Whole-Exome Sequencing Reveals Uncaptured Variation and Distinct Ancestry in the Southern African Population of Botswana

Gaone Retshabile,[1] Busisiwe C. Mlotshwa,[1] Lesedi Williams,[1] Savannah Mwesigwa,[2] Gerald Mboowa,[2,3] Zhuoyi Huang,[4] Navin Rustagi,[4] Shanker Swaminathan,[5,6] Eric Katagirya,[2] Samuel Kyobe,[2] Misaki Wayengera,[3] Grace P. Kisitu,[7] David P. Kateete,[2,3] Eddie M. Wampande,[2,8] Koketso Maplanka,[1] Ishmael Kasvosve,[9] Edward D. Pettitt,[10] Mogomotsi Matshaba,[10,11] Betty Nsangi,[7] Marape Marape,[10] Masego Tsimako-Johnstone,[1] Chester W. Brown,[5,12] Fuli Yu,[4,5] Adeodata Kekitiinwa,[7,11] Moses Joloba,[2] Sununguko W. Mpoloka,[1] Graeme Mardon,[5,13] Gabriel Anabwani,[10,11] Neil A. Hanchard,[5,6,*] and for the Collaborative African Genomics Network (CAfGEN) of the H3Africa Consortium

Large-scale, population-based genomic studies have provided a context for modern medical genetics. Among such studies, however, African populations have remained relatively underrepresented. The breadth of genetic diversity across the African continent argues for an exploration of local genomic context to facilitate burgeoning disease mapping studies in Africa. We sought to characterize genetic variation and to assess population substructure within a cohort of HIV-positive children from Botswana—a Southern African country that is regionally underrepresented in genomic databases. Using whole-exome sequencing data from 164 Batswana and comparisons with 150 similarly sequenced HIV-positive Ugandan children, we found that 13%–25% of variation observed among Batswana was not captured by public databases. Uncaptured variants were significantly enriched ($p = 2.2 \times 10^{-16}$) for coding variants with minor allele frequencies between 1% and 5% and included predicted-damaging non-synonymous variants. Among variants found in public databases, corresponding allele frequencies varied widely, with Botswana having significantly higher allele frequencies among rare (<1%) pathogenic and damaging variants. Batswana clustered with other Southern African populations, but distinctly from 1000 Genomes African populations, and had limited evidence for admixture with extra-continental ancestries. We also observed a surprising lack of genetic substructure in Botswana, despite multiple tribal ethnicities and language groups, alongside a higher degree of relatedness than purported founder populations from the 1000 Genomes project. Our observations reveal a complex, but distinct, ancestral history and genomic architecture among Batswana and suggest that disease mapping within similar Southern African populations will require a deeper repository of genetic variation and allelic dependencies than presently exists.

Introduction

Genomic studies have played a crucial role in enhancing knowledge of genetic variation between and within populations[1] and have presented a new lens through which to view the genetic basis of disease.[1–3] Such surveys have consistently indicated a broad genetic diversity and complex ancestry among African populations, characterized by migration out of Africa and subsequent back migration, leading to gene flow between and within both African and non-African populations.[3–9]

Language and geographical distance have been shown to have the strongest correlation with differences in genetic variation between populations.[4–6,9–12] For instance, Bantu-speaking populations—the largest language grouping in sub-Saharan Africa—appear separate from non-Bantu groups such as the Khoisan on principal component and F_{ST} analysis and have been shown to have differing levels of recent return-admixture events within their genomes, particularly among coastal populations such as those in Kenya.[3,6,7,9,11,13,14] Southern African populations also separate from—and exhibit varying patterns of admixture when contrasted with—East and West African populations.[3,6,9–11,15] Despite this marked diversity in African populations, ascertainment bias in genotyping technologies and limited sampling have meant that African populations remain vastly underrepresented and poorly characterized at the genome level, a stark oversight in an era of large-scale genomics.[3,9,16]

Genomic underrepresentation is particularly true of Southern African populations,[1,3,9,17] especially when compared to populations from West and East Africa.[1,18] This lack of data adversely impacts the interpretation of medical genomic and disease association studies in

[1]Department of Biological Sciences, University of Botswana, Gaborone, Botswana; [2]Department of Medical Microbiology, College of Health Sciences, Makerere University, Kampala, Uganda; [3]Department of Immunology and Molecular Biology, College of Health Sciences, Makerere University, Kampala, Uganda; [4]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; [5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [6]USDA/ARS/Children's Nutrition Research Center, Baylor College of Medicine, Houston, TX 77030, USA; [7]Baylor College of Medicine Children's Foundation, Kampala, Uganda; [8]Department of Bio-molecular Resources, College of Veterinary Medicine, Makerere University, Kampala, Uganda; [9]Department of Medical Laboratory Sciences, University of Botswana, Gaborone, Botswana; [10]Botswana-Baylor Children's Clinical Centre of Excellence, Gaborone, Botswana; [11]Pediatric Retrovirology, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA; [12]University of Tennessee Health Science Center, Memphis, TN 38105, USA; [13]Department of Pathology and Immunology, Baylor College of Medicine, Houston, TX 77030, USA
*Correspondence: hanchard@bcm.edu
https://doi.org/10.1016/j.ajhg.2018.03.010.

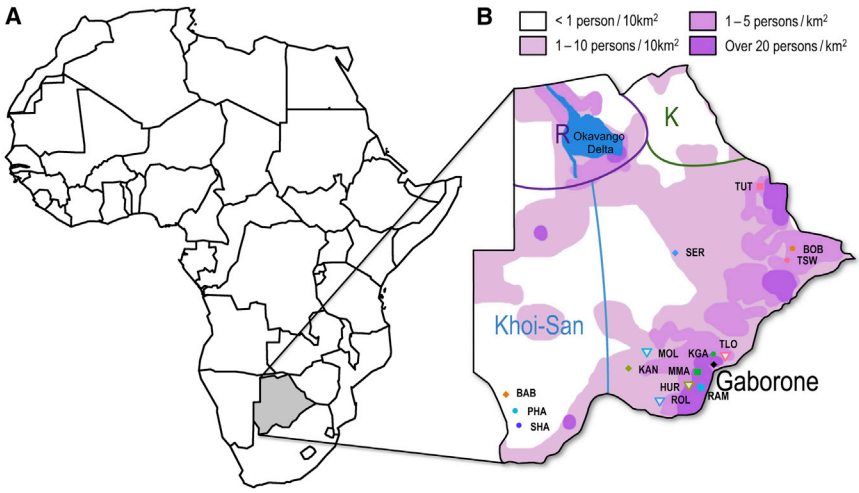

Figure 1. Geographical Location of Botswana in Africa

(A) The location of Botswana (gray-shaded) on the African continent.

(B) The approximate regions within Botswana where self-reported ethnicities represented in the study are traditionally located. Abbreviations: TUT, Bakalanga; SHO, Shona; BOB, Babirwa; TSW, Batswapong; SER, Bangwato; KKRA, Bakgatla-Ba-Kgafela; TLO, Batlokwa; MOL, Bakwena; MMA, Bakgatla-Ba-Mmanaana; HUR, Bahurutshe; KAN, Bangwaketse; ROL, Barolong; RAM, Balete; BAB, Babolaongwe; SHA, Bashaga; PHA, Baphaleng. Color shades give the approximate population density across the country. Regions marked 'K' and 'R' correspond to the primary concentrations of individuals belonging to the K and R Guthrie groups.

these groups.[2,3,19–21] The Southern African country of Botswana provides such an example; it is one of several Southern African countries occupying a large geographical region of Africa, sharing its borders with Namibia, Zimbabwe, Zambia, and South Africa (Figure 1). The people of Botswana—collectively referred to as Batswana—represent an amalgam of populations from multiple ethnicities, languages, and ancestry groups (Figure 1).[22–26] Historically, the southeastern and central areas of the country, which include the capital Gaborone, were largely populated by ethnic groups speaking Setswana, a term used to describe a cluster of Sotho-Tswana-derived, and closely related, Bantu languages.[23] Migrant Southeastern and Western Bantu groups interacted with Khoisan hunter-gatherers and pastoralists along migration routes northward and southward, respectively, into modern day Botswana.[25–27] The Batswana thus share deep ancestral roots with established African ethnic groups; however, a subsequent history of cultural customs that included polygynous marriage, often-disputed patrilineal succession, and tribal schisms[22,24,28,29] and the potential admixture of Bantu ancestry individuals with Khoisan[30,31] and, to a limited extent, Eurasian ancestry individuals[22] means that the genetic ancestry among Batswana is difficult to extrapolate from other African ethnic groups. Further complicating definitions of ethnic identity, anecdotal reports indicate that ethnic determination among Batswana is defined on a patrilineal basis, defaulting to the mother's ethnicity for births that occur out of wedlock. Notwithstanding the aforementioned cultural complexities, previous genetic surveys of Batswana have been limited to a small number of geographically limited individuals, mainly of Khoisan ancestry, and included only a handful of loci.[30–32] Therefore, the broader genetic variation and substructure among the Batswana remains largely undescribed.[33]

The Collaborative African Genomics Network (CAfGEN), under the auspices of the Human Heredity and Health in Africa (H3Africa) Consortium (Web Resources),[34,35] has among its primary goals the use of genomics to study pediatric HIV and TB disease progression in the sub-Saharan countries of Uganda, Botswana, and, more recently, Swaziland. The ability to utilize the wealth of genetic diversity within these countries to better understand phenotypic variability in HIV and other prevalent diseases is highly desirable;[1,3,16] however, the present dearth of available population-level genomic data[30–32] makes attaining such a goal challenging. Therefore, to fulfil the primary aims of CAfGEN, we sought to provide a reference framework for medically relevant genomics studies in our Southern African population of Botswana by characterizing genetic variation and assessing genetic substructure within our cohort. We used whole-exome sequencing (WES),[36] complemented in part by genome-wide SNP genotyping, from 164 pediatric HIV-affected case subjects from Botswana to assess genetic variation and population substructure based on self-reported ethnicity and Guthrie language groups. WES is attractive for variant characterization in underrepresented populations[37] due to its comparatively low cost,[38] its amenability to unbiased variant discovery, and the relative ease of interpretation of the molecular impact of discovered variants.[39,40] We compared the frequency and burden of rare variants within the Botswana population with data from public databases and 150 pediatric CAfGEN case subjects from the East African country of Uganda—recruited under the same protocol and sequenced on the same platform—and contextualized population-level variation and admixture with publicly available datasets.

Subjects and Methods

Study Samples

Samples were collected as part of a study of pediatric HIV disease progression within CAfGEN,[35] and all participants were HIV positive. The 164 Batswana participants were recruited through the Botswana-Baylor Children's Clinical Centre of Excellence—a tertiary pediatric HIV referral center in Gaborone, the capital of Botswana, which has the largest density of the country's population of ~2 million (Figure 1). The 150 Ugandan participants were recruited at the Baylor College of Medicine Children's Foundation in Kampala, Uganda, under the same consent and

protocol, and both cohort groups underwent the same genomic studies at the same time (Table S1). Approvals for the study were obtained from the institutional review boards of each of the participating institutions in CAfGEN. Genetic variation among the population of Uganda is well-represented among the ~2,000 samples from the African Genome Variation Project (AGVP)[3] and is more broadly proxied by the East African Luhya in Webuye, Kenya (LWK) from the 1000 Genomes project;[1] both of these datasets are currently included among public databases. For our analyses, therefore, the use of the Uganda dataset was predominantly as a sequencing control to account for potential differences in disease ascertainment and capture platform between the Botswana dataset and available public data.

Self-reported ethnicity data were recorded for all participants and this was used to infer language group for each ethnic group from Botswana using the Ethnologue website database (Web Resources) as per the Guthrie classification of Bantu languages.[23] The 164 Botswana participants included 18 self-reported ethnicities and 8 Guthrie Bantu language classes (Table S2). More than half of all Botswana participants self-reported belonging to ethnic groups associated with the Tswana language, with most groups falling into the Guthrie S-class of languages, which are associated with the Southern Bantu; one participant self-reported as Herero, which is associated with the Western Bantu.

DNA Processing and Sequencing

DNA was collected from whole blood after informed consent and was quantified for quality control (QC) using a Thermo Scientific NanoDrop 2000 and underwent Picogreen quantification on the Tecan Genios Basic plate reader. Sequencing was undertaken at the Baylor College of Medicine Human Genome Sequencing Center (HGSC). Paired-end 100 bp read libraries were prepared and sequenced on an Illumina HiSeq 2500 machine after whole-exome capture using the VCRome v.2.1 capture kit[41] with an in-house spike-in called Panel-Killer v.2 (PKv2) for 209 of the 314 samples. The final batch of 105 Uganda samples was captured using VCRome v.2.1 with a newer spike-in (PKv3). Both spike-in versions are designed to capture low-coverage targets in VCRome v.2.1. At least 96% of the bases targeted were covered at a depth >20× and the sequences were aligned to the human genome reference build 37 using BWA-mem (v0.75).[42] Filtered variants were then imported into Variant Tools v2.6.1[43] and were further annotated using data within publicly available databases: 1000 Genomes,[1] dbSNP (see Web Resources), Exome Aggregation Consortium (ExAC),[44] and ANNOVAR.[45]

Joint Calling of Sequence Data

Variants were jointly called across Botswana and Uganda samples using the Genome Analysis Toolkit (GATK v.3.5.0) Haplotype caller.[46,47] Sequence variant quality was assessed using SnpEff (v.4.22, build 2015-12-05)[48] and VCFTools (v.01.1.12).[49] Variants were initially filtered to have a minimum depth of 10×, Phred quality score > 30, and genotype quality score > 20; thereafter, we removed variants with >5% missingness. A total of 600,965 high-quality variants remained after QC and filtering, inclusive of 194,186 exonic variants. Variants utilized had a median of average depth of 92× and transition/transversion (TiTv) ratios within expected ranges:[38] 2.49 for non-exonic variants and 3.10 for exonic variants. In Uganda the median depth of sequencing coverage per-sample was 72×; similarly, TiTv ratios were within the expected range: 2.56 for non-exonic variants and 3.11 for exonic variants.

Exon Definition and Coding Variant Annotation

To define exons for the WES data, we used Variant Tools (v.2.6.1) and selected the UCSC KnownGene exon definition from the KnownGene database.[50] Exonic data were then annotated on the basis of exon coding sequence start and end sites as defined within the database. Additional annotation was conducted on the coding sequence variants with the online version of the Variant Effect Predictor (VEP) tool (Ensembl GRCh37 release 88),[51] using the following parameters: exon and exon position defined; biotype defined; APPRIS, which accounts for alternative splicing; Consensus Coding Sequence identifier; Condel,[52] which is a consensus score of deleteriousness from SIFT[53] and PolyPhen-2;[54] and canonical transcripts. Annotations were then visualized with R package ggplot2 v.2.10[55] within the R statistical software v.3.2.4 (see Web Resources). Uncaptured low-frequency variants were selected using Variant Tools and annotated using VEP (Ensembl GRCh37 release 88).

BeadChip Microarray Genotyping

Concurrently with exome sequencing, samples from both populations were also genotyped for quality control using the Illumina HumanCoreExome-24 BeadChip kit v.1.0, which has 547,644 markers of which 265,919 are exonic. There was at least 90% self-concordance in variant genotype calls between sequencing and array platforms in both country cohorts, which was the minimum self-concordance ratio QC threshold. SNP genotyping quality control was performed using PLINK v.1.90b3.36.[56] No samples were excluded on the basis of excessive heterozygosity, which had been defined as samples exceeding 5 standard deviations of the mean heterozygosity. A set of independent SNPs for ancestry inference was obtained using linkage disequilibrium (LD)-based pruning ($r^2 > 0.2$) and variants out of Hardy-Weinberg equilibrium ($p < 0.0001$) and samples with close familial relationship (PI_HAT > 0.1), limited to one of a pair of sisters (with PI_HAT > 0.5) from Botswana, were removed prior to analysis.

Principal Component Analysis

Principal component analyses (PCA) were used to distinguish cohort individuals among global populations represented in publicly available data. Data were merged with 1000 Genomes phase 3 data for 2,504 individuals, representing 5 continental ancestries, as well as genotyping and sequence data from the AGVP for 86 Sotho and 100 Zulu individuals, respectively. Shared markers between Batswana and 1000 Genomes populations were identified and extracted using Variant Tools (v.2.6.1)[43] and the resulting variants were exported to PLINK[57] v.1.9 to subset the datasets to the shared markers. QC was undertaken for each population before merging the datasets. Retained markers had a minor allele frequency (MAF) > 0.05, passed the Hardy-Weinberg equilibrium test at a p value of 0.0001, and were deemed independent on the basis of pairwise LD pruning using a window of 1,000 base pairs advanced by 100 SNPs at a time and an r-squared coefficient of 0.2 (see Table S4). The median PI_HAT, assessed in PLINK, for the Sotho and Zulu datasets were 0 and 0.0136, respectively, with no pair of individuals above a PI_HAT value of 0.1. Plots were then visualized with R package ggplot2 v.2.10[55] within the R statistical software v.3.2.4. Comparisons of the Botswana cohort with close African populations was undertaken after excluding Afro-Caribbean in Barbados (ACB) and the African American in Southwest USA (ASW) individuals from the 1000 Genomes African populations and retaining the Sotho and Zulu AGVP populations.

Substructure Analysis within the Botswana Population (Batswana)

To assess substructure within the Botswana cohort, we followed the same QC pipeline as the PCA analysis. Independent autosomal markers pruned by LD (r-squared coefficient of 0.2) in windows of 1,000 base pairs advanced 100 SNPs at a time were used for the analysis with the SNPRelate v.1.2.0 package[58] in R v.3.2.4. Botswana ethnic groups were defined per their self-reported ethnicities. The Guthrie classification of Bantu languages was used to infer language associated with the self-reported ancestry to have a secondary definition of ethnic groups[23] (see Ethnologue website).

Weir and Cockerham's Fst

Differentiation between the Botswana and the 1000 Genomes phase 3 populations using Weir and Cockerham's fixation index estimator was assessed with the SNPRelate v.1.2.0 package[58] in R v.3.2.4. Only biallelic autosomal SNPs that were shared between the Botswana population and the reference dataset (1000 Genomes) were used for this analysis. We used the default Weir and Cockerham 84 estimator option to run the analysis.

Admixture Analysis

ADMIXTURE v.1.3.0 (see Web Resources) was used to estimate ancestral clusters within the cohort given the potential of genetic admixture between the African, Khoisan, Asian, and European ancestries that are present within the population of Botswana. This software models K, the number of ancestral populations that best represent the data in the model. We used the default cross validation parameters for determining the best estimate of K using the 5-fold cross validation settings. We assessed structure between Batswana and African populations in the 1000 Genomes project and two Southern African populations from the African Genome Variation Project (AGVP) (Sotho and Zulu). The input was 26,763 WES autosomal markers at the default cross validation parameter. The estimation was from K = 1 to K = 8 and the CV error estimation minimized at K = 3. Markers were pruned for LD in PLINK[57] using a window of 50 SNPs advanced by 5 SNPs at a time and an r-squared coefficient of 0.2. Plots were visualized with the web tool Pophelper v.1.1.10.[59] We then assessed potential admixture within the Batswana using another unsupervised clustering model based on African, Asian, and European data from the 1000 Genomes using default parameters.

Comparison of Pathogenic Variants between Botswana and ExAC AFR Samples

The allele frequencies of pathogenic variants among Botswana samples were compared with that of samples in the Exome Aggregation Consortium (ExAC; Web Resources) denoted as "African" (AFR). To classify the pathogenic variants, we first filtered all variants found in Botswana samples with MAF > 5% in any super-population in 1000 Genomes, ExAC, or Exome Sequencing Project (ESP). After removing common variants, a variant was classified as pathogenic if it was annotated as "Pathogenic" in HGMD[60] (version March 2014) or "Disease-causing mutation" (DM) in Clinvar[61] (version September 2017). In total, 485 pathogenic variants were discovered in 164 Botswana samples, of which 64 were not captured among African samples in ExAC (n = 5,203). The number of captured pathogenic variants in the rare (with ExAC AFR allele frequency f < 0.1%), low frequency (0.1% < f < 1%), and intermediate frequency (1% < f < 5%) ranges were 67, 183, and 171, respectively.

Identical-by-Descent (IBD) and Inbreeding Analyses

To assess IBD across populations, variant call format (v*cf.*) files for each population were intersected with the target region bed files using BEDTools[62] v.2.16.2. Only autosomes were used in this analysis. Phasing information was removed from v*cf.* files of the 1000 Genomes Phase 3 populations, and each population under study was then phased using Beagle[63] 4.1 with default burn-in iterations and 15 phasing iterations. In the subsequent stage, 15 independent runs of the Beagle[64] 4.1 IBD calling algorithm were executed on the phased target region data. For each of the runs, a random seed was generated between 1 and 10,000. We limited the IBD length to a minimum of 3 cM in each run to reduce the effects of incorrect phasing and genotyping errors in IBD length estimation. Beagle's utilities tool *ibdmerge* was used to combine the output of the 15 Beagle IBD runs. To normalize the population-wide IBD length comparisons for differences in the sample sizes of each population, we used the normalization factor employed by Nakatsuka et al.,[65] in which the sum of all the IBD segment lengths in a population are divided by $(^{2n}C_2)$-n, where n is the sample size. VCFTools (v.0.1.12) was used to compute the sample wise inbreeding coefficients and LROH (long runs of homozygosity) analysis.

Correlation between Variant Classification and African Ancestry

To determine the correlation between Southern African admixture and the number of variants classified as pathogenic and/or deleterious in each of the Botswana samples, we used the classification criteria of Kessler et al.[66] Specifically, among 202,547 exonic variants called in 164 Botswana samples, we filtered 17,292 (8.5%) common variants (MAF > 5% in any super-population in 1000 Genomes, ExAC, and ESP) in genes related to diseases as reported in HGMD[60] (version March 2014) and Clinvar[61] (version September 2017). For variants not in these genes, we first filtered 10,982 (5.4%) variants that were not protein changing or RNA splicing from gene-based annotations (refGene, knownGene, ensGene) using Annovar (July 2017) and then further filtered 59,246 (29%) common variants with MAF > 2%. Variants passing the above filters were further classified as "pathogenic" if they were annotated "Pathogenic" in HGMD[60] or "Disease-causing mutation" in Clinvar,[61] and as "deleterious" if they were predicted as deleterious among at least 2 of 11 *in silico* predictors or if they were nonsense or splicing variants. We found 308 (0.15%) variants classified as pathogenic and deleterious (PAV), 223 (0.11%) pathogenic but non-deleterious (PAV Non-del.), 21,405 (11%) non-pathogenic but deleterious (NAV Del.), and 93,091 (46%) non-pathogenic and non-deleterious variants (NAV Non-Del.). We evaluated the correlation between the number of variants belonging to each classification in each Botswana samples with the proportion of Southern African ancestry estimated using ADMIXTURE with K = 2, which was the optimal cross-validation for number of sub-populations within the Botswana cohort.

Statistical Comparisons

The *prop.test()* function in R, which uses a chi-square test with a continuity correction, was used to test whether the proportion of ExAC uncaptured low-frequency variants were significantly different between the two CAfGEN groups. The cut-off for significance was set at 0.05.

Results

Uncaptured Low-Frequency Variation Is Characteristic of the Botswana Population

We observed that between 15% and 25% of sequence variants observed among Batswana were not represented in dbSNP141 (15.9%) or 1000 Genomes phase 3 (25.1%). By comparison, in the Uganda cohort, sequenced on the same platform, the proportion of uncaptured variants was smaller, particularly in the 1000 Genomes database (13.1% not in dbSNP; 12.4% not in 1000 Genomes); this presumably reflects the better overall representation of African genomes in dbSNP than 1000 Genomes and the comparatively better representation of East African versus Southern African variation in both databases. Among these uncaptured variants (n = 312,920), 2.6% were common (with MAF > 5%) and the majority (n = 257,562; 82.3%) were non-coding. Among the 194,186 coding variants in Batswana, 191,758 overlapped UCSC KnownGene[50] and ENSMBL[51] exon definitions, of which 19.0% (n = 36,432) were not represented within dbSNP 141 and approximately a quarter (26.6%; n = 50,955) were not observed in 1000 Genomes phase 3 (Table S3). When we compared our coding SNVs with those in the larger Exome Aggregation Consortium (ExAC) database, 16.7% of variants (n = 32,077) were not observed and the majority of these were rare, often singleton, SNVs; however, in Botswana 19.6% (n = 6,294) of ExAC uncaptured SNVs had minor allele frequencies between 1% and 5% (which we refer to as low-frequency variants); this was significantly more than that observed in Uganda (9.1%; n = 2,330; χ^2 $p < 2.2 \times 10^{-16}$) and vastly different than the comparable proportions of captured variants (database variants found in cohort datasets) in the two countries (Figures 2A and 2B).

More than half of these uncaptured, low-frequency, coding variants were either missense or putative loss-of-function (LOF; splice variants, stop-gains, stop-losses, frameshifts) variants (Figure 2C) and the majority were unique to Botswana (Figure 2D). Of the 2,458 uncaptured, missense, or putative LOF variants in Botswana, 193 (7.8%) were predicted to be both deleterious and probably/possibly damaging by SIFT[53] and PolyPhen-2,[54] respectively. Low-frequency putative LOF or predicted-damaging missense variants were found in 184 unique genes, of which 45 (24%) are associated with a known Mendelian phenotype, including *PYGM* and *PHKG2* (glycogen storage disease types V and IX [MIM: 232600 and 613027]) and *TGM6* (spinocerebellar ataxia, type 35 [MIM: 613908]) (Table S5). The majority of remaining genes were not associated with known human disease phenotypes.

When we focused on overlapping variation—SNVs observed in both public references and in our data—putative LOF variants among Batswana showed a similar pattern to uncaptured LOF variants: 44% were non-singleton, with one-quarter (25.2%) of these being observed in more than three individuals (Figure S1). In fact, when we compared the allele frequencies of uncommon (MAF < 5%) variants between two datasets, we found that while the allele frequencies were concordant between two datasets for MAF~1%, for rare predicted-pathogenic variants with MAF < 0.1% (mean allele frequency f = 0.034%) in ExAC AFR populations, the allele frequencies in Botswana samples were significantly elevated (mean f = 0.51%, $p = 1.24 \times 10^{-29}$) (Figure 2E). Across the full range of MAFs observed among ExAC "African" groups, the Botswana cohort also exhibited a broad range of corresponding MAFs (Figure 2F); for instance, among ExAC variants reported as rare (MAF < 1%) in African groups, 1.2% (n = 1,924) had a MAF > 5% in Botswana, whereas only 0.4% (n = 692) had MAF > 5% in Uganda ($p < 2.2 \times 10^{-16}$).

The Population of Botswana Illustrates Regional Distinctions among African Populations

Next, we sought to place the genetic variation observed in the Botswana cohort within the context of other global populations. Using common markers (MAF > 0.05) shared with 1000 Genomes phase 3 data as well as Sotho genotyping and Zulu sequence data from the African Genome Variation Project (AGVP), the Botswana population was found to cluster closely with the other African populations on principal components 1 and 2 (Figure 3A). When we focused on solely African populations in 1000 Genomes phase 3, however, the Botswana cohort and the other Southern African populations clearly separated from West and East African populations on principal component 1 (Figure 3B). Additionally, when we focused on the Southern African populations alone, we found that Botswana clustered separately from Zulu, with the Sotho being intermediate between the two (Figure 3C). Quantitative assessments of inter-population genetic distance using Weir and Cockerham's fixation index (F_{ST}) confirmed these observations, with higher F_{ST} values being observed between Botswana and 1000 Genomes non-African populations than between Botswana and other African populations (Figure S2).

Given the complex historical ancestry of the Batswana and the apparent genetic isolation in comparison with East and West African populations, we then assessed evidence of structure between our cohort and populations represented in 1000 Genomes under a maximum likelihood-based model using ADMIXTURE v.1.3.0 (see Subjects and Methods). Using 501,963 WES autosomal markers, we compared West African, East African, and Southern African population groups (Figure 4) from 1000 Genomes and AGVP. This cross-validation error for the model was minimized at K = 3 (see Subjects and Methods), with clusters K2 and K3 separating the populations into West and East, with Batswana and the other Southern African populations appearing closer to the East African Luhya (LWK) population. Cluster K4 distinguished West African populations and at cluster K5 Batswana and the other Southern African populations had a component distinct from both the LWK and the West African populations. Cluster K6 demonstrated the higher level of sub-structure within the

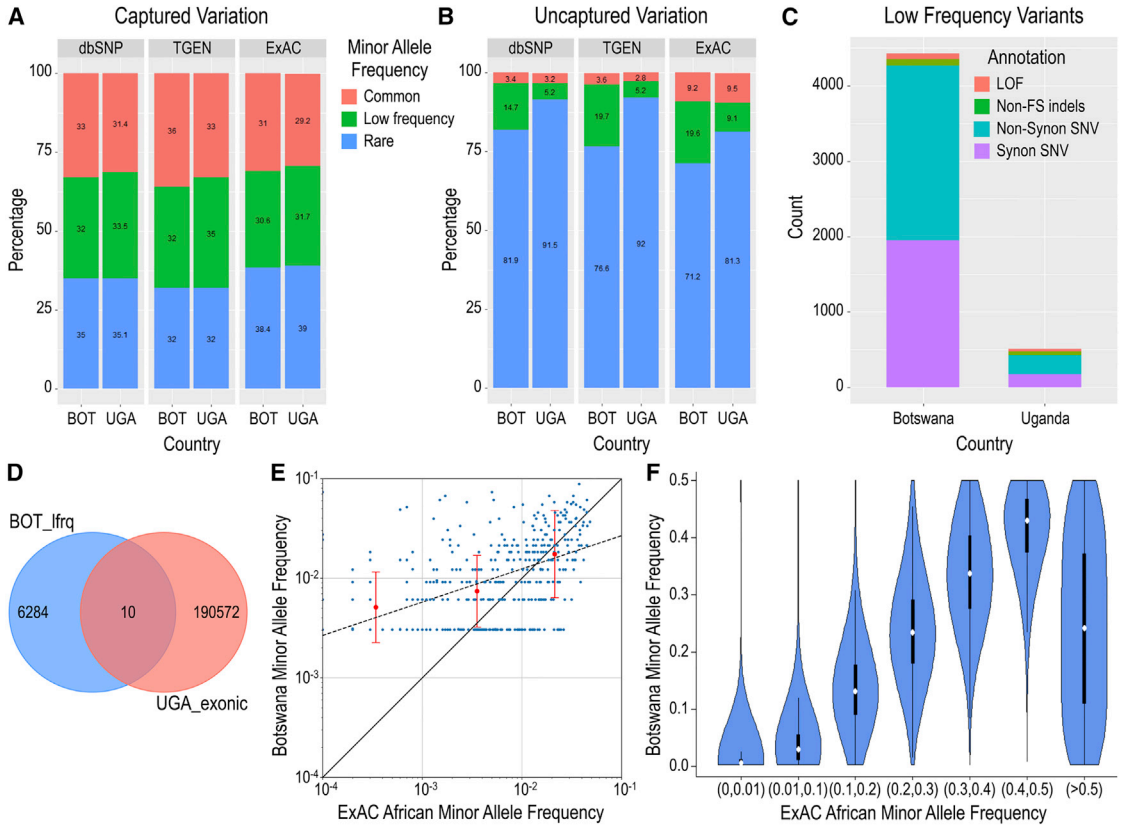

Figure 2. Genetic Variation within the CAfGEN Cohorts

(A) CAfGEN coding variant representation captured within public databases.

(B) CAfGEN coding variation uncaptured in public databases. Abbreviations: dbSNP, Database of Single Nucleotide Polymorphisms; TGEN, 1000 Genomes phase3.

(C) Annotation of uncaptured low-frequency variants (minor allele frequency [MAF] 0.01–0.05) within Botswana and Uganda populations. Abbreviations: LOF, putative loss-of-function; non-FS indels, non-frameshifting insert-deletions; non-synon SNV, non-synonymous single-nucleotide variants.

(D) Overlap of uncaptured low-frequency variants in Botswana (BOT_lowfrq) and all exonic variants in Uganda (UGA_exonic).

(E) Comparison of allele frequencies of ClinVar and HGMD pathogenic and damaging variants among Batswana versus ExAC Africans (AFR). Red bars represent the mean (central point) ± 3 standard deviations (whiskers) for allele frequencies of all Botswana pathogenic variants in each ExAC AFR frequency (x-axis) bin range (<0.001, 0.001–0.009, 0.01–0.05).

(F) Violin plot showing minor allele frequencies of ExAC African (AFR) variants (x axis) among Botswana (y axis). The median allele frequency is shown as a white circle, with the 25th and 75th centiles as black bars, and the 5th and 95th centiles as whiskers.

Southern African populations with respect to their West African and East African counterparts from 1000 Genomes; this highlights the level of structure that exists between our populations and the reference populations found within databases. When we included non-African ancestral groups (European, South Asian, and East Asian populations) in the model, we observed minimal Eurasian ancestral components within the Botswana population, suggesting minimal admixture between our cohort and these population groups (Figure S3).

Batswana Show Minimal Population Stratification, but a High Degree of Relatedness

We then assessed substructure within the Botswana population based on self-reported ethnicity and Guthrie language. Somewhat surprisingly, given strong self-identification of ethnic affiliations within the country, the resulting PCA plot showed little evidence for genetic substructure within the Botswana population on the first two principal components (PCs) (1.56% of variance; Figure 5A). Self-reported ethnicities were distributed in a cline along PC1 with non-Sotho-Tswana groups at the edges. When we repeated this analysis using the language associated with individuals' self-reported ancestry, there were again no easily distinguishable clusters (Figure S4A); the majority of individuals clustered on a cline along principal component 2, with individuals that speak Kalanga (a non-Sotho-Tswana language) tending toward the extremes of the cline, which was consistent with expectations based on self-reported ethnicity.

Given the lack of substructure within Botswana, we proceeded to assess segments of the genome shared identical-by-descent (IBD) using the exome data from our two cohorts. Consistent with the PCA analyses, we did not find evidence of stratification within Botswana using IBD; however, when we considered the degree of pairwise IBD

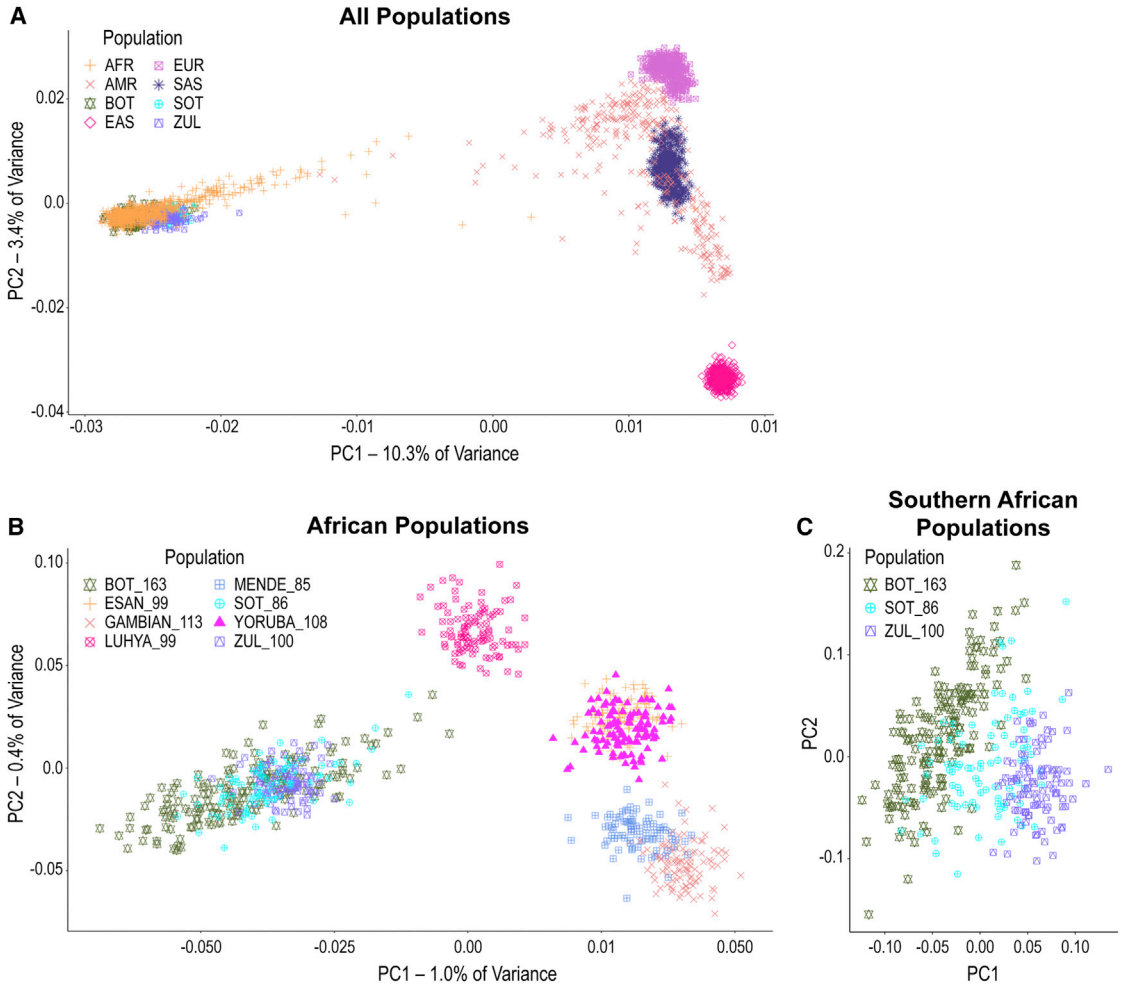

Figure 3. Principal Component Analysis of Batswana, 1000 Genomes, and Southern African Populations
(A) Botswana in the context of global 1000 Genomes and AGVP populations using common (MAF > 0.05) shared biallelic autosomal markers. Each symbol is an individual.
(B) Analysis of "African Populations" shows separation of Southern African populations from East and West African groups.
(C) "Southern African Populations" analysis was restricted to Batswana, Sotho, and Zulu (AGVP) and shows separation between the three groups with partial overlap of individuals at the margins of these clusters.
Abbreviations: BOT, Batswana; SOT, Sotho; ZUL, Zulu; AFR, African; AMR, Admixed American; EAS, East Asian; EUR, European; the number of individuals sampled follows the underscore (_).

sharing among Botswana samples alongside our Uganda cohort and the Finnish population from phase 3 of the 1000 Genomes project (FIN), we found that Botswana had the longest shared IBD tracks (total and normalized IBD lengths of 8.84×10^{10} and 1.65×10^{6}, respectively) of the three populations, with Uganda also having longer normalized shared segment lengths (total 3.01×10^{10} and mean 6.74×10^{5}) than the purportedly founder Finnish population of 1000 Genomes (total 8.30×10^{9} and mean 4.28×10^{5}) (Figure 5B). Even after normalizing for sample size (Subjects and Methods), IBD sharing in Botswana was still 4–5 times higher than that in FIN and was substantially greater than among the PUR, CLM, FIN, and LWK populations from 1000 Genomes (Figure 5C), which are known to have among the smallest effective population sizes among 1000 Genomes populations[1] (see Web Resources for IBD). IBD sharing in Uganda was closest to the CLM population, but was still greater than FIN.

To assess the possible contribution of consanguinity and inbreeding to these results, we also calculated the sample-wise inbreeding coefficient for the populations used in Figure 5C; however, we did not observe significant differences in sample-wise inbreeding coefficients (Figure S4B), and analysis of runs of homozygosity did not demonstrate any extended segments in the Botswana population. These results suggest that both of our HIV-positive pediatric cohorts, but Botswana in particular, have a smaller effective population size than African populations currently represented in 1000 Genomes.

Discussion

We provide data derived from high-depth sequencing coverage of the coding regions of a Southern African population, sampling more than 160 individuals (>320

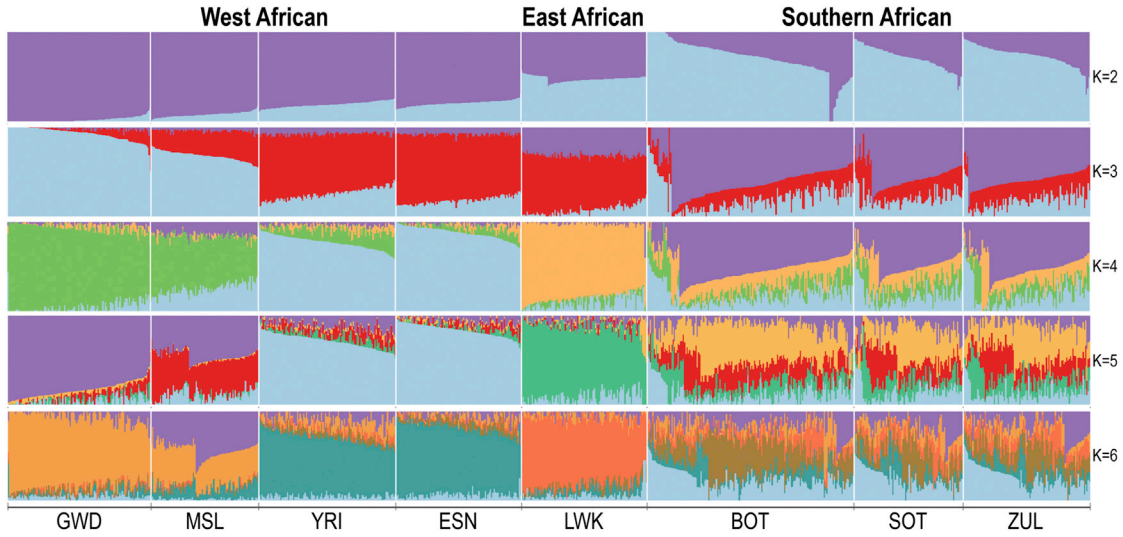

Figure 4. Admixture Analysis of Botswana and African Populations
The x axis shows the populations that were included in the ADMIXTURE unsupervised clustering maximum likelihood model. The width of each column represents the number of individuals within the listed population. Each row is independent of the other rows and the colors reflect the different genomic proportions that are derived from the African populations included in the model. The y axis represents a proportion ranging from 0 to 1. More uniform color within a given column suggests a genome composed of fewer contributing components, while increased color suggests an individual genome made up of multiple components from the surveyed populations. Abbreviations: GWD, Gambians in the Western Divisions of the Gambia; MSL, Mende in Sierra Leone; YRI, Yoruba in Nigeria; ESN, Esan in Nigeria; BOT, Batswana; SOT, Sotho from the African Genome Variation Project; ZUL, Zulu from the African Genome Variation Project.

chromosomes). In contrast to previous surveys of genetic variation from the region, which have often utilized smaller sample sizes and relatively isolated population groups, we evaluated individuals from a highly populous urban center and utilized exome-wide sequencing data;[10,13,67] this afforded us a view of uncaptured and rare variation that was not evident from previous characterizations.

Our Botswana population was found to harbor a significant proportion of variants that are not represented in public databases, and among variants that are represented, there are wide discrepancies with respect to minor allele frequencies. These observations present practical lessons for disease mapping efforts in Africans. Both rare-variant enrichment models (in which disease variants are expected to be relatively rare in the population but highly enriched among individuals at the extremes of disease[68]) and Mendelian disease mapping (in which very rare or novel variation is associated with causal variation) rely upon allele frequencies observed in large public databases[1,18,20,69–71] to define "rare" or "novel." Our study suggests that identifying and interpreting such variants in exomes of individuals with African, and particularly Southern African and Botswana, ancestry using current iterations of available public databases will be challenging;[1] however, viewed in the context of our population-level sequence data, many of the uncaptured variants occurred at frequencies that would make them less likely to be considered "damaging" or "deleterious." Further, among captured variants, several uncommon, putatively damaging variants in ClinVar[61] and HGMD[60] were found to have an appreciably higher MAF in Botswana, making their pathogenicity more suspect, at least under a dominant mode of inheritance.

African ancestry has been shown to be positively correlated with the number of damaging variants identified in a given individual over most variant classes except predicted-deleterious variants annotated as pathogenic.[66] We observed the same trend in our data (Subjects and Methods): significant correlation between Southern African admixture proportion and the number of neutral variants in the genome ($r = 0.831$, $p = 6.21 \times 10^{-43}$ for non-pathogenic, non-deleterious variants (NAV Non-del.) and $r = 0.689$, $p = 2.72 \times 10^{-24}$ for NAV Deleterious) and no significant correlation with the number of damaging variants ($r = 0.038$, $p = 0.631$ for pathogenic deleterious variants (PAV Del). and $r = 0.055$, $p = 0.484$ for PAV Non-deleterious) (Figure S5). Thus, at an individual exome level, current clinical filtering procedures will still result in multiple candidate variants to be reviewed and validated in persons of Southern African ancestry. These results bolster the assertion that the discovery of medically relevant genetic variants in African populations will likely require sequenced-based characterization of genetic variation in the respective, relevant populations.[17,20,37,70]

The successful mapping of complex disease traits over the past decade has exploited linkage disequilibrium-derived haplotype proxies to provide genome-wide coverage using common variants. Given more disparate patterns of LD among African populations, the more common uncaptured variants identified here are unlikely to be well represented on genotyping assays designed using

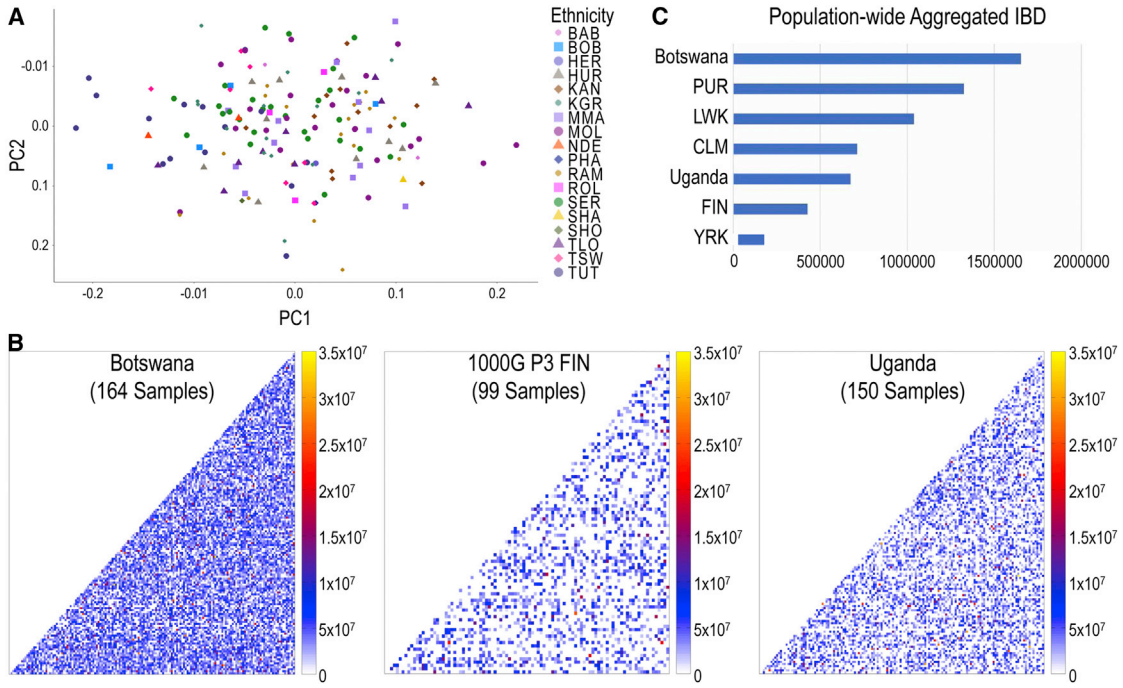

Figure 5. Population Stratification and IBD in the Botswana Population

(A) PCA analysis of Botswana population; each point indicates an individual defined by their self-reported ancestry.

(B) Mean length of pairwise IBD segments shared within Botswana, 1000 Genomes phase 3 FIN, and the Uganda populations. Each dot represents a pairwise IBD comparison between two individuals; scale has been normalized to the same value (based on FIN) for all the plots.

(C) Aggregate IBD sharing normalized by population sample size for 1000 Genomes populations (PUR, Puerto Ricans from Puerto Rico; LWK, Luhya in Webuye, Kenya; CLM, Colombians from Medellin, Colombia; FIN, Finnish in Finland; YRI, Yoruba in Ibadan, Nigeria) and CAfGEN populations (Botswana and Uganda). YRI populations serve as a control for low IBD sharing, as they have the largest effective population size among African populations in 1000 Genomes.

non-African populations, and the widely varying allele frequencies of common variants in our cohorts compared to public databases suggest that imputation of variants in Botswana may also be sub-optimal. Much has been recently made of the "genomic gap" between genome-wide studies among African (and other underrepresented) populations versus that in European groups.[72–74] Our results suggest that redressing this gap will require additional investments in under-studied, high-yield populations in order to ascertain variant markers within representative populations.[3,9,16,17,20,70] Ongoing efforts to produce representative genotyping platforms aimed at African[34] and African ancestry[75,76] populations are thus laudable and auger well for maximizing disease-variant genetic mapping within Africa and the wider genomics community.

The tremendous ethnic diversity within African countries is typically cited as a hindrance for conducting genomic studies within Africa,[77] as the potential for population stratification creates a challenge for adequately powering association studies.[78,79] By contrast, we observed minimal evidence for substructure in our Botswana cohort, despite multiple self-reported language groups and ethnicities; this suggests that conducting traditional genome-wide association studies (GWASs) within the Botswana population will be feasible even with more modest sample sizes. This lack of substructure is likely to reflect, in part, the complex social ancestry of the country, which includes a shared history of succession breakaways and intermarriage, particularly between Sotho-Tswana ethnic groups,[22,24,25,28] along with historical, political, and cultural practices of assimilating large groups from different Bantu language clusters under a single umbrella ethnicity.[26,80] For example, the Balete, despite their large size and Nguni origins, were assimilated[26] into Sotho-Tswana groups and now speak Tswana, which is a different language group from the Nguni Bantu language they spoke prior to their assimilation. The presence of non-Tswana-speaking ethnicities at the margins of the clines observed in our data may thus represent the remnants of past Sotho-Tswana endogamy.[22,24] These anthropological groups have since merged within modern-day Gaborone, a populous, urban city in which current marriage and mating practices[17] mean that self-reported ethnic affiliations are more social constructs than reflective of the highly stratified genetic ancestry found in more rural Botswana.[33] This separates our work from previous genetic surveys.

In fact, compared to other African populations, Batswana not only clustered distinctly from—and showed little evidence of admixture with—East and West African population genomes, but also had a much higher degree of relatedness than even known founder populations. This relatedness is likely to be a major factor contributing

to the significantly higher number of low-frequency variants observed in our cohort. Despite this, the data provided little evidence for excessive inbreeding; therefore, we postulate that demographic events and genetic drift are the main contributors to the distinctiveness of the Botswana cohort.[31] As noted above, our cohort included a large number of Southern Bantu, whose migration patterns,[11] shared history,[22,26,27,81] and assimilation of other ethnic groups[11,17,24] following their split from East and West African Bantu populations are likely to have contributed to the distinct Batswana ancestry. In addition, at its height in the mid 1990s, HIV prevalence in Botswana for the general population was close to 25% and closer to 40% in pregnant women (UNAIDS in Web Resources). Given the high mortality rate of HIV before the widespread use of antiretroviral therapy (ART), it is possible that the excessive lengths of the genome found to be IBD in this childhood cohort are the consequence of the HIV epidemic, with the effects of this population drift being more manifest in the relatively small population of Botswana (~2 million) than in the large population of Uganda (~40 million), even though HIV levels were higher in Uganda at the peak of the epidemic. Irrespective of the underlying factors driving this particular observation, the resulting extensive shared segments would be expected, paradoxically, to make disease mapping in Botswana significantly *easier* than in other African populations.

The clinical center in Botswana is the largest regional pediatric HIV referral center in the country; however, our local ascertainment meant that we did not have any participants from the genetically isolated Khoisan people[6,30,31] (who were also not represented in our ADMIXTURE analysis) or Western Bantu related ethnicities—both more populous in the western and northern regions of the country (Figure 1).[23] It is still plausible, however, that admixture with the Khoisan also contributes to the relative distinctness of the Botswana population[10,13,17,30] and that wider sampling of the population would reveal still greater isolation. Although our population is biased by HIV status, infection with HIV is known to cut across all strata of the population in Botswana; we thus do not expect our overall results to differ significantly if our study was conducted in a non-HIV-positive cohort. To the best of our knowledge, there is no other currently available similar sequence data from Botswana to provide further context for our findings, and we note that these observations were also unique with respect to our similarly recruited pediatric HIV cohort from Uganda.

The genetic architecture of the Botswana population described here underscores the complex ancestry of Southern African populations and reinforces recent suggestions that reliance on ethnic, tribal, or language group labels as indicators of study feasibility may be overstated, particularly among urban African populations.[17] As one of the first large-scale deep-sequencing studies in sub-Saharan Africa, our results also emphasize the need to characterize fine-scale genome variation among underrepresented African populations; this is imperative to better facilitate Mendelian and complex-trait mapping among those who harbor a significant burden of global disease. Doing so promises to both uncover the allelic architecture needed to interpret genomic studies on the continent and provide a deeper understanding of population movements that are fundamental to human history.

Accession Numbers

AGVP Datasets: EGAS00001000960/TBA (AGV curated all WGS v*cf*. files), GAS00001000960/EGAD00001001663 (AGV allele frequencies v*cf*. files). CAfGEN exome sequence datasets (BAMs and vcfs) are being made publicly available via the European Genome Archive (EGA) in accordance with guidelines agreed upon by the Human Health and Heredity in Africa (H3Africa) Consortium.

Supplemental Data

Supplemental Data include five figures and five tables and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.03.010.

Acknowledgments

This study makes use of data generated by the African Partnership for Chronic Disease Research. A full list of the investigators and funders who contributed to the generation of the data is available from APCDR website (see Web Resources). The authors would like to acknowledge the participating families, as well as the contributions of Kennedy Sichone in participant recruitment and sample acquisition, Nancy Hall and Roa Sadat in sample inventory and management, and Adam Gillum with manuscript figures. This work was supported by a Collaborative Center grant (Collaborative African Genomics Network [CAfGEN]) to G.A., G. Mardon, A.K., M.J., and S.W.M., from the National Institutes of Health (grant U54AI110398), and from the Center for Globalization at Baylor College of Medicine to G. Mardon. N.A.H. was also supported by a Clinical Scientist Development Award from the Doris Duke Charitable Foundation (grant 2013096).

Received: October 3, 2017
Accepted: February 26, 2018
Published: April 26, 2018

Web Resources

1000 Genomes, http://www.internationalgenome.org/
1000 Genomes IBD Segment Methods, http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/ibd_by_pair/20150129_IBD_segment_methods.pdf
Admixture Manual, https://www.genetics.ucla.edu/software/admixture/admixture-manual.pdf
APCDR, https://www.apcdr.org
dbNSFP, http://varianttools.sourceforge.net/Annotation/DbNSFP
dbSNP, https://www.ncbi.nlm.nih.gov/projects/SNP/
Ethnologue, https://www.ethnologue.com
European Genome-phenome Archive (EGA), https://www.ebi.ac.uk/ega
ExAC Browser, http://exac.broadinstitute.org/
H3Africa, https://h3africa.org/

OMIM, http://www.omim.org/
Pophelper, http://pophelper.com/SUWRS
R statistical software, https://www.r-project.org/
UNAIDS, http://aidsinfo.unaids.org

References

1. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
2. Baker, J.L., Shriner, D., Bentley, A.R., and Rotimi, C.N. (2017). Pharmacogenomic implications of the evolutionary history of infectious diseases in Africa. Pharmacogenomics J. *17*, 112–120.
3. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African Genome Variation Project shapes medical genetics in Africa. Nature *517*, 327–332.
4. Reed, F.A., and Tishkoff, S.A. (2006). African human diversity, origins and migrations. Curr. Opin. Genet. Dev. *16*, 597–605.
5. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.
6. Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G., et al. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science *338*, 374–379.
7. Currie, T.E., Meade, A., Guillon, M., and Mace, R. (2013). Cultural phylogeography of the Bantu languages of sub-Saharan Africa. Proc. Biol. Sci. *280*, 20130695.
8. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science *343*, 747–751.
9. Busby, G.B., Band, G., Si Le, Q., Jallow, M., Bougama, E., Mangano, V.D., Amenga-Etego, L.N., Enimil, A., Apinjoh, T., Ndila, C.M., et al.; Malaria Genomic Epidemiology Network (2016). Admixture into and within sub-Saharan Africa. eLife *5*, e15266.
10. Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., Petersen, D.C., Zhao, F., Qi, J., et al. (2010). Complete Khoisan and Bantu genomes from southern Africa. Nature *463*, 943–947.
11. Li, S., Schlebusch, C., and Jakobsson, M. (2014). Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proc. Biol. Sci. *281*, 1448.
12. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
13. Petersen, D.C., Libiger, O., Tindall, E.A., Hardie, R.A., Hannick, L.I., Glashoff, R.H., Mukerji, M., Fernandez, P., Haacke, W., Schork, N.J., Hayes, V.M.; and Indian Genome Variation Consortium (2013). Complex patterns of genomic admixture within southern Africa. PLoS Genet. *9*, e1003309.
14. Pickrell, J.K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. Proc. Natl. Acad. Sci. USA *111*, 2632–2637.
15. Kim, H.L., Ratan, A., Perry, G.H., Montenegro, A., Miller, W., and Schuster, S.C. (2014). Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. Nat. Commun. *5*, 5692.
16. Rotimi, C.N., Tekola-Ayele, F., Baker, J.L., and Shriner, D. (2016). The African diaspora: history, adaptation and health. Curr. Opin. Genet. Dev. *41*, 77–84.
17. May, A., Hazelhurst, S., Li, Y., Norris, S.A., Govind, N., Tikly, M., Hon, C., Johnson, K.J., Hartmann, N., Staedtler, F., and Ramsay, M. (2013). Genetic diversity in black South Africans from Soweto. BMC Genomics *14*, 644.
18. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
19. Mersha, T.B., and Abebe, T. (2015). Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Hum. Genomics *9*, 1.
20. Dopazo, J., Amadoz, A., Bleda, M., Garcia-Alonso, L., Alemán, A., García-García, F., Rodriguez, J.A., Daub, J.T., Muntané, G., Rueda, A., et al. (2016). 267 Spanish exomes reveal population-specific differences in disease-related genetic variation. Mol. Biol. Evol. *33*, 1205–1218.
21. Chapman, S.J., and Hill, A.V. (2012). Human genetic susceptibility to infectious disease. Nat. Rev. Genet. *13*, 175–188.
22. Sillery, A. (1974). Botswana: A Short Political History, A.H.M. Kirk-Greene, ed. (Bungay: Methuen & Co Ltd).
23. Batibo, H.M. (1998). A lexicostatistical survey of the Bantu language of Botswana. S. Afr. J. Afr. Lang. *18*, 22–28.
24. Tlou, T. (1998). The nature of Batswana states: towards a theory of Botswana traditional government - the Batawana case. In Botswana: Politics and Society, W.A. Edge and M.H. Lekorwe, eds. (Pretoria: J.L van Schaik), p. 22.
25. Gulbrandsen, Ø. (1993). The rise of the North-Western Tswana kingdoms: on the dynamics of interaction between internal relations and external forces. Africa *63*, 550–582.
26. Schapera, I., Comaroff, J., and Kuper, A. (1953). The Tswana, D. Forde, ed. (Plymoth: Clarke, Doble & Brendon).
27. van Waarden, C. (1998). The Late Iron Age. In Ditswammung, the Archeology of Botswana, P. Lane, A. Reid, and A. Segobye, eds. (Gaborone: Pula Press), pp. 115–160.
28. Ngcongo, L.D. (1978). Origins of the Tswana (Gaborone).
29. Matemba, Y.H. (2003). The pre-colonial political history of Bakgatla ba ga Mmanaana of Botswana, c. 1600-1881. Botsw. Notes Rec. *2003*, 53–67.
30. Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S.W., Nakagawa, H., Naumann, C., et al. (2012). The genetic prehistory of southern Africa. Nat. Commun. *3*, 1143.
31. González-Santos, M., Montinaro, F., Oosthuizen, O., Oosthuizen, E., Busby, G.B.J., Anagnostou, P., Destro-Bisol, G., Pascali, V., and Capelli, C. (2015). Genome-wide snp analysis of southern african populations provides new insights into the dispersal of bantu-Speaking groups. Genome Biol. Evol. *7*, 2560–2568.
32. Barbieri, C., Butthof, A., Bostoen, K., and Pakendorf, B. (2013). Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. Eur. J. Hum. Genet. *21*, 430–436.

33. Tau, T., Wally, A., Fanie, T.P., Ngono, G.L., Mpoloka, S.W., Davison, S., and D'Amato, M.E. (2017). Genetic variation and population structure of Botswana populations as identified with AmpFLSTR Identifiler short tandem repeat (STR) loci. Sci. Rep. *7*, 6768.
34. Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V.M., Adebamowo, C., Adebiyi, E., Ademola, A.D., Adeyemo, A., Adu, D., Affolabi, D., et al.; H3Africa Consortium (2014). Research capacity. Enabling the genomic revolution in Africa. Science *344*, 1346–1348.
35. Mlotshwa, B.C., Mwesigwa, S., Mboowa, G., Williams, L., Retshabile, G., Kekitiinwa, A., Wayengera, M., Kyobe, S., Brown, C.W., Hanchard, N.A., et al. (2017). The collaborative African genomics network training program: a trainee perspective on training the next generation of African scientists. Genet. Med. *19*, 826–833.
36. Belkadi, A., Pedergnana, V., Cobat, A., Itan, Y., Vincent, Q.B., Abhyankar, A., Shang, L., El Baghdadi, J., Bousfiha, A., Alcais, A., et al.; Exome/Array Consortium (2016). Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. Proc. Natl. Acad. Sci. USA *113*, 6713–6718.
37. Tang, D., Anderson, D., Francis, R.W., Syn, G., Jamieson, S.E., Lassmann, T., and Blackwell, J.M. (2016). Reference genotype and exome data from an Australian Aboriginal population for health-based research. Sci. Data *3*, 160023.
38. Carson, A.R., Smith, E.N., Matsui, H., Brækkan, S.K., Jepsen, K., Hansen, J.-B., and Frazer, K.A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinformatics *15*, 125.
39. Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., and Watson, M. (2015). Exome Sequencing: Current and Future Perspectives. G3 (Bethesda) *5*, 1543–1550.
40. Lelieveld, S.H., Veltman, J.A., and Gilissen, C. (2016). Novel bioinformatic developments for exome sequencing. Hum. Genet. *135*, 603–614.
41. Bainbridge, M.N., Wang, M., Wu, Y., Newsham, I., Muzny, D.M., Jefferies, J.L., Albert, T.J., Burgess, D.L., and Gibbs, R.A. (2011). Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. *12*, R68.
42. Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013). arXiv. 1303.3997.
43. Wang, G.T., Peng, B., and Leal, S.M. (2014). Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. Am. J. Hum. Genet. *94*, 770–783.
44. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
45. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
46. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
47. Van Der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, et al. (2014). From FastQ data to high confidence varant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 1–33.
48. Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (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.
49. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.
50. Hsu, F., Kent, W.J., Clawson, H., Kuhn, R.M., Diekhans, M., and Haussler, D. (2006). The UCSC known genes. Bioinformatics *22*, 1036–1046.
51. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.
52. González-Pérez, A., and López-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am. J. Hum. Genet. *88*, 440–449.
53. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.
54. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
55. Wickham, H. (2009). Elegant Graphics for Data Analysis (Springer-Verlag New York).
56. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
57. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
58. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328.
59. Francis, R.M. (2017). pophelper: an R package and web app to analyse and visualize population structure. Mol. Ecol. Resour. *17*, 27–32.
60. Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. *133*, 1–9.
61. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985.
62. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

63. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.
64. Browning, B.L., and Browning, S.R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics *194*, 459–471.
65. Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., Patterson, N., Bhavani, G.S., Girisha, K.M., Mustak, M.S., Srinivasan, S., et al. (2017). The promise of discovering population-specific disease-associated genes in South Asia. Nat. Genet. *49*, 1403–1407.
66. Kessler, M.D., Yerges-Armstrong, L., Taub, M.A., Shetty, A.C., Maloney, K., Jeng, L.J.B., Ruczinski, I., Levin, A.M., Williams, L.K., Beaty, T.H., et al.; Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) (2016). Challenges and disparities in the application of personalized genomic medicine to populations with African ancestry. Nat. Commun. *7*, 12521.
67. Fagny, M., Patin, E., MacIsaac, J.L., Rotival, M., Flutre, T., Jones, M.J., Siddle, K.J., Quach, H., Harmant, C., McEwen, L.M., et al. (2015). The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat. Commun. *6*, 10047.
68. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.
69. Wong, L.P., Ong, R.T.H., Poh, W.T., Liu, X., Chen, P., Li, R., Lam, K.K., Pillai, N.E., Sim, K.S., Xu, H., et al. (2013). Deep whole-genome sequencing of 100 southeast Asian Malays. Am. J. Hum. Genet. *92*, 52–66.
70. Scott, E.M., Halees, A., Itan, Y., Spencer, E.G., He, Y., Azab, M.A., Gabriel, S.B., Belkadi, A., Boisson, B., Clark, A.G., Greater Middle East Variome Consortium, Alkuraya, F.S., Casanova, J.L., and Gleeson, J.G. (2016). Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat. Genet. *48*, 1071–1076.
71. Petrovski, S., and Goldstein, D.B. (2016). Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. *17*, 157.
72. Need, A.C., and Goldstein, D.B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends Genet. *25*, 489–494.
73. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
74. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163–165.
75. Mathias, R.A., Taub, M.A., Gignoux, C.R., Fu, W., Musharoff, S., O'Connor, T.D., Vergara, C., Torgerson, D.G., Pino-Yanes, M., Shringarpure, S.S., et al.; CAAPA (2016). A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat. Commun. *7*, 12522.
76. Johnston, H.R., Hu, Y.-J., Gao, J., O'Connor, T.D., Abecasis, G.R., Wojcik, G.L., Gignoux, C.R., Gourraud, P.A., Lizee, A., Hansen, M., et al.; CAAPA Consortium (2017). Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. Sci. Rep. *7*, 46398.
77. Novembre, J., and Ramachandran, S. (2011). Perspectives on human population structure at the cusp of the sequencing era. Annu. Rev. Genomics Hum. Genet. *12*, 245–274.
78. Berger, M., Stassen, H.H., Köhler, K., Krane, V., Mönks, D., Wanner, C., Hoffmann, K., Hoffmann, M.M., Zimmer, M., Bickeböller, H., and Lindner, T.H. (2006). Hidden population substructures in an apparently homogeneous population bias association studies. Eur. J. Hum. Genet. *14*, 236–244.
79. Tian, C., Gregersen, P.K., and Seldin, M.F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. Hum. Mol. Genet. *17* (R2), R143–R150.
80. Wilmsen, E.N. (2002). Mutable identities: moving beyond ethnicity in Botswana. J. South. Afr. Stud. *28*, 825–841.
81. Morton, F. (2013). Settlements, landscapes and identities among the Tswana of the western transvaal and eastern Kalahari before 1820. S. Afr. Archaeol. Bull. *68*, 15–26.

The American Journal of Human Genetics, Volume *102*

Supplemental Data

Whole-Exome Sequencing Reveals Uncaptured Variation and Distinct Ancestry in the Southern African Population of Botswana

Gaone Retshabile, Busisiwe C. Mlotshwa, Lesedi Williams, Savannah Mwesigwa, Gerald Mboowa, Zhuoyi Huang, Navin Rustagi, Shanker Swaminathan, Eric Katagirya, Samuel Kyobe, Misaki Wayengera, Grace P. Kisitu, David P. Kateete, Eddie M. Wampande, Koketso Maplanka, Ishmael Kasvosve, Edward D. Pettitt, Mogomotsi Matshaba, Betty Nsangi, Marape Marape, Masego Tsimako-Johnstone, Chester W. Brown, Fuli Yu, Adeodata Kekitiinwa, Moses Joloba, Sununguko W. Mpoloka, Graeme Mardon, Gabriel Anabwani, Neil A. Hanchard, and for the Collaborative African Genomics Network (CAfGEN) of the H3Africa Consortium

SUPPLEMENTAL DATA

Supplemental data includes 5 figures and 5 tables

Figure S1 – Putative Loss-of-function variants in Botswana and Uganda

Figure S2 - Continental Weir and Cockerham's F_{ST} comparison

Figure S3 - Population admixture with 1000 genomes super-populations

Figure S4 – Principal components analysis and inbreeding coefficients

Figure S5 – Analysis of ClinVar damaging allele counts in Botswana

Table S1 - Demographics of the participants

Table S2 - Self-reported ancestry and Guthrie language classification

Table S3 - Number of variants identified in each cohort

Table S4 - Variants used in PCA analyses

Table S5 - Uncaptured Low-frequency Variants in Botswana (excel sheet)

Supplemental Figures

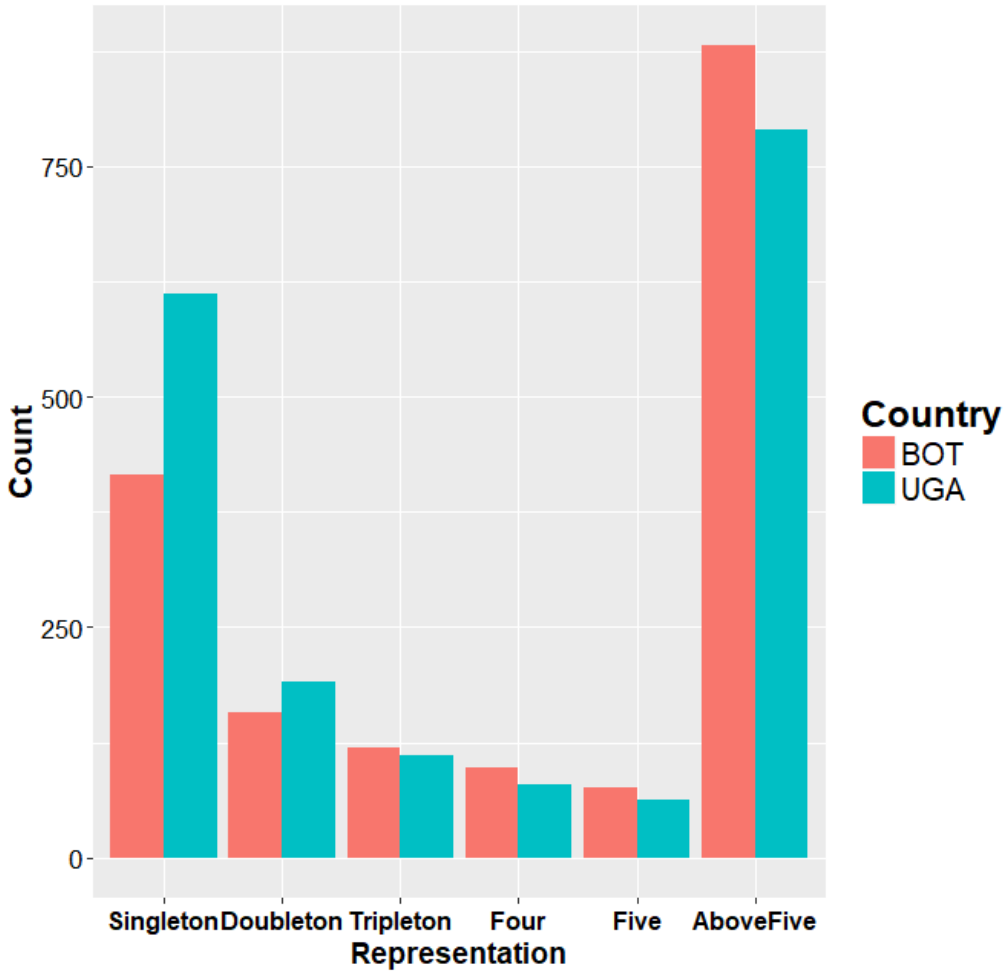

Figure S1. Putative loss-of-function variants in Botswana and Uganda population, annotated using ANNOVAR. Exons were defined using the exon start and end positions as defined within the UCSC KnownGene database using Variant Tools software (v2.6.1).

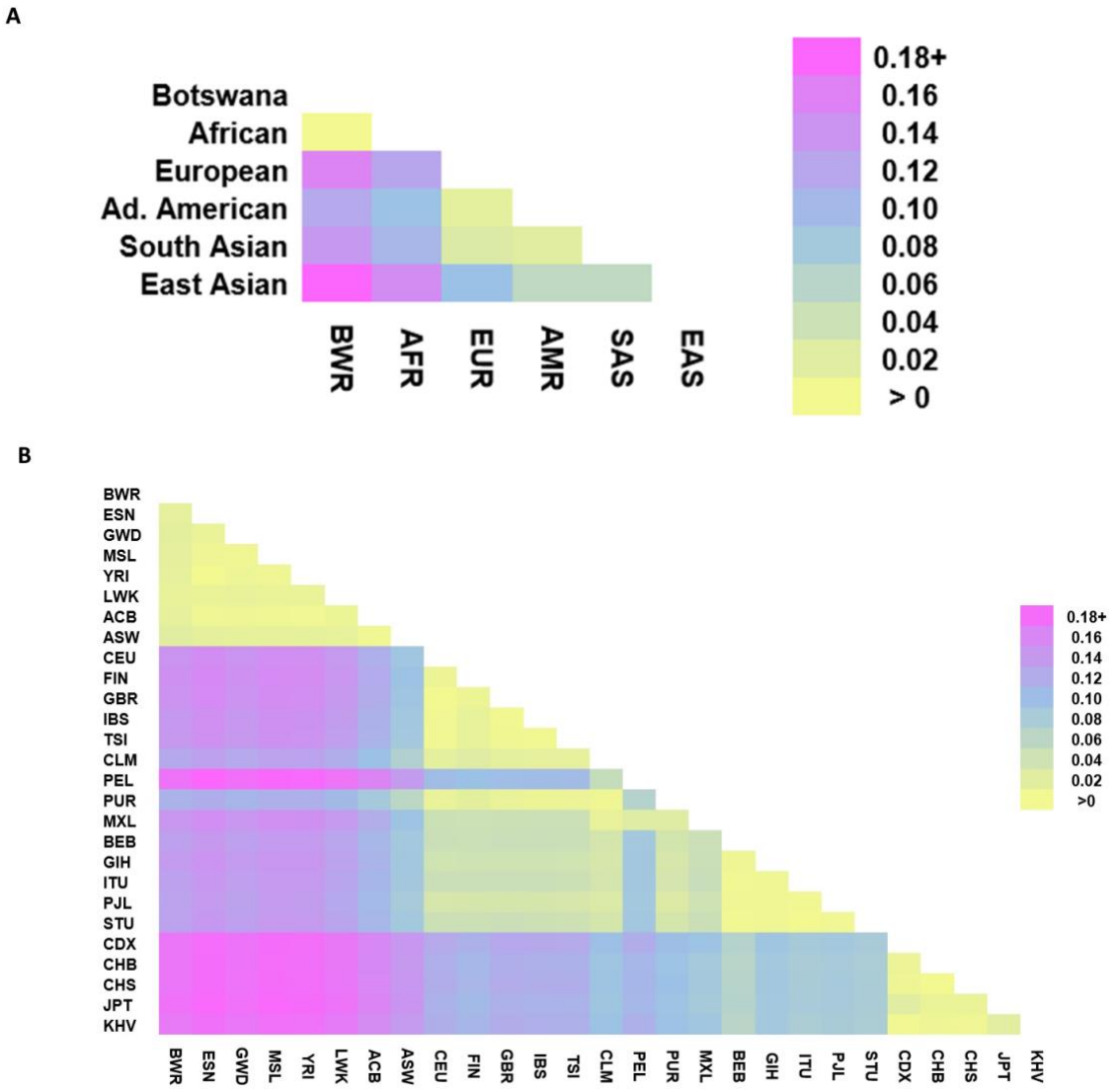

Figure S2 - Continental Weir and Cockerham's F_{ST} comparison. Closely related populations have smaller F_{ST} values, whilst populations with large F_{ST} values are taken to have higher genetic differentiation. **A -** F_{ST} comparison between Botswana and continental groupings of the 1000 Genomes data correlates with PCA results indicating affinity of Batswana with other African populations. **B -** Comparison between Batswana and other populations in the 1000 Genomes.

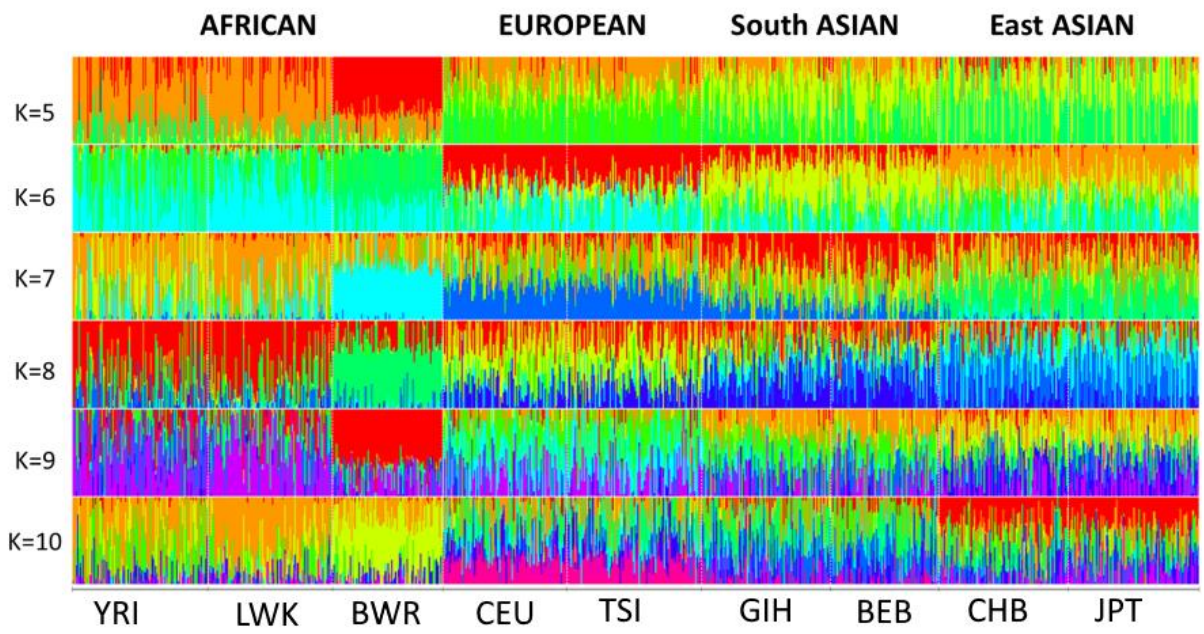

Figure S3 - Estimation of ancestral population admixture within Batswana and 1000 Genomes populations using unsupervised clustering. Runs for K5 - K10 are shown. At K=7 the Botswana population is best characterised by at least 3 ancestral populations for the majority of the individuals. Most ancestral contributions were African with a minimal Eurasian component as well as a component distinct from the two other African populations in the analysis. YRI – Yoruba in Nigeria; LWK - Luhya in Kenya; BWR – Batswana; CEU – Northern Europeans from Utah; TSI – Tuscans from Italy; GIH - Gujarati Indians from Houston, Texas; BEB - Bengali from Bangladesh; CHB - Han Chinese in Beijing, China; JPT - Japanese in Tokyo, Japan.

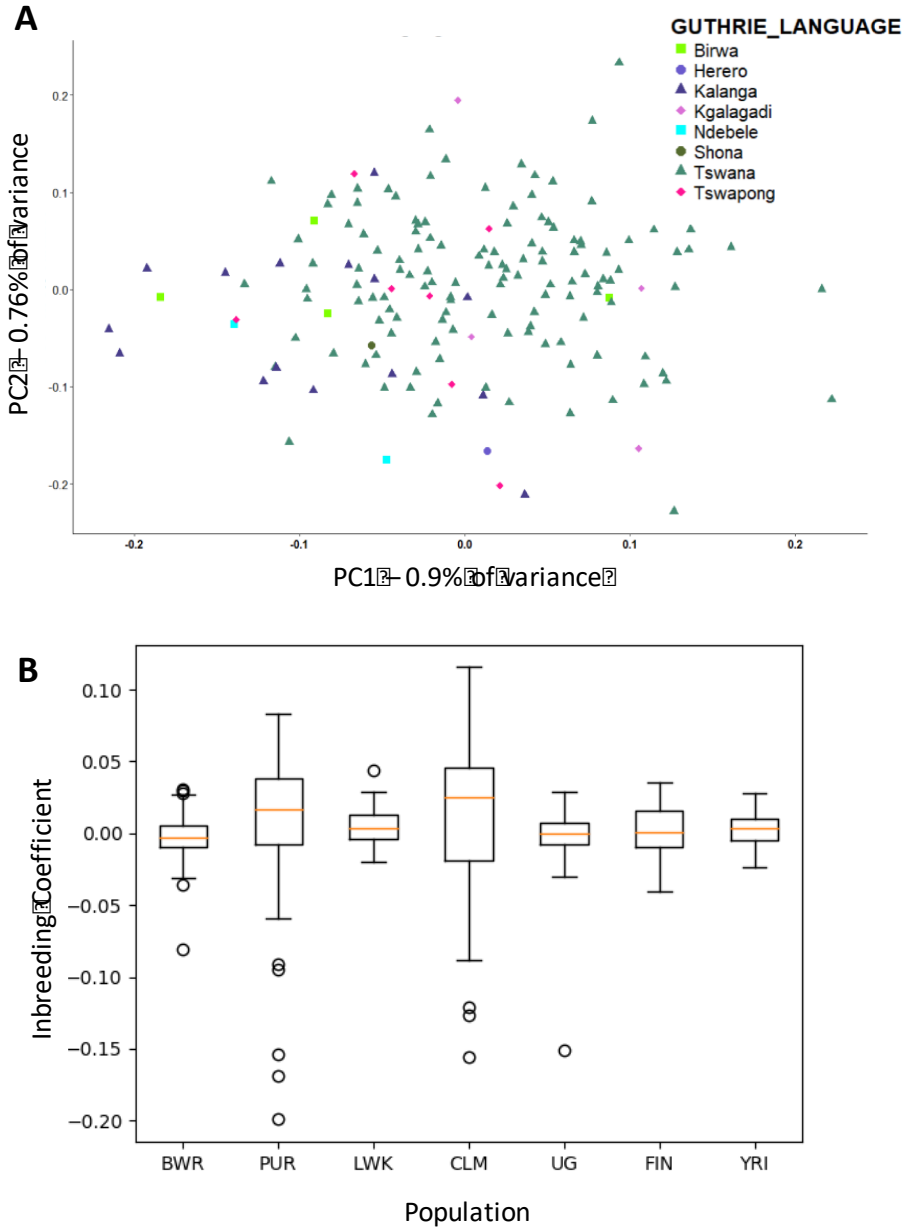

Figure S4 – **S4A** - population structure in Botswana by Guthrie language group; **S4B** - inbreeding coefficients for Botswana (BWR), Uganda (UG), and 1000 Genomes populations assessed in Figure 5.

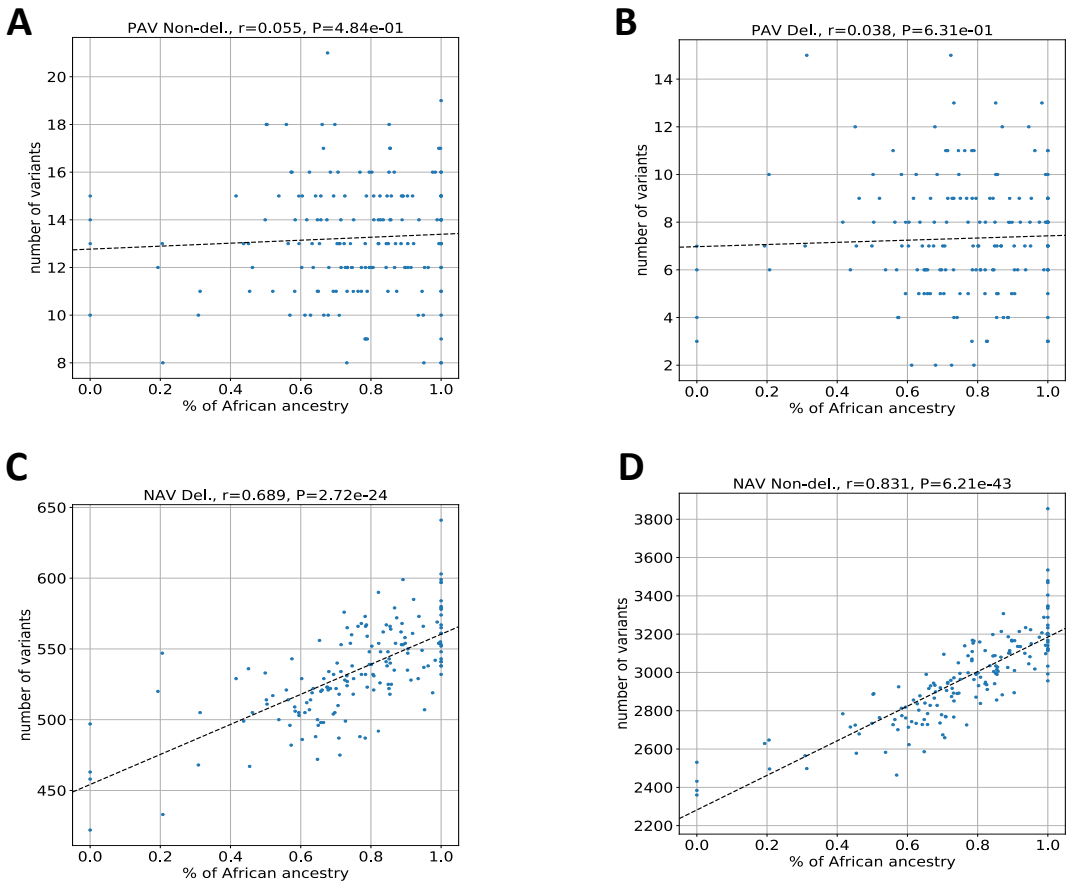

Figure S5 - The correlation between the number of variants per sample in each classification with the proportion of southern African ancestry. The variants were categorized into four groups: **A** - pathogenic and deleterious (PAV Del.); **B** - pathogenic but not deleterious (PAV Non-del.); **C** - non-pathogenic but deleterious (NAV Del.) and **D** - non-pathogenic and non-deleterious (NAV Non-del.) (see Methods for details of variant classification). The proportion of African ancestry was estimated using ADMIXTURE (K=2). Since the ADMIXTURE was performed over African samples from West, East and Southern African, the x-axis refers to the proportion of Southern African ancestry (see Figure 4).

Supplemental Tables

Table S1. Demographics of the participants. Gender and age distribution of the participants by country and HIV disease progression status.

Country	HIV Disease progression status	Male (N)	Female (N)	Median Age
Botswana	Rapid Progressors*	60	42	13
	Long-term Non-Progressors*	23	39	19
Uganda	Rapid Progressors	38	33	8
	Long-term Non-Progressors	33	46	15
	Total	154	160	

*Rapid Progressors – World Health Organization (WHO) clinical and immunological criteria for rapid progression: Anti-Retroviral treatment (ART) within 3 years of birth and/or an AIDS defining illness (WHO stage 3 or 4 OR Centers for Disease Control category 3); Two or more CD4 T cell percentage values below 15% within 3 years of birth. Long-term Non-Progressors – WHO clinical and immunological criteria for long term non-progression: Asymptomatic HIV-infection for 10 years or more after initial infection; not needing ART.

Table S2. Self-reported ancestry and Guthrie language classification of the Botswana participants

Self-Reported Ancestry	Guthrie Language Class	No. of participants
Babirwa	S32*	4
Bahurutshe	S31	10
Bakalanga	S16	15
Bakgatla_Kgafela	S31	15
Bakgatla_Mmanaana	S31	11
Bakwena	S31	24
Balete	S31	18
Bangwaketse	S31	13
Bangwato	S31	27
Barolong	S31	3
Batlokwa	S31	9
Batswapong	S32*	7
Babolaongwe	S311	2
Baphaleng	S311	1
Bashaga	S311	1
Shona	S10	1
Ndebele	S40	2
Herero	R31	1

*The Babirwa and Batswapong are described as speaking two distinct Northern Sotho languages that have come to resemble each other due their proximity rather than their origins[23].

Table S3. Database representation of all WES sequence variants from Botswana and Uganda samples.

	Botswana N (%)	**Uganda N (%)**
	On-target*	On-target*
Total number of Variants	191,758	190,584
dbSNP141_Uncaptured	36,432(14.4)	32,463(17.0)
ThouGen Uncaptured	50,955(26.6)	40,243(21.1)

*Vcrome v2.1 bed, KnownGene and ENSEMBL exon positions

Table S4. Quality control of autosomal biallelic variant markers used in PCA analysis comparing Batswana to data from 1000 Genomes and African Genome Variation Project.

Population	Variants	Variants removed	Post QC variants
Batswana	600,695	540,815	59,880
1000 Genomes	418,913	396,519	22,394
AGVP Sotho	2,139,912	2,124,068	15,844
AGVP Zulu	2,050,451	2,035,047	15,404