

Whole Genome Rare-Variant Association Study of HIV-1 Progression in a Southern African Population

Prisca K. Thami^{1,2}, Wonderful Choga¹, Delesa D. Mulisa¹, Collet Dandara¹, Andrey K. Shevchenko³, Melvin M. Leteane⁴, Vlad Novitsky⁵, Stephen J. O'Brien^{6,7}, Myron Essex⁵, Simani Gaseitsiwe^{2,5} and Emile R. Chimusa^{1*}

¹ Division of Human Genetics, Department of Pathology, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, 7925, South Africa

² Botswana Harvard AIDS Institute Partnership, Gaborone, Botswana.

³ Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg 199034, Russia

⁴ Department of Biological Sciences, University of Botswana, Gaborone, Botswana.

⁵ Harvard T.H. Chan School of Public Health AIDS Initiative, Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, 02115, USA.

⁶ Laboratory of Genomics Diversity, Center for Computer Technologies, ITMO University, St. Petersburg, 197101, Russia

⁷ Guy Harvey Oceanographic Center Halmos College of Natural Sciences and Oceanography Nova Southeastern University, Ft Lauderdale, Florida, 33004, USA

Corresponding author: Emile R. Chimusa^{1*}

Division of Human Genetics, Department of Pathology,
Institute of Infectious Disease and Molecular Medicine,
University of Cape Town,
Anzio Road, Observatory, 7925,
Cape Town, South Africa
Tel: +27 21 406 6425 Email: emile.chimusa@uct.ac.za

ABSTRACT

Despite the high burden of HIV-1 in Botswana, the population of Botswana is significantly underrepresentation in host genetics 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 evaluate the role of rare-variants in susceptibility to HIV-1 and progression through whole genome sequencing. Whole genome sequences (WGS) of 265 HIV-1 positive and 125 were HIV-1 negative unrelated individuals from Botswana were mapped to the human reference genome GRCh38. Population joint variant calling was performed using Genome Analysis Tool Kit (GATK) and BCFTools. Cumulative effects of rare variant sets on susceptibility to HIV-1 and progression (CD4+ T-cell decline) were determined with optimized Sequence Kernel

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

Association Test (SKAT-O). *In silico* functional analysis of the prioritized variants was performed through gene-set enrichment using databases in GeneMANIA and Enrichr. 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. Functional analysis revealed that these genes are involved in viral translation and transcription. These findings highlight the significance of whole genome sequencing in pinpointing rare-variants of clinical relevance. The research contributes towards a deeper understanding of the host genetics HIV-1 and offers promise of population specific interventions against HIV-1.

Key words: Rare-variant Association, Whole Genome Sequencing, host genetics, HIV-1, progression

INTRODUCTION

Human immunodeficiency virus (HIV) infection remains one of the world's worst pandemics. Over 37 million people lived with HIV by the end of 2018 globally. About half of these live in Southern and eastern Africa. Botswana is the third most affected country in Southern Africa, after eSwatini and South Africa in first and second positions respectively (1). 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 (2).

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 (3,4). HIV prevalence in Botswana has since lowered to 20.3% among adults (1). 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 (5–7). Treatment against HIV-1 does not offer cure and to date no effective vaccine has been found against HIV infection (8,9). 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

(10–12). 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 (13,14). This finding has successfully been translated into virus entry inhibitor antiretrovirals (15,16). Other genes identified through candidate gene method include *HLA-A*, *HLA-B* and *HLA-C*, *CCR2*, *SDF1*, *IL10* (17,18), *CCL5* (RANTES), *KIR* genes, *TRIM5* and *APOBEC3G* (10–12).

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 (19). 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* (20,21), *NOTCH4* (21,22), *C6orf48* (22,23) in European populations.

While the bulk of genome-wide association studies (GWAS) were carried out in European populations, few GWAS of African populations have been published to date (24–26). Of note, two novel variants within *HCG22* and *CCNG1* genes were found to be associated with progression and acquisition of HIV1 in an African (Botswana) population (26). We have previously discussed the results of GWAS of HIV-1 extensively here (27).

Genome-wide arrays revealed many variants of clinical significance to acquisition, control of HIV and progression (28). 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 (27–29).

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 (21,29). It has been proposed that this missing heritability of HIV-1 phenotypes may be hidden in rare variants among other factors (6,27). 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 may be of clinical significance to the acquisition of HIV-1.

RESULTS

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} . Substructure among the HIV-1 positive/negative individuals was not observed, indicating similar genetic exposure among HIV-1 samples (30). Since no population has an entirely homogeneous genetic architecture, hidden relatedness and population structure have been accounted for in the current genetic association study.

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 1**) in the GWAS of HIV-1.

Table 1. 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	<i>TET1</i>	Known	22.4	2.74×10^{-5}
6	164083809	164088302	6q26	3	<i>LOC105378106</i>	Novel	16.8	4.25×10^{-5}
11	94545330	94740356	11q21	22	<i>LOC105369438</i>	Novel	21.9	5.78×10^{-5}
11	42209293	42275240	11p12	24	<i>LOC100507205</i>	Novel	38.7	8.83×10^{-5}
2	6728177	6770311	2p25.2	7	<i>LINC00487</i>	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).

5.3.4.2 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)* genes 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 2, Table S1**). The λ_{GC} of the rare-variants was 0.99, which suggests that population structure and confounding was adequately controlled (**Figure 1**).

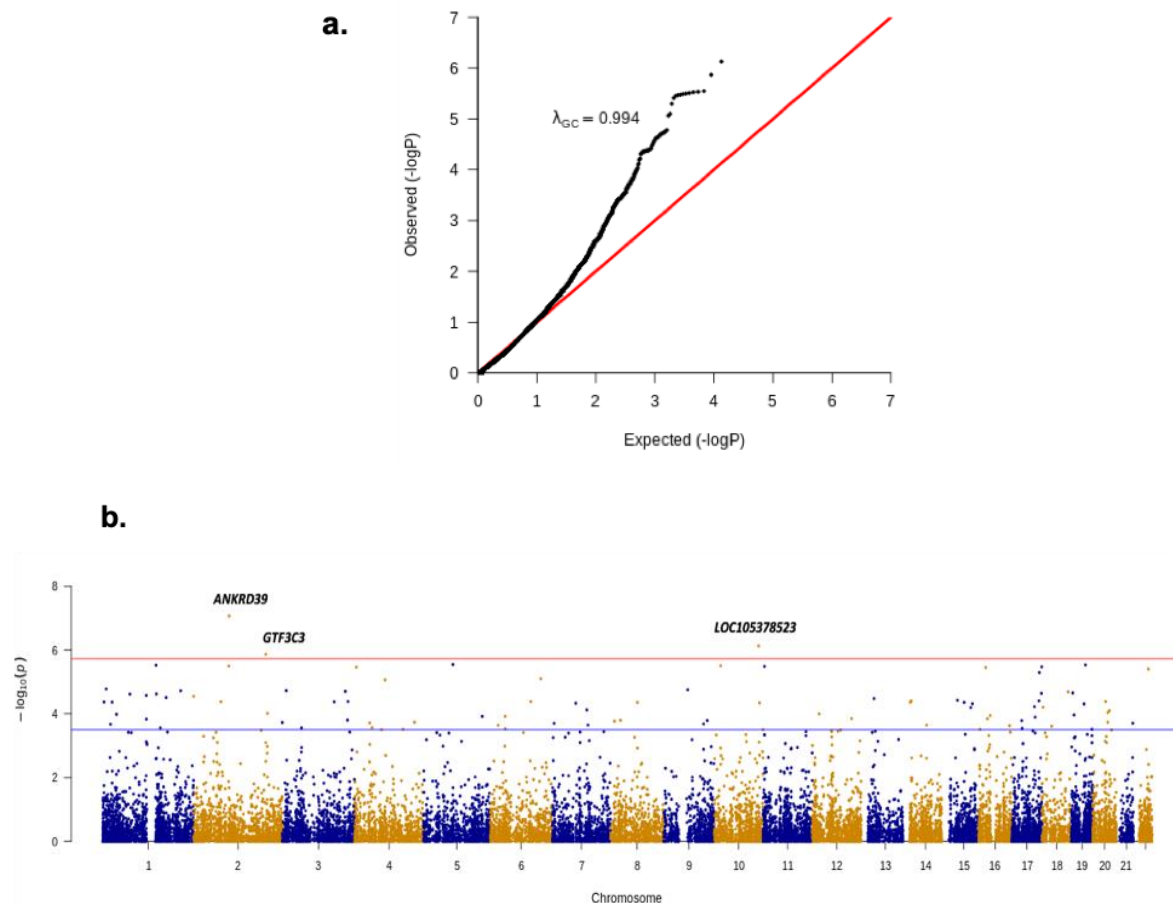


Figure 1. Quantile-quantile plot (with λ_{GC}) and Manhattan plot of HIV-1 rare-variant association with HIV-1 progression. a. Quantile-quantile plot and λ_{GC} of HIV-1 rare-variant association with CD4+ T-cell counts showing $-\log_{10}$ p-value of each variant is plotted against the expected null (the red line). b. Manhattan plot of HIV-1 rare-variant association with CD4+ T-cell counts showing $-\log_{10}$ p-value of each variant plotted against its genomic position. The red line is the $-\log_{10}$ p-value cut-off (1.86×10^{-6}).

Table 2. 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	<i>ANKRD39</i>	Novel	8.48×10^{-8}
10	121291608	121322945	10q26.12	3	<i>LOC105378523</i>	Novel	7.45×10^{-7}
2	196763035	196799725	2q33.1	7	<i>GTF3C3</i>	Novel	1.36×10^{-6}
5	79976716	79991262	5q14.1	1	<i>MTX3</i>	Novel	2.84×10^{-6}
19	38619082	38636955	19q13.2	1	<i>EIF3K</i>	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.

Gene-set enrichment of the strongest effects of genetic association of HIV-1 progression

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 2**). 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 = 2.51×10^{-6}].

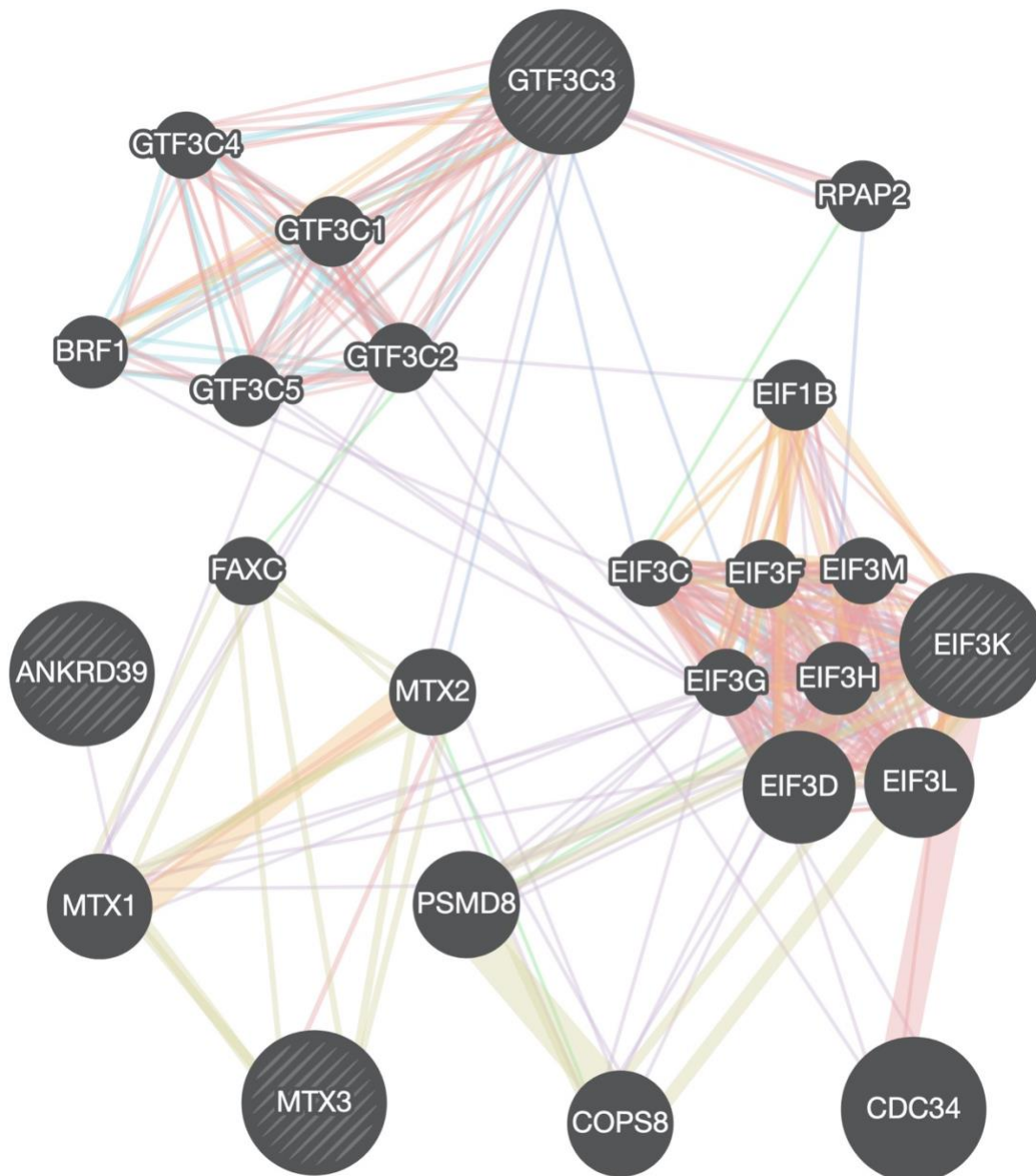


Figure 2. Gene interaction network of candidate genes identified from genetic association of HIV-1 progression. The different colours of branches represent how the genes are related; pink: physical interactions, purple: co-expression, orange: predicted, navy blue: co-localization, blue: Pathway, green: Genetic interactions, yellow: shared protein domains. Black and striped nodes: genes provided as input into GeneMANIA (**Table 3**). Black only nodes: genes predicted by GeneMANIA to interact with the input list.

Table 3. 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 x 10 ⁻¹⁹	9.10 x 10 ⁻¹⁶	Biological Process 2018 (31) http://www.informatics.jax.org/
Transcription from RNA polymerase III promoter (GO:0006383)	1.30 x 10 ⁻¹³	9.46 x 10 ⁻¹¹	
Cytoplasmic translational initiation (GO:0002183)	4.43 x 10 ⁻⁹	2.51 x 10 ⁻⁶	
RNA polymerase III transcription factor complex (GO:0090576)	1.21 x 10 ⁻¹⁸	5.38 x 10 ⁻¹⁶	Cellular Component 2018 (31) http://www.informatics.jax.org/
Transcription factor TFIIC complex (GO:0000127)	7.94 x 10 ⁻¹⁷	1.77 x 10 ⁻¹⁴	
Translation initiation factor activity (GO:0003743)	1.57 x 10 ⁻³⁷	1.81 x 10 ⁻³⁴	Molecular Function 2018 (31) http://www.informatics.jax.org/
RNA binding (GO:0003723)	8.89 x 10 ⁻⁷	3.41 x 10 ⁻⁴	
Pathways			
RNA polymerase III transcription initiation from type 2 promoter	2.27 x 10 ⁻¹¹	4.29 x 10 ⁻⁹	BioPlanet 2019 (32)
Translation Factors	2.84 x 10 ⁻²²	4.28 x 10 ⁻¹⁹	WikiPathways 2019 Human (Slenter et al, 2017)
			BioCarta 2016 http://www.biocarta.com

P-value_{adj}: adjusted P-value.

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 = 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) (33). In our study, the cumulative effects of three 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 (**Table 2**).

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 (34). An artificial ankyrin repeat domain was found to negatively interfere with HIV-1 Gag

assembly and budding (35). 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 (26). 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 2**). The significant association of *LOC105378523* with CD4+ T-cell counts 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 (36–39). Lastly, the *GTF3C3* gene (**Table 2**) encodes a transcription factor TFIIC that is involved in transcription of transfer ribonucleic acid (tRNA) and directly binds to virus-associated RNA promoters (34). The transcription factor TFIIC was observed to mediate HIV-1 transcription in HeLa and Jurkat T cells (40).

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 3**). The RNA polymerase III transcription initiation pathway is represented by the *GTF3C* genes cluster (**Figure 2**). Upon integration into the human genome, HIV-1 RNA transcription is mediated by RNA polymerase II (41). Though the product of the *GTF3C3* gene has been linked to upregulation of HIV-1 RNA (40), in addition to mRNA synthesis, HIV-1 specific small nuclear RNAs can also be produced from RNA polymerase III promoters (42–44). HIV-1 hijacks the host translation machinery to facilitate its viral proteins. The human *EIF3* gene products (**Figure 2**) are some of the targets for optimal HIV-1 translation (45). 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 statistical significance in the rare-variant association test of susceptibility to HIV-1 acquisition. The strongest effects were within the *TET1* (p-value = 2.74×10^{-5}), *LOC105378106* (p-value = 4.25×10^{-5}), *LOC105369438* (p-value = 5.78×10^{-5}), *LOC100507205* (p-value = 8.83×10^{-5}) and *LINC00487* (p-value = 9.61×10^{-5}) gene (**Table 1**). The product of the *TET1* gene is a demethylase that plays a role in DNA methylation and gene expression (34). TET (Ten-eleven translocation) family of proteins consists of three paralogs: TET1, TET2, and TET3 (46). It has been shown that the HIV-1 Vpr protein facilitates

the degradation of TET-2 to promote HIV-1 replication (47,48). 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 (48). 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 1**). There has been a recent discovery of the association of a *CCR5* dependent lncRNA with susceptibility to HIV-1 (49), therefore, though the variant sets did not reach statistical significance, the identification of rare variant sets within lncRNAs in the current study (**Table 1**) cannot be entirely overlooked.

The cumulative effects of rare variants within the ANKRD39, LOC105378523 and GTF3C3 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. Identifying these candidate disease genes warrants their further investigation as a potential target in HIV-1 pathogenesis in African populations. Although additional data was not available at the time of analysis, in future studies increasing the sample-size could improve the results of susceptibility to HIV-1 acquisition. There is also need for a development of statistical test rare-variant association that is designed to handle longitudinal phenotypic data as the current optimized rare-variant association tests are limited in this regard.

MATERIALS AND METHODS

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).

Selection of study participants and sample preparation

The study participants and sample preparation have been described previous in our previous

work (30). 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 (50).

Whole genome sequence Data Description

Quality assessment was performed on paired-end WGS (minimum of 30X depth) in FASTQ format (51) using FastQC (52). Low-quality sequence bases and adapters were trimmed using Trimmomatic with default parameters (53). The sequencing reads were aligned to the GRCh38 human reference genome using Burrows-Wheeler Aligner (BWA-MEM) (54,55) 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 (56) and Genome Analysis Toolkit (57).

We performed population joint calling (58,59) using two different methods to leverage the quality and accuracy of our results: GATK HaplotypeCaller (57,60) and BCFtools (56). The variant call format (VCF) dataset was filtered using VCFTOOLS (61), GATK Variant Quality Score Recalibration and BCFtools. The specific filtering parameters employed for both call-sets have been detailed (30). Downstream analyses were performed with GATK call-set and BCFtools call-set used as a validation set. Further quality control as required prior genetic association tests was performed using PLINK (62). 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.

Rare-variant association tests

To account for sample size and rare variants, (that standard GWAS could have missed or not reached the genome-wide significant level) we leverage possible small effects by aggregating SNPs effects at gene level with an optimal unified sequence kernel association test (SKAT-O) (63,64). This test combined burden and variance-component analyses using the SKAT package (65) 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. A

p-value of 0.05/genes was considered significant; where genes is the number of genes tested in the model.

For HIV-1 progression, we regressed CD4+ T-cell counts over several covariates (**Table S2**) using lmer function in the R lme4 package (66). We then used the regression coefficients (slopes) of each individual to infer CD4+ T-cell changes over a period of at least 18 months (236 young women). A number of models were tested to select confounding variables and interacting factors (Table S1). 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.

Pathways enrichment analysis and gene-gene interactions

Functional analysis was performed to elucidate variant effect mechanisms by identifying putatively affected biological processes and pathways. The genes identified in the rare-variant association of HIV-1 progression were subjected to gene-set enrichment analysis using GeneMANIA (67) and Enrichr (68) bioinformatics tools.

ACKNOWLEDGEMENTS

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].

FUNDING

PKT is funded by the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative [grant # DEL-15-006].

DECLARATION OF INTEREST

The authors declare that they have no competing interests.

AUTHORS CONTRIBUTIONS

PKT and ERC conceived and structured the content of the manuscript. PKT conducted data analysis and result interpretation. PKT drafted the manuscript. PKT, ERC, DDM, WTC, CD, MML, VN, SJO, ME and SG edited the manuscript.

REFERENCES

1. UNAIDS (2019) UNAIDS data 2019. UNAIDS data 2019 <https://www.unaids.org/en/regionscountries/countries/botswana> (accessed Mar 6, 2020).
2. Essex, M. (1999) Human immunodeficiency viruses in the developing world. *Adv. Virus Res.*, **53**, 71–88.
3. Escudero, D. J., Marukutira, T., McCormick, A., et al. (2019) Botswana should consider expansion of free antiretroviral therapy to immigrants. *J. Int. AIDS Soc.*, **22**, e25328.
4. Farahani, M., Vable, A., Lebelonyane, R., 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.
5. Carr, D. F., Bourgeois, S., Chaponda, 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.
6. Telenti, A. and Goldstein, D. B. (2006) Genomics meets HIV-1. Genomics meets HIV-1. *Nat. Rev. Microbiol.* (2006), **4**, 865–873.
7. Pereyra, F., Jia, X., McLaren, P. J., et al. (2010) The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*, **330**, 1551–1557.
8. 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.
9. Avert (2019) Searching for a cure for HIV and AIDS. Searching for a cure for HIV and AIDS <https://www.avert.org/professionals/hiv-science/searching-cure> (accessed Mar 23, 2020).
10. Hutcheson, H. B., Lautenberger, J. A., Nelson, G. W., et al. (2008) Detecting AIDS restriction genes: from candidate genes to genome-wide association discovery.

- Vaccine*, **26**, 2951–2965.
11. O’Brien, S. J. and Nelson, G. W. (2004) Human genes that limit AIDS. *Nat. Genet.*, **36**, 565–574.
 12. 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.
 13. Dean, M., Carrington, M., Winkler, C., 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.
 14. Ioannidis, J. P., Rosenberg, P. S., Goedert, J. J., 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.
 15. Henrich, T. J. and Kuritzkes, D. R. (2013) HIV-1 entry inhibitors: recent development and clinical use. *Curr. Opin. Virol.*, **3**, 51–57.
 16. 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.
 17. O’Brien, S. J., Nelson, G. W., Winkler, C. A., et al. (2000) POLYGENIC AND MULTIFACTORIAL DISEASE GENE ASSOCIATION IN MAN: Lessons from AIDS. *Annu. Rev. Genet.*, **34**, 563–591.
 18. Carrington, M. and O’Brien, S. J. (2003) The influence of HLA genotype on AIDS. *Annu. Rev. Med.*, **54**, 535–551.
 19. Hirschhorn, J. N. and Daly, M. J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.
 20. Fellay, J., Shianna, K. V., Ge, D., et al. (2007) A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science (80-.)*, **317**, 944–947.
 21. Fellay, J., Ge, D., Shianna, K. V, et al. (2009) Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.*, **5**, e1000791.
 22. Le Clerc, S., Coulonges, C., Delaneau, O., 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.
 23. Le Clerc, S., Limou, S., Coulonges, C., 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.
24. Lingappa, J. R., Petrovski, S., Kahle, E., 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.
 25. Petrovski, S., Fellay, J., Shianna, K. V, et al. (2011) Common human genetic variants and HIV-1 susceptibility: a genome-wide survey in a homogeneous African population. *AIDS*, **25**, 513–518.
 26. Xie, W., Agniel, D., Shevchenko, A., 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.
 27. 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.
 28. 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.
 29. Telenti, A. and Johnson, W. E. (2012) Host genes important to HIV replication and evolution. *Cold Spring Harb. Perspect. Med.*, **2**, a007203.
 30. Thami, P. K., Mulisa, D. D., Choga, W. T., et al. (2020) Whole Genome Sequencing-based Characterization of Human Genome Variation and Mutation Burden in Botswana. *bioRxiv*.
 31. Harris, M. A., Clark, J., Ireland, A., et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258-61.
 32. Huang, R., Grishagin, I., Wang, Y., 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.
 33. Mackelprang, R. D., Bamshad, M. J., Chong, J. X., 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.
 34. GeneCards (2020) GeneCards - Human Gene Database. GeneCards - Human Gene Database <https://www.genecards.org/> (accessed Aug 3, 2020).
 35. Nangola, S., Urvoas, A., Valerio-Lepiniec, M., et al. (2012) Antiviral activity of recombinant ankyrin targeted to the capsid domain of HIV-1 Gag polyprotein.

- Retrovirology*, **9**, 17.
36. Shen, L., Wu, C., Zhang, J., et al. (2020) Roles and potential applications of lncRNAs in HIV infection. *Int. J. Infect. Dis.*, **92**, 97–104.
 37. 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.
 38. Chao, T.-C., Zhang, Q., Li, Z., et al. (2019) The Long Noncoding RNA HEAL Regulates HIV-1 Replication through Epigenetic Regulation of the HIV-1 Promoter. *MBio*, **10**.
 39. Trypsteen, W., Mohammadi, P., Van Hecke, C., et al. (2016) Differential expression of lncRNAs during the HIV replication cycle: an underestimated layer in the HIV-host interplay. *Sci. Rep.*, **6**, 36111.
 40. 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 TFIIC. *J. Acquir. Immune Defic. Syndr.*, **5**, 1142–1147.
 41. Liu, R., Wu, J., Shao, R., et al. (2014) Mechanism and factors that control HIV-1 transcription and latency activation. *J. Zhejiang Univ. Sci. B*, **15**, 455–465.
 42. Boden, D., Pusch, O., Lee, F., et al. (2003) Promoter choice affects the potency of HIV-1 specific RNA interference. *Nucleic Acids Res.*, **31**, 5033–5038.
 43. Ratnasabapathy, R., Sheldon, M., Johal, L., et al. (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.
 44. Gunnery, S. and Mathews, M. B. (1995) Functional mRNA can be generated by RNA polymerase III. *Mol. Cell. Biol.*, **15**, 3597–3607.
 45. Guerrero, S., Batisse, J., Libre, C., et al. (2015) HIV-1 replication and the cellular eukaryotic translation apparatus. *Viruses*, **7**, 199–218.
 46. Akahori, H., Guindon, S., Yoshizaki, S., et al. (2015) Molecular Evolution of the TET Gene Family in Mammals. *Int. J. Mol. Sci.*, **16**, 28472–28485.
 47. Lv, L., Wang, Q., Xu, Y., 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.
 48. 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**.

49. Kulkarni, S., Lied, A., Kulkarni, V., et al. (2019) CCR5AS lncRNA variation differentially regulates CCR5, influencing HIV disease outcome. *Nat. Immunol.*, **20**, 824–834.
50. Novitsky, V., Woldegabriel, E., Wester, C., et al. (2008) Identification of primary HIV-1C infection in Botswana. *AIDS Care*, **20**, 806–811.
51. Cock, P. J. A., Fields, C. J., Goto, N., et al. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.*, **38**, 1767–1771.
52. Van Der Auwera, G. A., Carneiro, M. O., Hartl, C., et al. (2014) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma.*, **11**.
53. Bolger, A. M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
54. Li, H., Ruan, J., Durbin, R., et al. (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.
55. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
56. 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.
57. McKenna, A., Hanna, M., Banks, E., 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.
58. Pfeifer, S. P. (2017) From next-generation resequencing reads to a high-quality variant data set. *Heredity (Edinb.)*, **118**, 111–124.
59. Nielsen, R., Paul, J. S., Albrechtsen, A., et al. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**, 443–451.
60. DePristo, M. A., Banks, E., Poplin, R., et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–8.
61. Danecek, P., Auton, A., Abecasis, G., et al. (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
62. Purcell, S., Neale, B., Todd-Brown, K., et al. (2007) PLINK: a tool set for whole-

- genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
63. Lee, S., Emond, M. J., Bamshad, M. J., et al. (2012) 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.
 64. Lee, S., Wu, M. C. and Lin, X. (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775.
 65. Wu, M. C., Lee, S., Cai, T., et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.
 66. Bates, D., Mächler, M., Bolker, B., et al. (2015) Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, **67**, 1–48.
 67. Warde-Farley, D., Donaldson, S. L., Comes, O., et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214-20.
 68. Kuleshov, M. V, Jones, M. R., Rouillard, A. D., et al. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90-7.