of the identified genes were predicted to perform the following biological processes: gluconeogenesis, hexose and acyl-COA biosynthesis (**Table 4**). These gene products are localized within the oxoglutarate dehydrogenase complex and the mitochondria. The predicted molecular functions of these gene products were catalysis of peptidase, hydro-lyase, alcohol dehydrogenase and ATPase activities. The affected pathways included the glycolysis and gluconeogenesis, Krebs cycle, renal carcinoma, hypoxia-inducible factor 1 (HIF-1) signalling and folate biosynthesis pathways (**Table 4**). The identified genes were found to be associated with Pyruvate dehydrogenase complex deficiency (PDCD). One of the identified genes tumor susceptibility 101 (*TSG101*) was also found to be associated with human immunodeficiency virus 1 (HIV-1), albeit not statistically significant (**Table 4**).

Table 4. Enrichr gene-set enrichment of the genes harbouring the prioritized mutations.

Name	P-value	$P\text{-value}_{adj}$	Database
Gene Ontology			
Hexose biosynthetic process	2.59×10^{-6}	1.32×10^{-2}	Biological Process 2018 (Harris et al., 2004) http://www.informatics.jax.org/
Regulation of acyl-COA biosynthetic process	2.80×10^{-6}	7.14×10^{-3}	
Pyruvate metabolic process	3.11×10^{-6}	3.96×10^{-3}	
Glucose metabolic process	1.18×10^{-5}	1.2×10^{-2}	
Gluconeogenesis	9.99×10^{-5}	5.1×10^{-2}	
Oxoglutarate dehydrogenase complex	3.46×10^{-5}	1.54×10^{-4}	Cellular Component 2018 (Harris et al., 2004) http://www.informatics.jax.org/
Mitochondrial small ribosomal subunit	2.80×10^{-3}	6.24×10^{-3}	
Mitochondrial matrix	5.57×10^{-5}	8.28×10^{-2}	
Alcohol dehydrogenase (NADP+) activity	9.62×10^{-8}	5.54×10^{-5}	Molecular Function 2018 (Harris et al., 2004) http://www.informatics.jax.org/
Exopeptidase activity	1.32×10^{-4}	3.80×10^{-2}	
Hydro-lyase activity	1.32×10^{-4}	3.04×10^{-2}	
ATPase activity, coupled to movement of substances	2.54×10^{-4}	4.17×10^{-2}	
Pathways			
Glycolysis and Gluconeogenesis	2.84×10^{-6}	1.34×10^{-3}	WikiPathways 2019 Human (Slenter et al, 2017)
Hereditary leiomyomatosis and renal cell carcinoma pathway	1.10×10^{-5}	2.6×10^{-3}	
HIF-1 signalling pathway	2.08×10^{-9}	3.20×10^{-7}	
Citrate cycle (TCA cycle)	5.57×10^{-9}	5.72×10^{-7}	KEGG 2019 Human (Kanehisa and Goto, 2000)
RNA degradation	2.71×10^{-5}	2.09×10^{-3}	
Central carbon metabolism in cancer	3.95×10^{-4}	1.52×10^{-2}	Panther 2016 (Mi et al., 2017)
Histidine metabolism	1.16×10^{-3}	3.98×10^{-2}	
Folate biosynthesis	1.49×10^{-3}	4.58×10^{-2}	

Diseases			
Pyruvate dehydrogenase complex deficiency	4.71×10^{-5}	8.57×10^{-3}	ClinVar 2019 (Landrum et al., 2014) OMIM Disease (McKusick, 1998)
Human immunodeficiency virus 1	6.64×10^{-1}	1.0×10^{0}	VirusMINT (Chatr-aryamontri et al., 2009)

P-value$_{adj}$: adjusted P-value.

3.4 Discussion and Conclusion

Of the 27.7 million variants identified from 30X depth whole genomes of 390 individuals of Botswana. A critical and convenient QC metric to measure the quality and accuracy of genomic variation data is the TI/TV ratio (DePristo et al., 2011). The average TI/TV ratio of this set of variants was 2.1. This TI/TV ratio is within the expected range for human whole genome data which is ~2.0-2.1, meaning that the data has a very low frequency of false positive variant sites. As observed previously (Choudhury et al., 2017), intergenic variants had the highest frequency, followed by intronic variants and non-coding RNA (ncRNA) variants. Thirteen percent (2,789,599) of the discovered SNVs were novel. This number of previously uncaptured genetic variation highlights a potential of identifying population-specific variations through WGS. Whole genome sequencing also offers an opportunity to identify intronic variants and variants within non-coding regions. To this effect 1,066,166 intronic and 178,178 (ncRNA) novel variants were identified.

Recent human population expansion has resulted in a skewness towards excessive rare variants. This means that rare variants constitute a large part of the human genomic variations (Epi25 Collaborative, 2019; Hernandez et al., 2019; Johnston et al., 2015; Keinan and Clark, 2012; Nagasaki et al., 2015). Hence it is not surprising that a majority of the novel variants identified in the current study were very rare, occurring only once in the dataset (**Table A4**, **Figure 6a** and **c**). A substantial number of the exonic variants were nonsynonymous, stop gain and stop loss variants (**Figure 6d** and **Table A3**). These three types of mutations respectively cause a change in the amino acid and lead to an abnormal truncation or elongation of the protein, all leading to a change in the conformation or function of the encoded protein (Pagel et al., 2017). These changes have a potential to disturb normal biological processes and cause disease. In fact, a lot of genetic diseases are caused by nonsynonymous mutations.

Variants classified as probably damaging by at least 70% of the prediction tools were further prioritized with FATHHM score (**Table 3**). Here 5 variants within 5 genes were predicted to be the most deleterious (rs3795263 in the *Actin Related Protein T2* (*ACTRT2*) gene, rs200302685 in *homeobox D12* (*HOXD12*) gene, rs111647033 in ATP binding cassette subfamily B member 5 (*ABCB5*) gene, rs77004004 in *ATPase phospholipid transporting 8B4* (*ATP8B4*) gene and rs113496237 in *ATP Binding Cassette Subfamily C Member 12* (*ABCC12*) gene. The product of *ACTRT2* gene may be involved cytoskeletal organization (GeneCards, 2020). The rs3795263 variant was previously identified as harmful and associated with a severe form of tick-borne encephalitis virus infection (Ignatieva et al., 2019). The *HOXD12* gene belongs to the homeobox (*HOX*) family of genes that encode transcription factors involved in regulation of embryonic development (GeneCards, 2020; Lappin et al., 2006). The exact role of HOXD12 is unknown (GeneCards, 2020). The *HOX* genes have been implicated in maintenance and control of HIV-1 latency through epigenetic regulation (Khan et al., 2018).

The *ABCB5* gene belongs to the ATP-binding cassette (ABC) family that encodes proteins responsible for transmembrane transport of molecules including drugs such as doxorubicin (GeneCards, 2020). *ABCB5* is thought to also mediate chemoresistance of doxorubicin in malignant melanoma, (Whirl-Carrillo et al., 2012). The *ATP8B4* gene encodes an ATPase protein that is responsible for phospholipid translocation in the cell membrane (GeneCards, 2020). The *ABCC12* gene also encodes an ABC protein responsible for transmembrane transport of molecules. Overexpression of the *ABCC12* gene has been associated with breast cancer (GeneCards, 2020). Some members of the *ABC* family regulate the efflux of HIV-1 antiretrovirals from intracellular compartments (Eilers et al., 2008; Salvaggio et al., 2017). Biological pathways potentially affected by the products of these putatively deleterious genes and their interactome are discussed in subsequent paragraphs.

The minor allele frequencies of the *HOXD12* rs200302685, *ABCB5* rs111647033, *ATP8B4* rs77004004 and *ABCC12* rs113496237 variants in the Botswana data were generally higher when comparing to the gnomAD and the 1000 Genomes Project data. While the MAF for the *ACTRT2* rs3795263 variant was lower than in the gnomAD and the 1000 Genomes Project

data. This highlights that MAFs do vary per ethnicity which could affect the risk of disease differently between populations (**Table 3**).

Gene-set enrichment and functional analysis revealed the following pathways that were enriched for with the putatively deleterious genes: glycolysis and gluconeogenesis, Krebs cycle, renal carcinoma, Hypoxia-inducible factor 1 (HIF-1) signalling, RNA degradation, Histidine metabolism and folate biosynthesis pathways (**Table 4**). The *pyruvate dehydrogenase* (*PDH*), *enolase* (*ENO*) and *aldo-keto reductase* (*AKR1*) genes (**Figure 8**, **Table 4**) were significantly associated with glycolysis and gluconeogenesis (1.34×10^{-3}).

Both glycolysis and gluconeogenesis are glucose metabolism pathways; glycolysis is the catabolism of glucose (or glycogen) into pyruvate, while gluconeogenesis is the anabolism of pyruvate (from mainly proteins) into glucose (Bonora et al., 2012; Yang and Brunengraber, 2000). The PDH genes were also significantly associated with the Kreb's (Tricarboxylic Acid - TCA or Citric Acid) Cycle ($p = 5.72 \times 10^{-7}$). Glycolysis, gluconeogenesis and the TCA cycle are involved in energy (mostly in the form of Adenosine triphosphate - ATP) production, carried out in the cytoplasm or mitochondria of eukaryotes (**Table 4**) (Bonora et al., 2012; Yang and Brunengraber, 2000). In cells where there is low oxygen (a condition known as hypoxia), HIF-1 gets activated and triggers energy production through anaerobic glycolysis (Kim et al., 2006).

Glycolysis and TCA intermediates are used as precursors for macromolecule synthesis in hypoxic conditions such as cancer (DeBerardinis and Chandel, 2016; Huang et al., 2014). This may explain the association of the *PDH* genes with HIF-1 signalling and cancer pathways (**Table 4**). Amino acids are also made from intermediates of TCA, glycolysis and the pentose phosphate pathways (Berg et al., 2002b). This may explain the significant association of histidine biosynthesis with genes that are in connection with glycolysis and TCA pathways (**Figure 8**, **Table 4**).

The association of *AKR1* genes (**Figure 8**) with alcohol dehydrogenase (NADP+) activity and folate biosynthesis (**Table 4**) could be explained by that the alcohol dehydrogenases catalyse the reduction of NADP+ to NADPH (Penning, 2015). This reaction also takes place within

glycolysis, gluconeogenesis and pentose phosphate pathways (Berg et al., 2002b; Bonora et al., 2012). Furthermore, there is also evidence of NADPH being produced from folate metabolism (Fan et al., 2014). The human polynucleotide phosphorylase (hPNPase^{old-35}) is an evolutionary conserved RNA-degradation enzyme that has homologues in organisms such as *Escherichia coli* and yeast (Das et al., 2011; Leszczyniecka et al., 2002). In *E. coli* PNPase forms part of the degradosome with enolase and a helicase (Wilusz and Wilusz, 2008). This link between enolase and the evolutionary conserved PNPase may explain the association of the *ENO* genes with RNA degradation (**Table 4**). The degradation of HIV-1 mRNA in HIV-1 infected cells is important in suppressing HIV-1 replication (Hillebrand et al., 2019). Moreover, ENO-1 has been shown to prevent HIV-1 reverse transcription and ultimately decrease HIV-1 infectivity (Kishimoto et al., 2017).

Lower proportions of potentially pathogenic SNVs were observed for *SCFD1*, *HIST1H4B*, *HIST1H4A* and *ZDHHC19* genes, and higher proportions were observed for *IGSF21* and *NCBP2* genes in HIV-1 cases (**Figure 7**). The *IGSF21* gene encodes a cell receptor that is a member of the immunoglobulin superfamily (GeneCards, 2020). An intron variant rs2883821 within the *IGSF21* gene (chromosome 1) was reported to be associated with tenofovir pharmacokinetics and increased HIV-1 viral load (Buniello et al., 2019). The *NCBP2* gene encodes a protein that is part of the nuclear cap-binding protein complex (CBC). The CBC binds to the pre-mRNA and is involved in various processes such as splicing, transcription and nonsense-mediated mRNA decay (GeneCards, 2020). A splice site variant rs548853 within the *NCBP2* gene (chromosome 3) has been associated with decrease in viral load (Buniello et al., 2019). Since the proportion of pathogenic variants within the *IGSF21* and *NCBP2* genes in HIV-1 cases is higher than in HIV-1 control, this corroborates with the previous study, that *IGSF21* gene may harbour risk alleles.

The product of the *SCFD1* gene (chromosome 14) plays a role in SNARE-pin assembly and transport of molecules from the Golgi apparatus to the endoplasmic reticulum (GeneCards, 2020). SCFD1 interacts with other Golgi proteins and possibly affects HIV-1 replication through regulation of glycosylation. A reduction in the level of SCFD1 was observed to reduce HIV-1 infection (cell entry) (Zhu et al., 2014). The *HIST1H4A* and *HIST1H4B* genes (chromosome 6) encode histones, these are proteins that bind to DNA and assist in

compacting it into nucleosomes which are the basic repeating units of a chromatin. Therefore histones play critical roles in organizing chromatin structure and gene expression (Annunziato, 2008; Bednar et al., 1998; GeneCards, 2020). HIV-1 induced a modulation of the chromatin signalling network that involved HIST1H4A through epigenetic modifications (Deshiere et al., 2017).

The *ZDHHC19* gene (chromosome 3) encodes a palmitoyl acyltransferase, an enzyme that mediates palmitoylation of signal transducer and activator of transcription 3 (STAT3) (GeneCards, 2020). Palmitoylation is a lipid modification process in which a fatty acids such as palmitate are attached to a cysteine residue of a protein to regulate its attachment and localization to the cytoplasmic membrane (Conibear and Davis, 2010; Dietrich and Ungermann, 2004; Plain et al., 2020). In HIV-1 infection, STAT3 has been found to promote inflammation (Liu et al., 2013a) and also to promote antiviral immune responses (Del Cornò et al., 2014; Percario et al., 2003). According to the GWAS Catalog (Buniello et al., 2019) a variant rs11924930 within the *ZDHHC19* gene has been associated with HIV-1 susceptibility in a GWAS study of a Malawi population (Petrovski et al., 2011). However, the effect of this variant is not publicly available. In the current study, lower proportions of pathogenic SNVs were observed for the *SCFD1*, *HIST1H4B*, *HIST1H4A* and *ZDHHC19* genes in HIV-positive individuals. This might indicate that the minor alleles are protective against HIV-1 infection.

To conclude, this study is the first to use deep sequencing in efforts to delineate a complete genome map the human population of Botswana. Rare and low-frequency variants constituted the bulk of novel variants that were identified in this study. This was made possible by the unique potential of deep sequencing that offers an opportunity to discover rare variants. This is important because unlike Mendelian conditions, complex traits are influenced by many small-effect variants from different genetic loci, a concept known as polygenicity (Visscher et al., 2012). The cumulative effect of rare variants plays an important role in the expression of complex traits such as HIV-1.

Glycolysis, TCA and hexo-pentose pathways emerged to be the most affected by the putatively deleterious variants. These are critical physiological pathways responsible for energy production, amino-acid biosynthesis, immunity and tumorigenesis among other roles.

There were disparities in proportions of pathogenic variants within previously HIV-1 associated genes: *SCFD1*, *HIST1H4B*, *HIST1H4A*, *ZDHHC19*, *IGSF21* and *NCBP2* in HIV-1 infected versus uninfected individuals. Of interest in these genes is the *ZDHHC19* gene that encodes palmitoyl acyltransferase involved in pathways such as viral immunity. The *ZDHHC19* gene was previously identified in a GWAS of HIV-1 susceptibility in Malawi. Though the effect of this gene in the previous study is known, the identification of the gene in the current study confirms that the gene may have an implication in the genetics of HIV-1 in African populations. Although the candidate genes have been linked to HIV-1 infection, the same genes may also confer risk towards other health complications. This implies that the results may give insights into the potential interplay of genetic co-morbidities in the population of Botswana.

Chapter 4. Admixture and Population Structure of Botswana

Chapter 3 and chapter 4 have been published in the following original paper:

Prisca K. Thami, Wonderful T. Choga, Delesa Damena Mulisa, Collet Dandara, Andrey K. Shevchenko, Melvin M. Leteane, Vlad Novitsky, Stephen J. O'Brien, Myron Essex, Simani Gaseitsiwe and Emile R. Chimusa. 2020. Whole Genome Sequencing-based Characterization of Human Genome Variation and Mutation Burden in Botswana. *bioRxiv*. https://doi.org/10.1101/2020.12.15.422821

4.1 Introduction

Genetic variations can contribute in the understanding and reconstruction of population histories through the inference of genetic relatedness (or divergence) of individuals (Rosenberg et al., 2002). In genetic epidemiology, population structure is assessed and corrected to minimize spurious genetic associations and unmask signals of association (Hirschhorn and Daly, 2005; McLaren and Carrington, 2015; Price et al., 2006; Tishkoff et al., 2009).

The evaluation of population genetic structure can be considered as the estimation of genetic ancestry. Ancestry can affect genetic association results if there are significant differences in allele frequencies between the populations under investigation or if proportions of ancestry significantly differ between cases and controls (Chimusa et al., 2013b; Daya et al., 2014). Genetic ancestry can be classified as global and local ancestry with global ancestry being the average of the genome-wide ancestry of an individual while local ancestry would be estimating ancestry of a chromosomal segment. Global ancestry, which is the focus of this thesis, can be estimated through model-based or algorithmic estimations (Alexander et al., 2009).

Algorithmic approaches are multivariate statistical methods that identify clusters of ancestries by analysis of genetic similarities or distances between populations (Alexander et

al., 2009; Liu et al., 2013c). These approaches, also known as distance-based approaches, include PCA (Patterson et al., 2006; Price et al., 2006), analysis of molecular variance (AMOVA) (Excoffier et al., 1992), population graphs (Dyer and Nason, 2004), network theory (Greenbaum et al., 2016) and the F-statistics including F_{ST} (McQuillan et al., 2008; Weir and Cockerham, 1984; Wright, 1951). Model-based approaches, such as ADMIXTURE (Alexander et al., 2009), use explicit genetic models and evaluate the likelihood of the genetic data based on a model (Greenbaum et al., 2016). In this thesis we evaluated the genetic structure of the Botswana population using global ancestry estimation methods PCA, the F-statistics and ADMIXTURE. Brief explanations of the estimation formulae of these methods are subsequently provided.

4.1.1 Principal components analysis (PCA)

PCA is a dimension reduction technique. Here genetic data (such as SNPs) from each individual is projected along continuous axes of the variation in a way that reduces the data into a lesser number of dimensions. Usually the first 3 or 2 axes (principal components) can define most of the variability in the data. The axes of variation can be interpreted as the geographic location of a population (Price et al., 2006). We hereby describe PCA according to the authors (Patterson et al., 2006; Price et al., 2006).

Supposing that we have the genotype matrix ***G***, in which rows are indexed by individuals and columns by SNPs, then G_{is} would be the value representing the SNP bi-allele s for the individual i, where i=1 to N and s=1 to M. The allele can take values 0, 1 or 2 depending on the number of the reference allele copies. The column mean

$$\mu_s = \frac{\sum_{i=1}^{N} G_{is}}{N}$$

is subtracted from each entry in column s. Each column is then normalized by $\sqrt{p_s(1-p_s)}$, where p_s is the estimate of the SNP s allele frequency. The resulting matrix ***Z*** is an N by M normalized matrix and each i,s-th entry would then be

$$Z_{is} = \frac{G_{is}-\mu_s}{\sqrt{p_s(1-p_s)}}.$$

An N x N covariance matrix $\boldsymbol{\Psi}$ of individuals is then computed. This is the genetic relationship matrix (GRM).

$$\Psi = \frac{1}{S} ZZ^T$$

The i,j-th entry of the covariance matrix is the measure of the average genetic relatedness for individuals i and j

$$\Psi_{ij} = \frac{1}{S} \sum_{s=1}^{S} \frac{(G_i^s - p_s)(G_j^s - p_s)}{p_s(1 - p_s)}$$

PCA is then performed by obtaining the eigendecomposition of the matrix $\boldsymbol{\Psi}$ for dimension reduction. The eigenvectors are the linear combinations of SNPs that form the new dimensions. The axes of variation are defined as the eigenvectors with the largest eigenvalues. The eigendecomposition of $\boldsymbol{\Psi}$ is then

$$\Psi = VDV^T$$

where the columns of V are the eigenvectors and D is a diagonal matrix of eigenvalues in decreasing order. The *k*th eigenvector (principal component) corresponds to the *k*th eigenvalue that explains the variability explained by the eigenvector. In PCA, the principal components (PCs) are considered as axes of genetic variation with individuals of the same ancestry having similar values for the top PCs.

4.1.2 ADMIXTURE

Admixture analysis infers the genome-wide proportion of ancestry contributed by each source population (Alexander et al., 2009; Novembre, 2016). Genetic diversity in form of population structure and admixture have been documented in Southern Africa. The populations of Southern Africa have historically undergone admixture mainly from Niger-Congo, Khoisan population and ancient Eurasian ancestries (Busby et al., 2016; Choudhury et al., 2017, 2020; Gurdasani et al., 2015; Pickrell et al., 2012, 2014; Retshabile et al., 2018; Tishkoff et al., 2009). Here, we briefly describe the statistical model of the ADMIXTURE programme according to the authors (Alexander et al., 2009).

Let g_{is} represent the observed bi-allele s for the individual i, where g_{is} can be 0, 1 or 2 depending on the number of the reference allele copies at that SNP s. The likelihood of the data considering unrelated individuals is then

$$L(Q,F) = \sum_i \sum_s \left\{ g_{is} ln \left[\sum_k q_{ik} f_{ks} \right] + (2 - g_{is}) ln \left[\sum_k q_{ik} (1 - f_{ks}) \right] \right\}$$

where k is the source population of admixture that contributes q_{ik} proportion of the individual i's genome. The frequency of allele 1 (the minor allele) in the postulated ancestral population k would be f_{ks}. The parameters q_{ik} and f_{ks} are unknown, therefore the algorithm serves to estimate their values through maximization of the likelihood $L(Q,F)$.

For fast convergence and high accuracy of the point estimation, optimization is achieved by block relaxation which is a sequential quadratic programming algorithm where the increment $\Delta = x - x^n$ optimizes the quadratic approximation

$$f(x) = f(x^n) + df(x^n)\Delta + \frac{1}{2}\Delta^t d^2 f(x^n)\Delta$$

in which $df(x)$ and $d^2 f(x)$ denote the first and second differential of $f(x)$. The optimization alternates in updating the values of Q and F which are parameter matrices $Q = \{q_{ik}\}$ and $F = \{f_{ks}\}$.

4.1.3 Population-based genetic distance (F_{ST})

The Wright's fixation index (F_{ST}) is a measure of genetic differentiation between populations in comparison to within population diversity (Jakobsson et al., 2013; Tishkoff and Williams, 2002). F_{ST} belongs to the F-statistics family that were first introduced by Sewall Wright (Wright, 1951, 1965). Other statistical frameworks of F_{ST} have since been developed to improve the its estimations (Excoffier et al., 1992; Nei, 1973; Slatkin, 1995; Weir and Cockerham, 1984). The most convenient expression of F_{ST} can be expressed as

$$F_{ST} = \frac{h_T - h_S}{h_T} = \frac{H_S - H_T}{1 - H_T}$$

where h_T is the heterozygosity of the total population, h_S is the mean heterozygosity across subpopulations, $H_T = 1 - h_T$ is the homozygosity of the total population and $H_S = 1 - h_s$ the mean homozygosity across subpopulations. The values of F_{ST} are bound between 0 and 1 with values close to 1 indicating no shared genetic variation, while values close to 0

indicating no differentiation (Jakobsson et al., 2013; Tishkoff and Kidd, 2004). The inbreeding coefficient F_{IS} signifies the average proportion of autosomal genome that is inherited from a common ancestor. Related individuals share stretches of homologous chromosome segments that are identical by descent (McQuillan et al., 2008). These "long runs of homozygosity" (ROH) could infer population isolation and parental consanguinity (Li et al., 2006; McQuillan et al., 2008; Pemberton et al., 2012).

In this chapter population substructure and admixture of Botswana were assessed using PCA, admixture, F_{ST} and runs of homozygosity. Botswana is a landlocked country at the centre of Southern Africa (SADC, 2020). The population of Botswana is made up of mainly Bantu-speakers, an ethnolinguistic group of the Niger-Congo phylum (Batibo, 1999; Berman, 2017; Heine and Nurse, 2000). The languages spoken in Botswana as of the latest population census are Setswana (77.3%), Sekalanga (7.4%), Sekgalagadi/Sengologa (3.4%), English (2.8%), Zezuru/Shona (2.0%), Sesarwa/Khoe-San (1.7%), Sembukushu (1.6%), Herero (1.0%), Ndebele (1.0%) and others (such as Seyei, Setswapong, Sebirwa, Sesubiya, Sekgothu, Afrikaans, Indian, other African, other European and Asian) languages (Statistics Botswana, 2015). Botswana is also one of the few countries that carries the largest numbers of the Khoe-San in the Southern African region. Since admixture occurs when previously isolated populations interbreed, it is possible to observe admixture of the aforementioned populations in Botswana.

4.2 Materials and Methods

4.2.1 Population description and data acquisition

We obtained a joint-call VCF file containing 2,504 samples from 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010, 2012) and 2,428 sample from the African Genome Variation Project (Gurdasani et al., 2015) which has recently characterised the admixture across 18 ethno-linguistic groups from sub-Saharan Africa. We assembled a total of 4,932 samples. Based on initial sample description (population or country labels), we used the population ethno-linguistic information (Gudykunst and Schmidt, 1987; Michalopoulos, 2012) to categorize the obtained data per ethnic group as described in (**Table 5**) and we defined 20 world-wide ethnic groups (**Table 5**). We merged these 4,932 samples with our 390 Botswana samples resulting in a final total of 5,322 samples. The merge of our dataset and 1000

Genomes Project + African Genome Variation Project was based on overlapped SNPs using **PLINK** (Purcell et al., 2007)**.**

Table 5. Variants data from the 1000 Genomes Project (1KGP) and the African Genome Variation Project (AGVP) used for population structure and admixture analysis.

Population label	Ethnic group	Population description	Total Samples
AFR	Afro-Asiatic-Semitic	Amhara of Ethiopia	22
	African-American	Americans of African Ancestry in SW USA (ASW)	60
	African-Caribbean	African Caribbeans of 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 of Ethiopia	47
	Afro-Asiatic-Omotic	Wolayta of Ethiopia	24
	Khoe-San	Khoe-San	84
	Niger-Congo-Bantu	Baganda, Banyarwanda, Barundi, Rwandese, Ugandan, Banyankole of 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
EUR	European Center	British of England and Scotland (GBR)	91
	European North	Finnish of Finland (FIN)	99
	European South	Iberian Population in Spain (IBS), Toscani of Italia (TSI)	214
	European USA	Utah Residents with Northern and Western European Ancestry (CEU)	99
EAS	East Asian	Southern Han Chinese (CHS), Chinese Dai in Xishuangbanna, China (CDX), Kinh in Ho Chi	504

		Minh City, Vietnam (KHV), Han Chinese in Beijing, China (CHB), Japanese in Tokyo, Japan (JPT) *table continued from the previous page*	
	South Asian	Punjabi from Lahore, Pakistan (PJL), Bengali from Bangladesh (BEB)	180
SAS	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

4.2.2 Principal components analysis (PCA) and admixture analysis

For admixture analysis, we analysed the merged dataset of a total of 5,322 samples including Botswana HIV-1 positive/negative individuals and the 20 world-wide ethnic groups (see **Table 5**). The ADMIXTURE (Alexander et al., 2009) algorithm was used to estimate the ancestry proportions of the Botswana HIV-1 positive/negative groups. To evaluate the extent of substructure in the Botswana HIV-1 positive/negative population and whether stratification can be accounted for in the GWAS, PCA implemented in the EIGENSTRAT/smartpca programme of the EIGENSOFT package (Patterson et al., 2006; Price et al., 2006) was applied to the merged data set. We also assessed structure between the population of Botswana and the 20 world-wide ethnic groups.

Population structure and admixture were visualized by PCA plots generated using Genesis software (Buchmann and Hazelhurst, 2015) and R (R Core Team, 2019) with the pca3d package (Weiner, 2019).

4.2.3 Population-based genetic distance (F_{ST})

Pairwise genetic distance was estimated between the Botswana population and the 20 world-wide ethnic populations using the Weir and Cockerham's F_{ST} (Weir and Cockerham, 1984) in PLINK. A heatmap of the genetic distances was generated using ComplexHeatmap package (Gu et al., 2016) in R (R Core Team, 2019).

4.2.4 Genetic relatedness and runs of homozygosity

We assessed cryptic relatedness in the population of Botswana using PLINK. Pairwise allele sharing (identity-by-descent, IBD) was determined using pi_hat threshold of 0.2 (--genome --min 0.2). We further used PLINK to calculate homozygosity by keeping some of the default parameters while adjusting the window length and number of heterozygous SNVs allowed in the window (--homozyg-kb 150 and --homozyg-window-het 3). We compared the median lengths and segments of the runs of homozygosity (ROH) between the Botswana individuals and other world ethnic groups using Mann-Whitney U test in R.

4.2.5 Distribution of genetic ancestry proportions by HIV-1 positive/negative status

The mean proportions of the three ancestries were compared between HIV-1 positive and HIV-1 negative individuals using Student's T-test in R. The accurate admixture cluster was identified from model inference with lowest cross-validation (CV) error and the genome-wide admixture proportion estimations of that model inference were used as accurate genetic ancestry contribution (Figure A1). From these, and also basing on the population history of Southern Africa, we chose the best 3 proxy ancestral populations that had the highest genome-wide ancestry proportions from admixture analysis: Niger-Congo, Khoe-San and European.

4.3 Results

4.3.1 Population description and data acquisition

Variants were pruned to remove those with minor allele frequency < 5%, > 2% missingness, those that deviated from Hardy-Weinberg Equilibrium (HWE $p > 1.0 \times 10^{-5}$), and those in linkage disequilibrium (LD) $r^2 > 0.85$ within 1000kb window size, incrementing with 50 bases step (--indep-pairwise 1000 50 0.15). This resulted in 258,773 variants retained for assessing population diversity.

4.3.2 Principal components analysis (PCA) and admixture analysis

Population sub-structure was not observed within the Botswana study population. The plots of the first 3 PCs show a homogeneous mix of individuals from the HIV positive and the HIV negative groups with 3 outliers (**Figure 9**).

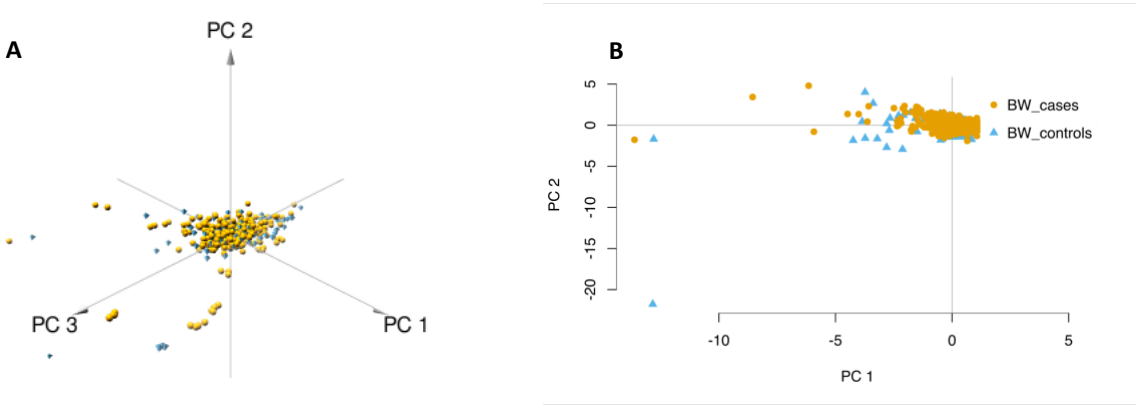


Figure 9. Principal component plot depicting population substructure of HIV-1 positive/negative individuals from Botswana.

A depiction of population substructure of Botswana with a 3D plot PCs 1,2 and 3 (**A**) and 2D plot of PC2 against PC1 (**B**) showing cases (HIV-1 positive) in bright brown and controls (HIV-1 negative in blue).

The Botswana population formed a cluster with other African populations of the Niger-Congo ethnolinguistic phylum, away from the other ethnicities (**Figure 10**). We also assessed the genetic relationship between Batswana, other Niger-Congo populations and the Khoe-San. We see in **Figure 11** that Batswana and the Niger-Congo Bantu South formed a separate cluster from other Niger-Congo populations, with a dispersion towards the Khoe-San.

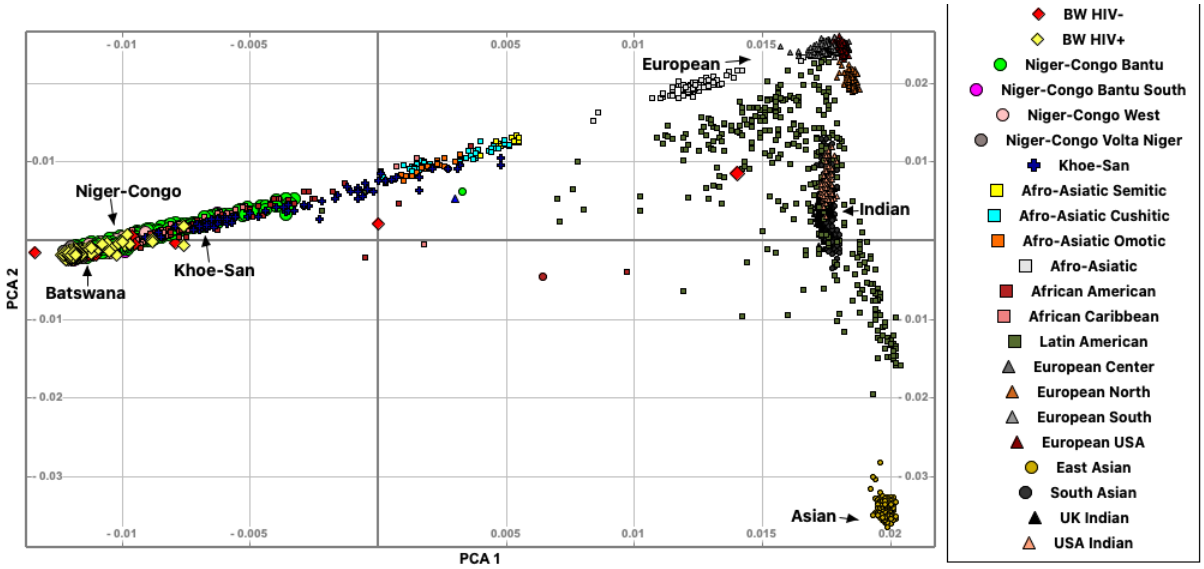


Figure 10. A PCA plot of the genetic relationship of the Botswana population with 20 world-wide ethnicities.

The points on the PCA plot represent each individual. Botswana individuals (known as Batswana) are shown in diamond. Botswana HIV-1 negative (BW HIV-) individuals are shown with red diamonds, while Botswana HIV-1 positive (BW HIV+) individuals are shown with yellow diamonds.

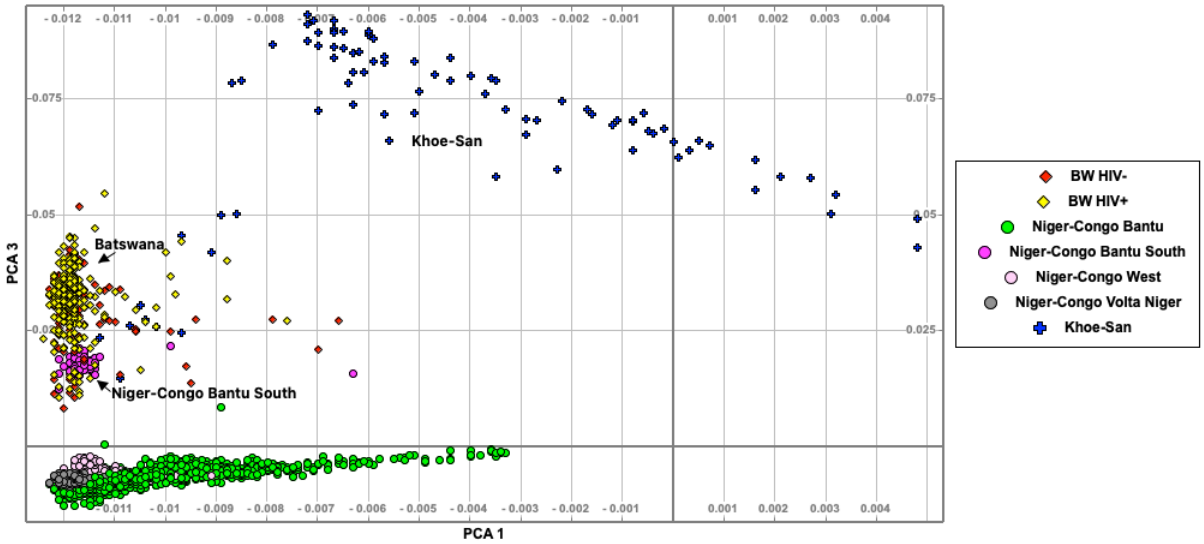


Figure 11. A PCA plot of the genetic relationship of Batswana, other Niger-Congo populations and the Khoe-San.

Botswana samples are in the convex of Khoe-San and Bantu, confirming the genetic contribution of both Bantu and Khoe-San in Botswana.

Given the results in Figure 10, we performed admixture analysis to estimate the individual fraction of genetic ancestry. We estimated the hypothesized the number of source populations (K) from literature (Busby et al., 2016; Chimusa et al., 2013b; Choudhury et al., 2017, 2020; Gurdasani et al., 2015; Pickrell et al., 2012, 2014; Retshabile et al., 2018; Tishkoff et al., 2009) and by assessing the raw matrices of the estimates of ancestry proportions in the optimal admixture model (**Figure A1**). Batswana assessed in this study show admixture of the following ancestry proportions: Niger-Congo (65.9%), Khoe-San (32.9%) and Europeans (1.1%) (**Figure 12**).

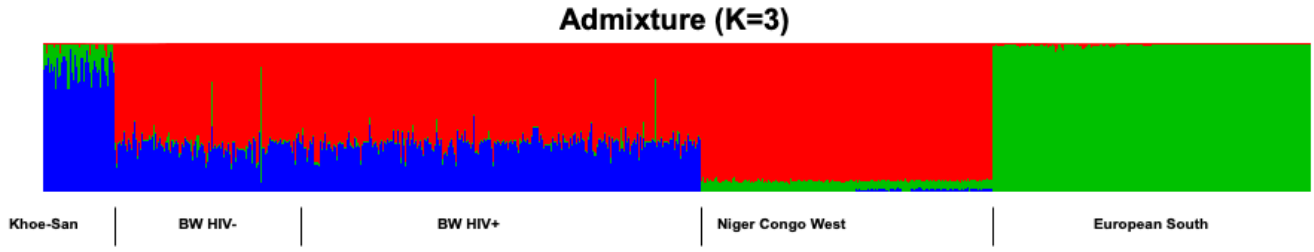


Figure 12. Genome-wide admixture proportions of Botswana.
Khoe-San, Niger-Congo and European populations were used as proxy ancestral populations that may have potentially contributed to the genetic architecture of Botswana.

4.3.3 Population-based genetic distance (F_{ST})

The pairwise F_{ST} results accentuates what was observed in assessment of global population structure. The heatmap and hierarchical clustering shows two distinct clusters separating into the Eurasian and African clades. A sub-clade that branches into the Niger-Congo populations and the Khoe-San population was observed. An inner sub-clade that separates Southern Bantu-speakers (including the Botswana population) from other Niger-Congo population is also observed (**Figure 13**).

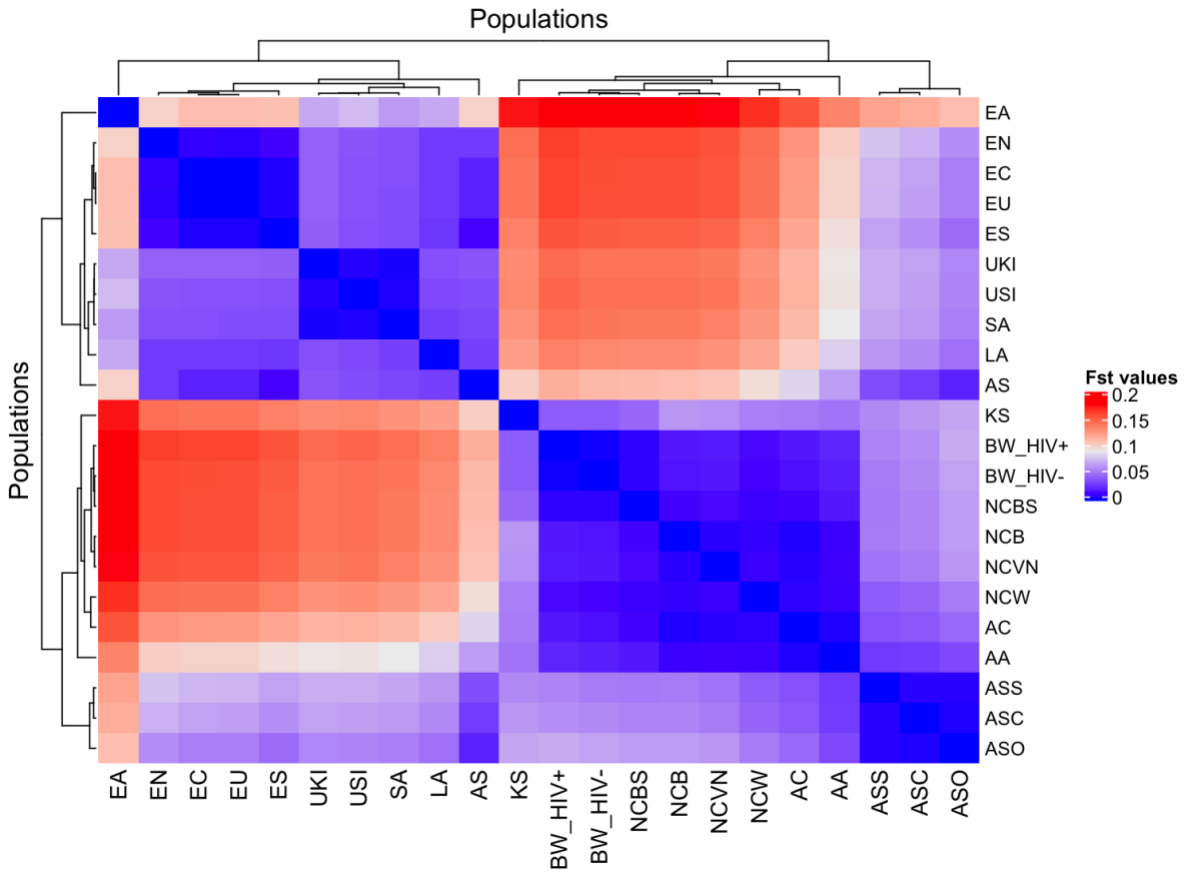


Figure 13. Pairwise genetic distance between the Botswana HIV-1 positive/negative population and 20 world-wide ethnicities.

This is a heatmap and dendrogram of F_{ST} values showing pairwise genetic divergence between populations. The blue shade represents similarity while the red shade represents divergence between the populations. The populations are AA: African-American, AC: African-Caribbean, AS: Afro-Asiatic, ASC: Afro-Asiatic Cushitic, ASO: Afro-Asiatic Omotic, ASS: Afro-Asiatic Semitic, LA: Latin American, KS: Khoe-San, BW_HIV+: Botswana HIV-1 positive, BW_HIV-: Botswana HIV-1 negative, NCB: Niger-Congo Bantu, NCBS: Niger-Congo Bantu South, NCVN: Niger-Congo Volta Niger, NCW: Niger-Congo West, EN: European North, ES: European South, EU:USA European, EC: European center, EA: East Asian, SA: South Asian, UKI: UK Indian and USI: USA Indian.

4.3.4 Genetic relatedness and runs of homozygosity

The IBD analysis revealed that none of the study participants were related. Our results further showed diversity in ROH segments among African populations, and between the African populations and non-African populations (**Figure 14**). Generally, the Niger-Congo populations

(including the Botswana HIV-1 positive/HIV-1 negative cohort) had lower ROH lengths and less abundant ROH segments than the European, Asian, Indian, Latin-American and Khoe-San populations (**Figure 14**).

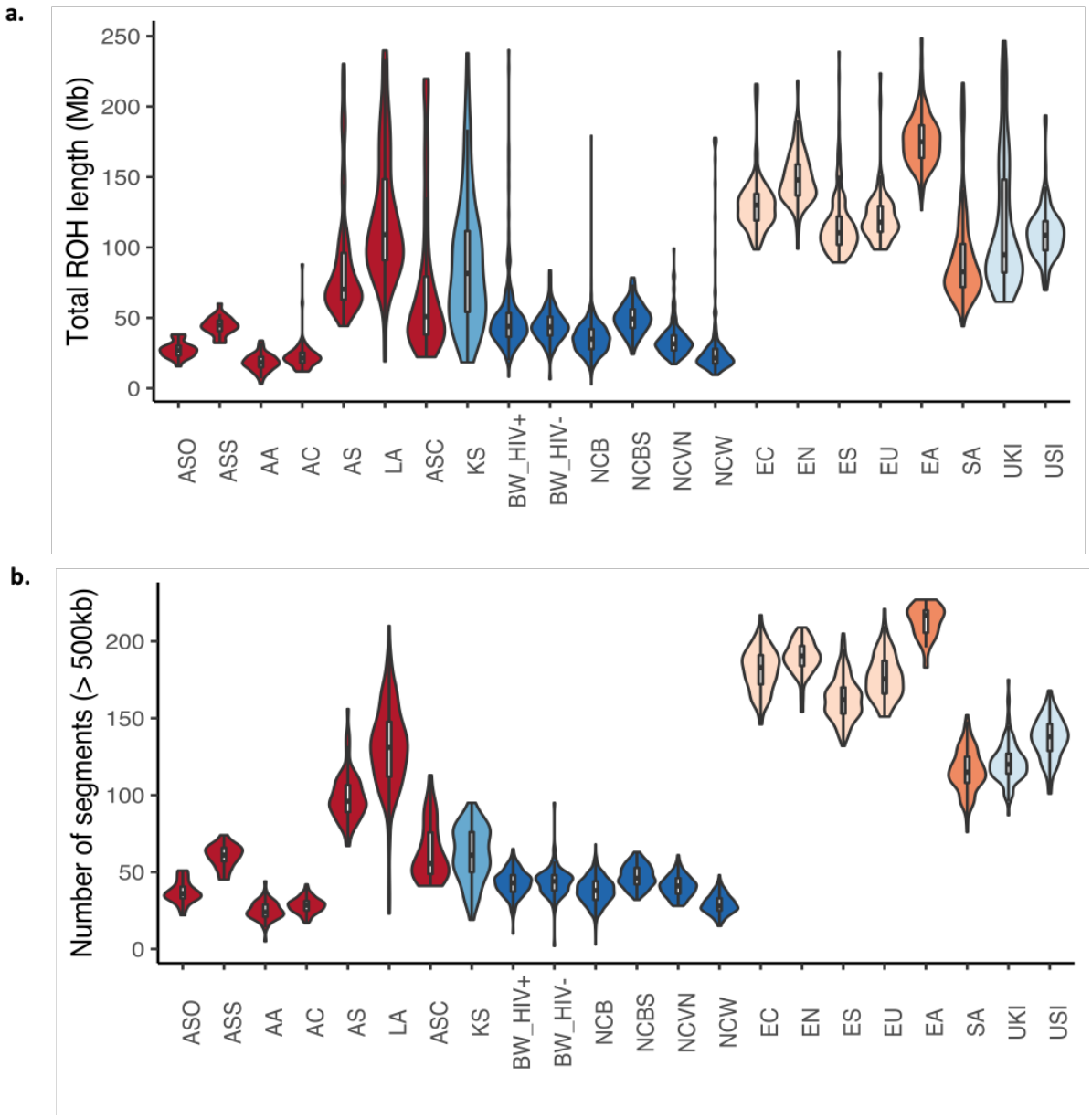


Figure 14. The lengths and number of runs of homozygosity (ROH) segments across different global ethnic groups.

Violin plots showing median the lengths (in Mb) and number of ROH. The colours represent different super-groups: Mixed populations (African-American (AA), African-Caribbean (AC), Afro-Asiatic (AS), Afro-Asiatic Cushitic (ASC), Afro-Asiatic Omotic (ASO), Afro-Asiatic Semitic (ASS) and Latin American (LA)) in dark-red, Khoe-San (KS) in light blue, Niger-Congo in navy blue (Botswana HIV-1 positive (BW_HIV+), Botswana HIV-1 negative (BW_HIV-), Niger-Congo

Bantu (NCB), Niger-Congo Bantu South (NCBS), Niger-Congo Volta Niger (NCVN) and Niger-Congo West (NCW)), Europeans (European North (EN), European South (ES), USA European (EU), European center (EC)) in light orange, Asians (East Asian: EA and South Asian: SA) in orange and Indians (UK Indian (UKI) and USA Indian (USI)) in very light blue.

4.3.5 Comparison of genome-wide admixture proportions between HIV-1 positive and HIV-1 negative groups

The genome-wide genetic proportions of Khoe-San ancestry in Botswana cases (HIV positive individuals) was significantly higher (0.336 ± 0.003 vs 0.315 ± 0.005, p-value = 0.002) than that observed in Botswana controls (HIV negative individuals) (**Table 6**). There was no significant difference in the genome-wide genetic proportions of the Niger-Congo and European ancestries when comparing Botswana cases to Botswana controls (**Table 6**).

Table 6. Comparison of the mean genetic ancestry proportions of Botswana estimated with ADMIXTURE between HIV-1 positive and HIV-1 negative groups.

	Genome-wide ancestry contribution (mean ± SE)			
Ancestry	All samples	Cases (HIV+)	Controls (HIV-)	Mean comparisons of cases vs controls (P-value)
Khoe-San	0.329 ± 0.003	0.336 ± 0.003	0.315 ± 0.005	0.002
Niger-Congo	0.659 ± 0.003	0.657 ± 0.003	0.665 ± 0.007	0.328
European	0.0114 ± 0.003	0.00727 ± 0.002	0.0202 ± 0.007	0.076

Estimation of the mean and SE (standard error of the mean) of ancestry proportions from each of the 3 populations contributing to the admixture of Botswana.

4.4 Discussion and conclusion

The PCA plots revealed that the Botswana study population is overall largely homogenous, making it suitable for genetic association studies (**Figures 9-11**). Although the Botswana HIV-1 positives and negatives almost completely overlap, there is a considerable spread in the data points (**Figures 10 and 11**). The spread is even more than that of European samples combined, this signifying a higher genetic diversity in the Botswana study population and African populations generally. The study participants were recruited from three districts in the southern part (Southern, Kweneng and South-East) of Botswana. Although the sampling

site does not span the whole of Botswana, the current findings have positive implications for genetic epidemiology in the southern part of Botswana.

Population substructure can mask true genetic associations and also lead to false discovery of causal (or modifier) variants (Hirschhorn and Daly, 2005; McLaren and Carrington, 2015; Price et al., 2006; Tishkoff et al., 2009). To find minimal or no substructure in the study population will minimize false positives in subsequent genetic association analyses. Furthermore, population-specific interventions against HIV-1 can be employed for this part of Botswana which will minimize costs that may arise in personalized medicine.

The PCA (**Figure 10**) and F_{ST} values (**Figure 13**) show that there was a clear distinction between African populations and European populations. The population of Botswana clustered with other Niger-Congo populations and showed a dispersion towards the Khoe-San population (**Figure 11**). The Botswana population showed a closer affinity with the Niger-Congo Bantu South (Zulu) population. This is expected as a close affinity of the Sotho with the Niger-Congo Bantu South (Zulu) has previously been reported (Choudhury et al., 2017). Batswana are members of the Sotho-Tswana clan of Southern Africa that includes the Sotho (of Lesotho and South Africa) and Batswana (of Botswana and South Africa) (Batibo, 1999; Berman, 2017).

Major events such as the "Bantu expansion" and Eurasian migration into Southern Africa have shaped the genetic landscape of the region. These events have led to varying degrees of admixture of the migrant groups and indigenous population (Chimusa et al., 2013a; Choudhury et al., 2017; Gurdasani et al., 2015; Montinaro et al., 2017; Petersen et al., 2013; Pickrell et al., 2014; Thami and Chimusa, 2019; Tishkoff et al., 2009). These previous findings are congruent with the current study that reports a 3-way admixture of Niger-Congo (65.9%), Khoe-San (32.9%) and European (1.1%) populations observed in the Botswana population (**Figure 12** and **Table 6**).

We found no evidence of consanguinity in the Botswana HIV-1 positive/HIV-1 negative cohort as defined by less abundant segments and lower lengths of ROH in comparison to non-African populations and the Khoe-San (**Figure 14**). This finding is supported by the previous observation of no extended ROH lengths in a Botswana HIV positive cohort (Retshabile et al.,

2018). Among the Niger-Congo populations, the median ROH length in the Botswana HIV-1 positive/HIV-1 negative and the Niger-Congo Bantu South were significantly higher (p-value = 2.2 x 10^{-16}) than of the Niger-Congo Bantu, Niger-Congo West and the Niger-Congo Volta Niger (**Figure 14**). These results are consistent with what was observed by Choudhury *et*. *al*., who observed that the Niger-Congo Bantu population of Southern Africa had the highest lengths of ROH compared to Niger-Congo populations of East, Central West and West Africa (Choudhury et al., 2017).

The mean proportion of Khoe-San ancestry was higher in HIV-1 cases than in controls (**Table 6**). Host genetics studies have previously linked Khoe-San ancestry to susceptibility to tuberculosis (TB) (Chimusa et al., 2013b). Could this be yet another association of the Khoe-San ancestry with an infectious disease as with TB? If the answer is an affirmation then from a public health standpoint, this could mean that efforts of HIV-1 prevention should be elevated among people of Khoe-San ancestry. However, this conclusion seems to be implausible since when we regressed HIV-1 status against the ancestry proportions none of the ancestries showed association with HIV-1 status.

It may be concluded that no overall substructure was observed in Botswana. This is good because if present, population structure cannot be fully corrected in genetic association studies. Nonetheless, the observed spread in the data that makes appropriate correction in association studies mandatory. Admixture of Niger-Congo, Khoe-San and European populations was observed in the Botswana population. Genetic association results can also be affected if cases and controls have significant differences in ancestry proportions from the source populations of admixture. This would warrant statistical adjustment for ancestry in the association model (Chimusa et al., 2013b; Daya et al., 2014). Although the cases and controls of our study showed a significant difference in Khoe-San genome-wide ancestry proportions, we did not observe an association between HIV-1 status and ancestry proportions.

It was not surprising to observe a considerable proportion of the Khoe-San ancestry as Botswana is one of the countries with the largest number of the Khoe-San. The Khoe-San are known to be the indigenous people of Southern Africa. Overtime the Khoe-San are expected

to have mingled and interbred with the Niger-Congo people of Botswana. Hence, this work shows the pivotal role played by genetics in the reconstruction of population histories. A limitation of this study is that the Botswana population had no ethnolinguistic labels, as such ethnicity inferences cannot be drawn from this study. Nevertheless, population structure and admixture could still be assessed as the algorithms used in this study are unsupervised machine learning methods and therefore can still give meaningful results.

Chapter 5. Whole Genome Association Study of HIV-1 in a Southern African Population

This chapter forms part of the following original publication:

Prisca K. Thami, Wonderful T. Choga, Delesa Damena Mulisa, Collet Dandara, Andrey K. Shevchenko, Melvin M. Leteane, Vlad Novitsky, Stephen J. O'Brien, Myron Essex, Simani Gaseitsiwe and Emile R. Chimusa. 2020. Whole Genome Rare-Variant Association Study of HIV-1 progression in a Southern African population. *medRxiv*. https://doi.org/10.1101/2020.12.16.20248307

Nature of publication: Original research

Journal: MedRxiv

Journal link: https://doi.org/10.1101/2020.12.15.422821

Candidate's contribution: Conceived the structure, conducted all bioinformatics analyses, drafted the manuscript, incorporated comments from the primary supervisor, submitted the manuscript to the journal.

Co-author contribution: ERC supervised the research and edited the manuscript. WTC, DDM, CD, AKS, MML, VN, SJO, ME, SG edited the manuscript.

Synopsis of paper 3: This paper addresses the following objective of this thesis: to identify HIV-1 associated genetic loci in the human whole genome sequences by aggregating the effects of rare-variants. The results presented in this chapter were a "spin-off" from the results of the findings from the "Whole Genome Sequencing-based Characterization of Human Genome Variation and Mutation Burden in Botswana" paper in Chapter 3. One of the findings from Chapter 3 of this thesis was that rare-variants constituted a large portion of the human genomic variations. From literature, we established that the cumulative effects of these rare-variants can contribute to diseases. Therefore, in Chapter 5, we evaluated the role of rare-variants in the control of HIV-1 progression and HIV-1 acquisition. We discovered 3 candidate genes (*ANKRD39*, *LOC105378523* and *GTF3C3*) that could potentially control HIV-1 replication which subsequently leads to the control of HIV-1 progression.

5.1 Introduction

Southern and eastern Africa carry the highest prevalence of Human immunodeficiency virus (HIV) infection globally. Botswana is the third most affected country in Southern Africa, after eSwatini and South Africa in first and second positions respectively (UNAIDS, 2019). The country is affected predominantly by HIV-1C. The HIV epidemic became severe in Botswana by the late 1990s at a prevalence of 30-40% in pregnant women (Essex, 1999).

Botswana was the first country in Southern Africa to offer free antiretroviral therapy (ART) to people infected with HIV. Due to the rapid scaleup of anti-HIV drugs, there has been a sharp decline in HIV-related morbidity and mortality (Escudero et al., 2019; Farahani et al., 2014). HIV prevalence in Botswana has since lowered to 20.3% among adults (UNAIDS, 2019). Nonetheless Botswana still remains one of the most affected countries globally due to the high baseline HIV prevalence and a successful national ART programme.

There is a remarkable interpersonal heterogeneity of HIV-1 phenotypes (acquisition, progression and drug metabolism). This heterogeneity in host phenotypes of HIV-1 infection has been attributed to several factors including host genetics (Carr et al., 2017; Pereyra et al., 2010; Telenti and Goldstein, 2006). Treatment against HIV-1 does not offer cure and to date no effective vaccine has been found against HIV infection (Avert, 2019; Siliciano and Siliciano, 2016). Therefore, identifying population specific genetic factors can catapult the invention of effective strategies against HIV-1 in African populations.

Candidate disease gene methods revealed a number of genes associated with HIV-1 infection (Hutcheson et al., 2008; O'Brien and Nelson, 2004; Winkler, 2008). A momentous achievement in the HIV-1 candidate gene research was the discovery of *CCR5*-Δ32 variant within the *CCR5* gene. When present the mutation confers resistance to HIV acquisition or slow progression in carriers (Dean et al., 1996; Ioannidis et al., 2001). This finding has successfully been translated into virus entry inhibitor antiretrovirals (Henrich and Kuritzkes, 2013; Woollard and Kanmogne, 2015). Other genes identified through candidate gene method include *HLA-A, HLA -B* and *HLA -C*, *CCR2*, *SDF1*, *IL10* (Carrington and O'Brien, 2003;

O'Brien et al., 2000), *CCL5* (RANTES), *KIR* genes, *TRIM5* and *APOBEC3G* (Hutcheson et al., 2008; O'Brien and Nelson, 2004; Winkler, 2008).

The advent of genome-wide arrays uncovered several more loci associated with HIV-1 acquisition and progression. Genome-wide arrays did not require *a priori* knowledge of genomic region to be tested which rendered the method nearly unbiased (Hirschhorn and Daly, 2005). These methods once again bolstered the finding of *HLA-B* and *HLA-C* loci being the major determinants of HIV-1 control. Although the HLA region accounted for most of the variability in HIV-1 progression and control, GWAS of HIV-1 have uncovered new genes such as *ZNRD1* (Fellay et al., 2007, 2009), *NOTCH4* (Fellay et al., 2009; Le Clerc et al., 2011), *C6orf48* (Le Clerc et al., 2009, 2011) in European populations.

While the bulk of genome-wide association studies (GWAS) were carried out in European populations, there has been few GWAS of African populations (Lingappa et al., 2011; Petrovski et al., 2011; Xie et al., 2017). Of note, two novel variants within *HCG22* and *CCNG1* genes were found to be associated with progression and acquisition of HIV1 in a Botswana population (Xie et al., 2017). We have previously discussed the results of GWAS of HIV-1 extensively here (Thami and Chimusa, 2019).

Genome-wide arrays revealed many variants of clinical significance to acquisition, control of HIV and progression (Limou and Zagury, 2013). However, the method has limitations such as that 1) the tag-SNPs embedded in the arrays are not really causal, 2) the arrays harbour common variations so rarer variations can be missed and 3) due to higher diversity and lower linkage disequilibrium in African populations, the arrays are less robust in capturing variations in African populations (Limou and Zagury, 2013; Telenti and Johnson, 2012; Thami and Chimusa, 2019).

Variants associated with HIV-1 in both candidate disease gene methods and GWAS accounted for just above 20% of the variability in the phenotypes of HIV (Fellay et al., 2009; Telenti and Johnson, 2012). This missing heritability of HIV-1 phenotypes may be hidden in rare-variants among other factors (Telenti and Goldstein, 2006; Thami and Chimusa, 2019). Whole genome (or exome) sequencing offers a more effective method to identify rare variants within African

populations. We present here, whole genome sequencing (WGS) in a Southern African population of Botswana in efforts to pinpoint rare-variants which could be of clinical significance to the susceptibility to HIV-1 acquisition and progression.

5.2 Materials and methods

5.2.1 Ethical approval

Ethical approval was obtained as explained in **Chapter 3**.

5.2.2 Patients and controls

Study participants recruitment has been explained in **Chapter 3**. Additionally, the HIV-1 negative participants enrolled in the current study were those that self-reported that they were highly exposed to HIV-1 and required HIV-1 testing in the Tshedimoso study. The Tshedimoso study sought to identify acute and recent HIV-1C infection in Botswana. These participants remained negative for the duration of the study (Novitsky et al., 2008).

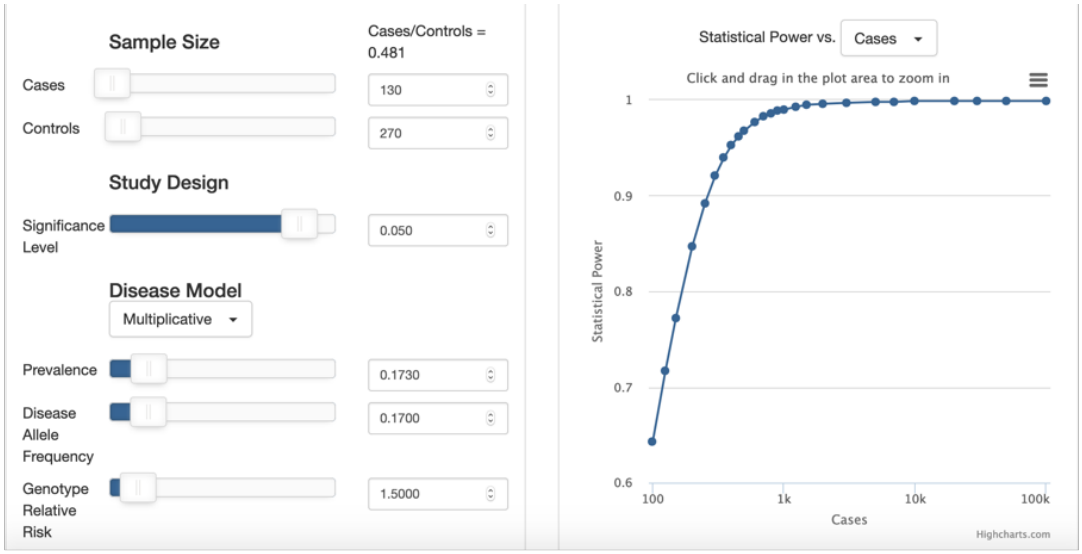
5.2.3 DNA and Genomic characterisation

The processing of DNA and generation of sequencing reads has been described in **Chapter 3**.

5.2.4 Sample size calculation for genetic association tests

Given (a) a prevalence of HIV-1 (17.3% up to 21.8%) in Botswana; which is about 50-fold higher than in Non-Africa populations; (b) assuming 17% up to 47% of HIV-1 risk allele frequency [rs17762192], (c) the expected genotypes relative risk in [1.5, 1.6], modelled with variation determined by a Poisson distribution with mean equal to the above prevalence (Purcell et al., 2003). We performed a power analysis (Johnson et al., 2016) assuming expected sample sizes of 130 cases / 270 controls, from **Figure 15** it is anticipated that this study has 73% up to 98% predicted power to identify candidate risk polymorphisms for HIV-1 and to replicate HIV-1 susceptibility at nominal p-values < 0.05.

a.



b.

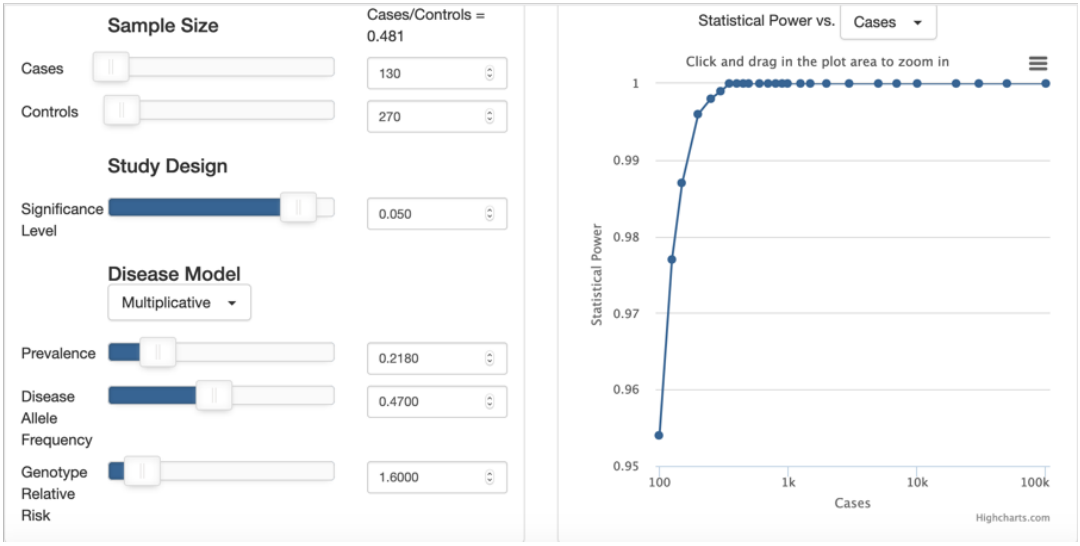


Figure 15. Power estimate of the genome-wide association study.

The above study power (**Figure 15**) was increased by combining two GWAS datasets through meta-analysis. Given current limitation of GWAS, to not account for rare variants and polymorphisms with small effect, and given the diversity in Southern African populations, we expect that polymorphisms with small effects and rare variants may contribute to genetic variation of HIV-1. Therefore, we conducted gene sets and burden and rare-variant analyses using SKAT (Lee et al., 2012a), an robust genetic association method that accounts for small sample size and rare-variants.

5.2.5 Variant Calling and Downstream Data Description

Variant calling was performed and explained in **Chapter 3**. Further quality control as required prior genetic association tests was performed using PLINK (Purcell et al., 2007). Variants that had genotype missingness and MAF of 5% or more, and those that deviated from Hardy-Weinberg Equilibrium (HWE $p > 1.0 \times 10^{-5}$) were filtered out.

5.2.6 Cross population meta-analysis of GWAS of susceptibility to HIV-1 acquisition

Genetic association testing was performed on the Botswana HIV-1 positive/negative dataset using the Efficient Mixed-Model Association eXpedited (EMMAX) method (Kang et al., 2010) which corrects for populations stratification and cryptic relatedness. First, EMMAX-kin was applied to compute pair-wise relatedness matrix from the dataset, which is representative of the structure of the samples. EMMAX estimated the contribution of the sample structure to Botswana HIV-1 positive/negative using a variance component model, resulting in an estimated covariance matrix of phenotypes that models the effect of genetic relatedness on the HIV-1 genotype-phenotype. EMMAX was run on the HIV-1 genotype-phenotype using the covariance matrix to detect possible association.

To identify associations with small effect sizes which aren't usually detected by standard GWAS methods, GWAS summary statistics from Botswana was combined with that from Uganda (Lingappa et al., 2011) in a single GWAS dataset. The Uganda GWAS participants were 798 HIV-1 serodiscordant couples from East and Southern Africa with a median age of 31 years (range: 18-70) (Lingappa et al., 2011). A fixed effects model (Han and Eskin, 2011) based on inverse-variance weighted effect size was used to combine the log odds ratio and stand error from the combined GWAS summary statistics dataset. Random effects and binary-effects models described in MetaSoft program (Han and Eskin, 2011) were applied; and the P-values and m-values (the posterior probability that the effect exist in the study), the mean effect and heterogeneity statistics were produced to interpret the association results showing high heterogeneity.

5.2.7 Rare-variant association test

To account for sample size and rare variants that standard GWAS could have missed, we leveraged possible small effects by aggregating SNPs effects at gene level with an optimal unified sequence kernel association test (SKAT-O) (Lee et al., 2012a, 2012b). This test combined burden and variance-component analyses using the SKAT package (Wu et al., 2011) in order to appropriately discriminate 265 HIV-1 positive and 125 HIV-1 negative individuals. SKAT-O has been optimized to control for type I error and increase power for small sample sizes. Moreover, the SKAT_NULL_emmaX() module was used with kinship relatedness matrix to adjust for genetic relatedness and population structure within the data. The MAF cut-off for the SKAT-O was set at 0.05. The SNV sets were created using a custom Python script that mapped SNV positions from our study to the reference SNV data (dbSNP151) to retrieve gene names that were then used as SetIDs. We used all the variants available per gene. A p-value of < 0.05/genes was considered significant; where genes is the number of genes tested in the model. A brief description of the SKAT-O statistical model is provided in section **5.2.7.1**.

For HIV-1 progression, we regressed CD4+ T-cell counts over several covariates (**Table A6**) using lmer function in the R lme4 package (Bates et al., 2015). We then used the regression coefficients (slopes) of HIV-1 positive individuals (236 young women) to infer CD4+ T-cell changes over a period of at least 18 months. A number of models were tested to select confounding variables and interacting factors (**Table A6**). We performed model selection using the following independent factors: age, education level, feeding strategy, and the interaction of HAART with time. There was no correlation between CD4+ T-cells and age, education level nor feeding strategy. As expected, CD4+ T-cell counts were significantly correlating with the presence of highly active antiretroviral therapy (HAART), this justifying the inclusion of HAART as a covariate to model CD4+ T-cell changes over time. The CD4+ T-cell slopes were used as a quantitative phenotype for genetic subsequent association tests.

5.2.7.1 A statistical model of SKAT-O

SKAT-O is a combination of the SKAT and burden tests (Lee et al., 2012a, 2012b; Wu et al., 2011). In a rare-variant association study, let G_i be a vector of J alleles for an individual i having a phenotype y_i and covariates denoted as a vector X_i . When dealing with a continuous phenotype the following linear regression model is considered

$$y_i = X_i'\alpha + G_i'\beta + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

while for binary phenotypes the following logistic regression model is considered

$$logitP(y_i = 1) = X_i'\alpha + G_i'\beta.$$

The regression coefficients for covariates and allele effects are the α and β vectors respectively. The score statistic of the j-th variant then becomes $S = \sum_{i=1}^{N} g_{ij}(y_i - \hat{\mu}_i)/\hat{\phi}_k$ where $\hat{\mu}_i$ is the estimated mean of y_i under the null linear or logistic regression models, and $\hat{\phi}_k = \hat{\sigma}^2$ for continuous phenotypes or $\hat{\phi}_k = 1$ for binary phenotypes. The SKAT statistic is then denoted as

$$Q_{SKAT} = \sum_{j=1}^{J} w_j^2 S_j^2$$

where w_j is the weight for the j-th allele which is usually a function of the MAF. The burden test uses the same weighting as SKAT, but works by aggregating the set of rare-variants and then regressing the trait y_i on the weighted total number of variants:

$$Q_{BURDEN} = \left(\sum_{j=1}^{J} w_j S_j\right)^2.$$

The SKAT-O test the optimal weighted average of the SKAT and the burden test statistics which is denoted as

$$Q_\rho = (1-\rho)Q_{SKAT} + \rho Q_{BURDEN}.$$

The asymptotic Q_ρ is a mixture of Chi-square distributions. The ρ parameter can be considered as the pair-wise correlation of the genetic effects coefficients.

5.2.8 Pathways enrichment analysis and gene-gene interactions

Functional analysis through gene-set enrichment was performed as in **Chapter 3** to elucidate variant effect mechanisms by identifying pathways affected by the variants with the strongest effects.

5.3 Results

5.3.1 Admixture and Population Structure

As reported in **Chapter 4**, PCA revealed that there were no systematic differences between the HIV-1 positive and negative subpopulations of Botswana. The genetics of Batswana showed a three-way admixture with genetic contributions from Khoe-San, Bantu-speakers

and European populations. However, no population has entirely homogeneous genetic architecture, therefore hidden relatedness and population structure have been accounted for in the current genetic association study.

5.3.2 Cross population meta-analysis of GWAS of susceptibility to HIV-1 acquisition

In total 265 HIV-1 cases (HIV-1 positive) and 125 healthy controls (HIV-1 negative) passed the QC procedure. The demographics of the study population are presented in **Table 7**.

Table 7. Study participants demographics.

Variable	HIV-1 negative	HIV-1 positive
Female (n, %)	93 (74.4)	264 (99.6)
Age (median, IQR)	*38.07 (36 - 40.00)	27.45 (23.72-31.85)
CD4+ T-cells at baseline (median, IQR)	NA	319.0 (205.0 - 474.5)
HAART (n, %)	NA	125 (52.97)
Education level (n, %)	NA	
Primary	NA	63 (26.69)
Junior	NA	123 (52.12)
Senior	NA	39 (16.53)
Tertiary	NA	5 (2.12)
Feeding	NA	
Breast feeding	NA	129 (54.67)
Formula feeding	NA	106 (44.92)
Marital status	NA	
Cohabiting	NA	20 (8.47)
Married	NA	24 (10.17)
Single	NA	185 (78.39)

* Age was available for 41/125 controls only.

A meta-analysis of the genetic association of susceptibility to HIV-1 acquisition was performed using Botswana and Uganda GWAS summary statistics. A total of 2,642,533 common variants between the Botswana and Uganda data were tested, resulting in a cut-off p-value of 1.89 x 10^{-8}. We transformed the effect sizes from EMMAX to odds ratio (OR) using LMOR (Lloyd-Jones et al., 2018). None of the tested variants reached the statistical significance (**Table 10**). The top effects were found within the *SV2B*, *IL20RA**IL22RA2* and *LINC00578**LINC02015*.

Table 8. The strongest effects of a meta-analysis of GWAS of susceptibility to HIV-1 acquisition.

CHR	BP	Variant	MAF_{BW}	GENE	OR	SE	P-value
15	91295826	rs7169918	0.328	SV2B	2.88	0.036	5.07×10^{-7}
6	137068291	rs56707550	0.235	IL20RA\ IL22RA2	0.15	0.041	7.34×10^{-7}
3	177789618	rs1477000145	0.383	LINC00578\ LINC02015	2.81	0.034	1.33×10^{-6}

CHR: chromosome, BP: genomic location in base pairs, MAF_{BW}: minor allele frequency in the current study of the Botswana population, SE: standard error of the odds ratio.

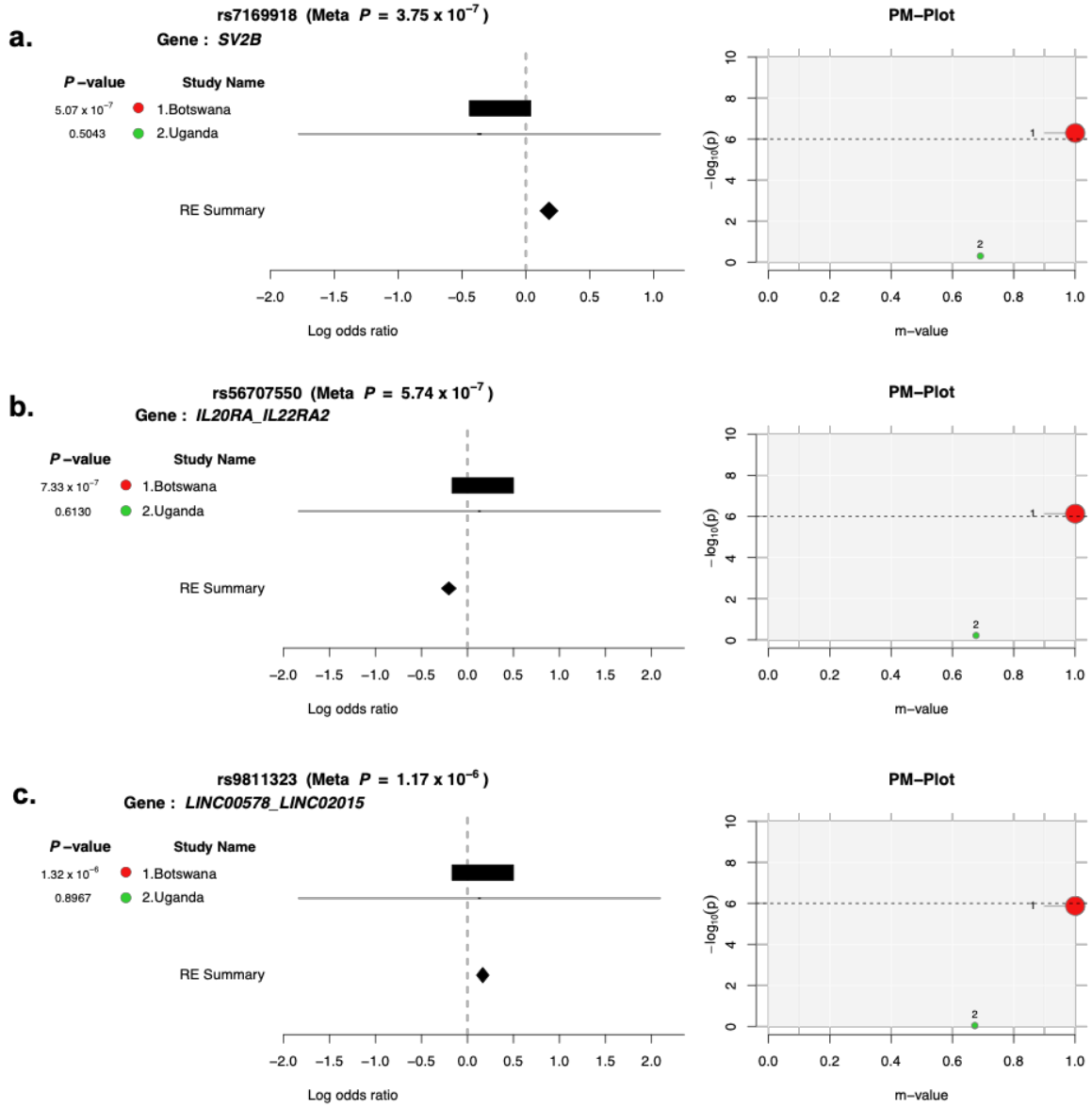


Figure 16. Meta-analysis results of rs7169918, rs56707550 and rs9811323 variants displayed in a ForestPMPlot that shows Forest Plot (Left) and PM Plot (Right).
The Forest Plot shows the p-value, study name, log odds ratio and its standard error and random effect summary statistic. The PM Plot displays -log(p-value) plotted against m-values

of the studies. An m-value > 0.9 means that the variant has an effect in the particular study. The size of the dot represents the sample size of the study.

5.3.3 Burden and rare-variant association test

A total of 26,935 genes were used in the SKAT-O test, leading to a multiple test correction p-value cut-off of 1.86×10^{-6}.

5.3.3.1 Aggregate rare-variant association of susceptibility to HIV-1 acquisition

No variant set reached statistical significance when controlling for possible confounding. With a SKAT univariate model including only HIV-1 status against the variant sets, we identified 194 variant sets with a p-value less than 0.01. The top effects included *Tet Methylcytosine Dioxygenase 1* (*TET1*) and 4 RNA genes which had not been reported previously (**Table 9**) in the GWAS of HIV-1.

Table 9. The strongest effects of the rare-variant association test of susceptibility to HIV-1 acquisition.

CHR	START	END	BAND	Markers	GENE	CLASS	Q	P-value
10	68560337	68694487	10q21.3	33	*TET1*	Known	22.4	2.74×10^{-5}
6	164083809	164088302	6q26	3	*LOC105378106*	Novel	16.8	4.25×10^{-5}
11	94545330	94740356	11q21	22	*LOC105369438*	Novel	21.9	5.78×10^{-5}
11	42209293	42275240	11p12	24	*LOC100507205*	Novel	38.7	8.83×10^{-5}
2	6728177	6770311	2p25.2	7	*LINC00487*	Novel	1.59	9.61×10^{-5}

CHR: chromosome, START: base pair position at the beginning of the gene region, END: base pair position at the end of the gene region, BAND: cytogenic band. Markers: number of variants used in the test, Q: SKAT test statistic. CLASS: gene previously associated with HIV-1 (known) or not (novel).

5.3.3.2 Aggregate rare-variant association of HIV-1 progression

Three sets of rare variants within the *Ankyrin Repeat Domain 39* (*ANKRD39*), *LOC105378523* and *General Transcription Factor IIIC Subunit 3* (*GTF3C3*) were statistically significant. Among the top 5 effects were the Metaxin (*MTX3*) and *Eukaryotic Translation Initiation Factor 3 subunit K* (*EIF3K*) genes, though not statistically significant. These genes have not been previously reported in the GWAS of HIV-1 (**Table 10**).

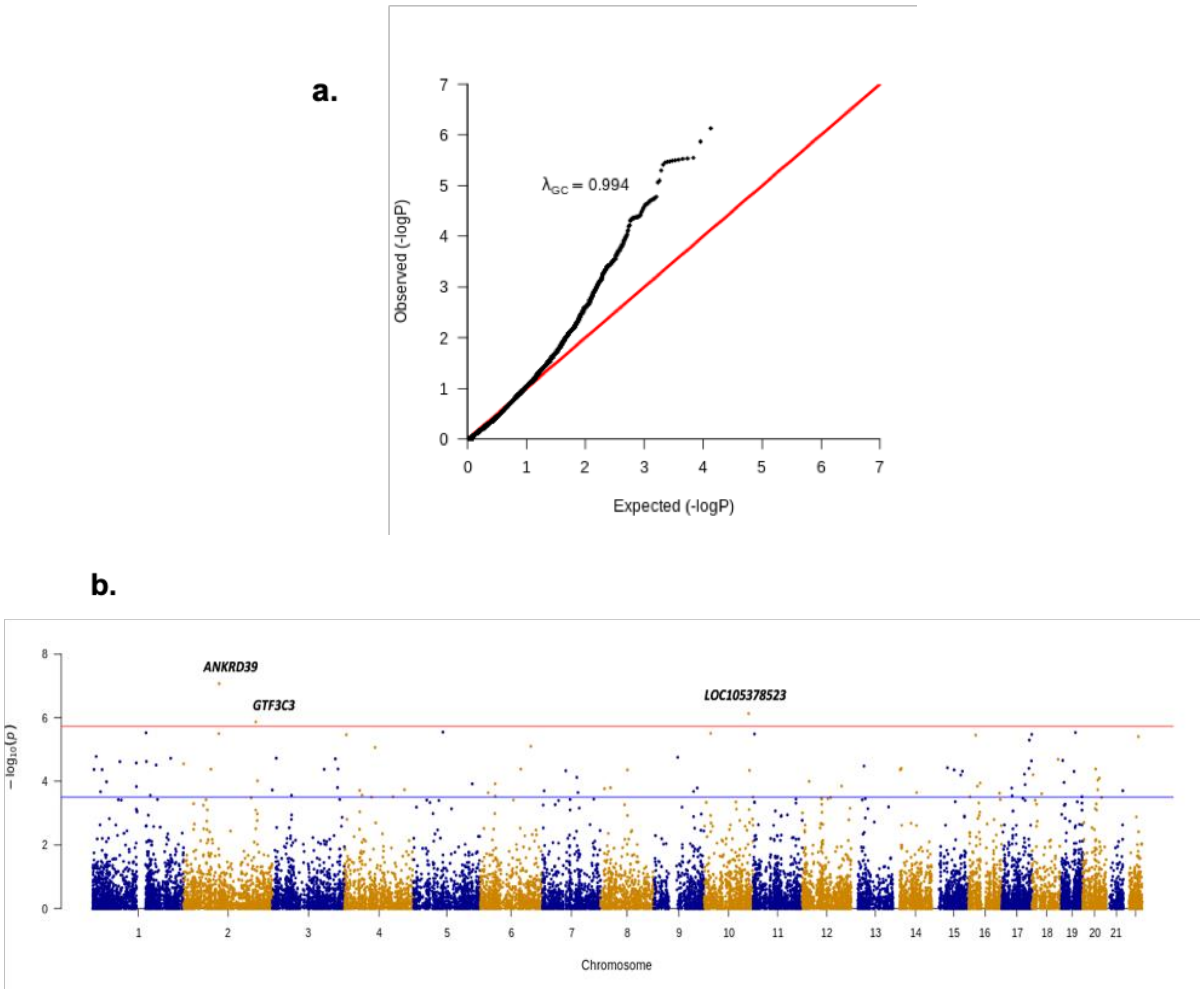


Figure 17. Quantile-quantile plot (with lambdaGC) and Manhattan plot of HIV-1 rare-variant association of HIV-1 progression.

a. Quantile-quantile plot and λ_{GC} of HIV-1 rare-variant association of HIV-1 progression showing -$\log_{10}$ p-value of each variant is plotted against the expected null (the red line). **b.** Manhattan plot of HIV-1 rare-variant association of HIV-1 progression showing -$\log_{10}$ p-value of each variant plotted against its genomic position. The red line is the -$\log_{10}$ p-value cut-off ($1.86x\ 10^{-6}$).

Table 10. The strongest effects of the rare-variant association of HIV-1 progression.

CHR	START	END	BAND	Markers	GENE	CLASS	P-value
2	96836611	96858016	2q11.2	2	*ANKRD39*	Novel	8.48×10^{-8}
10	121291608	121322945	10q26.12	3	*LOC105378523*	Novel	7.45×10^{-7}
2	196763035	196799725	2q33.1	7	*GTF3C3*	Novel	1.36×10^{-6}
5	79976716	79991262	5q14.1	1	*MTX3*	Novel	2.84×10^{-6}
19	38619082	38636955	19q13.2	1	*EIF3K*	Novel	2.92×10^{-6}

CHR: chromosome, START: base pair position at the beginning of the gene region, END: base pair position at the end of the gene region, BAND: cytogenic band. Markers: number of variants used in the test, Q: SKAT test statistic. CLASS: previously associated with HIV-1 (known) or not (novel) in a GWAS.

5.3.4 In-silico functional analysis of prioritized variants

To enrich for biological processes and pathways affected by the top identified variants from genetic association of progression to HIV-1. The genetic association candidate genes were used to retrieve more related genes that are predicted to interact with the candidate genes using GeneMANIA. The putatively affected biological processes and pathways were then identified using a suite of databases in Enrichr (**Figure 18**, **Table 11**).

5.3.4.2 Functional analysis of the strongest effects of genetic association of HIV-1 progression through gene-set enrichment

In addition to the 5 candidate genes identified in common and rare-variant association of HIV-1 progression, 20 more related genes (grey circles without stripes) were retrieved through a gene-gene network (**Figure 18**). These genes were significantly (p-value < 0.05) enriched for the following biological processes: viral translation [(GO:0019081), p-value = 9.10×10^{-16}], transcription from RNA polymerase III promoter [(GO:0006383), 9.46×10^{-11}], and Cytoplasmic translational initiation [(GO:0002183), p-value = 2.51 x 10-6].

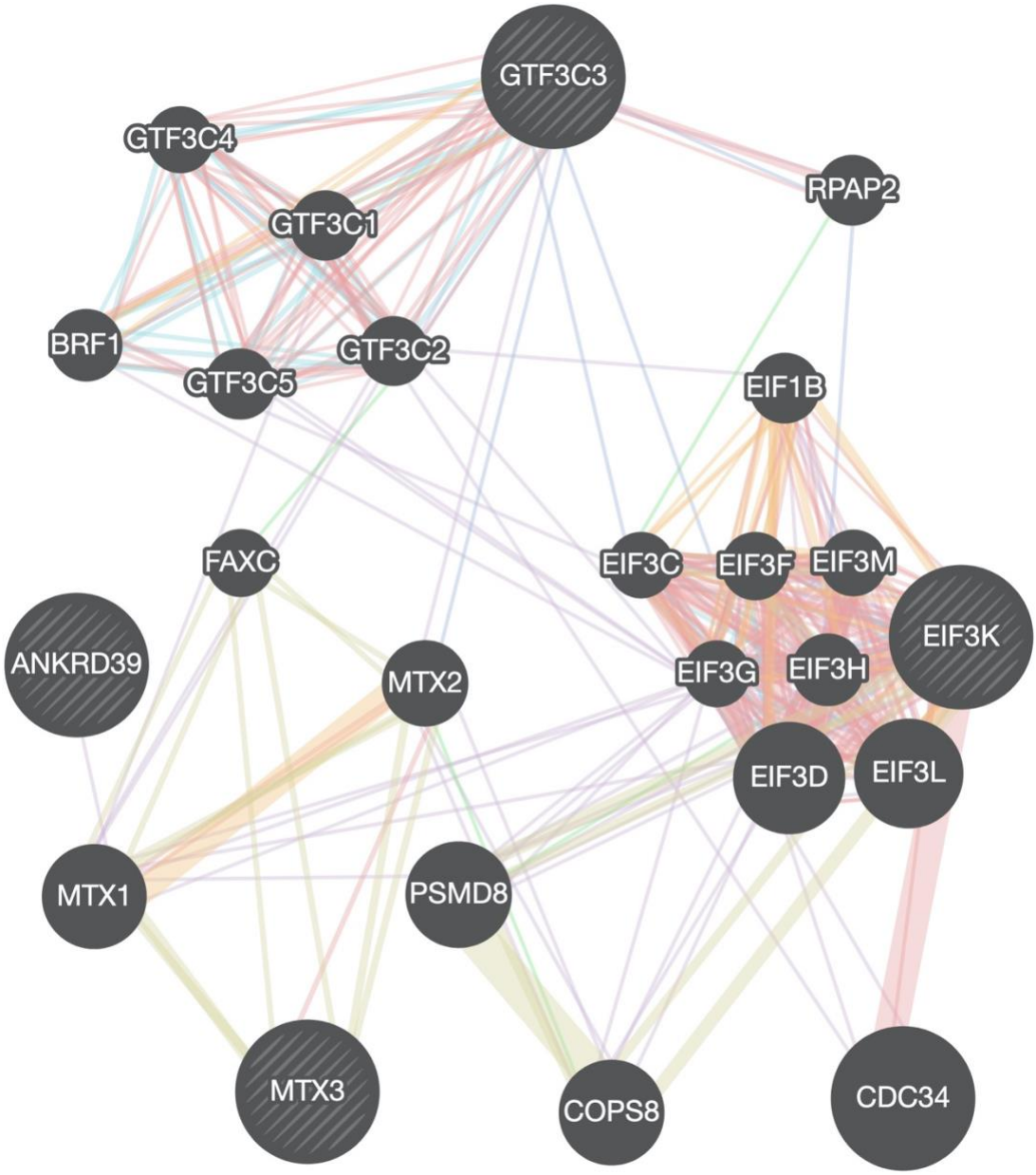


Figure 18. Gene interaction network of candidate genes identified from genetic association of HIV-1 progression.

The different colours of branches represent how the genes are related; pink: physical interactions, purple: co-expression, orange: predicted, navy blue: co-localization, blue: Pathway, green: Genetic interactions, yellow: shared protein domains. Black and stripped nodes: genes provided as input into GeneMANIA (Table 10). Black only nodes: genes predicted by GeneMANIA to interact with the input list.

Table 11. Enrichr gene-set enrichment of the candidate genes of HIV-1 progression.

	P-value	P-value$_{adj}$	Database
Gene Ontology			
Viral translation (GO:0019081)	3.56×10^{-19}	9.10×10^{-16}	Biological Process 2018 (Harris et al., 2004) http://www.informatics.jax.org/
Transcription from RNA polymerase III promoter (GO:0006383)	1.30×10^{-13}	9.46×10^{-11}	
Cytoplasmic translational initiation (GO:0002183)	4.43×10^{-9}	2.51×10^{-6}	
RNA polymerase III transcription factor complex (GO:0090576)	1.21×10^{-18}	5.38×10^{-16}	Cellular Component 2018 (Harris et al., 2004) http://www.informatics.jax.org/
Transcription factor TFIIIC complex (GO:0000127)	7.94×10^{-17}	1.77×10^{-14}	
Translation initiation factor activity (GO:0003743)	1.57×10^{-37}	1.81×10^{-34}	Molecular Function 2018 (Harris et al., 2004) http://www.informatics.jax.org/
RNA binding (GO:0003723)	8.89×10^{-7}	3.41×10^{-4}	
Pathways			
RNA polymerase III transcription initiation from type 2 promoter	2.27×10^{-11}	4.29×10^{-9}	BioPlanet 2019 (Huang et al., 2019)
Translation Factors	2.84×10^{-22}	4.28×10^{-19}	WikiPathways 2019 Human (Slenter et al, 2017) BioCarta 2016 (http://www.biocarta.com)

P-value$_{adj}$: adjusted P-value.

5.4 Discussion and conclusion

Our study is the first to evaluate the role of rare-variants in susceptibility to HIV-1 and progression to disease in Botswana using a larger sample size (n $\geq$ 236), comparing to a previous study that had a cohort of 100 participants from Southern and Eastern African combined (at most 10 samples from Botswana) (Mackelprang et al., 2017). In our study, the cumulative effects of three sets of novel rare variants within the *ANKRD39* (8.48×10^{-8}), *LOC105378523* (7.45×10^{-7}) and *GTF3C3* (1.36×10^{-6}) genes were significantly associated with HIV-1 progression (**Table 10**).

The *ANKRD39* gene is not well characterized, however, it is a paralog of the *Cyclin Dependent Kinase Inhibitor 2D* (*CDKN2D*) that plays a role in regulation of cyclin and cell growth cycle (GeneCards, 2020). An artificial ankyrin repeat domain was found to negatively interfere with HIV-1 Gag assembly and budding (Nangola et al., 2012). In addition, a common variant within a cyclin (*CCNG1*) gene was found to be associated with HIV-1 disease progression in a previous GWAS of Botswana (Xie et al., 2017). This supports the finding of the current study and possibly means that the rare variants within the *ANKRD39* gene might contribute to the control of DNA synthesis, cell division and consequently replication of HIV-1 in the host's cells.

The *LOC105378523* gene is an uncharacterized lncRNA gene located within chromosome 10 (**Table 10**). The significant association of *LOC105378523* with HIV-1 progression in the current study is expected as lncRNAs are known to be involved in regulation HIV replication, transcription and post-transcription, this makes them potential biomarkers for HIV-1 progression and targets for HIV treatment (Chao et al., 2019; Lazar et al., 2016; Shen et al., 2020; Trypsteen et al., 2016). Lastly, the *GTF3C3* gene (**Table 10**) encodes a transcription factor TFIIIC that is involved in transcription of transfer ribonucleic acid (tRNA) and directly binds to virus-associated RNA promoters (GeneCards, 2020). The transcription factor TFIIIC was observed to mediate HIV-1 transcription in HeLa and Jurkat T cells (Jang et al., 1992).

The functional analysis results accentuates that the candidate genes that potentially associate with HIV-1 progression affect RNA polymerase III transcription initiation, viral transcription and translation (**Table 11**). The RNA polymerase III transcription initiation pathway is

represented by the *GTF3C* genes cluster (**Figure 18**). Upon integration into the human genome, HIV-1 RNA transcription is mediated by RNA polymerase II (Liu et al., 2014). Though the product of the *GTF3C3* gene has been linked to upregulation of HIV-1 RNA (Jang et al., 1992), in addition to mRNA synthesis, HIV-1 specific small nuclear RNAs can also be produced from RNA polymerase III promoters (Boden et al., 2003; Gunnery and Mathews, 1995; Ratnasabapathy et al., 1990). HIV-1 hijacks the host translation machinery to facilitate its viral proteins. The human *EIF3* gene products (**Figure 18**) are some of the targets for optimal HIV-1 translation (Guerrero et al., 2015). These findings therefore suggest that the genes are involved in the regulation of HIV-1 transcription and translation which are markers of progression.

No variant set reached statistically significance in the rare-variant association test of susceptibility to HIV-1 acquisition. The strongest effects were within the *TET1* (p-value = 2.74 x 10^{-5}), *LOC105378106* (p-value = 4.25 x 10^{-5}), *LOC105369438* (p-value = 5.78 x 10^{-5}), *LOC100507205* (p-value = 8.83 x 10^{-5}) and *LINC00487* (p-value = 9.61 x 10^{-5}) gene (**Table 9**). The product of the *TET1* gene is a demethylase that plays a role in DNA methylation and gene expression (GeneCards, 2020). TET (Ten-eleven translocation) family of proteins consists of three paralogs: TET1, TET2, and TET3 (Akahori et al., 2015). It has been shown that the HIV-1 Vpr protein facilitates the degradation of TET-2 to promote HIV-1 replication (Lv et al., 2018; Wang and Su, 2019).

Furthermore, the HIV-1 Vpr protein also promotes HIV-1 Env processing and infectivity of macrophages which are known to be less susceptible to HIV-1 than CD4+ T-cells (Wang and Su, 2019). This may imply that *TET1* gene plays a role in both HIV-1 progression and acquisition which supports the results of the current study that suggest that the *TET1* gene might be implicated in susceptibility to HIV-1 acquisition (**Table 9**). There has been a recent discovery of the association of a *CCR5* dependent lncRNA with susceptibility to HIV-1 (Kulkarni et al., 2019), therefore, though the variant sets did not reach statistical significance, the identification of rare variant sets within lncRNAs in the current study (**Table 9**) cannot be entirely overlooked.

To improve the power of the study, we performed a meta-analysis of susceptibility to HIV-1 acquisition results using 2,642,533 common variants between the Botswana and the Uganda datasets. Although the meta-analysis did not identify statistically significant results, several variants had an m-value greater than 0.9 in the current study (**Table 8, Figure 16**) which indicated that the variants had an effect in the current study. The strongest effects of the meta-analysis were rs7169918G (OR ± SE = 2.88 ± 0.036, p-value = 5.07 x 10^{-7}), rs56707550G (OR ± SE = 0.15 ± 0.041, p-value = 7.34 x 10^{-7}) and rs9811323C (OR ± SE = 2.81 ± 0.034, p-value = 1.33 x 10^{-7}) (**Table 8**). The results suggest that rs7169918G and rs9811323C alleles confer a 3-fold increase in the risk of acquiring HIV-1, while having rs56707550G reduces the risk by 85%.

The rs7169918 is found within the UTR-3' region of the *SV2B* gene in chromosome 15. The product of the *SV2B* gene (synaptic vesicle protein) is a neurotransmitter transporter that may play a role in the regulation of synaptic vesicle trafficking and exocytosis (GeneCards, 2020). SV2B is commonly known as a co-receptor for botulinum neurotoxins (Gustafsson et al., 2018) that mediate the toxin's entry into neurons. During HIV-1 infection, the HIV-1 protein transactivator of transcription (Tat) neurotoxin causes a disruption in neuronal synapsis through possibly misfolding of the SV proteins leading to autophagy and protein quality control (Mohseni Ahooyi et al., 2019). The *SV2B* gene has also been associated with neurodegenerative disorders and reduced neurotransmission in the retina (Morgans et al., 2009).

The rs56707550 variant is found within chromosome 6 in the intergenic region of the *IL20RA* and *IL22RA2* genes that encode type II cytokine receptors (**Table 8**). The *IL20RA* gene encodes a subunit of the interleukin (IL) 20 receptor and is highly expressed in skin (GeneCards, 2020). The *IL22RA2* gene encodes a protein that is a natural antagonist of IL-22 in regulation of inflammation (Xu et al., 2001). IL-22 belongs to the IL-10 family of cytokines and plays a major role in maintaining mucosal integrity and immunity (Aujla and Kolls, 2009). The cells that produce IL-22, the Th22 cells, express a high number of HIV co-receptors which means that IL-22 may also play a role in susceptibility to HIV-1. A reduction in IL-22 production has been linked to resistance to HIV infection in HIV exposed African and Chinese individuals (Chege et al., 2012; Hu et al., 2016). Conversely elevated levels of IL-22 have been linked to resistance

to HIV-1 in a European cohort (Missé et al., 2007). The protective effect of rs56707550 minor allele might indicate that this intergenic variant plays a role in enhancing the inhibition of IL-22 which leads to resistance to HIV-1 infection.

The rs9811323 variant intergenic to *LINC00578* and *LINC02015* genes had conflicting effects between the Botswana and the Uganda study (**Table 8**, **Figure 16**). However, the overall effect predicted by meta-analysis was that rs9811323C is a risk allele. The *LINC00578* and *LINC02015* genes are long intergenic ncRNAs located in chromosome 3. ncRNAs are often the target of copy number aberrations (Volders et al., 2018). The rs9811323 variant is located within a region that is characterized by a lot of structural variation and copy number variations in chromosome 3 (Karczewski et al., 2020; MacDonald et al., 2014). Therefore, it is possible that the rs9811323C points to some structural variations within this genomic region.

The current study has pinpointed some candidate lncRNA genes which are often linked to structural variation. Future research could also evaluate the effect of structural variations on HIV-1. The cumulative effects of rare variants within the *ANKRD39*, *LOC105378523* and *GTF3C3* genes were found to be significantly associated with HIV-1 progression. To the best of our knowledge, none of these genes have been previously associated with HIV-1 through GWAS. The sample size for this study was limited, this rendering the study underpowered for both the single-variant tests and the rare-variant aggregate test. Although additional data was not available at the time of analysis, in future larger studies or functional studies of the identified variants may provide evidence for the involvement of these variants on the aetiology of HIV-1 in the Botswana population. The genomic inflation factor (λ_{GC}) for the rare-variant association results was slightly below 1, however, the quantile-quantile plot suggested an inflation. This may indicate that population substructure may had been higher for rare-variants and the efforts to control for it with linear mixed models in SKAT-O were not completely successful. There is also need for a development of statistical test of rare-variant association that is designed to handle longitudinal phenotypic data as the current optimized rare-variant association tests are limited in this regard.

Chapter 6. General Discussion and Conclusion

6.1 Motivation

The unveiling of the human genome was a momentous milestone that contributed invaluable insights towards understanding of human physiology, disease causality and human evolution (International Human Genome Sequencing Consortium, 2001, 2004; Venter et al., 2001). The initial reference genome was 74.3% of a single individual, meaning this genome is not representative of the global genetic variation (Ballouz et al., 2019; International Human Genome Sequencing Consortium, 2001; Reich et al., 2009). Therefore there is on-going work to incorporate global diversity in the human reference genome (Ballouz et al., 2019; E pluribus unum, 2010). We used the current reference genome, GRCh38 (an update of GRCh37 or hg19), which is a mosaic of about 13 individuals, has more structural variation and genomic diversity, less gaps and 26.9% more exome coverage (Guo et al., 2017; Schneider et al., 2017). This means increased mapping rate and more accurate sequence analysis.

The human genome facilitated the uncovering of genetic diversity among populations (Schneider et al., 2017; Sudmant et al., 2015; The 1000 Genomes Project Consortium, 2010, 2012; The International HapMap Consortium, 2005). One of the insights gained from this endeavour was that global genomes are divergent with the African genomes having the highest diversity (Campbell and Tishkoff, 2008; Choudhury et al., 2018; The 1000 Genomes Project Consortium, 2010). The bulk of population genetics studies were performed in non-African regions where burden of disease is lower (Awany et al., 2018; Choudhury et al., 2017; Sirugo et al., 2019). Considering the underrepresentation of African populations in medical population genetics studies and the high burden of complex genetic diseases such as HIV-1, undertaking the current study could not be more apt.

6.2 Discussion of research highlights

It has been 15 years since the first genome-wide association study (GWAS) and these studies have proved to be a valuable tool in prediction of disease outcome. Stupendous progress has been made in the understanding of HIV-1 infection and pathogenesis through GWAS. In order to identify gaps to be filled within the field of HIV-1 GWAS, a comprehensive review of

literature from 2007 when the first HIV-1 GWAS was published was performed (**Chapter 2**). In the same publication, the implications of population genetic structure on HIV-1 phenotypes were discussed. We found out that although Southern Africa carries the highest burden of HIV-1, very few GWAS of HIV-1 have been conducted in the region. There is population structure and complex patterns of admixture between the indigenous populations and migrant populations of Southern Africa. This genetic diversity has to be considered for the region to also benefit from the advances in personalized medicine.

The three GWAS of HIV-1 discussed in **Chapter 2** were all conducted using the common-disease common-variants approach. Only one of these, the latest, was performed in a population of Botswana, identified 2 novel variants within the *HCG22* and *CCNG1* genes that were associated with HIV-1 acquisition and progression. One of the reasons that stood out from this review is that the common variants that were evaluated in the previous GWAS of HIV-1 probably could not pick causal variants in African populations due to the discrepancy in linkage disequilibrium (LD) patterns between populations. Due to this, we proposed to explore an unbiased method that can unravel population specific variation that may be of clinical relevance to HIV-1.

The analysis plan that was followed to analyse WGS data of Batswana and the main findings of this project are presented in **Figure 19**.

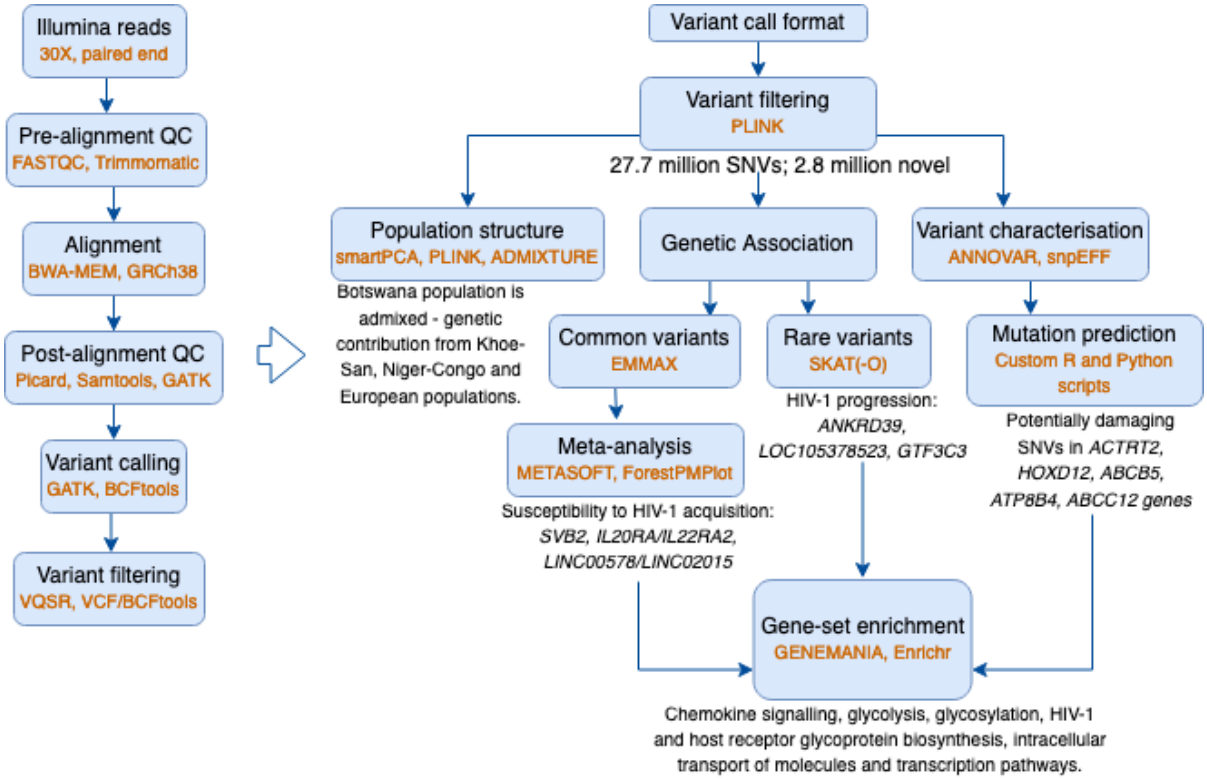


Figure 19. Whole genome sequence analysis plan of the Botswana population.
Computational analysis of 390 human whole genome sequences from Botswana was performed following the depicted analysis plan. The figure was generated using https://app.diagrams.net/.

Prior to undertaking a genetic association study, we endeavoured to map a complete population genome of Botswana and to characterize the genomic variations within the population. In **Chapter 3**, we presented the findings of analysing whole genome sequences (WGS) of 390 Batswana (the people of Botswana) using bioinformatics methods. WGS offers an opportunity to uncover every base in the genome, thereby circumventing the challenge of masked causal variants. Upon mapping the genomes to the human reference genome GRCh38, we determined genomic variations from raw WGS of high sequence depth (30X) using two state-of-the-art methods; Genome Analysis Toolkit (GATK) and BCFtools.

We present 27.7 million single nucleotide variations (SNVs) of which approximately 2.8 million were novel. To identify the novel variants we searched our discovered variants in dbSNP build 151, 1000 Genomes Project, African Genome Variation Project (AGVP) and The Genome Aggregation Database (gnomAD). Compared to previous studies prior ours, we added the most recently released gnomAD database. This ensured that we increase the accuracy of

identifying novel variants as AGVP and gnomAD databases carry the largest number in history of publicly available African genomic variation. Majority of the novel variants were low-frequency and rare-variants (minor allele frequency (MAF) < 0.05). This is expected as population-specific variants tend to be of low-frequency.

We also elucidated the general mutation burden in the genomes of Batswana. Here we identified the most deleterious variants within the *ACTRT2*, *HOXD12*, *ABCB5*, *ATP8B4* and *ABCC12* genes according at least 70% of the prediction tools. Since the *ACTRT2* is involved in cytoskeleton organization, the gene might be a target for remodelling by HIV-1 to facilitate the virus' entry into the host cells. Though the exact role of the *HOXD12* gene is unknown, other *HOX* genes are known to regulate development and have been linked to maintenance of HIV-1 latency. The products of *ABCB5*, *ATP8B4* and *ABCC12* genes are transport molecules, including drugs. Some *ABC* genes are involved in regulation of anti-HIV drug efflux, therefore the variants within the *ABCB5* and *ABCC12* genes may be of have pharmacogenetics relevance.

We constructed gene-gene networks of the most deleterious genes and retrieved additional genes that are predicted to interact with the most deleterious genes. We performed gene-set enrichment to identify biological processes and pathways that may be affected by the most deleterious genes and their interactomes. The most affected pathways were glycolysis, gluconeogenesis, tri-carboxylic acid cycle and pentose-hexose pathways. Pyruvate produced from glycolysis is a substrate for TCA where acetyl-CoA, a precursor of cholesterol, is produced (Berg et al., 2002a). Cholesterol is required for plasma membrane formation and integrity. Furthermore cholesterol is required for viral fusion to the host's cell membrane for entry and virus release following assembly and maturation (egress) (Coomer et al., 2020). Oxidation of cholesterol to 25-hydroxycholesterol can block HIV-1 cell entry (Liu et al., 2013b). In addition, HIV-1 infection upregulates glycolysis to meet the demands of viral replication (Hegedus et al., 2014; Palmer et al., 2016; Valle-Casuso et al., 2019). Still in **Chapter 3** we also found out that cancer was among affected pathways. This is expected as increased glycolysis is also a "hallmark" of cancer (DeBerardinis and Chandel, 2016; Huang et al., 2014). Our results are timely and can contribute to the emerging field of immunometabolism in which

therapy against HIV-1 infection is being evaluated through reduction of glycolysis and inhibition of cholesterol (Liu et al., 2013b; Taylor and Palmer, 2020; Valle-Casuso et al., 2019).

In polygenic and complex diseases such as HIV-1, pathogenic variants usually occur at a very low frequency and these are usually population specific. To determine the distribution of potentially pathogenic SNVs in previously reported HIV-1 associated genes (**Chapter 3**), we observed lower proportions of SNVs in the *SCFD1*, *HIST1H4B*, *HIST1H4A* and *ZDHHC19* genes, and higher proportions of SNVs in *IGSF21* and *NCBP2* genes in HIV-1 positive individuals. What was remarkable with this observation is that the *ZDHHC19* gene has been linked to HIV-1 susceptibility in a population of Malawi. The *ZDHHC19* gene is involved in palmitoylation of STAT3 that has been implicated in HIV-1 related inflammation and immune responses (Del Cornò et al., 2014; Liu et al., 2013a; Percario et al., 2003). Identifying the *ZDHHC19* gene for the second time as a HIV-1 associated candidate gene warrants its further investigation as a potential target in HIV-1 pathogenesis in African populations.

We observed a clustering of the Botswana population with other Niger-Congo populations, with a clear distinction from non-African populations. Batswana had closer affinity affinity with the Zulu population of South Africa. This finding was expected as Batswana are part of the Sotho-Tswana clan and a close affinity of the Sotho with the Zulu of South Africa has been observed previously (Choudhury et al., 2017). Our results of **Chapter 4** suggested that there was no overall substructure within Botswana population. However, the genetics of Batswana are not completely homogeneous as we observed a considerable spread and a level of the Khoe-San, Niger-Congo and European admixture within the population. Our results echo the knowledge of the historical relationships of the populations of Southern Africa. It is not surprising that we observed Khoe-San admixture in the population of Botswana as the country carries a large number of the people of Khoe-San ancestry. Moreover a recent study postulated that modern humans come from a Khoe-San woman who inhabited the prehistoric wetlands (Makgadikgadi-Okavango) in the northern part of Botswana (Chan et al., 2019). Although there is controversy around this recent finding, the Chan et. al. study may support our previous hypothesis of that the Northern San originated in Botswana as evidenced by the rock paintings at Tsodilo Hills in Northwestern Botswana (Thami and Chimusa, 2019).

We sought to evaluate the role of genetic variations in susceptibility to HIV-1 and progression to disease through meta-analysis of common-variants and rare-variant association test in **Chapter 5**. For the rare-variant association test we used a cumulative effects method implemented in SKAT(-O) (Lee et al., 2012a, 2012b), that control for confounding from covariates such as sex, age, population structure and cryptic relatedness. We further performed a meta-analysis of the Botswana (current study) and Uganda (Lingappa et al., 2011) HIV-1 GWAS.

The top effects from meta-analysis of the Botswana (current study) and Uganda (Lingappa et al., 2011) in **Chapter 5** were rs7169918G (ß ± SE = 2.88 ± 0.036, p-value = 5.07 x 10^{-7}), rs56707550G (ß ± SE = 0.15 ± 0.041, p-value = 7.34 x 10^{-7}) and rs9811323C (ß ± SE = 2.81 ± 0.034, p-value = 1.33 x 10^{-7}) within the *SV2B*, *IL20RA**IL22RA2* and *LINC00578**LINC02015* genes respectively. The results suggested that the rs56707550G variant confers 85% reduced susceptibility to HIV-1 infection, while rs7169918G and rs9811323C increase the risk of acquiring HIV-1 by a 3-fold. The *SV2B* gene encodes a well characterized co-receptor of botulinum neurotoxins that promotes the toxin's entry into neurons. In HIV-1 infection, synaptic vesicle proteins are involved in Tat-induced neurotoxicity.

Further investigations are needed to delineate whether SV2B acts as a co-receptor for HIV-1 induced toxins for neuron entry as well. Since a reduction in IL-22 production has been linked to resistance to HIV infection in HIV exposed African and Chinese individuals, we postulate that the rs56707550G variant prevents HIV-1 acquisition through the inhibition of IL-22. The rs9811323C intergenic to *LINC00578* and *LINC02015* genes is located within a genomic region that is characterized by numerous copy number variations (CNVs). Moreover, the *LINC00578* and *LINC02015* genes are located within chromosome 3 which harbours chemokine (C-C motif) receptor 5 (CCR5). CCR5 is a predominant co-receptor of HIV-1. A 32 base pair deletion of the *CCR5* gene confers resistance to HIV-1 acquisition. Interestingly, lower copy number of the C-C Motif Chemokine Ligand 3 Like 1 (*CCL3L1*; MIP-1αP) gene, that encodes a natural ligand of CCR5, has been associated with increased risk of acquiring HIV-1 (GeneCards, 2020; Gonzalez et al., 2005; Liu et al., 2010). Therefore, the rs9811323C variant might contribute to HIV-1 resistance through indirect regulation of the expression of the *CCR5* gene.

We have previously posited that some of the missing heritability of HIV-1 might be solved through rare-variant association test (Thami and Chimusa, 2019; Verma and Ritchie, 2018). As mentioned in Chapter 3, the majority of the novel variants that we identified were rare-variants. Since rare-variants are usually of small effect, still in **Chapter 5**, we evaluated the cumulative effects of sets of rare-variants in relation to susceptibility to HIV-1 and progression. We used a stringent multiple-test correction cut-off and no variant reached statistical significance in the rare-variant association test of HIV-1 susceptibility. In assessing the role of genomic variations on HIV-1 progression we used depletion of CD4+ T-cells over at least 18 months as a marker of progression. For the rare-variant association test of HIV-1 progression (decreasing CD4+ T-cell counts), we identified sets of novel rare variants within the *ANKRD39* (8.48×10^{-8}), *LOC105378523* (7.45×10^{-7}) and *GTF3C3* (1.36×10^{-6}) genes that reached statistical significance. We postulate that the *ANKRD39* gene might be involved in the cell growth cycle as its paralog *CDKN2D*. If this is true, then the rare variants within the *ANKRD39* gene contribute to the regulation of the division of HIV-1 infected cells and consequently HIV-1 replication. Similarly, lncRNAs and the *GTF3C3* gene that encodes a transcription factor TFIIIC have been implicated in the regulation of HIV-1 transcription and translation which are markers of progression.

In general, the candidate genes prioritized in this thesis could potentially have a bearing in the following:

1. ***HIV-1 susceptibility***: The chemokine and the IL-22 receptor signalling pathways. Chemokines and interleukin (such as IL-22) receptors facilitate or block HIV-1 entry into the host's cells.
2. ***HIV-1 progression***: RNA polymerase III transcription initiation and translation. HIV-1 hijacks the host's translation and transcription for its optimal replication.
3. ***HIV-1 pharmacovigilance***: Variants within ABC transporters and the immunoglobulin superfamily may affect the efflux of HIV-1 drugs.

6.3 Study limitations

1. The samples of this project were collected in the southern part of Botswana therefore the results may not be extrapolated to the rest of the country. However, sample

collection was done in the capital city, a town and major villages. We expect that these locations, particularly the capital city, to be cosmopolitan and we assume that it is an amalgam of different ethnicities.

2. Computational infrastructure: the size of our WGS data was initially 40TB and increased exponentially when we commenced the analysis. Our data often exceeded our memory allocation, this compelled us to delete data as we advanced with the analysis (which could be inconvenient when you have to revisit certain sections of the workflow) and also delayed some stages of the data analysis.

6.4 Future perspectives and recommendations

1. Identifying an SNV rs9811323C located within lncRNA genes in a copy number variations (CNVs) dense region of the genome, hinted to us that structural variations (SVs) may be involved in the control of HIV-1 entry in our study population. We could not ascertain this suggestion since detection of SVs was beyond the scope of our study. Therefore, it would be beneficial to mine SVs from the whole genome sequences of Batswana in order to interrogate their relationship with HIV-1. However, identification of SVs from WGS remains challenging until methods developed for this purpose are improved.
2. Although additional data was not available at the time of analysis, in future larger studies or functional studies of the identified variants may provide evidence for the impact of these variants on the aetiology of HIV-1 in the Botswana population.
3. A genomic reference panel assembled entirely of African population variations is needed to improve the accuracy of variant calling in future studies of African genomics.
4. There is need for a development of statistical test of rare-variant association that is designed to handle longitudinal phenotypic data as the current optimized rare-variant association tests are limited in this regard.
5. There is need to acquire larger and faster computing infrastructure that would foster the success of future large-scale genomics studies in Africa.

6.5 Concluding remarks

This thesis commenced with a review that aimed to fill the knowledge gaps in the understanding of the genetic diversity of Southern African people and the implications of the genetics on susceptibility to HIV-1 and progression. We narrowed our scope to characterize the whole genome sequences of individuals from Botswana and evaluate the role of genomic variations on susceptibility to HIV-1 and progression through bioinformatics methods. We identified novel variants within the population of Botswana, some of which were significantly associated with HIV-1 progression. These novel findings contribute a deeper understanding of the African genomic variations and their implications in the medical genetics field towards combating HIV-1.

Appendices

Appendix A
Supplementary information

Supplementary Table 1. Genetic polymorphisms significantly associated with HIV-1 acquisition, set point and progression in GWAS

SNP	Chr	Gene	HIV-1 phenotype	Effect	Effect size	(P-value)	Population (n)	References
rs2395029	6	*HCP5*	VL set-point	Low VL	β = -1.0	(9.36E-12)	Europeans Caucasians (n = 486)	(Fellay et al., 2007)
rs9264942	6	*HLA-C*	VL set-point	Low VL	β = -0.39	(3.77E-9)		
rs9261174	6	*ZNRD1*	VL set point	Low VL		(7.11E-3)		
rs9261174	6	*ZNRD1*	Progression	Delayed progression		(3.89E-7)		
rs2395029	6	*HCP5*	Progression Reservoir	Low VL Low reservoir	β = -0.540	(672E-07)	Europeans (n = 605)	(Dalmasso et al., 2008)
rs13199524	6	*TNXB*	Progression Reservoir	Low VL	β = 0.255	(5.70E-05)		
rs3093662	6	*TNF*	Progression Reservoir	Low VL reservoir	β = -0.247	2.18E-05		
rs12198173	6	*TNXB*	Progression Reservoir	Low VL Low reservoir	β = 0.240	1.28E-04		
rs6503919	17	*DDX40 YPEL2*	Reservoir	Low reservoir	β = -0.211	2.00E-06		
rs2575735	8	*SDC2*	Reservoir	Low reservoir	β = -0.176	1.34E-06		
rs2395029	6	*HCP5*	VL set-point	Low VL	-	4.48E-35	Caucasian (n = 2554)	(Fellay et al., 2009)
rs9264942	6	*HLA-C*	VL set-point	Low VL	-	5.85E-32		
rs259919	6	*C6orf12*	VL set-point	-	-	5.3E-04		
rs9468692	6	*TRIM10*	VL set-point	Low VL	-	7.6E-04		
rs9266409	6	*HLA-B*	VL set-point	Low VL	-	4.86E-14		
rs8192591	6	*NOTCH4*	VL set-point	Low VL	-	9.02E-09		
rs2395029	6	*HCP5*	Progression	Delayed progression	-	1.20E-11		
rs9264942	6	*HLA-C*	Progression	Delayed progression	-	6.40E-12		
rs9261174	6	*ZNRD1*	Progression	Delayed progression	-			
rs3869068	6	*ZNRD1*	Progression	Delayed progression	-	1.80E-08		
rs2074480	6	*RNF39*	Progression	Delayed progression	-	1.80E-08		
rs7758512		*ZNRD1*	Progression	Delayed progression	-	1.80E-08		
rs9261129	6	*HCG8, ZNRD1*	Progression	Delayed progression	-	1.80E-08		
rs2301753	6	*RNF39*	Progression	Delayed progression	-	1.80E-08		

rs2074479	6	*RNF39*	Progression	Delayed progression	-	1.80E-08		
rs2395029	6	*HCP5*	Set-point	Low VL	OR = 3.47	6.79E-10	Europeans (n = 275 nonprogressors, 1352 controls)	(Limou et al., 2009)
rs1245371	6	*RNF39*	Progression	Delayed progression	-	9.21E-7		
rs9368699	6	*C6orf48*	Set-point	Low VL	-	1.84E-11		
rs3823418	6	*PSORS1C 1*	Set-point	Low VL	-	1.4E-8		
rs2248462	6	*MICB*	Set-point	Low VL	-	4.26E-8		
rs2516509	6	*MICB*	Set-point	Low VL	-	4.95E-8		
rs10484554	6	*HLA-C*	Set-point	Low VL	-	6.27E-8		
rs3815087	6	*PSORS1C 1*	Set-point	Low VL	-	1.46E-7		
rs259940	6	*ZNRD1*	Progression	Delayed progression	-	2.04E-6		
rs4118325	1	*Intergenic*	Progression	Delayed progression	OR = 0.24	6.09E-7	Europeans (n = 1352 controls; 85 rapid progressors)	(Le Clerc et al., 2009)
rs1522232	12	*SOX5*	Progression	Delayed progression	OR = 0.45	1.80E-6		
rs1360517	9	*Intergenic*	Progression	Rapid progression	OR = 3.09	3.27E-6		
rs3108919	8	*Intergenic*	Progression	Rapid progression	OR = 2.13	3.86E-6		
rs10800098	1	*RXRG*	Progression	Rapid progression	OR = 3.29	3.86E-6		
rs10494056	1	*Intergenic*	Progression	Delayed progression	OR = 0.27	4.29E-6		
rs12351740	9	*Intergenic*	Progression	Rapid progression	OR = 3.46	6.63E-6		
rs1020064	2	*TGFBRAP1*	Progression	Delayed progression	OR = 0.34	7.04E-6		
Haplotype of rs17762192, rs17762150 and rs1367951	1	*PROX1*	Progression	Delayed progression	RH = 0.69	6.23E-7	European Americans (n = 156)	(Herbeck et al., 2010)
rs16899646	6	*HLA-B*	Progression	Delayed progression	-	1.33E-5		
rs2248462	6	*Intergenic*	Set-point	Low VL	-	1.16E-2		
rs2516422	6	*MICB*	Set-point	Low VL	-	5.12E-4		
rs2395034	6	*MICB*	Set-point	Low VL	-	5.13E-4		
HLA-B*5703	6	*HLA-B*	Set-point	Low VL	-	5.60E-10	African Americans (n = 515)	(Pelak et al., 2010)
rs2523608	6	*HLA-B*	Set-point	Low VL	-	2.29E-6		
rs9264942	6	*HLA-C*	VL set-point	Low VL	OR[a] = 2.9	2.8E-35	Europeans (n = 1712)	(Pereyra et al., 2010)
rs2395029	6	*HCP5*	VL set-point	Low VL	OR[a] = 5.3	9.7E-26		
rs4418214	6	*MICA*	VL set-point	Low VL	OR[a] = 4.4	1.4E-34		
rs3131018	6	*PSORS1C 3*	VL set-point	Low VL	OR[a] = 1.5	4.2E-16		
rs2523608	6	*HLA-B*5703*	VL set-point	Low VL	OR[a] = 2.6	8.9E-20	African Americans (n = 1233)	

rs2255221	6	*HCP5*	VL set-point	Decreased VL	OR[a] = 2.7	3.5E-14		
rs2523590	6	*DHFRP2*	VL set-point	Low VL	OR[a] = 2.3	1.7E-13		
rs9262632	6	*HCG22*	VL set-point	Low VL	OR[a] = 3.1	1.0E-8		
rs2523590	6	*DHFRP2*	VL set-point	Low VL	OR[a] = 2.5	8.3E-8	Hispanics (n = 667)	
rs2395029	6	*HCP5*	Progression	Delayed progression	OR[a] = 3.14	8.54E-15	Europeans and Caucasians (n variable in cases, not > 270)	(Le Clerc et al., 2011)
rs9368699	6	*C6orf48*	Progression	Delayed progression	OR[a] = 2.9	3.03E-10		
rs8192591	6	*NOTCH4*	Progression	Delayed progression	OR[a] = 2.32	9.08E-27		
rs2072255	17	*RICH2*	Progression	Rapid progression	OR = 0.43	3.30E-26		
rs11884476	2	*PARD3B*	Progression	Delayed progression	RH = 0.3	3. 37E-9	European Americans (n = 755)	(Troyer et al., 2011)
CCR5-Δ32	3	*CCR5*	Acquisition	Resistance to HIV-1	OR = 0.2	5.0E-9	(n = 2,173)	(McLaren et al., 2013)
rs2535307	6	*HCG22*	Progression & acquisition	Rapid progression and Increased susceptibility	-	3.72E-7	Southern Africans (n = 556)	(Xie et al., 2017)
kgp22385164	5	*CCNG1*	Progression	Rapid progression	-	1.88E-6		(Xie et al., 2017)

Chr: chromosome, -: information not available, VL: viral load, OR: odds ratio, RH: relative hazard, ORa : odds ratio given for the minor allele where OR > 1 is means a protective effect.

Table A1. Human genes previously associated with HIV-1

HIV-1 associated genes
A4GALT, ABCB1, ABCF2, ABO, ABTB2, AC005062.1, AC006305.1, AC009132.1, AC009229.1, AC009271.7, AC010476.2, AC010519.1, AC010998.3, AC011306.1, AC013727.1, AC018880.2, AC021573.1, AC022483.1, AC023798.16, AC023950.6, AC025614.2, AC026341.1, AC026371.1, AC034229.1, AC037459.4, AC044836.1, AC062039.1, AC063952.2, AC073475.1, AC087190.5-2, AC087854.1, AC091905.4, AC092745.2, AC092745.5, AC093801.2, AC095058.3, AC096719.1, AC097652.1, AC099795.1, AC100812.1, AC103881.1, AC104232.2, AC105916.1, AC114322.1, AC115283.1, AC116362.1, AC126407.1, ACTR3BP6, ADAM10, ADAM18, ADAMTS1, ADH5P4, AE01, AF233439.1, AGAP2, AGBL5, AJ239318.1, AJ239321.1, AKR7A2, AKT1, AL008729.1, AL031315.1, AL035246.1, AL035461.3, AL096854.1, AL136363.2, AL138752.2, AL138889.1, AL161781.1, AL353743.2, AL354694.1, AL358787.1, AL359382.2, AL390763.1, AL391500.13, AL391832.4, AL451007.2, AL451127.2, AL512329.2, AL513164.1, AL591509.5, AL671762.1, AL671883.2, AL671883.3, ALK, ALKBH8, ANKRD22, ANKRD30A, ANKRD43, ANKRD6, ANKRD9, ANXA1, AOAH, AP001021.1, AP001021.3, AP003398.2, AP003464.1, AP2M1, APOBEC3, APOBEC3B, APOBEC3G, ARF1, ARGLU1, ARHGAP32, ARHGEF12, ARHGEF19, ARPC1A, ASXL2, ATG12, ATG16L2, ATG7, ATP6V0A1, BAHD1, BCL9, BICRA, BIRC4BP, BOD1P, BRWD2, BSDC1, BTNL2, BUD13, BX323046.2, C10orf11, C10orf71, C1ORF103, C21orf96, C3ORF56, C4orf17, C6ORF1, C6orf106, C6orf15, C6orf48, C7orf58, CACNG1, CADM1, CADPS, CAPN6, CARD16, CAV2, CBS, CCBE1, CCDC134, CCL11, CCL17, CCL18, CCL2, CCL3, CCL3L1, CCL4, CCL5, CCNG1, CCNT1, CCR2, CCR5, CCRL2, CCT8L2, CD209, CD247, CD33, CD4, CDCA7L, CDSN, CHORDC1, CHRNA3, CHRNA5, CIG-5, CLDND1, CLEC18B, CLN3, CLNS1A, CMPK2, CMTM8, COG2, COG3, COG4, COLEC11, COX10-AS1, COX6A1P3, CREB5, CREBBP, CRIPAK, CRTC2, CRTC3, CSPP1, CTDP1, CTLA4, CUL5, CUX2, CX3CR1, CXCL10, CXCL11, CXCL12, CXCL9, CXCR4, CXCR6, CYP1B1-AS1, CYP3A4, CYP7B1, CYPA, CYYR1, DAB1, DARC, DC-SIGN, DCK, DDEF2, DDOST, DDR1, DDX10, DDX3X, DDX40, DDX53, DDX55, DEFB1, DEPDC5, DEPDC6, DEPTOR, DHFRP2, DHX33, DIMT1L, DISC1FP1, DMXL1, DNAJB1, DNAJC18, DNAJC27, DNAJC5B, DNAL1, DOK6, DPCR1, DPM1, DPY19L3, DPYD, DRGX, DYRK1A, DYSF, ECR777, EDNRA, EFEMP1, EFHC2, EGF, EGFR, EIF2C3, EIF3H, EPHA5, EPS8, ERCC3, ERI2, ERP27, ETF1, ETHE1, EVI5L, EXOSC3, EXOSC5, FA2H, FAM174B, FAM200B, FAM229A, FAM5B,

FAM76B, FANCL, FAS, FASLG, FBN3, FBXO10, FBXO18, FBXO21, FBXW11, FCGR2A, FGD6, FGF1, FHL3, FKBP1AP4, FLII, FNTA, FOXCUT, FOXN3, FRMPD1, FUT2, FUT9, GABARAP, GALNT14, GAPVD1, GBAS, GBP1, GCK, GFRAL, GLRX3, GLTSCR1, GNPDA2, GOLPH3, GOSR2, GPC5, GPR156, GRIN2A, GRM5, GRTP1, GZMH, H3F3A, HAP1, HB1, HCG22, HCP5, HCP5HCP5, HCP5P2, HCRTR2, HEATR1, HGS, HIBCH, HIP1R, HIST1H3B, HIST1H3C, HIST1H4A, HIST1H4B, HIV-1, HLA-A, HLA-B, HLA-B57, HLA-C, HLA-DPA1, HLA-DQA1, HLA-DQB1, HLA-DRA, HLA-E, HLA-G, HMGXB3, HNRNPF, HS3ST3A1, HS6ST2, HSP90AB3P, HTATSF1, HUWE1, IDH1, IER3, IFI44, IFI6, IFIT3, IFNAR1, IFNG, IFNGR1, IGHMBP2, IGHV1-12, IGHV1OR21-1, IGHV3-13, IGHV3-52, IGHV3-53, IGSF21, IKBKG, IL10, IL10RA, IL12A, IL12B, IL13, IL2, IL2RA, IL32, IL4, IL4R, IL6, INTS7, IPO8, IQCA1L, IQUB, IRF4, ISG43, ITPKA, JAK1, JHDM1D, JUP, KAT2A, KB-67B5.12, KBTBD7, KCNIP1, KCNIP3, KCNK9, KCNMB3P1, KCNQ5, KDM3B, KDM4C, KDM4D, KEL, KERA, KIAA1012, KIF3C, KIF4B, KIR, KIR3DL1, KLHDC2, KLHL1, KTN1, LAPTM5, LARS, LARS2, LCP2, LEFTY1, LENG1, LINC00836, LINC00937, LINC00992, LINC01150, LINC01556, LINC01985, LINC02177, LINC02240, LINC02364, LINC02374, LINC02477, LINC02646, LINC02647, LINC02667, LINC02702, LINC02748, LNX2, LOC100129699, LOC375190, LPL, LRMDA, LRP4, LRRC58, LRRC8D, LSM3, LY6D, LYPD4, MAD2L1, MAGI1, MAP4, MBL2, MBNL2, MCM8, MDN1, MED14, MED28, MED4, MED6, MED7, MEPE, MESTP3, MGAT1, MICA, MICB, MID1IP1, MIR1275, MIR6891, MIR8074, MKI67, MKRN2, MKRN3, MMADHC, MND1, MOB1B, MOS, MPHOSPH6, MR1, NAV2, NBEA, NBPF13P, NCBP2, NCBP2-AS1, NCBP2-AS2, NCOR2, NDUFB7, NEDD9, NF2, NGLY1, NIPSNAP3B, NKG7, NLRP1, NMT1, NOS3, NOTCH4, NR0B2, NTM, NTM-AS1, NUFIP1P1, NUP107, NUP133, NUP153, NUP155, NUP160, NUP85, OAS1, OAS2, OASL, ODZ4, OTUD3, PABPC1P2, PANK1, PARD3B, PAX5, PBX1, PC, PCDH11X, PCNT, PDCD5, PDE7A, PDIA6, PHACTR1, PHF12, PHF3, PIGH, PIGK, PIGY, PIGZ, PIP5K1C, PKD1L2, PLD5, PLEKHA3, PLOD3, PM20D1, PNRC1, POLR3A, POLR3F, POU1F1, POU2F1, POU5F1, PP2672, PPIA, PPIAP33, PPIB, PPP1CB, PPP2R2A, PPP3CC, PQLC2, PRDM10, PRDM14, PRDM7, PRF1, PRKG2, PRKX, PROX1, PROX1-AS1, PSME2, PSORS1C1, PSORS1C3, PURA, RAB1B, RAB28, RAB2A, RAB6A, RAB6B, RAB6C, RABEPK, RANBP1, RANBP2, RAP1B, RAP1GAP2, RAPGEF1, RAPGEF2, RELA, RGCC, RGP1, RHOH, RICH2, RIMS4, RN7SKP199, RN7SL492P, RNA5SP407, RNA5SP408, RNF130, RNF170, RNF212, RNF26, RNF39, RNU105C, RNU6-1133P, RNU6-737P, RNU6-931P, RP11-100A16.1, RP11-707M13.1, RPL13AP15, RPL15P15, RPL19P16, RPL21P119, RPL21P126, RPL21P75, RPL23AP96, RPL28P3, RPL32P3, RPL4P5, RPS6KA2, RPSAP40, RPTN, RRAGB, RSAD2, RSL1D1, RTN2, RUNX1, RUSC2, RXRG, SAMD5, SCFD1, SCGB1D4, SCGB2A2, SDC1, SDC2, SDF1, SEC14L1, SERPINA1, SESTD1, SFT2D1, SGCD, SIGLEC17P, SIGLEC22P, SILC1, SIP1, SLC2A1, SLC2A1-AS1, SLC35F4, SLC46A1, SLC9A9, SLCO5A1, SMC4P1, SMIM12, SNHG32, SNN, SNRPEP6, SORBS3, SOX11, SOX5, SP110, SPAST, SPATS2L, SPCS3, SPDYA, SPOCK1, SPRY4-AS1, SPTAN1, SPTBN1, SSB, ST14, ST3GAL5, ST8SIA3, STAC2, STARD3NL, STAT1, STT3A, STX11, STX5, SUGCT, SUV420H1, TAOK1, TAP2, TBC1D7, TCEB3, TFAP4, TFDP2, TFE3, TGFBRAP1, THAP3, THOC2, THRAP3P1, THSD7B, TIAM2, TIMM8A, TLR7, TLR8, TLR9, TM9SF2, TMC4, TMED2, TMEFF2, TMEM132C, TMEM163, TMEM181, TMEM182, TMEM230, TMTC1, TNF, TNPO3, TNS1, TNXB, TOMM70A, TOR2A, TRAPPC1, TRIB1, TRIM10, TRIM27, TRIM5, TRIM55, TRIM58, TRMT5, TRPM6, TSBP1-AS1, TSSK3, TUBAL3, TXNL3, UBQLN4, UBQLN4P1, UGT1A11P, UGT1A12P, USP18, USP26, USP6, VANGL2, VAV2, VEGFC, VPRBP, VPS53, VWC2L, WASF5P, WDR27, WDR59, WDTC1, WNK1, WNT1, XAF1, XKR4, YPEL2, YTHDC2, ZBTB2, ZBTB20, ZBTB7C, ZCCHC7, ZDHHC19, ZFP90, ZNF12, ZNF182, ZNF354A, ZNF385D, ZNF436, ZNF512B, ZNF536, ZNF704, ZNF720, ZNF785, ZNF791, ZNF804A, ZNF831, ZNF90, ZNRD1, ZNRD1AS, ZNRD1ASP

A total of 27,662,062 variants were discovered in the whole genomes of 390 individuals from Botswana (**Table A2**).

Table A2. Total variants discovered in the genomic data of Batswana

Functional class	Number of variants
UTR3	237690
UTR5	23702
UTR5 and UTR3	100
downstream	169730
exonic	171348
exonic and splicing	69
intergenic	15116866
intronic	10085022
ncRNA exonic	89096

ncRNA exonic and splicing	20
ncRNA intronic	1646163
ncRNA splicing	578
splicing	2868
upstream	114776
upstream and downstream	4034
Exonic variants	
nsSNV	91994
sSNV	72040
FS indel	2964
stopgain	1635
nonFS indel	1635
stoploss	123
unknown	957

A total of 2,789,599 novel variants were discovered in the genomes of Batswana. The categories of these variants are presented in **Table A3**.

Table A3. Novel variants discovered in the genomic data of Batswana

Functional class	Number of variants
UTR	29904
flanking	30731
exonic	21665
intergenic	1461193
intronic	1066166
ncRNA	178178
splicing	1762
Exonic variants	
nsSNV	13769
sSNV	5670
FS indel	1313
stopgain	412
nonFS indel	380
stoploss	25
unknown	96

The distribution of novel variants was compared among four classes of minor allele counts (MAC) in **Table A4**.

Table A4. Novel variants per MAF category

Functional class	MAF			
	>0-0.01	>0.01-0.05	>0.05	Total
UTR	28370	1503	31	29904
flanking	28962	1725	44	30731
exonic	21160	500	5	21665
intergenic	1374111	85337	1745	1461193
intronic	1007198	57925	1043	1066166
ncRNA	167663	10331	184	178178
splicing	1758	3	1	1762
Total	2629222	157324	3053	2789599

Table A5. The most deleterious (damaging) variants as predicted by at least 10 databases in ANNOVAR

CHR	POS	SNV	REF/ALT	cDNA change	AA change	Gene	FATHHM	Siphy_29	CADD_phred	LOF_rare*	Freq_BW	gnomAD_AFR	Freq_EU
1	1184029	rs111377341	C/G,T	exon8:c.C979G	p.R327G	*TTLL10*	T	11.689	20.6	.	0.033,0.005141	0.000309112,0.00076089	0,1.54851e-05
1	3022425	rs3795263	G/A,C	exon1:c.G739A	p.G247R	*ACTRT2*	D	11.17	16.11	.	0.001289,0.003866	0.0441905	0.236254
1	8863242	.	G/A,T	exon7:c.C890A	p.T297N	*ENO1*	T	18.297	27.7	stopgain	0.001282,0.001282	0	1.55E-05
1	47140895	rs61507155	A/T,C	exon2:c.A311T	p.Y104F	*CYP4A22*	T	5.486	17.08	.	0.035,0.001282	0.0645445	0.000325198
1	205844819	rs374784539	C/A,G	exon4:c.G568C	p.D190H	*PM20D1*	T	19.649	18.79	.	0.001282,0.002564	4.76E-05	0
2	176100737	rs200302685	G/C,A	exon2:c.G790C	p.E264Q	*HOXD12*	D	19.976	34	.	0.032,0.002564	0.000641818,0.000380337,2.3771e-05	0.000139336,9.28908e-05,0.000216745
2	195906730	rs168192	G/T,C	exon27:c.C4264G	p.P1422A	*DNAH7*	T	17.963	17.81	stopgain	0.094,0.003846	0.049666	0.000372151
4	95841276	rs17024795	C/G,A	exon1:c.C1126G	p.R376G	*PDHA2*	T	10.855	14.07	.	0.027,0.001282	0.0407337,0	0.0000309809,1.54904e-05
4	150599048	rs111883007	G/C,A	exon38:c.C6005G	p.S2002W	*LRBA*	T	19.872	22.2	FS insertion	0.005128,0.001282	0	0
4	154234762	rs61746111	G/A,T	exon25:c.C8525A	p.P2842Q	*DCHS2*	T	20.346	17.9	FS insertion	0.032,0.003846	0.0370177,0.00335395	0.000154895,1.54895e-05
5	83512337	rs183969050	T/C,G	c.T983G	p.F328S	*VCAN*	T	10.479	18	FS insertion	0.013,0.001282	0.000356888	0
5	90652411	rs147062294	C/G,T	exon19:c.C3482G	p.S1161C	*ADGRV1*	T	20.672	34	FS insertion	0.014,0.001282	0.0149374	0.000108416
6	43675226	rs35480020	G/C,A	exon5:c.C422G	p.P141R	*MRPS18A*	.	12.36	17.34	.	0.022,0.001282	0,0.0353417	0.0000154866,0.000278759
7	20727068	rs111647033	G/A,C	exon13:c.G1319C	p.R440P	*ABCB5*	D	15.611	20.1	.	0.026,0.002564	0.000523585,0.00218954	0,0
7	134538257	rs142388608	C/G,T	exon8:c.C805G	p.R269G	*AKR1B10*	T	11.869	15.57	FS deletion	0.007692,0.001282	0.000214133,0.000118963	0.000139371,0.000185828
8	24487298	rs146451180	G/A,T	exon11:c.G1072A	p.V358M	*ADAM7*	T	9.477	14.63	splicing	0.001282,0.002564	0.000499762	0.00142455
11	4946887	rs2412467	C/T,G	exon1:c.G214A	p.D72N	*OR51A4*	T	14.304	18.96	.	0.002564,0.001282	0.0168363	0.038437
11	6891758	.	A/C,T	exon1:c.T743A	p.V248D	*OR2D2*	T	13.213	19.01	.	0.002564,0.003846	.	.
12	52775400	rs149569624	T/A,C	exon2:c.A803T	p.D268V	*KRT76*	T	14.612	22.8	.	0.01,0.012	0.00497004,0.00998811	0,4.64626e-05
14	20641145	rs148281603	C/G,A	exon1:c.G547C	p.D183H	*OR6S1*	T	13.924	19.52	.	0.013,0.007692	0.00107112,7.14082e-05	0,0
15	49972713	rs77004004	G/C,T	exon13:c.C1112A	p.P371H	*ATP8B4*	D	8.864	14.25	.	0.04,0.019	0.0000238243,0.0132463	0,0
16	48139198	rs113496237	C/T,G	exon5:c.G796C	p.G266R	*ABCC12*	D	15.741	19.24	.	0.021,0.013	0.0000714184,2.38061e-05	0,0
17	3292209	rs703903	C/T,A	exon1:c.G374T	p.R125L	*OR3A1*	T	17.71	29.8	.	0.16,0.001282	0.0132319,0.295096	0.000077553,0.542611
20	35801566	.	T/A,G	exon2:c.T44A	p.V15E	*PHF20*	T	15.004	23.8	.	0.001282,0.001282	.	.

*The function of an additional rare, novel variant. CHR: chromosome, SNV: single nucleotide variant, REF/ALT: reference allele/alternative allele, AA change: amino acid change, Freq_BW: frequency in the Botswana data, gnomAD_AFR: frequency in the gnomAD African/African-American population, Freq_EU: frequency in the European population.

For GWAS and rare-variant association test, model 3 was chosen as the overall best model; with a relatively low AIC, higher likelihood and a lower p-value. The CD4+ T-cell slopes predicted by this model were used in subsequent analyses (**Table A6**).

Table A6. Model selection of CD4+ T-cell changes of 236 HIV-positive participants over a period of at least 18 months

Model	LogL	Comparison to	LRT X^2	P-value	AIC
0. Constant	-7153.5	-	-	$< 2 \times 10^{-16}$	14313
1. HAART	-7129	model0	41.402	1.53×10^{-10}	14266
2. HAART + HAART *time	-7107.4	model1	41.616	3.70×10^{-7}	14227
2a. HAART + age + HAART *time	-7105.7	model2	0.1458	0.705	14225
3. HAART + HAART *time +(time|pid)	-7093.2	model2a	27.372	8.60×10^{-14}	14202
4. HAART + education level + HAART *time +(time|pid)	-6954.7	model3	-	0.229	13934
5. HAART + education level + feeding+ HAART *time +(time|pid)	-6954.7	model4	0.0058	0.0118	13935

LogL, log-likelihood of the model. LRT X^2, Chi-square statistic of the likelihood ratio test. AIC, Akaike information criterion. HAART, highly active antiretroviral therapy. Models 0-2a assess varying intercept only, while models 3-5 assess both varying intercept and slope for each individual. The p-values are for the following variables: model0, intercept; model1; model2, HAART *time, model2a, age; model3, HAART; model4, HAART, model3 is of HAART.

Admixture analysis

Genome-wide ancestry proportions in the Botswana population were estimated using the ADMIXTURE programme (Figure A1). Upon assessment of all the matrices of raw estimations of ancestry proportions, we concluded that the optimal model was K=6 model that also lowest available CV. Figure A1 shows that consistently from K4 to K6, the Botswana genetic data showed ancestry contribution from the Khoe-San and Niger-Congo-West (and Niger-Congo-Volta-Niger) populations, with the Khoe-San segments showing a little genetic contribution from the Europeans.

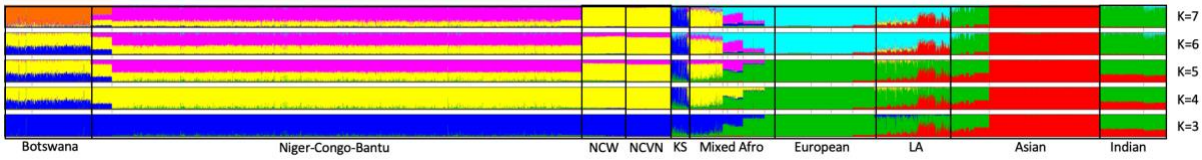


Figure A1. Genome-wide ancestry estimations of global populations. NCW: Niger-Congo-West, NCVN: Niger-Congo-Volta-Niger, KS: Khoe-San, LA: Latin American. For each hypothesised number of ancestral populations (K), the cross-validation (CV) of the models were as follows: K=3; CV=0.49241, K=4; CV=0.49001, K=5; CV=0.48902, K=6; CV=0.48688.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249.

Akahori, H., Guindon, S., Yoshizaki, S., and Muto, Y. (2015). Molecular Evolution of the TET Gene Family in Mammals. *Int. J. Mol. Sci.* 16, 28472–28485. doi:10.3390/ijms161226110.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi:10.1101/gr.094052.109.

An, P., and Winkler, C. A. (2010). Host genes associated with HIV/AIDS: advances in gene discovery. *Trends Genet.* 26, 119–131.

Annunziato, A. (2008). DNA Packaging: Nucleosomes and Chromatin. *Nat. Educ.* 1, 26.

Antonarakis, S. E., and Cooper, D. N. (2019). "6 - Human Genomic Variants and Inherited Disease: Molecular Mechanisms and Clinical Consequences," in, eds. R. E. Pyeritz, B. R. Korf, and W. W. B. T.-E. and R. P. and P. of M. G. and G. (Seventh E. Grody (Academic Press), 125–200. doi:https://doi.org/10.1016/B978-0-12-812537-3.00006-8.

Aujla, S. J., and Kolls, J. K. (2009). IL-22: a critical mediator in mucosal host defense. *J. Mol. Med. (Berl).* 87, 451–454. doi:10.1007/s00109-009-0448-1.

Avert (2019). Searching for a cure for HIV and AIDS. *Avert*. Available at: https://www.avert.org/professionals/hiv-science/searching-cure [Accessed March 23, 2020].

Awany, D., Allali, I., Dalvie, S., Hemmings, S., Mwaikono, K. S., Thomford, N. E., et al. (2018). Host and Microbiome Genome-Wide Association Studies: Current State and Challenges. *Front. Genet.* 9, 637. doi:10.3389/fgene.2018.00637.

Ballana, E., and Este, J. A. (2013). Insights from host genomics into HIV infection and

disease: identification of host targets for drug development. *Antiviral Res.* 100, 473–486.

Ballouz, S., Dobin, A., and Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biol.* 20, 159. doi:10.1186/s13059-019-1774-4.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* 67, 1–48. doi:10.18637/jss.v067.i01.

Batibo, H. M. (1999). A lexicostatistical survey of the Setswana dialects spoken in Botswana. *South African J. African Lang.* 19, 2–11.

Bednar, J., Horowitz, R. A., Grigoryev, S. A., Carruthers, L. M., Hansen, J. C., Koster, A. J., et al. (1998). Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14173–14178. doi:10.1073/pnas.95.24.14173.

Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002a). "Cholesterol Is Synthesized from Acetyl Coenzyme A in Three Stages," in *Biochemistry* (New York: W H Freeman). Available at: https://www.ncbi.nlm.nih.gov/books/NBK22350/.

Berg, J., Tymoczko, J., and Stryer, L. (2002b). "Amino Acids Are Made from Intermediates of the Citric Acid Cycle and Other Major Pathways," in *Biochemistry* (New York: W H Freeman). Available at: https://www.ncbi.nlm.nih.gov/books/NBK22459/.

Berman, S. K. (2017). A Bible translation inspired look at the history and ethnography of the Batswana. *die Skriflig* 51, 1–6. Available at: http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S2305-08532017000100003&nrm=iso.

Boden, D., Pusch, O., Lee, F., Tucker, L., Shank, P. R., and Ramratnam, B. (2003). Promoter choice affects the potency of HIV-1 specific RNA interference. *Nucleic Acids Res.* 31, 5033–5038. doi:10.1093/nar/gkg704.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120.

Bonora, M., Patergnani, S., Rimessi, A., De Marchi, E., Suski, J. M., Bononi, A., et al. (2012). ATP synthesis and storage. *Purinergic Signal.* 8, 343–357. doi:10.1007/s11302-012-9305-8.

Buchanan, C. C., Torstenson, E. S., Bush, W. S., and Ritchie, M. D. (2012). A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *J. Am. Med. Informatics Assoc.* 19, 289–294.

Buchmann, R., and Hazelhurst, S. (2015). The 'Genesis' Manual. Available at: http://www.bioinf.wits.ac.za/software/genesis/Genesis.pdf [Accessed November 28, 2018].

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi:10.1093/nar/gky1120.

Busby, G. B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V. D., et al. (2016). Admixture into and within sub-Saharan Africa. *Elife* 5. doi:10.7554/eLife.15266.

Campbell, M. C., and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu. Rev. Genomics Hum. Genet.* 9, 403–433. doi:10.1146/annurev.genom.9.081307.164258.

Carr, D. F., Bourgeois, S., Chaponda, M., Takeshita, L. Y., Morris, A. P., Cornejo Castro, E. M., et al. (2017). Genome-wide association study of nevirapine hypersensitivity in a sub-Saharan African HIV-infected population. *J. Antimicrob. Chemother.* 72, 1152–1162. doi:10.1093/jac/dkw545.

Carrington, M., and O'Brien, S. J. (2003). The influence of HLA genotype on AIDS. *Annu. Rev. Med.* 54, 535–551. doi:10.1146/annurev.med.54.101601.152346.

Chan, E. K. F., Timmermann, A., Baldi, B. F., Moore, A. E., Lyons, R. J., Lee, S.-S., et al. (2019). Human origins in a southern African palaeo-wetland and first migrations. *Nature* 575,

185–189. doi:10.1038/s41586-019-1714-1.

Chao, T.-C., Zhang, Q., Li, Z., Tiwari, S. K., Qin, Y., Yau, E., et al. (2019). The Long Noncoding RNA HEAL Regulates HIV-1 Replication through Epigenetic Regulation of the HIV-1 Promoter. *MBio* 10. doi:10.1128/mBio.02016-19.

Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S., Sacco, F., et al. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Res.* 37, D669-73. doi:10.1093/nar/gkn739.

Chege, D., Chai, Y., Huibner, S., Kain, T., Wachihi, C., Kimani, M., et al. (2012). Blunted IL17/IL22 and pro-inflammatory cytokine responses in the genital tract and blood of HIV-exposed, seronegative female sex workers in Kenya. *PLoS One* 7, e43670. doi:10.1371/journal.pone.0043670.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14, 128. doi:10.1186/1471-2105-14-128.

Chimusa, E. R., Daya, M., Moller, M., Ramesar, R., Henn, B. M., van Helden, P. D., et al. (2013a). Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PLoS One* 8, e73971. doi:10.1371/journal.pone.0073971.

Chimusa, E. R., Zaitlen, N., Daya, M., Moller, M., van Helden, P. D., Mulder, N. J., et al. (2013b). Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Hum. Mol. Genet.* 23, 796–809. doi:10.1093/hmg/ddt462.

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., et al. (2020). High-depth African genomes inform human migration and health. *Nature* 586, 741–748.

Choudhury, A., Aron, S., Sengupta, D., Hazelhurst, S., and Ramsay, M. (2018). African genetic diversity provides novel insights into evolutionary history and local adaptations. *Hum. Mol. Genet.* 27, R209–R218. doi:10.1093/hmg/ddy161.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., et al. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* 8. doi:10.1038/s41467-017-00663-9.

Chun, S., and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Res.* 19, 1553–1561.

Chundru, V. K., Marioni, R. E., Prendergast, J. G. D., Vallerga, C. L., Lin, T., Beveridge, A. J., et al. (2019). Examining the Impact of Imputation Errors on Fine-Mapping Using DNA Methylation QTL as a Model Trait. *Genetics* 212, 577–586. doi:10.1534/genetics.118.301861.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* 6, 80–92. doi:10.4161/fly.19695.

Cirulli, E. T., Singh, A., Shianna, K. V, Ge, D., Smith, J. P., Maia, J. M., et al. (2010). Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol.* 11, R57.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771.

code by Richard A. Becker, O. S., version by Ray Brownrigg. Enhancements by Thomas P Minka, A. R. W. R., and Deckmyn., A. (2018). maps: Draw Geographical Maps. Available at: https://cran.r-project.org/package=maps.

Conibear, E., and Davis, N. G. (2010). Palmitoylation and depalmitoylation dynamics at a glance. *J. Cell Sci.* 123, 4007 LP – 4010. doi:10.1242/jcs.059287.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature* 464, 704–712.

Coomer, C. A., Carlon-Andres, I., Iliopoulou, M., Dustin, M. L., Compeer, E. B., Compton, A. A., et al. (2020). Single-cell glycolytic activity regulates membrane tension and HIV-1 fusion. *PLoS Pathog.* 16, e1008359.

Cooper, G. M., Stone, E. A., Asimenos, G., Program, N. C. S., Green, E. D., Batzoglou, S., et al. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15, 901–913. doi:10.1101/gr.3577405.

Dalmasso, C., Carpentier, W., Meyer, L., Rouzioux, C., Goujard, C., Chaix, M.-L., et al. (2008). Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study. *PLoS One* 3, e3907. doi:10.1371/journal.pone.0003907.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158.

Das, S. K., Bhutia, S. K., Sokhi, U. K., Dash, R., Azab, B., Sarkar, D., et al. (2011). Human polynucleotide phosphorylase (hPNPaseold-35): an evolutionary conserved gene with an expanding repertoire of RNA degradation functions. *Oncogene* 30, 1733–1743. doi:10.1038/onc.2010.572.

Daya, M., Van Der Merwe, L., Van Helden, P. D., Möller, M., and Hoal, E. G. (2014). The role of ancestry in TB susceptibility of an admixed South African population. *Tuberculosis* 94, 413–420. doi:10.1016/j.tube.2014.03.012.

Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Smith, M. W., Allikmets, R., et al. (1996). Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE . *Science* 273, 1856–1862.

DeBerardinis, R. J., and Chandel, N. S. (2016). Fundamentals of cancer metabolism. *Sci. Adv.* 2, e1600200. doi:10.1126/sciadv.1600200.

Del Cornò, M., Donninelli, G., Varano, B., Da Sacco, L., Masotti, A., and Gessani, S. (2014).

HIV-1 gp120 Activates the STAT3/Interleukin-6 Axis in Primary Human Monocyte-Derived Dendritic Cells. *J. Virol.* 88, 11045 LP – 11055. doi:10.1128/JVI.00307-14.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–8. doi:10.1038/ng.806.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi:10.1101/gr.132159.111.

Deshiere, A., Joly-Beauparlant, C., Breton, Y., Ouellet, M., Raymond, F., Lodge, R., et al. (2017). Global Mapping of the Macrophage-HIV-1 Transcriptome Reveals that Productive Infection Induces Remodeling of Host Cell DNA and Chromatin. *Sci. Rep.* 7, 5238. doi:10.1038/s41598-017-05566-9.

Dietrich, L. E. P., and Ungermann, C. (2004). On the mechanism of protein palmitoylation. *EMBO Rep.* 5, 1053–1057. doi:10.1038/sj.embor.7400277.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* 24, 2125–2137. doi:10.1093/hmg/ddu733.

Dyer, R. J., and Nason, J. D. (2004). Population Graphs: the graph theoretic shape of genetic structure. *Mol. Ecol.* 13, 1713–1727. doi:10.1111/j.1365-294X.2004.02177.x.

E pluribus unum (2010). *Nat. Methods* 7, 331. doi:10.1038/nmeth0510-331.

Eichler, E. E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *N. Engl. J. Med.* 381, 64–74. doi:10.1056/NEJMra1809315.

Eilers, M., Roy, U., and Mondal, D. (2008). MRP (ABCC) transporters-mediated efflux of anti-HIV drugs, saquinavir and zidovudine, from human endothelial cells. *Exp. Biol. Med.* 233, 1149–1160.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi:10.1038/nature11247.

Epi25 Collaborative (2019). Ultra-Rare Genetic Variation in the Epilepsies: A Whole-Exome Sequencing Study of 17,606 Individuals. *Am. J. Hum. Genet.* 105, 267–282. doi:10.1016/j.ajhg.2019.05.020.

Escudero, D. J., Marukutira, T., McCormick, A., Makhema, J., and Seage, G. R. I. I. I. (2019). Botswana should consider expansion of free antiretroviral therapy to immigrants. *J. Int. AIDS Soc.* 22, e25328. doi:10.1002/jia2.25328.

Essex, M. (1999). Human immunodeficiency viruses in the developing world. *Adv. Virus Res.* 53, 71–88. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10582095.

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

Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131, 479–491.

Fan, J., Ye, J., Kamphorst, J. J., Shlomi, T., Thompson, C. B., and Rabinowitz, J. D. (2014). Quantitative flux analysis reveals folate-dependent NADPH production. *Nature* 510, 298–302. doi:10.1038/nature13236.

Farahani, M., Vable, A., Lebelonyane, R., Seipone, K., Anderson, M., Avalos, A., et al. (2014). Outcomes of the Botswana national HIV/AIDS treatment programme from 2002 to 2010: a longitudinal analysis. *Lancet. Glob. Heal.* 2, e44-50. doi:10.1016/S2214-109X(13)70149-9.

Fellay, J., Ge, D., Shianna, K. V, Colombo, S., Ledergerber, B., Cirulli, E. T., et al. (2009). Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.* 5, e1000791. doi:10.1371/journal.pgen.1000791.

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., et al. (2007). A

Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science (80-.).* 317, 944–947. doi:10.1126/science.1143767.

Fellay, J., Shianna, K. V, Telenti, A., and Goldstein, D. B. (2010). Host genetics and HIV-1: the final phase? *PLoS Pathog.* 6, e1001033. doi:10.1371/journal.ppat.1001033.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 43, D805-11. doi:10.1093/nar/gku1075.

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 25, i54–i62. doi:10.1093/bioinformatics/btp190.

GeneCards (2020). GeneCards - Human Gene Database. Available at: https://www.genecards.org/ [Accessed August 3, 2020].

Gonzalez-Garay, M. L. (2014). The road from next-generation sequencing to personalized medicine. *Per. Med.* 11, 523–544.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science (80-.).* 307, 1434–1440.

Greenbaum, G., Templeton, A. R., and Bar-David, S. (2016). Inference and Analysis of Population Structure Using Genetic Data and Network Theory. *Genetics* 202, 1299–1312. doi:10.1534/genetics.115.182626.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*.

Gudykunst, W. B., and Schmidt, K. L. (1987). Language and ethnic identity: An overview and prologue. *J. Lang. Soc. Psychol.* 6, 157–170.

Guerrero, S., Batisse, J., Libre, C., Bernacchi, S., Marquet, R., and Paillart, J.-C. (2015). HIV-1 replication and the cellular eukaryotic translation apparatus. *Viruses* 7, 199–218.

doi:10.3390/v7010199.

Gunnery, S., and Mathews, M. B. (1995). Functional mRNA can be generated by RNA polymerase III. *Mol. Cell. Biol.* 15, 3597–3607. doi:10.1128/mcb.15.7.3597.

Guo, Y., Dai, Y., Yu, H., Zhao, S., Samuels, D. C., and Shyr, Y. (2017). Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* 109, 83–90. doi:https://doi.org/10.1016/j.ygeno.2017.01.005.

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

Gustafsson, R., Zhang, S., Masuyer, G., Dong, M., and Stenmark, P. (2018). Crystal Structure of Botulinum Neurotoxin A2 in Complex with the Human Protein Receptor SV2C Reveals Plasticity in Receptor Binding. *Toxins (Basel).* 10. doi:10.3390/toxins10040153.

Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* 88, 586–598. doi:10.1016/j.ajhg.2011.04.014.

Handley, L. J. L., Manica, A., Goudet, J., and Balloux, F. (2007). Going the distance: human population genetics in a clinal world. *TRENDS Genet.* 23, 432–439.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258-61. doi:10.1093/nar/gkh036.

Hegedus, A., Kavanagh Williamson, M., and Huthoff, H. (2014). HIV-1 pathogenicity and virion production are dependent on the metabolic phenotype of activated CD4+ T cells. *Retrovirology* 11, 98. doi:10.1186/s12977-014-0098-4.

Heine, B., and Nurse, D. (2000). *African languages: An introduction*. 1st ed. Cambridge: Cambridge University Press.

Henrich, T. J., and Kuritzkes, D. R. (2013). HIV-1 entry inhibitors: recent development and

clinical use. *Curr. Opin. Virol.* 3, 51–57. doi:10.1016/j.coviro.2012.12.002.

Herbeck, J. T., Gottlieb, G. S., Winkler, C. A., Nelson, G. W., An, P., Maust, B. S., et al. (2010). Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. *J. Infect. Dis.* 201, 618–626. doi:10.1086/649842.

Hernandez, R. D., Uricchio, L. H., Hartman, K., Ye, C., Dahl, A., and Zaitlen, N. (2019). Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* 51, 1349–1355. doi:10.1038/s41588-019-0487-7.

Hillebrand, F., Ostermann, P. N., Müller, L., Degrandi, D., Erkelenz, S., Widera, M., et al. (2019). Gymnotic delivery of LNA mixmers targeting viral SREs induces HIV-1 mRNA degradation. *Int. J. Mol. Sci.* 20, 1088.

Hirschhorn, J. N., and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108. doi:10.1038/nrg1521.

Hu, J., Li, Y., Chen, L., Yang, Z., Zhao, G., Wang, Y., et al. (2016). Impact of IL-22 gene polymorphism on human immunodeficiency virus infection in Han Chinese patients. *J. Microbiol. Immunol. Infect.* 49, 872–878. doi:10.1016/j.jmii.2014.09.002.

Huang, D., Li, C., and Zhang, H. (2014). Hypoxia and cancer cell metabolism. *Acta Biochim. Biophys. Sin. (Shanghai).* 46, 214–219. doi:10.1093/abbs/gmt148.

Huang, R., Grishagin, I., Wang, Y., Zhao, T., Greene, J., Obenauer, J. C., et al. (2019). The NCATS BioPlanet - An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. *Front. Pharmacol.* 10, 445. doi:10.3389/fphar.2019.00445.

Hutcheson, H. B., Lautenberger, J. A., Nelson, G. W., Pontius, J. U., Kessing, B. D., Winkler, C. A., et al. (2008). Detecting AIDS restriction genes: from candidate genes to genome-wide association discovery. *Vaccine* 26, 2951–2965. doi:10.1016/j.vaccine.2007.12.054.

Ignatieva, E. V, Yurchenko, A. A., Voevoda, M. I., and Yudin, N. S. (2019). Exome-wide search and functional annotation of genes associated in patients with severe tick-borne

encephalitis in a Russian population. *BMC Med. Genomics* 12, 61. doi:10.1186/s12920-019-0503-x.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.

Ioannidis, J. P., Rosenberg, P. S., Goedert, J. J., Ashton, L. J., Benfield, T. L., Buchbinder, S. P., et al. (2001). Effects of CCR5-Delta32, CCR2-64I, and SDF-1 3'A alleles on HIV-1 disease progression: An international meta-analysis of individual-patient data. *Ann. Intern. Med.* 135, 782–795. doi:10.7326/0003-4819-135-9-200111060-00008.

Jakobsson, M., Edge, M. D., and Rosenberg, N. A. (2013). The relationship between F(ST) and the frequency of the most frequent allele. *Genetics* 193, 515–528. doi:10.1534/genetics.112.144758.

Jang, K. L., Collins, M. K., and Latchman, D. S. (1992). The human immunodeficiency virus tat protein increases the transcription of human Alu repeated sequences by increasing the activity of the cellular transcription factor TFIIIC. *J. Acquir. Immune Defic. Syndr.* 5, 1142–1147.

Jeong, S., Patel, N., Edlund, C. K., Hartiala, J., Hazelett, D. J., Itakura, T., et al. (2015). Identification of a Novel Mucin Gene HCG22 Associated With Steroid-Induced Ocular Hypertension. *Invest. Ophthalmol. Vis. Sci.* 56, 2737–2748. doi:10.1167/iovs.14-14803.

Johnson, J. L., Clark, C. C., Li, K. W., Caron, S., and Abecasis, G. (2016). Genetic Association Study (GAS) Power Calculator. *Available csg. sph. umich. edu/abecasis/cats/gas_power_calculator/index. html. Accessed Novemb.* 16, 2015.

Johnston, H. R., Hu, Y., and Cutler, D. J. (2015). Population genetics identifies challenges in analyzing rare variants. *Genet. Epidemiol.* 39, 145–148. doi:10.1002/gepi.21881.

Joubert, B. R., Lange, E. M., Franceschini, N., Mwapasa, V., North, K. E., and Meshnick, S. R. (2010). A whole genome association study of mother-to-child transmission of HIV in

Malawi. *Genome Med.* 2, 17. doi:10.1186/gm138.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi:10.1038/ng.548.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi:10.1038/s41586-020-2308-7.

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., et al. (2017). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* 45, D840–D845. doi:10.1093/nar/gkw971.

Keinan, A., and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740–743. doi:10.1126/science.1217283.

Khan, S., Iqbal, M., Tariq, M., Baig, S. M., and Abbas, W. (2018). Epigenetic regulation of HIV-1 latency: focus on polycomb group (PcG) proteins. *Clin. Epigenetics* 10, 14. doi:10.1186/s13148-018-0441-z.

Kim, J., Tchernyshyov, I., Semenza, G. L., and Dang, C. V (2006). HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metab.* 3, 177–185. doi:10.1016/j.cmet.2006.02.002.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi:10.1038/ng.2892.

Kishimoto, N., Iga, N., Yamamoto, K., Takamune, N., and Misumi, S. (2017). Virion-incorporated alpha-enolase suppresses the early stage of HIV-1 reverse transcription. *Biochem. Biophys. Res. Commun.* 484, 278–284.

Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., et al. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* 15, e1008500. doi:10.1371/journal.pgen.1008500.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* 58, 1347–1363.

Kuleshov, M. V, Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90-7. doi:10.1093/nar/gkw377.

Kulkarni, S., Lied, A., Kulkarni, V., Rucevic, M., Martin, M. P., Walker-Sperling, V., et al. (2019). CCR5AS lncRNA variation differentially regulates CCR5, influencing HIV disease outcome. *Nat. Immunol.* 20, 824–834. doi:10.1038/s41590-019-0406-1.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. doi:10.1093/nar/gkt1113.

Lappin, T. R. J., Grier, D. G., Thompson, A., and Halliday, H. L. (2006). HOX genes: seductive science, mysterious mechanisms. *Ulster Med. J.* 75, 23–31. Available at: https://pubmed.ncbi.nlm.nih.gov/16457401.

Lazar, D. C., Morris, K. V, and Saayman, S. M. (2016). The emerging role of long non-coding RNAs in HIV infection. *Virus Res.* 212, 114–126. doi:10.1016/j.virusres.2015.07.023.

Le Clerc, S., Coulonges, C., Delaneau, O., Van Manen, D., Herbeck, J. T., Limou, S., et al. (2011). Screening low-frequency SNPS from genome-wide association study reveals a new risk allele for progression to AIDS. *J. Acquir. Immune Defic. Syndr.* 56, 279–284. doi:10.1097/QAI.0b013e318204982b.

Le Clerc, S., Limou, S., Coulonges, C., Carpentier, W., Dina, C., Taing, L., et al. (2009).

Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). *J. Infect. Dis.* 200, 1194–1201. doi:10.1086/605892.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., et al. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91, 224–237. doi:10.1016/j.ajhg.2012.06.007.

Lee, S., Wu, M. C., and Lin, X. (2012b). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13, 762–775. doi:10.1093/biostatistics/kxs014.

Leszczyniecka, M., Kang, D., Sarkar, D., Su, Z., Holmes, M., Valerie, K., et al. (2002). Identification and cloning of human polynucleotide phosphorylase, hPNPase^{old-35}, in the context of terminal differentiation and cellular senescence. *Proc. Natl. Acad. Sci.* 99, 16636 LP – 16641. doi:10.1073/pnas.252643699.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi:10.1093/bioinformatics/btr509.

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

Li, H., Ruan, J., Durbin, R., Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores Mapping short DNA sequencing reads and calling variants using mapping quality scores. 1851–1858. doi:10.1101/gr.078212.108.

Li, L.-H., Ho, S.-F., Chen, C.-H., Wei, C.-Y., Wong, W.-C., Li, L.-Y., et al. (2006). Long contiguous stretches of homozygosity in the human genome. *Hum. Mutat.* 27, 1115–1121.

Limou, S., Le Clerc, S., Coulonges, C., Carpentier, W., Dina, C., Delaneau, O., et al. (2009).

Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J. Infect. Dis.* 199, 419–26. doi:10.1086/596067.

Limou, S., and Zagury, J.-F. (2013). Immunogenetics: Genome-Wide Association of Non-Progressive HIV and Viral Load Control: HLA Genes and Beyond. *Front. Immunol.* 4, 118. doi:10.3389/fimmu.2013.00118.

Lingappa, J. R., Petrovski, S., Kahle, E., Fellay, J., Shianna, K., McElrath, M. J., et al. (2011). Genomewide association study for determinants of HIV-1 acquisition and viral set point in HIV-1 serodiscordant couples with quantified virus exposure. *PLoS One* 6, e28632. doi:10.1371/journal.pone.0028632.

Liu, R., Huang, L., Li, J., Zhou, X., Zhang, H., Zhang, T., et al. (2013a). HIV Infection in Gastric Epithelial Cells. *J. Infect. Dis.* 208, 1221–1230. doi:10.1093/infdis/jit314.

Liu, R., Wu, J., Shao, R., and Xue, Y. (2014). Mechanism and factors that control HIV-1 transcription and latency activation. *J. Zhejiang Univ. Sci. B* 15, 455–465. doi:10.1631/jzus.B1400059.

Liu, S.-Y., Aliyari, R., Chikere, K., Li, G., Marsden, M. D., Smith, J. K., et al. (2013b). Interferon-Inducible Cholesterol-25-Hydroxylase Broadly Inhibits Viral Entry by Production of 25-Hydroxycholesterol. *Immunity* 38, 92–105. doi:https://doi.org/10.1016/j.immuni.2012.11.005.

Liu, S., Yao, L., Ding, D., and Zhu, H. (2010). CCL3L1 copy number variation and susceptibility to HIV-1 infection: a meta-analysis. *PLoS One* 5, e15778.

Liu, Y., Nyunoya, T., Leng, S., Belinsky, S. A., Tesfaigzi, Y., and Bruse, S. (2013c). Softwares and methods for estimating genetic ancestry in human populations. *Hum. Genomics* 7, 1. doi:10.1186/1479-7364-7-1.

Lloyd-Jones, L. R., Robinson, M. R., Yang, J., and Visscher, P. M. (2018). Transformation of Summary Statistics from Linear Mixed Model Association on All-or-None Traits to Odds Ratio. *Genetics* 208, 1397–1408. doi:10.1534/genetics.117.300360.

Lv, L., Wang, Q., Xu, Y., Tsao, L.-C., Nakagawa, T., Guo, H., et al. (2018). Vpr Targets TET2 for Degradation by CRL4(VprBP) E3 Ligase to Sustain IL-6 Expression and Enhance HIV-1 Replication. *Mol. Cell* 70, 961-970.e5. doi:10.1016/j.molcel.2018.05.007.

MacArthur, D. G., and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* 19, R125-30. doi:10.1093/hmg/ddq365.

MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., and Scherer, S. W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42, D986-92. doi:10.1093/nar/gkt958.

Mackelprang, R. D., Bamshad, M. J., Chong, J. X., Hou, X., Buckingham, K. J., Shively, K., et al. (2017). Whole genome sequencing of extreme phenotypes identifies variants in CD101 and UBE2V1 associated with increased risk of sexually acquired HIV-1. *PLoS Pathog.* 13, e1006703. doi:10.1371/journal.ppat.1006703.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402.

May, A., Hazelhurst, S., Li, Y., Norris, S. A., Govind, N., Tikly, M., et al. (2013). Genetic diversity in black South Africans from Soweto. *BMC Genomics* 14, 644. doi:10.1186/1471-2164-14-644.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res. Sep;20(9)1297-303. doi 10.1101/gr.107524.110* Sep 20, 1297–303. doi:10.1101/gr.107524.110.

McKusick, V. A. (1998). *Mendelian inheritance in man: a catalog of human genes and genetic disorders*. JHU Press.

McLaren, P. J., and Carrington, M. (2015). The impact of host genetic variation on infection with HIV-1. *Nat. Immunol.* 16, 577–583. doi:10.1038/ni.3147.

McLaren, P. J., Coulonges, C., Bartha, I., Lenz, T. L., Deutsch, A. J., Bashirova, A., et al. (2015). Polymorphisms of large effect explain the majority of the host genetic contribution to

variation of HIV-1 virus load. *Proc. Natl. Acad. Sci.* 112. doi:10.1073/pnas.1514867112.

McLaren, P. J., Coulonges, C., Ripke, S., van den Berg, L., Buchbinder, S., Carrington, M., et al. (2013). Association study of common genetic variants and HIV-1 acquisition in 6,300 infected cases and 7,200 controls. *PLoS Pathog.* 9, e1003515. doi:10.1371/journal.ppat.1003515.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., et al. (2008). Runs of homozygosity in European populations. *Am. J. Hum. Genet.* 83, 359–372.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., et al. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* 45, D183–D189. doi:10.1093/nar/gkw1138.

Michalopoulos, S. (2012). The origins of ethnolinguistic diversity. *Am. Econ. Rev.* 102, 1508–1539.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., et al. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature* 470, 59–65. doi:10.1038/nature09708.

Missé, D., Yssel, H., Trabattoni, D., Oblet, C., Lo Caputo, S., Mazzotta, F., et al. (2007). IL-22 Participates in an Innate Anti-HIV-1 Host-Resistance Network through Acute-Phase Protein Induction. *J. Immunol.* 178, 407 LP – 415. doi:10.4049/jimmunol.178.1.407.

Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., et al. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* 25, 869–876. doi:10.1038/ejhg.2017.51.

Mohseni Ahooyi, T., Torkzaban, B., Shekarabi, M., Tahrir, F. G., Decoppet, E. A., Cotto, B., et

al. (2019). Perturbation of synapsins homeostasis through HIV-1 Tat-mediated suppression of BAG3 in primary neuronal cells. *Cell Death Dis.* 10, 473. doi:10.1038/s41419-019-1702-2.

Montinaro, F., Busby, G. B. J., Gonzalez-Santos, M., Oosthuitzen, O., Oosthuitzen, E., Anagnostou, P., et al. (2017). Complex Ancient Genetic Structure and Cultural Transitions in Southern African Populations. *Genetics* 205, 303–316. doi:10.1534/genetics.116.189209.

Morgans, C. W., Kensel-Hammes, P., Hurley, J. B., Burton, K., Idzerda, R., McKnight, G. S., et al. (2009). Loss of the Synaptic Vesicle Protein SV2B results in reduced neurotransmission and altered synaptic vesicle protein expression in the retina. *PLoS One* 4, e5230. doi:10.1371/journal.pone.0005230.

Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y., et al. (2015). Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* 6, 1–13. doi:10.1038/ncomms9018.

Nangola, S., Urvoas, A., Valerio-Lepiniec, M., Khamaikawin, W., Sakkhachornphop, S., Hong, S.-S., et al. (2012). Antiviral activity of recombinant ankyrin targeted to the capsid domain of HIV-1 Gag polyprotein. *Retrovirology* 9, 17. doi:10.1186/1742-4690-9-17.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci.* 70, 3321–3323.

Ng, P. C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi:10.1093/nar/gkg509.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi:10.1038/nrg2986.

Novembre, J. (2016). Pritchard, Stephens, and Donnelly on Population Structure. *Genetics* 204, 391–393. doi:10.1534/genetics.116.195164.

Novitsky, V., Woldegabriel, E., Wester, C., McDonald, E., Rossenkhan, R., Ketunuti, M., et al. (2008). Identification of primary HIV-1C infection in Botswana. *AIDS Care* 20, 806–811.

O'Brien, S. J., and Nelson, G. W. (2004). Human genes that limit AIDS. *Nat. Genet.* 36, 565–574. doi:10.1038/ng1369.

O'Brien, S. J., Nelson, G. W., Winkler, C. A., and Smith, M. W. (2000). POLYGENIC AND MULTIFACTORIAL DISEASE GENE ASSOCIATION IN MAN: Lessons from AIDS. *Annu. Rev. Genet.* 34, 563–591. doi:10.1146/annurev.genet.34.1.563.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733-45. doi:10.1093/nar/gkv1189.

Pagel, K. A., Pejaver, V., Lin, G. N., Nam, H.-J., Mort, M., Cooper, D. N., et al. (2017). When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. *Bioinformatics* 33, i389–i398. doi:10.1093/bioinformatics/btx272.

Palmer, C. S., Henstridge, D. C., Yu, D., Singh, A., Balderson, B., Duette, G., et al. (2016). Emerging role and characterization of immunometabolism: relevance to HIV pathogenesis, serious non-AIDS events, and a cure. *J. Immunol.* 196, 4437–4444.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2, e190. doi:10.1371/journal.pgen.0020190.

Pavlopoulos, G. A., Oulas, A., Iacucci, E., Sifrim, A., Moreau, Y., Schneider, R., et al. (2013). Unraveling genomic variation from next generation sequencing data. *BioData Min.* 6, 13.

Pelak, K., Goldstein, D. B., Walley, N. M., Fellay, J., Ge, D., Shianna, K. V, et al. (2010). Host determinants of HIV-1 control in African Americans. *J. Infect. Dis.* 201, 1141–1149. doi:10.1086/651382.

Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., and Li, J. Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* 91, 275–292.

Penning, T. M. (2015). The aldo-keto reductases (AKRs): Overview. *Chem. Biol. Interact.* 234,

236–246. doi:https://doi.org/10.1016/j.cbi.2014.09.024.

Percario, Z., Olivetta, E., Fiorucci, G., Mangino, G., Peretti, S., Romeo, G., et al. (2003). Human immunodeficiency virus type 1 (HIV-1) Nef activates STAT3 in primary human monocyte/macrophages through the release of soluble factors: involvement of Nef domains interacting with the cell endocytotic machinery. *J. Leukoc. Biol.* 74, 821–832.

Pereyra, F., Jia, X., McLaren, P. J., Telenti, A., de Bakker, P. I. W., Walker, B. D., et al. (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 330, 1551–1557. doi:10.1126/science.1195271.

Petersen, D. C., Libiger, O., Tindall, E. A., Hardie, R.-A., Hannick, L. I., Glashoff, R. H., et al. (2013). Complex patterns of genomic admixture within southern Africa. *PLoS Genet.* 9, e1003309. doi:10.1371/journal.pgen.1003309.

Petrovski, S., Fellay, J., Shianna, K. V, Carpenetti, N., Kumwenda, J., Kamanga, G., et al. (2011). Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population. *AIDS* 25, 513–518. doi:10.1097/QAD.0b013e328343817b.

Pfeifer, S. P. (2017). From next-generation resequencing reads to a high-quality variant data set. *Heredity (Edinb).* 118, 111–124.

Pickrell, J. K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., et al. (2012). The genetic prehistory of southern Africa. *Nat. Commun.* 3, 1143. doi:10.1038/ncomms2140.

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., et al. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proc. Natl. Acad. Sci. U. S. A.* 111, 2632–2637. doi:10.1073/pnas.1313787111.

Plain, F., Howie, J., Kennedy, J., Brown, E., Shattock, M. J., Fraser, N. J., et al. (2020). Control of protein palmitoylation by regulating substrate recruitment to a zDHHC-protein acyltransferase. *Commun. Biol.* 3, 411. doi:10.1038/s42003-020-01145-3.

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding

RNAs. *Cell* 136, 629–641. doi:10.1016/j.cell.2009.02.006.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi:10.1038/ng1847.

Purcell, S., Cherny, S. S., and Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19, 149–150.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795.

R Core Team (2019). R: A Language and Environment for Statistical Computing. Available at: https://www.r-project.org/.

Ratnasabapathy, R., Sheldon, M., Johal, L., and Hernandez, N. (1990). The HIV-1 long terminal repeat contains an unusual element that induces the synthesis of short RNAs from various mRNA and snRNA promoters. *Genes Dev.* 4, 2061–2074. doi:10.1101/gad.4.12a.2061.

Reed, F. A., and Tishkoff, S. A. (2006). African human diversity, origins and migrations. *Curr. Opin. Genet. Dev.* 16, 597–605.

Reich, D., Nalls, M. A., Kao, W. H. L., Akylbekova, E. L., Tandon, A., Patterson, N., et al. (2009). Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet* 5, e1000360.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 47, D886–D894. doi:10.1093/nar/gky1016.

Retshabile, G., Mlotshwa, B. C., Williams, L., Mwesigwa, S., Mboowa, G., Huang, Z., et al. (2018). Whole-Exome Sequencing Reveals Uncaptured Variation and Distinct Ancestry in the Southern African Population of Botswana. *Am. J. Hum. Genet.* 102, 731–743. doi:10.1016/j.ajhg.2018.03.010.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 39, e118--e118.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science (80-.).* 273, 1516–1517.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., et al. (2002). Genetic Structure of Human Populations. 298, 2381–2385.

SADC (2020). Member states. *South. African Dev. Community - SADC*. Available at: https://www.sadc.int/member-states [Accessed August 7, 2020].

Salvaggio, S. E., Giacomelli, A., Falvella, F. S., Oreni, M. L., Meraviglia, P., Atzori, C., et al. (2017). Clinical and genetic factors associated with kidney tubular dysfunction in a real-life single centre cohort of HIV-positive patients. *BMC Infect. Dis.* 17, 396. doi:10.1186/s12879-017-2497-3.

Sauna, Z. E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* 12, 683–691. doi:10.1038/nrg3051.

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P. A., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 27, 849–864.

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., et al. (2010). Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463, 943–947. doi:10.1038/nature08795.

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* 11, 361–362. doi:10.1038/nmeth.2890.

Servin, B., and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* 3, e114. doi:10.1371/journal.pgen.0030114.

Shapiro, R. L., Thior, I., Gilbert, P. B., Lockman, S., Wester, C., Smeaton, L. M., et al. (2006). Maternal single-dose nevirapine versus placebo as part of an antiretroviral strategy to prevent mother-to-child HIV transmission in Botswana. *Aids* 20, 1281–1288.

Shen, L., Wu, C., Zhang, J., Xu, H., Liu, X., Wu, X., et al. (2020). Roles and potential applications of lncRNAs in HIV infection. *Int. J. Infect. Dis.* 92, 97–104. doi:https://doi.org/10.1016/j.ijid.2020.01.006.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–11. doi:10.1093/nar/29.1.308.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., et al. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* 34, 57–65. doi:10.1002/humu.22225.

Shihab, H. A., Gough, J., Mort, M., Cooper, D. N., Day, I. N. M., and Gaunt, T. R. (2014). Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum. Genomics* 8, 11. doi:10.1186/1479-7364-8-11.

Siliciano, J. D., and Siliciano, R. F. (2016). Recent developments in the effort to cure HIV infection: going beyond N = 1. *J. Clin. Invest.* 126, 409–414. doi:10.1172/JCI86047.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457. doi:10.1093/nar/gks539.

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell* 177. doi:10.1016/j.cell.2019.02.048.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139, 457–462.

Statistics Botswana (2015). *Population and Housing Census 2011: National Statistical Tables*. Gaborone: Statistics Botswana Available at:

http://www.statsbots.org.bw/sites/default/files/publications/national_statisticsreport.pdf.

Steward, C. A., Parker, A. P. J., Minassian, B. A., Sisodiya, S. M., Frankish, A., and Harrow, J. (2017). Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome Med.* 9, 49. doi:10.1186/s13073-017-0441-1.

Strauch, K., Fimmers, R., Kurz, T., Deichmann, K. A., Wienker, T. F., and Baur, M. P. (2000). Parametric and Nonparametric Multipoint Linkage Analysis with Imprinting and Two-Locus–Trait Models: Application to Mite Sensitization. *Am. J. Hum. Genet.* 66, 1945–1957. doi:https://doi.org/10.1086/302911.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526. doi:10.1038/nature15394.

Tabor, H. K., Risch, N. J., and Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat. Rev. Genet.* 3, 391–397.

Tau, T., Davison, S., and D'Amato, M. E. (2015). Polymorphisms at 17 Y-STR loci in Botswana populations. *Forensic Sci. Int. Genet.* 17, 47–52. doi:10.1016/j.fsigen.2015.03.001.

Tau, T., Wally, A., Fanie, T. P., Ngono, G. L., Mpoloka, S. W., Davison, S., et al. (2017). Genetic variation and population structure of Botswana populations as identified with AmpFLSTR Identifiler short tandem repeat (STR) loci. *Sci. Rep.* 7, 6768. doi:10.1038/s41598-017-06365-y.

Taylor, H. E., and Palmer, C. S. (2020). CD4 T Cell Metabolism Is a Major Contributor of HIV Infectivity and Reservoir Persistence. *Immunometabolism* 2.

Telenti, A., and Goldstein, D. B. (2006). Genomics meets HIV-1. *Nat. Rev. Microbiol.* 4, 865–873. doi:10.1038/nrmicro1532.

Telenti, A., and Johnson, W. E. (2012). Host genes important to HIV replication and evolution. *Cold Spring Harb. Perspect. Med.* 2, a007203. doi:10.1101/cshperspect.a007203.

Thami, P. K., and Chimusa, E. R. (2019). Population Structure and Implications on the Genetic Architecture of HIV-1 Phenotypes Within Southern Africa. *Front. Genet.* 10, 905. doi:10.3389/fgene.2019.00905.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi:10.1038/nature09534.

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi:10.1038/nature11632.

The 1000 Genomes Project Consortium, Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., et al. (2015). A global reference for human genetic variation. *Nature* 526. doi:10.1038/nature15393.

The International HapMap Consortium (2003). The International HapMap Project. *Nature* 426, 789–796. doi:10.1038/nature02168.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature* 437, 1299.

Thior, I., Lockman, S., Smeaton, L. M., Shapiro, R. L., Wester, C., Heymann, S. J., et al. (2006). Breastfeeding plus infant zidovudine prophylaxis for 6 months vs formula feeding plus infant zidovudine for 1 month to reduce mother-to-child HIV transmission in Botswana: a randomized trial: the Mashi Study. *Jama* 296, 794–805.

Tishkoff, S. A., and Kidd, K. K. (2004). Implications of biogeography of human populations for "race" and medicine. *Nat. Genet.* 36, S21–S27. doi:10.1038/ng1438.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., et al. (2009). The genetic structure and history of Africans and African Americans. *Science* 324, 1035–1044. doi:10.1126/science.1172257.

Tishkoff, S. A., and Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* 3, 611–621. doi:10.1038/nrg865.

Torkamani, A., Scott-Van Zeeland, A. A., Topol, E. J., and Schork, N. J. (2011). Annotating

individual human genomes. *Genomics* 98, 233–241. doi:https://doi.org/10.1016/j.ygeno.2011.07.006.

Tough, R. H., and McLaren, P. J. (2019). Interaction of the Host and Viral Genome and Their Influence on HIV Disease. *Front. Genet.* 9, 720. doi:10.3389/fgene.2018.00720.

Troyer, J. L., Nelson, G. W., Lautenberger, J. A., Chinn, L., McIntosh, C., Johnson, R. C., et al. (2011). Genome-wide association study implicates PARD3B-based AIDS restriction. *J. Infect. Dis.* 203, 1491–1502. doi:10.1093/infdis/jir046.

Trypsteen, W., Mohammadi, P., Van Hecke, C., Mestdagh, P., Lefever, S., Saeys, Y., et al. (2016). Differential expression of lncRNAs during the HIV replication cycle: an underestimated layer in the HIV-host interplay. *Sci. Rep.* 6, 36111. doi:10.1038/srep36111.

Tsai, Z. T.-Y., Lloyd, J. P., and Shiu, S.-H. (2017). Defining Functional Genic Regions in the Human Genome through Integration of Biochemical, Evolutionary, and Genetic Evidence. *Mol. Biol. Evol.* 34, 1788–1798. doi:10.1093/molbev/msx101.

UNAIDS (2019). UNAIDS data 2019. *Unaids*. Available at: https://www.unaids.org/en/regionscountries/countries/botswana [Accessed March 6, 2020].

UNAIDS (2020). Global HIV & AIDS statistics — 2020 fact sheet. Available at: https://www.unaids.org/en/resources/fact-sheet [Accessed August 10, 2020].

Valle-Casuso, J. C., Angin, M., Volant, S., Passaes, C., Monceaux, V., Mikhailova, A., et al. (2019). Cellular Metabolism Is a Major Determinant of HIV-1 Reservoir Seeding in CD4+ T Cells and Offers an Opportunity to Tackle Infection. *Cell Metab.* 29, 611-626.e5. doi:https://doi.org/10.1016/j.cmet.2018.11.015.

Van Der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Levy-moonshine, A., Jordan, T., et al. (2014). From FastQ data to high confidence varant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma.* 11. doi:10.1002/0471250953.bi1110s43.From.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science (80-.).* 291, 1304–1351.

Verma, S. S., and Ritchie, M. D. (2018). Another Round of “Clue” to Uncover the Mystery of Complex Traits. *Genes (Basel).* 9. doi:10.3390/genes9020061.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five Years of GWAS Discovery. *Am. J. Hum. Genet.* 90, 7–24. doi:https://doi.org/10.1016/j.ajhg.2011.11.029.

Volders, P.-J., Lefever, S., Baute, S., Nuytens, J., Vanderheyden, K., Menten, B., et al. (2018). Targeted Genomic Screen Reveals Focal Long Non-Coding RNA Copy Number Alterations in Cancer Cell Lines. *Non-coding RNA* 4. doi:10.3390/ncrna4030021.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603.

Wang, Q., and Su, L. (2019). Vpr Enhances HIV-1 Env Processing and Virion Infectivity in Macrophages by Modulating TET2-Dependent IFITM3 Expression. *MBio* 10. doi:10.1128/mBio.01344-19.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214-20. doi:10.1093/nar/gkq537.

Weiner, J. (2019). pca3d: Three Dimensional PCA Plots. Available at: https://cran.r-project.org/package=pca3d.

Weir, B. S., and Cockerham, C. C. (1984). ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution* 38, 1358–1370. doi:10.1111/j.1558-5646.1984.tb05657.x.

Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., et al. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome.

Nat. Commun. 10, 5241. doi:10.1038/s41467-019-13212-3.

Whirl-Carrillo, M., McDonagh, E. M., Hebert, J. M., Gong, L., Sangkuhl, K., Thorn, C. F., et al. (2012). Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* 92, 414–417. doi:10.1038/clpt.2012.96.

Wilusz, C. J., and Wilusz, J. (2008). New ways to meet your (3′) end—oligouridylation as a step on the path to destruction. *Genes Dev.* 22, 1–7.

Winkler, C. A. (2008). Identifying Host Targets for Drug Development with Knowledge from Genome-wide Studies: Lessons from HIV-AIDS. *Cell Host Microbe* 3, 203–205. doi:10.1016/j.chom.2008.04.001.

Wonkam, A., Chimusa, E. R., Mnika, K., Pule, G. D., Ngo Bitoungui, V. J., Mulder, N., et al. (2020). Genetic modifiers of long-term survival in sickle cell anemia. *Clin. Transl. Med.* 10, e152.

Woollard, S. M., and Kanmogne, G. D. (2015). Maraviroc: a review of its use in HIV infection and beyond. *Drug Des. Devel. Ther.* 9, 5447–5468. doi:10.2147/DDDT.S90580.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.* 15, 323–354. doi:10.1111/j.1469-1809.1949.tb02451.x.

Wright, S. (1965). The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution (N. Y).* 19, 395–420. doi:10.2307/2406450.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi:10.1016/j.ajhg.2011.05.029.

Xie, W., Agniel, D., Shevchenko, A., Malov, S. V, Svitin, A., Cherkasov, N., et al. (2017). Genome-Wide Analyses Reveal Gene Influence on HIV Disease Progression and HIV-1C Acquisition in Southern Africa. *AIDS Res. Hum. Retroviruses* 33, 597–609. doi:10.1089/AID.2016.0017.

Xu, W., Presnell, S. R., Parrish-Novak, J., Kindsvogel, W., Jaspers, S., Chen, Z., et al. (2001). A

soluble class II cytokine receptor, IL-22RA2, is a naturally occurring IL-22 antagonist. *Proc. Natl. Acad. Sci. U. S. A.* 98, 9511–9516. doi:10.1073/pnas.171303198.

Yang, D., and Brunengraber, H. (2000). Glutamate, a window on liver intermediary metabolism. *J. Nutr.* 130, 991S–4S. doi:10.1093/jn/130.4.991S.

Zhang, W., and Dolan, M. E. (2010). Impact of the 1000 genomes project on the next wave of pharmacogenomic discovery. *Pharmacogenomics* 11, 249–256.

Zhu, J., Davoli, T., Perriera, J. M., Chin, C. R., Gaiha, G. D., John, S. P., et al. (2014). Comprehensive Identification of Host Modulators of HIV-1 Replication using Multiple Orthologous RNAi Reagents. *Cell Rep.* 9, 752–766. doi:https://doi.org/10.1016/j.celrep.2014.09.031.