Whole-genome sequencing across 449 samples spanning 47 ethnolinguistic groups provides insights into genetic diversity in Nigeria

Esha Joshi[1], Arjun Biddanda[1], Jumi Popoola[1], Aminu Yakubu[1], Oluyemisi Osakwe[1], Delali Attipoe[1], 54gene Team, NCD-GHS Consortium, Estelle Dogbo[1], Babatunde Salako[2], Oyekanmi Nash[3,4], Omolola Salako[6], Olubukunola Oyedele[1], Golibe Eze-Echesi[1], Segun Fatumo[4,5], Abasi Ene-Obong[1], Colm O'Dushlaine[1*]

[1] 54gene Inc.,1100 15th St NW, Washington, DC 20005, United States

[2] Nigerian Institute of Medical Research, Lagos, Nigeria

[3] Center for Genomics Research and Innovation, National Agency for Biotechnology Development, Abuja, Nigeria

[4] H3Africa Bioinformatics Network (H3ABioNet) Node, Centre for Genomics Research and Innovation, NABDA/FMST, Abuja, Nigeria

[5] The African Computational Genomics (TAGC) Research Group, MRC/UVRI and London School of Hygiene and Tropical Medicine, London, UK

[6] College of Medicine University of Lagos, Lagos, Nigeria

* Corresponding author: Colm O'Dushlaine, colm@54gene.com

Abstract

African populations have been drastically underrepresented in genomics research and failure to capture the genetic diversity across the numerous ethnolinguistic groups (ELGs) found in the continent has hindered the equity of precision medicine initiatives globally. Here, we describe the whole-genome sequencing of 449 Nigerian individuals across 47 unique self-reported ELGs. Population structure analysis reveals genetic differentiation amongst our ELGs, consistent with previous findings. From the 36 million SNPs and indels discovered in our dataset, we provide a

high-level catalog of both novel and medically-relevant variation present across the ELGs. These results emphasize the value of this resource for genomics research, with added granularity by representing multiple ELGs from Nigeria. Our results also underscore the potential of using these cohorts with larger sample sizes to improve our understanding of human ancestry and health in Africa.

Introduction

Recent advances in human genomics research have provided compelling insights into how genetic variation plays a role in disease predisposition, and its impact on disease pathogenesis and treatment. Whole-genome sequencing (WGS) in particular can be used to identify known and novel variation in disease-associated genes, and elucidate differences in disease prevalence across diverse geographic regions and ethnolinguistic groups. However, the lack of adequate representation of diverse, non-European, genomes in human genomics research may limit insights that can be made about variants influencing disease susceptibility and trait variability across populations.

Large-scale sequencing efforts such as the 1000 Genomes Project [1], the HapMap Project [2], and TOPMed [3] have contributed to our understanding of genetic variation on a global scale and helped to narrow the gap in representation of diverse populations. In particular, these datasets have uncovered valuable insights into the distribution of novel and rare variation that exists in African populations, relative to Europeans. Despite being the most genetically diverse continent, the extent to which variation has been characterized across the numerous ethnolinguistic groups found in African countries has been limited [4]. Nigeria represents one of the most diverse and populous regions in Africa, with a population of over 200 million [5] and over 250 unique ethnolinguistic groups [6]. Genomics research involving Nigerian individuals and comprehensive cataloging of genetic variation in this diverse region can allow us to use these data as a proxy for variation on the continent. These data can subsequently inform the development of precision medicine initiatives for non-communicable diseases (NCDs).

We established the Non-Communicable Diseases Genetic Heritage Study (NCD-GHS) consortium to catalog the human genetic variation in 100,000 adults in Nigeria. In this first study, we performed germline whole-genome sequencing of an initial 449 samples from the NCD-GHS. Here, we describe the methods used to generate a whole-genome sequencing dataset of 449 Nigerian individuals spanning 47 self-reported ethnolinguistic groups (ELGs) generated using the GATK Best Practices workflow [7]. We explore the benchmarking of variant filtering strategies used to strike a balance between sensitivity and specificity by leveraging sequenced control samples. We provide a population genetics summary of the broad patterns in the data and a high-level characterization of variants, complementing that of results reported previously in the 1000 Genomes Project [8]. While sample size limits our ability to make any definitive statements about the clinical actionability of variants enriched or private to specific ELGs, we do summarize the extent to which these variants differ in prevalence within our ELGs, compared to global populations.

Our work represents a first effort to add more granularity to sample collections in Africa, with representation of several distinct ELGs from Nigeria. We provide initial insights into the relative genetic distances between ELGs and the extent to which they vary in the number of rare or common variants they contain. Our findings have implications for precision medicine across global populations, such as prioritization of more at-risk groups for screening or population-specific drug dose calibration.

Results

Samples were collected from several locations across Nigeria, with some of the larger collections based in cities and larger healthcare settings (Figures 1 and 2). The majority of subjects (68%) were female and a median age of 51 (Table 1). Approximately 60% of the dataset consists of individuals referred to healthcare settings with cardiovascular disease (Table 1). A range of ELGs are represented across the 449 samples, with 68% of the dataset being described by 15 ELGs (Table 1, Table S7).

Given that current variant calling approaches have been largely benchmarked using populations of European descent, we incorporated a non-GIAB, Yoruba sample (NA19238) as a control amongst our sequencing cohort to evaluate whether our variant calling pipeline is able to achieve a high rate of sensitivity and precision on a reference dataset that more closely relates to our population of study. When comparing the NA19238 control with Filter B applied to its corresponding HiFi dataset, we were able to achieve a precision/recall/F1 score of 97.9/91.4/94.5% for SNPs and 79.6/57.6/66.9% for Indels (Figure S1, Table S2). It is possible that using higher coverage NA19238 data would improve this performance. Combined with our findings of variant counts (Table 2) across our cohort and NYGC's AFR dataset, our results demonstrate our variant calling pipeline and post-processing filtering strategies are well-suited for variant discovery in this dataset.

Patterns of variation across ethnolinguistic groups in Nigeria

We compared the properties of observed genetic diversity in the 54gene dataset to the subset of 650 African-ancestry subjects from the New York Genome Center 1000 Genomes Project high-coverage dataset (Table 2) [1]. We found that both transition/transversion ratio (T_i/T_v) and the median number of variants per subject are comparable between datasets. We observed an increase in the overall count of SNPs and indels within the 54gene dataset relative to the African-ancestry subset of the 1000 Genomes Project (Table 2). This increase in overall variant counts are observed across all functional annotation categories (Table 3). We hypothesize that this effect could be driven by an increase in the abundance of rare variants from a wider range of ELGs in the 54gene dataset relative to the 1000 Genomes dataset. However, we cannot rule out that there may be other reasons for this observation due to sampling design, variant calling strategy, or experimental noise.

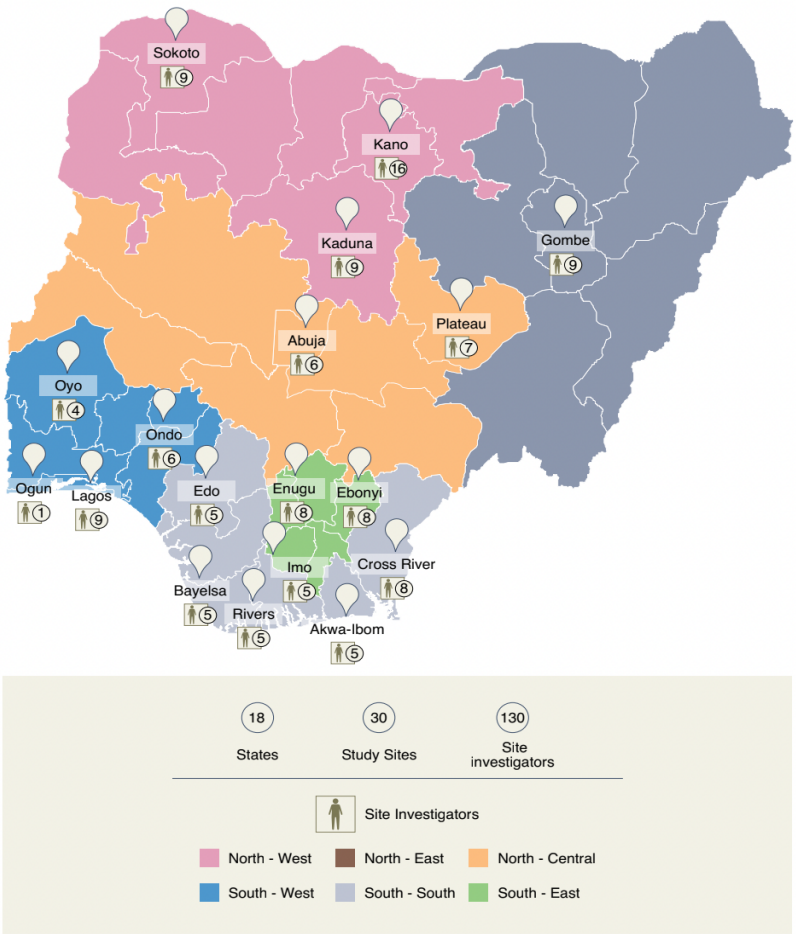


Figure 1: Overview of collection locations and regional designations within Nigeria. Additional details of the sample collection framework are discussed in Fatumo et al 2022 [9].

(a) (b)

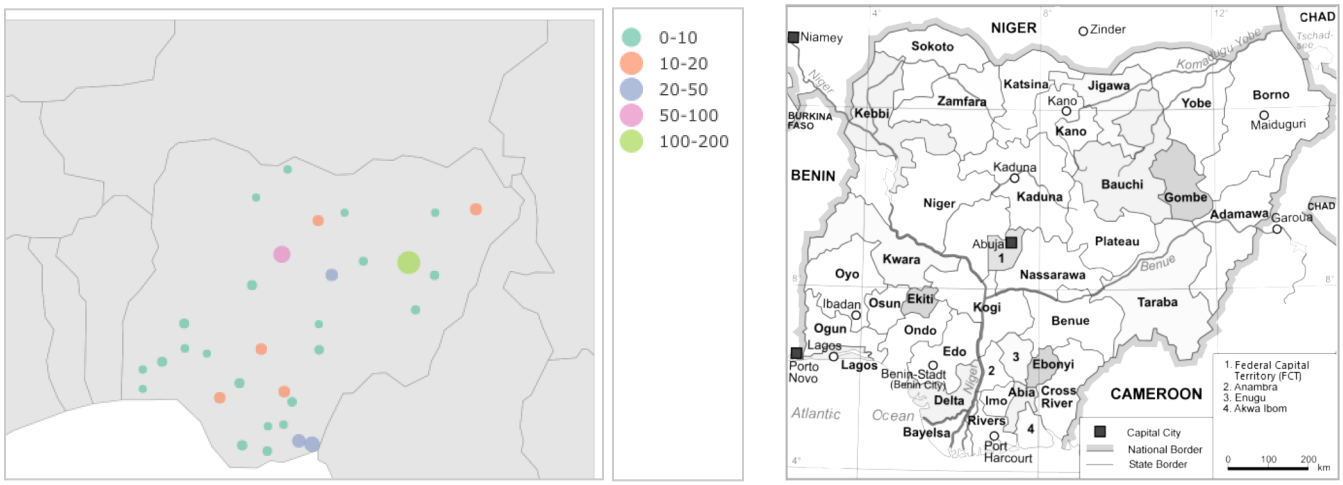


Figure 2: (a) States of origin for collected samples. Size of markers are proportional to the number of individuals collected. All states are listed in Table S1. An alternate map showing reported states for top 15 ancestries by subject count is shown in Figure S3. (b) Map of all Nigerian states (adapted from https://en.wikipedia.org/wiki/States_of_Nigeria).

	N	%
Female	307	68
Male	142	32
Median age (IQR)	51 (20)	
Median BMI (IQR)	26.7 (8.7)	
Ascertained phenotypic groups		
Cardiovascular disease	268	60
Diabetes	88	20
Solid oncology	32	7
Sickle cell disease	30	7
Neurological diseases	14	3
Endocrine thyroid disorders	11	2
Hemato-oncology	6	1
Top 15 self-reported ethnolinguistic groups (N=307, 68% of sample)		
Tangale	45	10
Fulani	33	7.3
Igbo	32	7.1
Hausa	27	6
Ibibio	26	5.8
Yoruba	26	5.8
Tera	20	4.5
Kanuri	19	4.2
Atyap	14	3.1
Bura-Pabir	13	2.9
Bini	11	2.4
Ham	11	2.4
Izon	10	2.2
Nupe	10	2.2
Waja	10	2.2

Table 1: Demographics and clinical characteristics of the 54gene dataset. Samples were ascertained across 7 phenotypic groups and 47 unique self-reported ELGs (mean 9, median 5 samples per group, IQR 7.5). The top 15 groups are shown.

Dataset	Variant subset	Variant counts	Ti/Tv	Median variants per subject	IQR
Pre-filtering 54gene cohort (N=543)	all	53360383	1.81	5675617	82912.5
	SNPs	45709746	N/A	4886715	64853.5
	indels	7650637	N/A	787822	17059.0
Post-filtering 54gene cohort (N=451)	all	36822733	2.09	4580531	62566.0
	SNPs	32496712	N/A	4002707	52381.5
	indels	4326021	N/A	576997	12900.0
Pre-filtering NYGC AFR cohort (N=661)	all	63816296	1.72	6053976	82416.0
	SNPs	55258388	N/A	4937900	66728.0
	indels	8557908	N/A	1116645	16421.0
Post-filtering NYGC AFR cohort (N=650)	all	43845824	2.08	4628977	71259.3
	SNPs	40451499	N/A	3978564	56881.0
	indels	3394325	N/A	650738	14111.3

Table 2: Counts of variants observed across all individuals by type. Counts are shown for both 54gene (449 subjects and 2 controls) and NYGC AFR datasets.

Impact category (all transcripts)	54gene cohort (N=451)	NYGC AFR cohort (N=650)
HIGH	47,627	36,910
MODERATE	732,194	529,975
LOW	844,195	781,205
MODIFIER	241,825,031	166,434,458

Table 3: Counts of variants in high-level classes of functional impact, for 54gene and NYGC datasets. Impact categories are defined as indicated here: https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html.

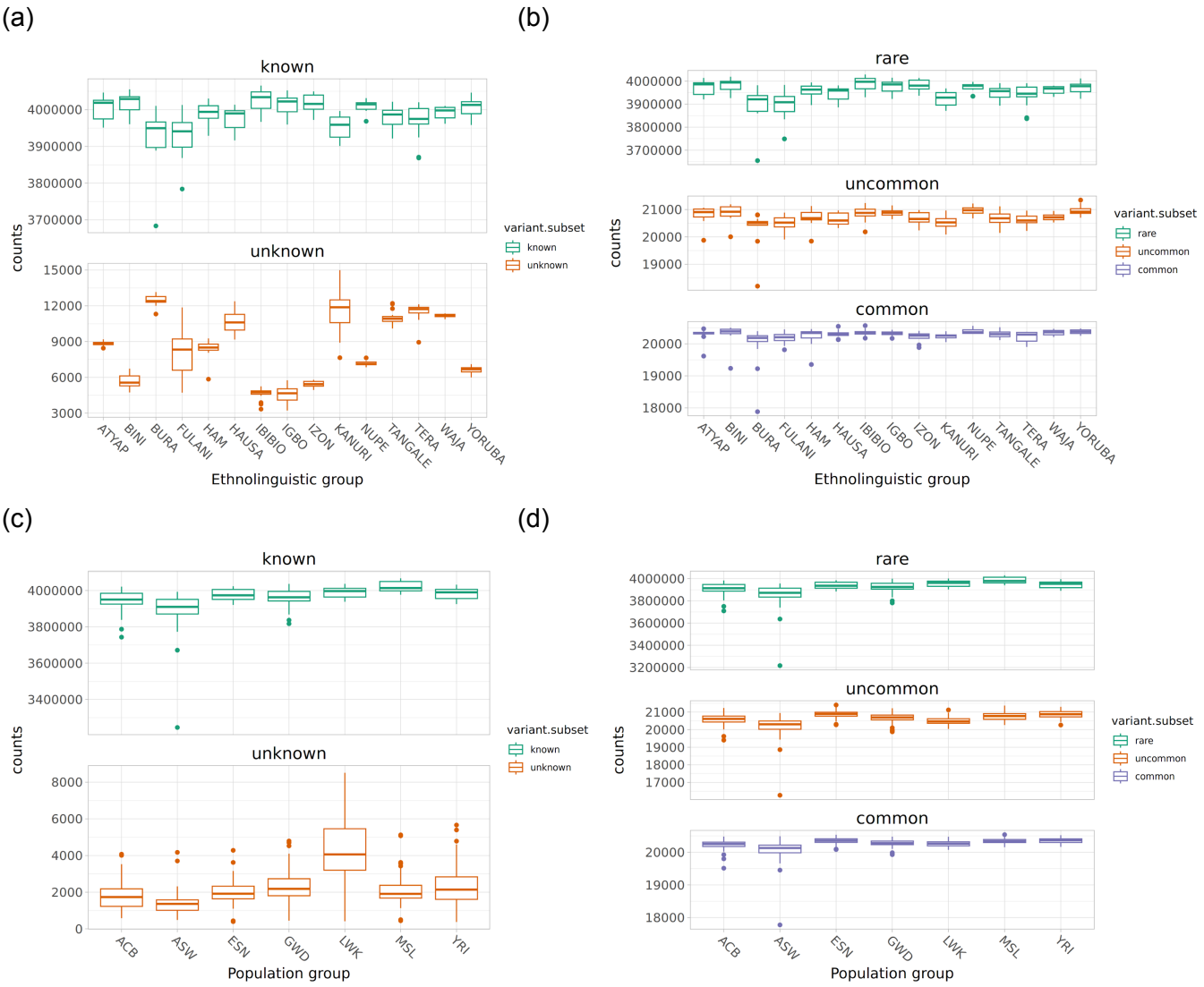


Figure 3: Variant counts across ELGs. (a) 54gene cohort, top 15 ancestries by subject count, known (present in dbsnp154) vs. unknown (not present in dbsnp154), (b) 54gene cohort, top 15 ancestries by subject count, rare (MAF < 0.1 %)/uncommon (MAF >= 0.1% and MAF < 0.5 %)/common (MAF >=0.5 %) in GnomAD AFR, (c) NYGC cohort, known (in dbsnp154) vs. unknown, (d) NYGC cohort, rare/uncommon/common in gnomad AFR (bounds the same as in (b)).

We examined the proportions of rare and novel variation across ELGs within our dataset, with the hypothesis that undersampled ELGs may harbor variation unobserved in broader catalogs of human genetic variation. Specifically, we compared counts of known and unobserved variants across the top 15 ELGs in the 54gene cohort (Figure 3). We observe that ELGs where the majority of samples are from southern Nigerian states (see Table S7) qualitatively have lower counts of unknown variants (e.g. Bini from Edo state, Ibibio from Akwa Ibom state, Igbo from Enugu state, Izon from Bayelsa state) relative to individuals from northern and northeastern states (e.g. Bura from Borno, Kanuri from

Gombe, Tera from Gombe) who tend to have higher numbers of novel variants (Figure 1, Figure 3A). However, these results remain to be corroborated by larger sample sizes across ELGs in Nigeria.

Comparing the number of rare, uncommon, and common variants across ELGs within our dataset, we see most variation in the rare category as expected [10,11]. Several ELGs show a qualitative decrease in the number of rare variants, particularly the Bura, Fulani and Kanuri groups (Figure 3B). For the latter two groups at least, we see evidence of Northern African or European admixture (Figure 5), which we hypothesize may play a role in this observation of a decrease in rare variation overall [12]. For the NYGC data, LWK (Luhya from Kenya) had the highest number of novel variants (Figure 3C). An excess of variants common in this population but rare in other populations have been reported previously, attributed to a larger long-term effective population size [8].

Population structure across ethnolinguistic groups across Nigeria

We applied Principal components analysis (PCA) to investigate patterns of population structure across the ELGs in Nigeria. For example, we noted three clusters of genetically similar ELGs (Figure 4, Figure S4). The first consists of colocalized groups of Yoruba, Ibibio, Bini, Igbo and Izon. A second group consists of Ham and Atyap. A third cluster consists of Tangale, with some overlap with Waja, Bura-Pabir and Tera. The remaining samples were substantially more heterogeneous, consisting of Hausa, Fulani and Kanuri where a major axis in PCA is dictated by the degree of admixture from populations with Eurasian ancestry [12–14]. For twelve ELGs sequenced by both Yale and MGO sequencing centers, we did not find a strong bias on genome-wide estimates of genetic ancestry (Figure S5).

Admixture clustering had the lowest cross-validation error between K=2 and K=3 (Figure S6). We found similar patterns of ancestry between the Yoruba and Esan ELGs within our dataset and YRI and ESN populations from 1000 Genomes respectively (Figure 5). Individuals reported as Yoruba,

Esan, Igbo, Ibibio, Bini, and Izon showed evidence of similar ancestral composition (Figure 5). The states of origin for individuals from these ELGs tended to be South Western (Oyo), South-South (Bayelsa, Akwa Ibom, Edo) and South Eastern (Enugu) (Figure 1, Table S7). Individuals self-reported as Nupe, Ham and Atyap differed somewhat from the first group and reflected origins from states that were largely central or central-western (Kaduna, Niger). A third group - Waja, Tangale, Bura-Pabir and Tera - corresponded to central-western and north-western states (Gombe, Borno). Lastly, Fulani, Hausa, and Kanuri stood out as having shared ancestry with northern-African or European groups (using TSI as a proxy for this ancestry), corroborating results from PCA (Figure 4, Figure S4).

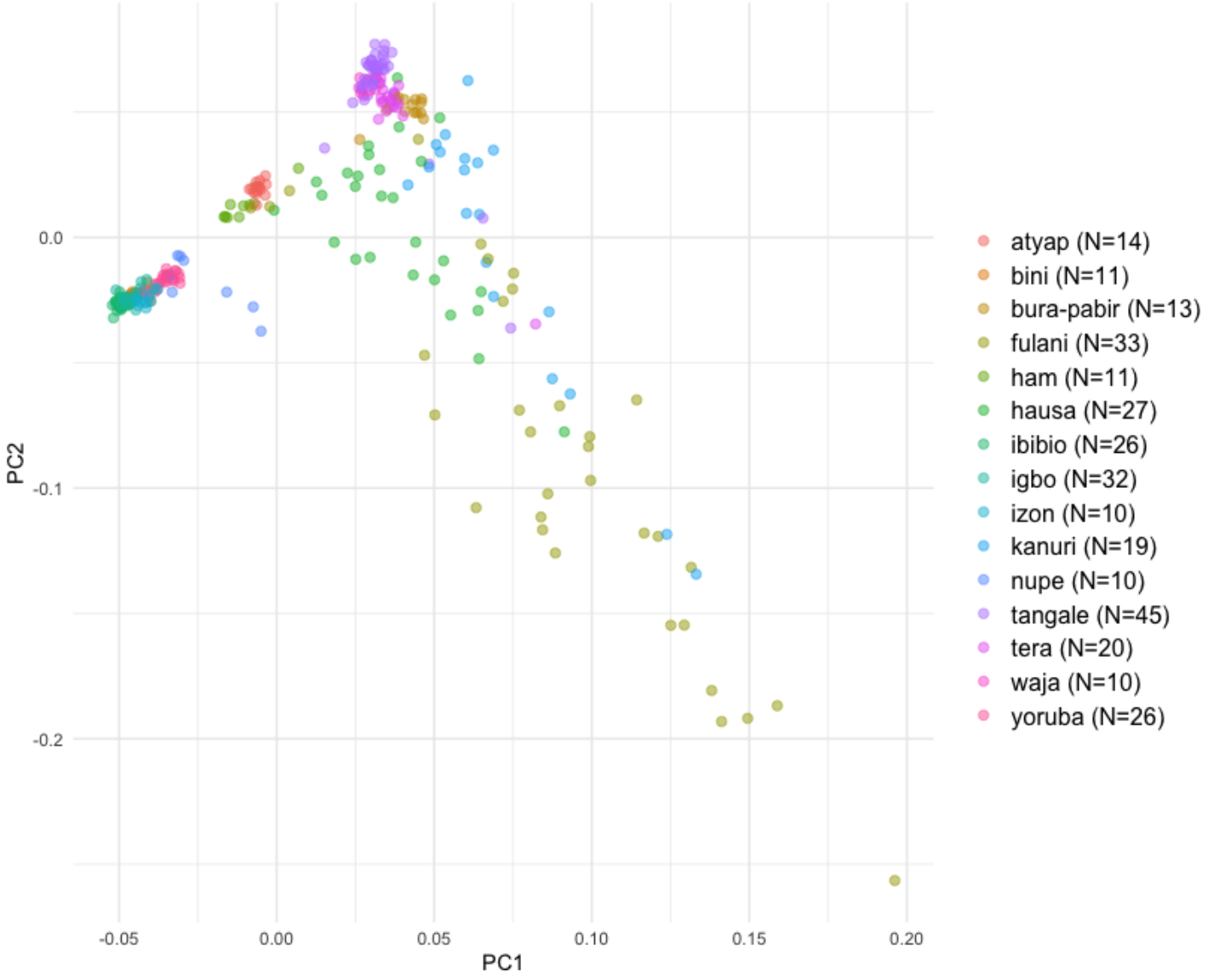


Figure 4: Principal components plot of ethnolinguistic groups listed in Table 1. An additional version of this plot with all ethnolinguistic groups is shown in Figure S3.

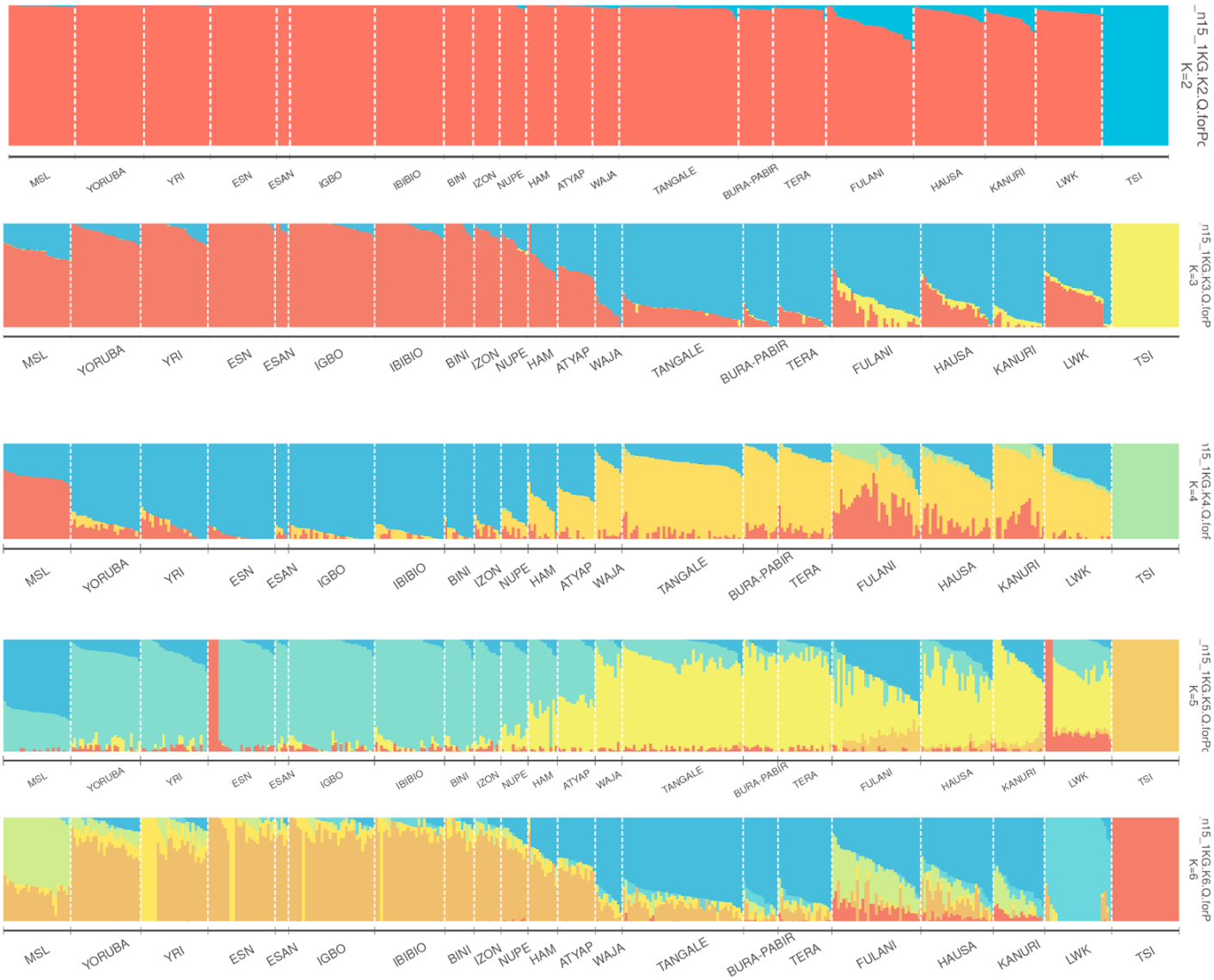


Figure 5: Population structure analysis using ADMIXTURE of ethnolinguistic groups listed in Table 1, alongside select populations from 1000 genomes (25 random samples from Gambia (GWD), Yoruba (YRI), Italy (TSI), Luhya (LWK), Esan (ESN), Mende (MSL)). Admixture plots were generated using Pophelper v1.0.10 [15].

Variation of Clinical Importance

To get a broad understanding of the genetic variation that is clinically relevant to our cohort, we further annotated and subsetted our dataset to identify variants designated as 'Pathogenic' in the ClinVar Database, with additional annotation on whether these variants belonged to genes from the

American College of Medical Genetics and Genomics (ACMG) recommended list of 73 genes with reportable variants [16].

We identified a total of 134 variants categorized as 'Pathogenic' in our cohort (Supplemental Table S3). Fourteen individuals from our cohort carried at least one reportable ACMG variant, three carried a variant in *BRCA2* (associated with breast and ovarian cancer), four carried a variant in *BTD* (associated with biotinidase deficiency), and two carried a variant in *GAA* (associated with glycogen storage disease) (Supplemental Table S4).

Of the 134 variants identified as 'Pathogenic', eight were found to have a MAF > 5% in at least one of the self-reported ELGs in our cohort (Supplemental Table S5). These eight variants were further compared to observed allele frequencies available for global populations and African population subsets in GnomAD [10], and the 1000 Genomes project. Similar to previous comparisons performed [17], we observed several of these variants with disease associations to rare disorders with a MAF < 5% across all populations in GnomAD and the 1000 Genomes project. Larger sample sizes across these ELGs would be helpful to better understand the frequencies of these variants across multiple regions in Nigeria. This falls within a larger effort to re-examine alleles associated with rare diseases in more comprehensive population reference datasets.

Allele frequencies of known variants associated with response to indicated drugs

To assess the value of underrepresented ancestries on the interpretation of genetic variants related to drug response, we surveyed frequencies of variants in key pharmacogenes within the ELGs from the 54gene dataset. Specifically, we compared the frequencies of variants in these key pharmacogenes across ELGs to selected ancestry groups from the 1000 Genomes project (Table S6) [18].

We note elevated frequencies of pharmacogenomic variants for ELGs where the majority of samples are from northern states (Hausa, Fulani) relative to other ELGs sampled from the 54gene dataset as well as ancestry groups from the 1000 Genomes Project (Table S6). For example, the variant rs3093105 which affects warfarin response has a frequency of approximately 40% in the Fulani and Hausa ELGs, but is closer to 30% frequency in the Yoruba [19]. We also observed an elevated frequency of rs2108622 in the Fulani and Hausa ELGs (15-17% vs. ~5% in YRI), which is also associated with warfarin metabolism (Table S6). These observations suggest there may be an increased risk for adverse drug reactions within these ELGs. However, there are additional factors to consider beyond the differing allele distributions such as socioeconomic factors, sampling strategy, and the geographic location and environment of these populations. The analysis performed here only applies to a limited subset of known variants within these genes, and further studies are needed to characterize novel variants in pharmacogenes and their effects on drug efficacy.

Discussion

This report represents an initial assessment using whole-genome sequencing to understand variation within, and the population structure of, some of the predominant ethnolinguistic groups in Nigeria. This report also demonstrates the capacity for conducting large-scale genome analyses in the region, speaking to the promise of building research capacity on the African continent [17,20]. We present results for several ethnolinguistic groups that have not been previously sequenced, or for which there is very little existing publicly available sequence data. We demonstrate that we can observe a discernible population structure amongst closely related populations, even with limited sample sizes across groups. Our results are consistent with results for populations already sequenced as part of previous efforts, e.g. Yoruba from 1000 genomes. We have added to this by sampling from a wider set of ethnolinguistic groups across Nigeria.

By using the NA19238 African control in addition to the gold-standard NA12878 to perform benchmarking using field standard strategies [21], we show that we were able to well calibrate our variant calling pipeline for variant discovery and generate comparable variant counts between the 54gene dataset and NYGC's African sample data. The sample sizes across the self-reported ethnolinguistic groups in our cohort and their depths of coverage limit the discovery of clinically relevant variants, and potential conclusions that can be made about the distribution of pathogenic disease-associated variants in Nigeria. Nevertheless, our findings of relative counts of ACMG-reportable variants and broad comparisons of pathogenic variant frequency can serve as a template for cataloging variation at the level of ethnolinguistic group, helping to add more granularity to future medical genetics research in Africa.

This report provides an approach for conducting further population genomic studies in Nigeria using whole-genome sequencing with larger sample sizes to provide more definitive insights into novel or rare variation in certain ethnolinguistic groups and provide a high-level summary of population structure. Our results also emphasize the utility of publicly available whole-genome sequencing data from under-sampled African populations as a resource to enable better cataloging of genetic variation to drive initiatives in precision medicine, improvement of human reference genomes, and the elucidation of population histories.

Methods

Study Participants

Individuals for this study were recruited from numerous study sites across Nigeria (Figure 2) [9]. The hospital study sites cared for patients with non-communicable diseases including cardiovascular diseases, neurological diseases, thyroid disorders, diabetes mellitus, solid and hematological cancers, and sickle cell disease. Patients within these diseases of interest were introduced to the

study by their attending physician and subsequently recruited if they met the following inclusion criteria: (1) participants aged 18 years or older, and (2) participants voluntarily provided informed consent. Upon obtaining informed consent, study-specific Research Assistants (RAs) administered questionnaires collecting basic demographic, behavioral, and medical history information from the participant. Thereafter, a laboratory technician collected and processed the requisite blood biospecimen.

Sample processing, DNA extraction, and Whole Genome Sequencing

Whole blood aliquots at a volume of four (4) milliliters were collected from participants using sterile Vacutainer blood collection kits (BD Vacutainer®) and subsequently biobanked at -80°C for DNA extraction. DNA was extracted from peripheral whole blood using the automated KingFisher Extractor and using the MagMax DNA multi-sample extraction kit, according to manufacturer protocol recommendations (Thermo Fisher, U.S.A.). Resulting genomic DNA was assessed for concentration and purity using Promega DNA quantification kit on the Promega Quantus Fluorometer (Promega, Germany) and measurement of A260/A280 ratio on the MultiSkan Sky High spectrophotometer (Thermo Fisher, U.S.A.). DNA libraries were prepared using the Illumina DNA PCR-Free Library Preparation kit, Tagmentation, and following Illumina recommended protocol (Illumina, U.S.A.).

Resulting libraries were subjected to whole-genome sequencing (WGS) on the Illumina Novaseq 6000 sequencer, and pooled to achieve a desired target of 30x genome coverage, accepting a minimum of 20x. WGS was carried out at the 54gene Nigeria Molecular Genetics Laboratory. A portion of samples (256/449 subjects) were sequenced at our partner laboratory at Yale Center for Genome Analysis (YCGA), at a target of minimum 30x coverage using paired end sequencing as well (Figure S2). This was carried out using the Illumina Novaseq 6000 instrument and the Lotus DNA library preparation kit.

FASTQ generation

Raw FASTQs were provided for samples sequenced by YCGA. Samples sequenced in-house were converted from raw binary base call (BCL) files to FASTQ using Illumina's bcl2fastq2 utility available on their BaseSpace Sequence Hub platform, using the following parameters: minimum trimmed read length and masking of short adapter reads set to 35, barcode mismatches set to 0, and adapter stringency set to 0.9. Additional flags applied to the BCL conversion were: `--find-adapters-with-sliding-window`, `--ignore-missing-bcls`, `--ignore-missing-filter`, `--ignore-missing-positions`, and `--ignore-missing-controls`.

Variant calling and QC

The 54gene WGS germline pipeline (see Data and Code Availability) was used to process the raw sequencing data. FastQC reports were generated for all raw FASTQs, followed by read trimming and adapter removal with fastp (see Data and Code Availability) [22]. A second pass of FastQC was performed to confirm effective adapter removal and trimming. Reads were aligned to the GRCh38 reference genome using bwa-mem (see Data and Code Availability) [23], followed by deduplication. GATK (v4.2.5.0) [24] BaseRecalibrator was used to generate a recalibration table for base quality scores using the VCFs for known SNP and indel sites in dbSNP (build 138) from the Broad's genome references on Amazon Web Services, then applied to each BAM file. Samtools stats were then generated for all BAMs. Variant calling was performed using GATK's HaplotypeCaller. Joint genotyping was performed using GATK's GenomicsDBImport tool to generate database stores for each sample, parallelized across fifty (50) regions using interval lists of approximately 59Mb each in size made available by the Broad Institute (see Data and Code Availability). The database stores for each of these regions were subsequently passed to GenotypeGVCFs. Variant normalization was applied and multiallelic variants were split into multiple records using `bcftools norm`. Hard-filtering was performed using GATK's VariantFiltration tool.

Post-calling subject-level QC consisted of the following steps: contamination checks were performed using VerifyBamID; subject relatedness was estimated using Somalier in order to identify unexpected genetic duplicates; sex discordances were detected using two orthogonal techniques, Somalier and the `bcftools guess-ploidy` plugin (see Data and Code Availability). Samples were excluded based on the following thresholds: het/hom ratio above 2.5, average depth less than 20x, and a contamination estimate above 0.03.

More details on the variant calling pipeline used can be found in the following documentation: (https://54gene-wgs-germline.readthedocs.io/en/latest/).

Variant annotation

We used VEP [25] to annotate variants using the *Homo sapiens* database version 106. Annotations were performed with the following flags: `--sift b --polyphen b --variant_class --symbol --canonical --check_existing --af --max_af --af_1kg --af_esp --af_gnomad`. Additionally, the following regions were acquired from the hg38 UCSC Genome Browser track and applied as custom annotations: repeatmasker, simpleRepeat, microsatellite, segdups, windowmaskerSdust, centromeres, telomeres, and gaps. Finally, a custom annotation was applied for variant presence in the ENCODE blacklist, hg38 version 2 (see Data and Code Availability) [26]. Annotations for variant classifications of clinical importance using the ClinVar Database (v.20221113), ACMG v3.0 list of genes from secondary findings, and variant labels (rs IDs) from dbSNP (v.154) were also included.

Post-calling variant filtering

We explored several different variant filtering strategies to apply to our dataset and to evaluate sensitivity and specificity (Table S2, Supplemental Notes). We applied the filtering strategies to two control samples, the Genome in a Bottle (GIAB) NA12878 genome (benchmarked against the NIST high confidence call set [27]) and NA19238, a Yoruba reference sample from the 1000 Genomes

project (benchmarked against a publicly available HiFi call set) [28]. We selected these controls to assess sensitivity and specificity of our variant calling and filtering in both a well-characterized European-ancestry sample and a sample more representative of the ancestries we are studying. We used hap.py (see Data and Code Availability) to assess performance; see Supplemental Notes for details.

Preparing publicly available data

We acquired publicly available data produced by the New York Genome Center [1] (see Data and Code Availability). We subsetted the data to 661 subjects from the populations in the African-ancestry regional grouping (ACB, ASW, ESN, GWD, LWK, MSL, and YRI) and removed all second degree relatives, leaving 650 subjects available for merging with the 54gene dataset. Prior to merging, we applied the annotation and filtering criteria as described above for the 54gene dataset.

Principal Components Analysis and analysis of population structure

Using PLINK [29] version 1.9, a starting call set of 36,987,934 variants across 451 subjects (Table 2) were filtered to 0.5% minor allele frequency within the cohort, cleaned using common variant filters, two control samples were removed, and data were subjected to principal component analysis: variants were filtered to genotype missingness less than 5% and Hardy-Weinberg Equilibrium exact p-value greater than 0.001. Variants were also filtered out if they exhibited patterns of non-random missing genotypes based on a `plink --test-mishap` p-value less than $1e^{-5}$. The resulting set of 19,359,756 million variants across 449 subjects was further filtered for linkage disequilibrium with successive passes through `plink --indep 50 5 2` and `plink --indep-pairwise 50 5 0.2`. Variants in the MHC regions (25-35Mb on chromosome 6) were removed, leaving 2,627,713 variants for population genetic analyses. Principal components were estimated using the default settings of `plink --pca` [29]. We used ADMIXTURE [30] to characterize the population structure of 54gene samples, alongside 150 African and European (TSI) samples from the 1000 Genomes Project [1], 25 randomly selected

samples per group summing to 150 individuals and LD pruned to 636,076 variants. We applied the clustering approach in ADMIXTURE across a range of cluster counts (K), from K=1 to K=7, on a merged set of 42,219 variants across 599 54gene and 1000 Genomes samples.

Data and Code Availability

Our data will be made available to approved researchers through the NCG-GHS consortium and through the European Genome-Phenome Archive (EGA).

The code for the pipeline used to perform germline variant calling can be found on GitLab at https://gitlab.com/data-analysis5/dna-sequencing/54gene-wgs-germline. Code for figure generation can be found at https://gitlab.com/data-analysis5/dna-sequencing/generate_figures.

Tools and datasets used

- Fastp: https://github.com/OpenGene/fastp
- Somalier: https://github.com/brentp/somalier
- VerifyBamID: https://github.com/statgen/verifyBamID/
- Encode blacklist: https://github.com/Boyle-Lab/Blacklist/tree/master/lists
- Hap.py: https://github.com/Illumina/hap.py
- NYGC dataset: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/CCDG_13607_B01_GRM_WGS_2019-02-19_chr[1-22,X,Y].recalibrated_variants.vcf.gz.
- Broad resources for known sites VCFs and scattered calling intervals: https://s3.amazonaws.com/broad-references/broad-references-readme.html

Acknowledgements

This work represents a significant effort led from Nigeria. We thank the members of NCD-GHS Scientific Advisory Board Members: A. Adeyemo, F. I. Olopade, M. Nyirenda, N. Mulder, M. Owolabi and D. Nitsch and members of the NCD-GHS Steering Committee: A. Adenipekun, A. Banjo and S. Saidu. Biobanking and sequencing were conducted within our labs in Lagos. We also thank the Yale Center for Genome Analysis (YCGA) for additional sequencing support. Investigator engagement and patient recruitment were conducted by field teams over the course of several months. We thank the site investigators and the 54gene site quality assurance officers for effectively coordinating the study at the various sites. We also thank biobank governance, ethics and data access committee members Martin Meremikwu, Aminu Zakari, Aisha Mamman, Alatise Olusegun, Ike Anya, Moji Makanjuola, Adeola Fowotade, Nnamdi Ilodiuba. Funding for this study was provided by the ACTG with support from 54gene Inc. Above all, we thank participants for their support, participation and engagement as we seek to promote genomics at scale on the continent of Africa.

Author information

Authors and Affiliations

NCD-GHS Consortium, African Center for Translational Genomics (ACTG), Ikate, Lekki, Lagos, Nigeria

Segun Fatumo, Aminu Yakubu, Abdullahi Musa, Abdulrasheed M. Mujtaba, Abiodun Popoola, Abubakar M. Bello, Anthony Anyanwu, Ashiru Yusuf, Gesiye EL Bozimo, Goddy Bassey, Hadiza Bala, Istifanus Bala Bosan, Jemimah Edah, Mutiu Alani Jimoh, Kenneth Nwankwo, Olalekan Ojo, Marcus Inyama, Maryam Apanpa, Mohammed Inuwa Mustapha, Musa Ali-Gombe, Olubukola Ojo, Oludare F Adeyemi, Samuel Ajayi, Sanusi Bala, Temitope Ojo, Usman Malami Aliyu, Yemi Raji, Zainab Tanko, Amina Mohammed, David Oladele, Muhammed Hamzat, Emmanuel Agaba, Emeka Nwankwo, Ifeoma Ulasi, Jonah Musa, Umeora Odidika, Omolola Salako, Oyekanmi Nash, Babatunde L Salako, Nwankwo Kenneth Chima, Marcus Inyama Asuquo, Timothy Ekwere, Ezechukwu Aniekwensi, Chidi

Ezeude, Olayemi Awopeju, Tolutope Kolawole, Olubiyi Adesina, Vandi Ghyi, Olaolu Oni, Zumnan Gimba, Colm O'Dushaline, Abasi Ene-Obong

The African Computational Genomics (TACG) Research Group, MRC/UVRI and LSHTM, Entebbe, Uganda

Segun Fatumo

The Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK

Segun Fatumo

54gene Team, Washington, DC, USA

Ogochukwu Francis Osifo, Zahra Isa Moddibo, Aisha Nabila Ado-Wanka, Aminu Yakubu, Olubukunola Oyedele, Jumi Popoola, Delali Attiogbe Attipoe, Golibe Eze-Echesi, Fatima Z Modibbo, Nabila Ado-Wanka, Oluyemisi Osakwe, Onome Braimah, Eramoh Julius-Enigimi, Terver Mark Akindigh, Bolutife Kusimo, Chinenye Akpulu, Chiamaka Nwuba, Ofonime Ebong, Chinyere Anyika, Oluwatimilehin Adewunmi, Yusuf Ibrahim, Janet Kashimawo, Ogochukwu Francis Osifo, Chidi Nkwocha, Peter Iyitor, Temi Abiwon, Adeola Adeleye, Abayomi Ode, Anjola Ayo-Lawal, Kasiena Akpabio, Emame Edu, Chiemela Njoku, Bari Ballew, Cameron Palmer, Esha Joshi, Arjun Biddanda, Colm O'Dushlaine, Abasi Ene-Obong, Teresia L Bost, Olubukunola Oyedele

Department of Radiation Biology, Radiotherapy and Radiodiagnosis, College of Medicine, University of Lagos, Lagos, Nigeria

Omolola Salako

Centre for Genomics Research and Innovation, National Biotechnology Development Agency NABDA, Abuja, Nigeria

Oyekanmi Nashiru

The Nigerian Institute of Medical Research (NIMR), Lagos, Nigeria

Babatunde L Salako

Corresponding author

Colm O'Dushlaine

Declaration of interests

E.J., A.B., A.Y., O.O., D.A., O.O., G.E., and C.O'D. are currently employed by 54gene Inc. at the time of preprint release.

Supplementary Materials

Supplementary Notes

Benchmarking variant filtering strategies

We benchmarked our variant filtering strategies using NA12878 as a control. DNA was obtained from Coriell and was run concomitant with our sequencing samples in the 54gene Nigeria Molecular Genetics Laboratory. We assessed specificity and sensitivity using the NIST platinum truth set [27].

Additionally, we wanted to determine if variant calling accuracy would be systematically different in African genomes, or whether NA12878 was a reasonable representative control even for genomes of quite distinct ancestry with different LD patterns. Therefore, we acquired DNA from Coriell for an African subject (NA19238) for whom there is also publicly available PacBio HiFi sequencing data [28].

We assessed a variety of filtering strategies by evaluating (1) NA12878 with the NIST truth set [27], (2) NA12878 using a HiFi truth set [28], and (3) NA19238 using a HiFi truth set (Table S2) [28]. We used the hap.py package to evaluate performance compared to truth data (see Data and Code Availability).

We found that in benchmarking our hard-filtered calls for NA12878 against the NIST truth set, within the NIST-defined high-confidence regions, precision/recall/F1 for SNPs was 99.4/98.1/98.7%, and for indels was 98.9/80.0/88.4%, across all stratification regions. This indicated that our variant calling pipeline was performing as expected, and evaluated using field-standard benchmarking techniques. We then switched to using the HiFi data as our truth set, across all calls within the NIST high-confidence regions. In this benchmark comparison, we found that precision/recall/F1 for SNPs was 97.6/98.1/97.8%, and for indels was 88.7/78.7/83.4%, across all stratification regions. This gave us a reasonable basis for expectations when comparing NA19238 to HiFi data.

For the next steps, we evaluated all variant calls across the genome, instead of exclusively within the NIST high-confidence regions. Using the variant filtering strategies outlined in Table S2, we performed benchmarking of either NA12878 or NA19238 against the relevant HiFi dataset, and evaluated performance across all stratification regions defined by GIAB [21] (https://github.com/genome-in-a-bottle/genome-stratifications, GRCh38 v3.0). We tuned our filtering to prioritize precision over recall for this evaluation. We chose Filter E outlined in Table S3 as the final filter for our dataset.

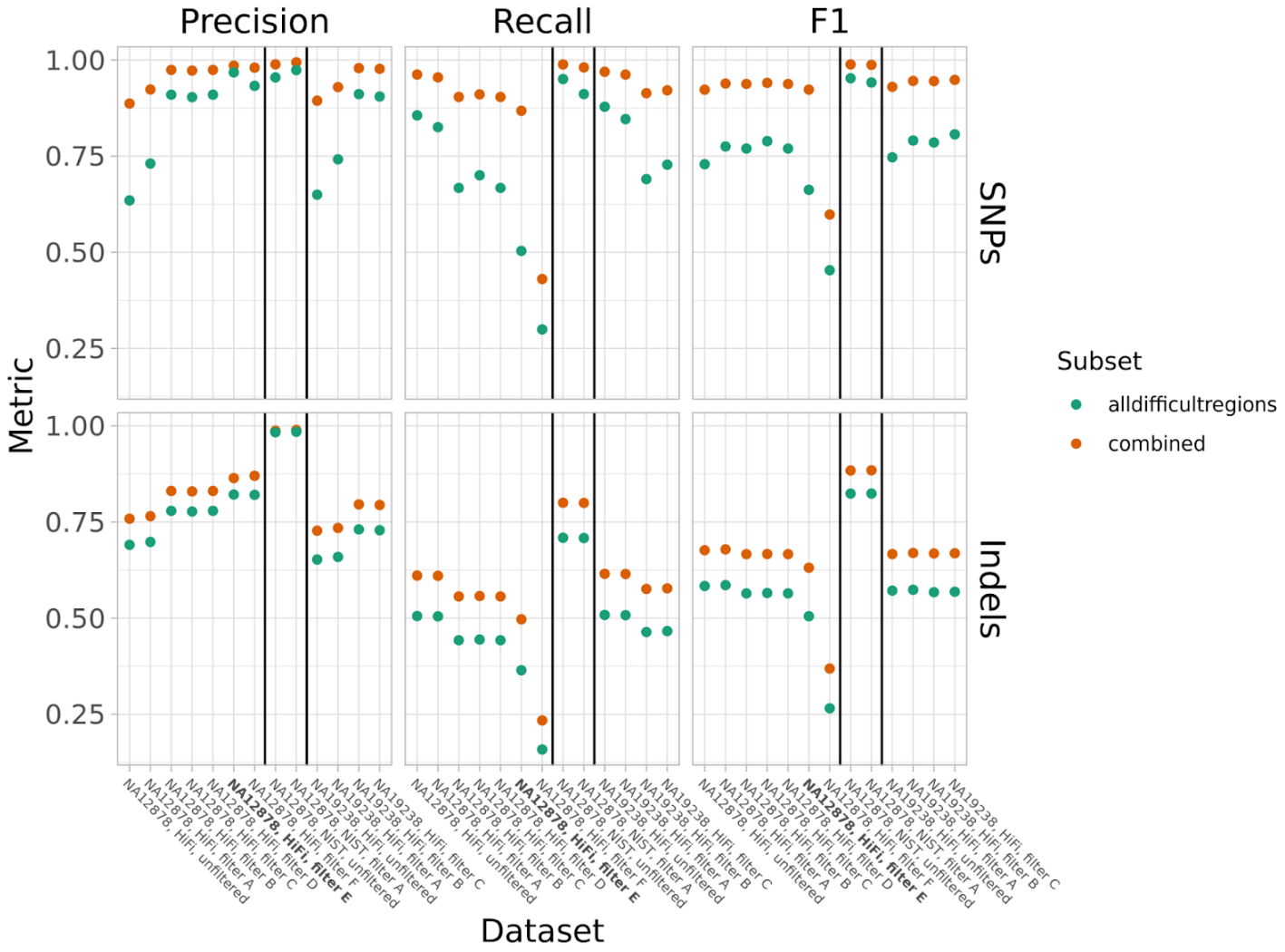


Figure S1. Precision, recall and F1 for SNPs and indels across a range filtering paradigms. X-axis labels correspond to filters detailed in Table S2. “alldifficultregions” are defined at https://github.com/genome-in-a-bottle/genome-stratifications/blob/v3.0/GRCh38/union/GRCh38-union-README.md (GRCh3X_alldifficultregions.bed.gz, corresponding to the union of all tandem repeats, all homopolymers >6bp, all imperfect homopolymers >10bp, all difficult to map regions, all segmental duplications, GC <25% or >65%, "Bad Promoters", and "OtherDifficult" regions (including regions from the T2T-consortium for GRCh38 only)). Unfiltered sets are indicated at the vertical black lines. We used Filter E as the final filter for downstream analysis, shown in bold text on the X-axis.

Supplementary Tables

Table S1: States of origin for participants

Table S2: Evaluation of variant and region filtering strategies
All regions are defined based on hg38. Centromere bed file: UCSC table browser group “Mapping and Sequencing”, track “Centromeres”. Telomere bed file: UCSC table browser group “Mapping and Sequencing”, track “Gap”, type=Telomere. Simple repeats bed file: UCSC table browser group “Repeats”, track “Simple Repeats”. Segdups bed file: UCSC table browser group “Repeats”, track “Segmental Dups”. Microsatellites bed file: UCSC table browser group “Repeats”, track “Microsatellite”. RepeatMasker bed file: UCSC table browser group “Repeats”, track “RepeatMasker”. ENCODE blacklist bed file:https://github.com/Boyle-Lab/Blacklist/blob/master/lists/hg38-blacklist.v2.bed.gz

Table S3: ClinVar pathogenic and likely pathogenic variants in our cohort

Table S4: ACMG reportable variants (v3.0) found in our cohort

Table S5: Clinvar pathogenic variants identified in the 47 ELGs with MAF > 5%

Table S6: Frequencies of key pharmacogenomic variant alleles found within ELGs compared to 1000 Genomes

Table S7: Counts of individuals in each ELGs by state

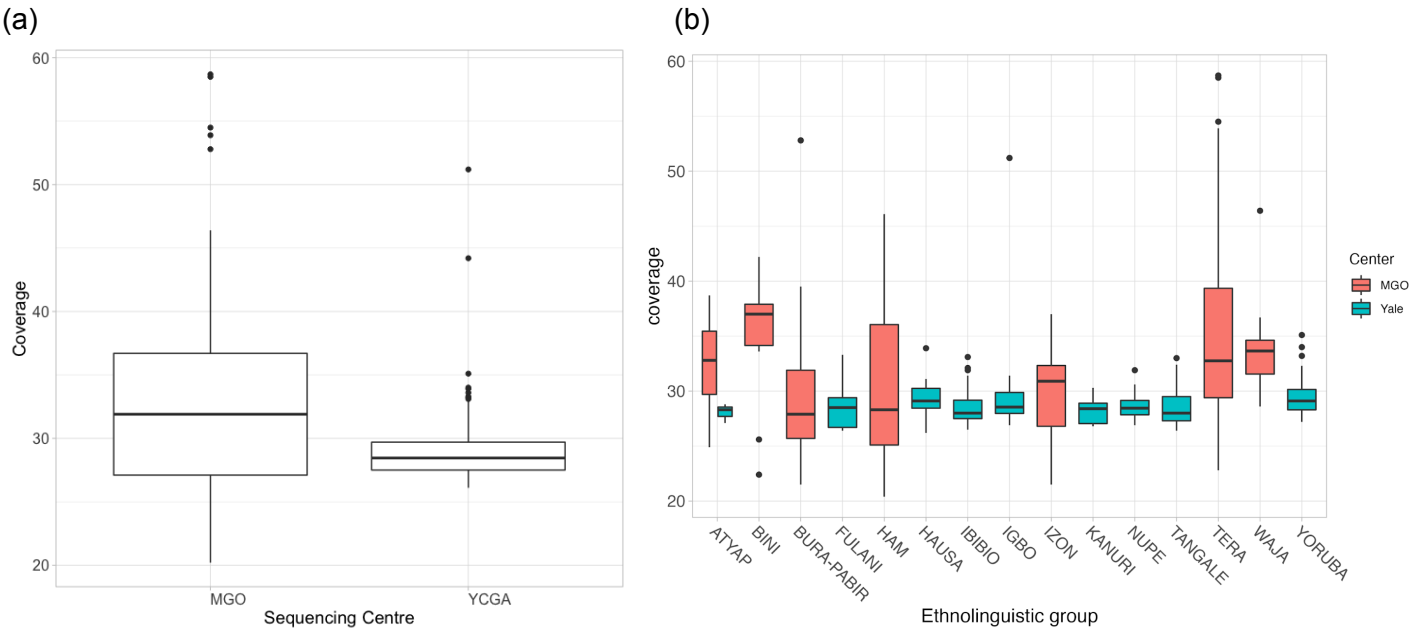


Figure S2: Depth of coverage of all N=449 subjects. (a) Coverage shown by sequencing center for subjects sequenced at 54gene's Molecular Genetics Laboratory (denoted as MGO), and at our partner laboratory Yale Centre for Genome Analysis (YCGA), (b) Coverage shown for each of the top 15 ethnolinguistic groups by sequencing center.

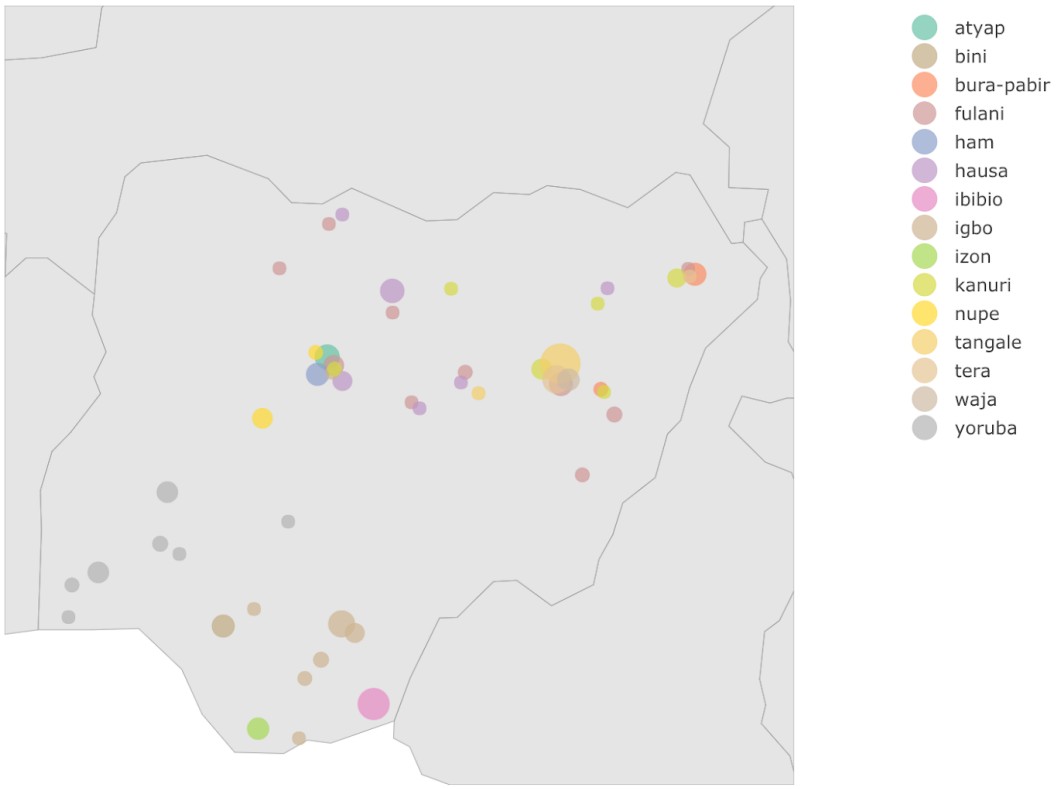
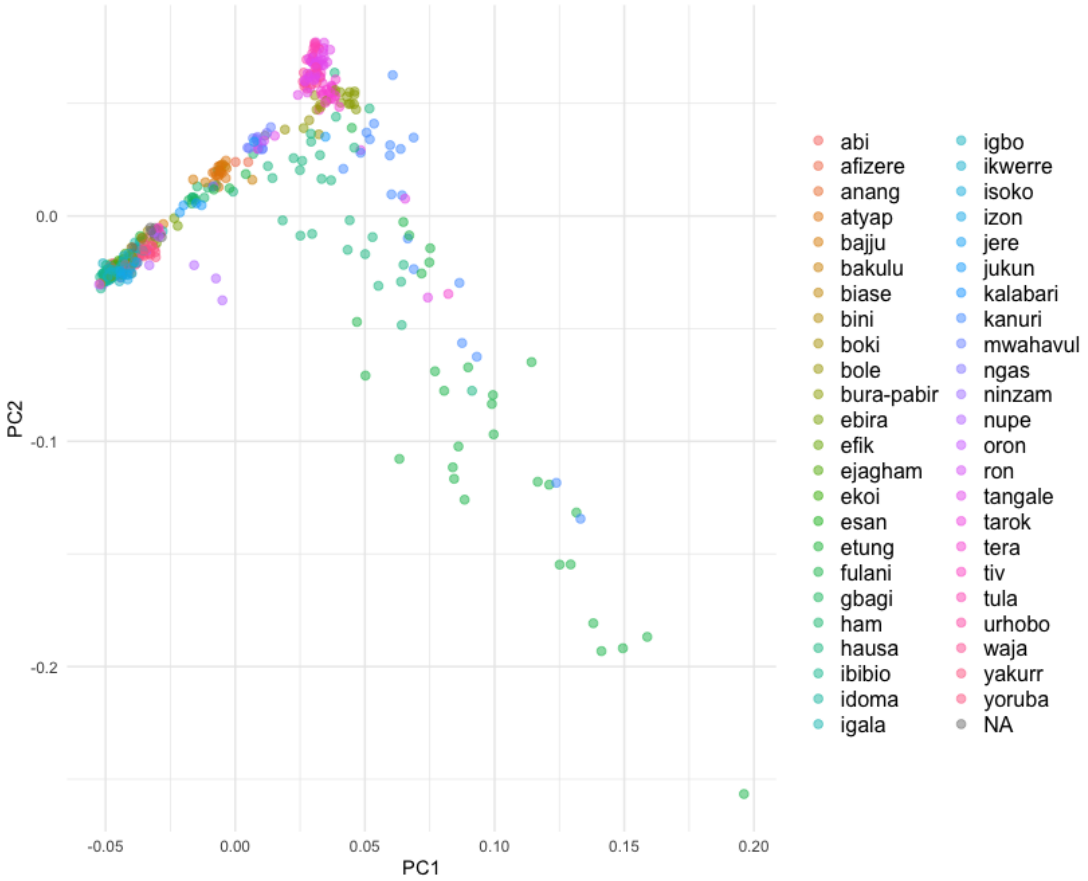


Figure S3: Reported ethnolinguistic group and state of origin for top 15 most prevalent groups.

A)



(B)

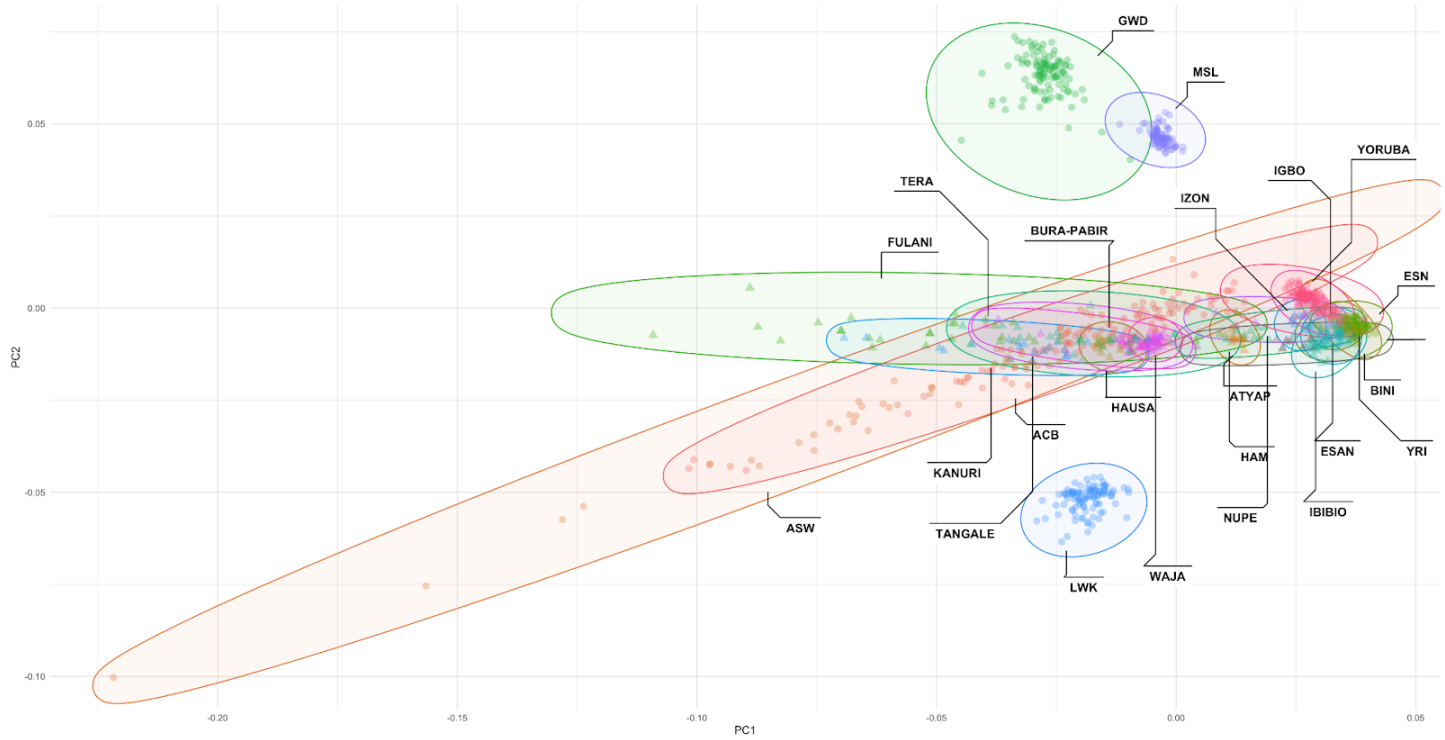
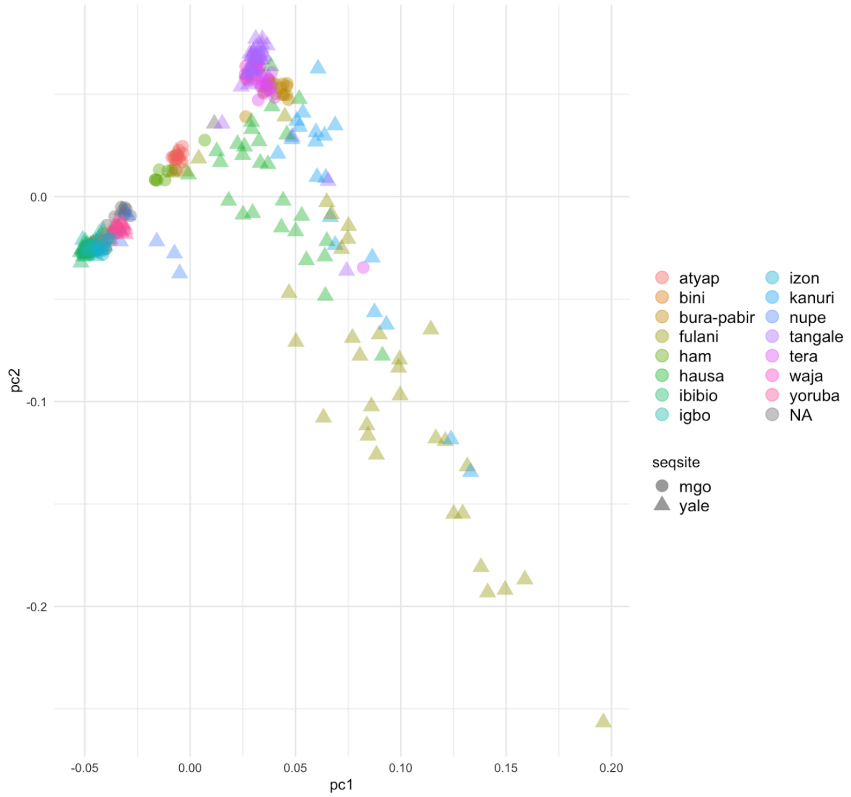


Figure S4: (A) Principal components plot of all ethnolinguistic groups (N=449). (B) Principal components plot of all ethnolinguistic groups listed in Table 1, in addition to 5 Esan 54gene-sequenced individuals, alongside 650 African-ancestry subjects from the New York Genome Center 1000 Genomes Project high-coverage dataset.

(A)



(B)

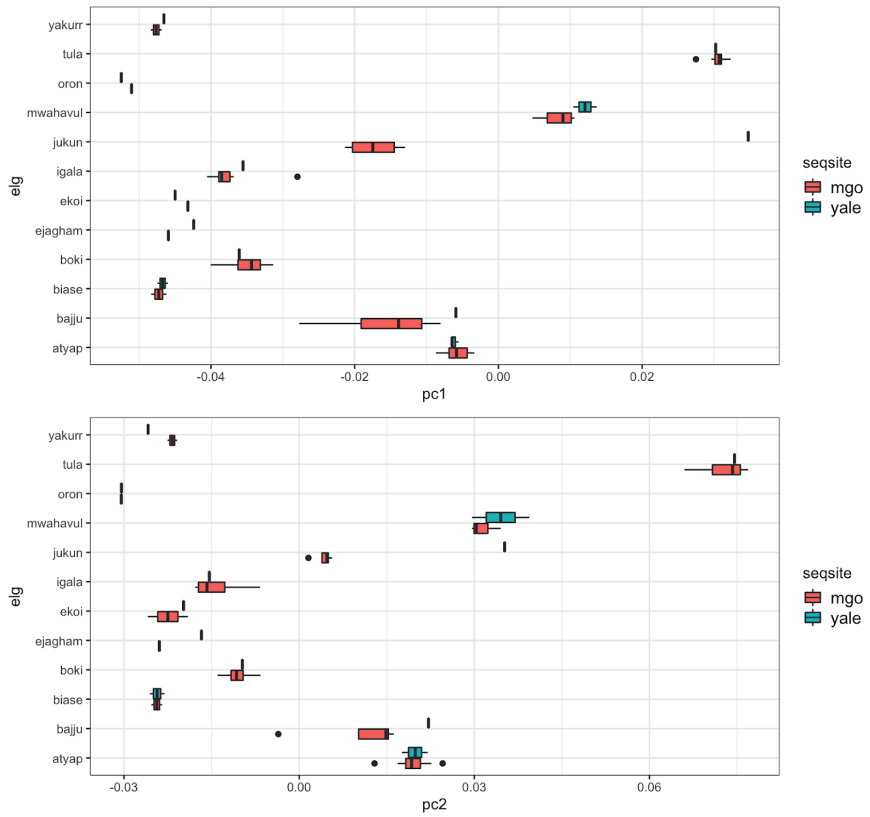


Figure S5. Principal components analysis of top 15 ethnolinguistic groups, annotated by sequencing center. (A) summary of PCs 1 and 2 by ethnolinguistic group and sequencing center for 8 groups that were sequenced in both centers, (B)

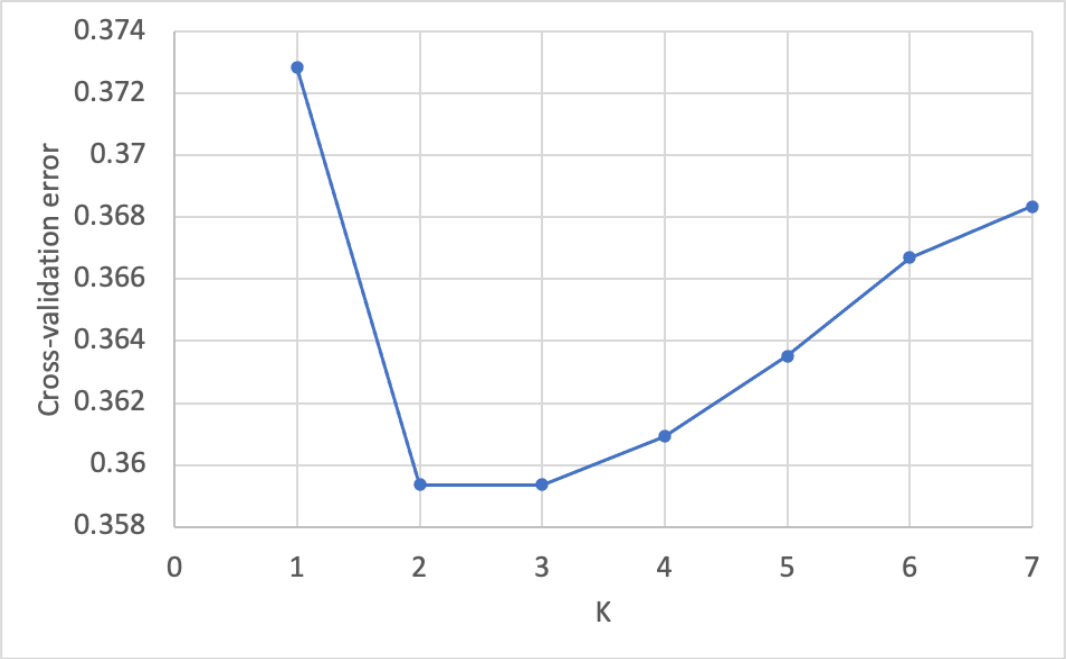


Figure S6: Cross-validation error from Admixture.

References

1. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
2. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
3. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
4. Tishkoff, S. A. *et al.* The genetic structure and history of Africans and African Americans. *Science* **324**, 1035–1044 (2009).
5. United Nations Department of Economic and Social Affairs, Population Division. World Population Prospects. (2022).
6. United States. Department of State. Bureau of Public Affairs. Nigeria. *Backgr. Notes Ser.* 1–8 (1987).
7. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
8. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
9. Fatumo, S. *et al.* Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. *Nat. Genet.* **54**, 531–536 (2022).
10. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
11. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
12. Vicente, M. *et al.* Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. *BMC Genomics* **20**, 915 (2019).

13. Cuadros-Espinoza, S., Laval, G., Quintana-Murci, L. & Patin, E. The genomic signatures of natural selection in admixed human populations. *Am. J. Hum. Genet.* **109**, 710–726 (2022).
14. Busby, G. B. *et al.* Admixture into and within sub-Saharan Africa. *Elife* **5**, (2016).
15. Francis, R. M. pophelper: an R package and web app to analyse and visualize population structure. *Mol. Ecol. Resour.* **17**, 27–32 (2017).
16. Miller, D. T. *et al.* ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 1381–1390 (2021).
17. Choudhury, A. *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
18. GenomeAsia100K Consortium. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
19. Caldwell, M. D. *et al.* CYP4F2 genetic variant alters required warfarin dose. *Blood* **111**, 4106–4112 (2008).
20. Adoga, M. P., Fatumo, S. A. & Agwale, S. M. H3Africa: a tipping point for a revolution in bioinformatics, genomics and health research in Africa. *Source Code Biol. Med.* **9**, 10 (2014).
21. Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
22. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
24. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
25. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
26. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic

Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

27. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* vol. 3 160025 (2016).
28. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, (2021).
29. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
30. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).