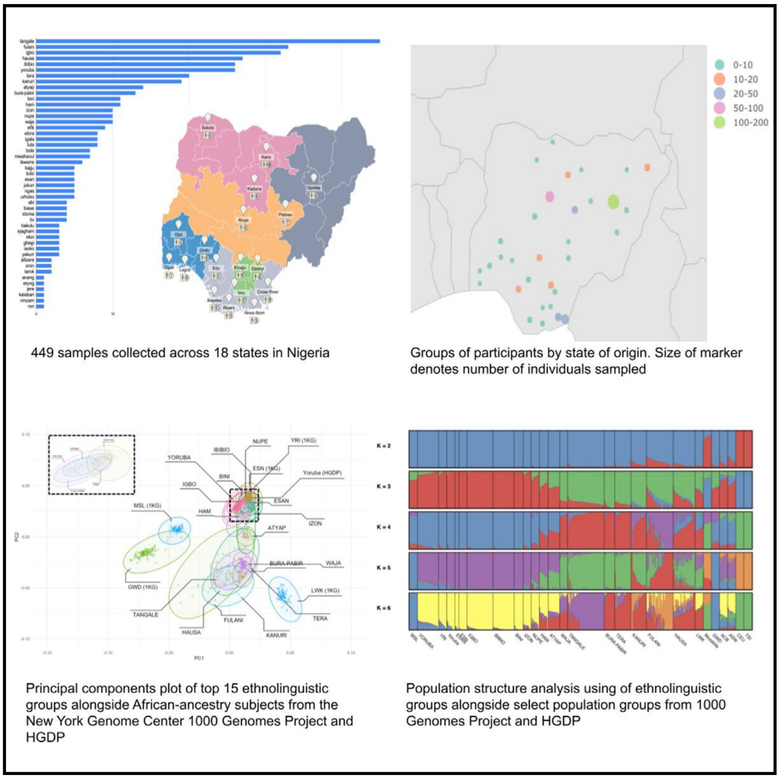
Cell Genomics

Resource

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

Graphical abstract



449 samples collected across 18 states in Nigeria

Groups of participants by state of origin. Size of marker denotes number of individuals sampled

Principal components plot of top 15 ethnolinguistic groups alongside African-ancestry subjects from the New York Genome Center 1000 Genomes Project and HGDP

Population structure analysis using of ethnolinguistic groups alongside select population groups from 1000 Genomes Project and HGDP

Highlights

- Whole-genome sequencing dataset from Nigeria with 449 individuals
- 47 unique self-reported ethnolinguistic groups from across Nigeria
- Fine-scale population structure and cataloguing of novel genetic variation to Nigeria
- Diverse cohort yields new insights into genomic variation in understudied populations

Authors

Esha Joshi, Arjun Biddanda,
Jumi Popoola, ..., Segun Fatumo,
Abasi Ene-Obong, Colm O'Dushlaine

Correspondence

eshrjoshi@gmail.com (E.J.),
codushlaine@gmail.com (C.O.)

In brief

Joshi et al. present a whole-genome sequencing dataset from 449 samples spanning 47 unique self-reported ethnolinguistic groups in Nigeria, briefly characterizing the genetic structure of the population and exploring novel and medically relevant variation across the groups. These findings emphasize the need for more inclusive cataloguing of human genetic variation through increased representation of African genomes.

Joshi et al., 2023, Cell Genomics *3*, 100378
September 13, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.xgen.2023.100378

Cell Genomics

CellPress
OPEN ACCESS

Resource

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] and Colm O'Dushlaine[1,7,*]

[1]54gene, Inc., 1100 15th St. NW, Washington, DC 20005, USA
[2]Nigerian Institute of Medical Research, Lagos 101245, Nigeria
[3]Center for Genomics Research and Innovation, National Agency for Biotechnology Development, Abuja 09004, Nigeria
[4]H3Africa Bioinformatics Network (H3ABioNet) Node, Centre for Genomics Research and Innovation, NABDA/FMST, Abuja 09004, Nigeria
[5]The African Computational Genomics (TAGC) Research Group, MRC/UVRI and London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
[6]College of Medicine University of Lagos, Lagos 101233, Nigeria
[7]Lead contact
*Correspondence: eshrjoshi@gmail.com (E.J.), codushlaine@gmail.com (C.O.)
https://doi.org/10.1016/j.xgen.2023.100378

SUMMARY

African populations have been drastically underrepresented in genomics research, and failure to capture the genetic diversity across the numerous ethnolinguistic groups (ELGs) found on 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 among our ELGs, consistent with previous findings. From the 36 million SNPs and insertions or deletions (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 to 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 have 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) such as type 2 diabetes, cancers, and cardiovascular disease, which are expected to be the leading cause of mortality in Africa within the next decade.[7] We established the Non-Communicable Diseases Genetic Heritage Study (NCD-GHS) consortium to assess the burden of NCDs, characterize their etiological characteristics, and catalog the human genetic variation in 100,000 adults in Nigeria.[8] We aim to contribute to prevention, treatment, and control strategies addressing NCDs through development of a resource that is comprehensive of purposeful sampling, deep phenotyping, and genomic studies centered around WGS/whole-exosome sequencing (WES) and genotyping with arrays. The NCD-GHS also aims to empower further genomics research initiatives in Africa through data sharing that

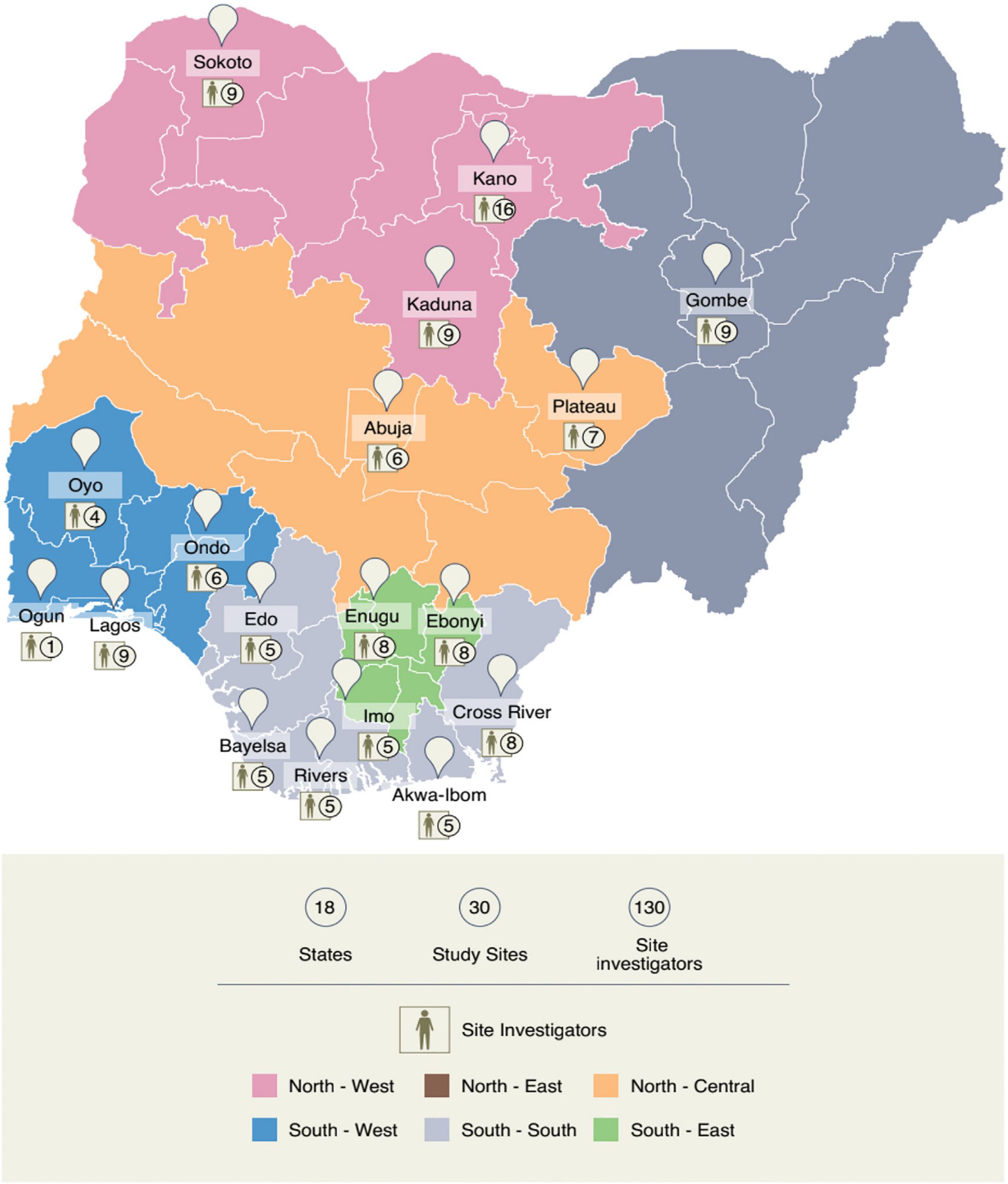


Figure 1. Overview of collection locations and regional designations within Nigeria
Additional details of the sample collection framework are discussed elsewhere.[8]

promotes scientific reproducibility but is conscious of ethical and legal standards.

In this first study, we performed germline WGS of an initial 449 samples from the NCD-GHS. Here, we describe the methods used to generate a WGS dataset of 449 Nigerian individuals spanning 47 self-reported ethnolinguistic groups (ELGs) generated using the GATK Best Practices workflow.[9] 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.[10] 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 with global populations.

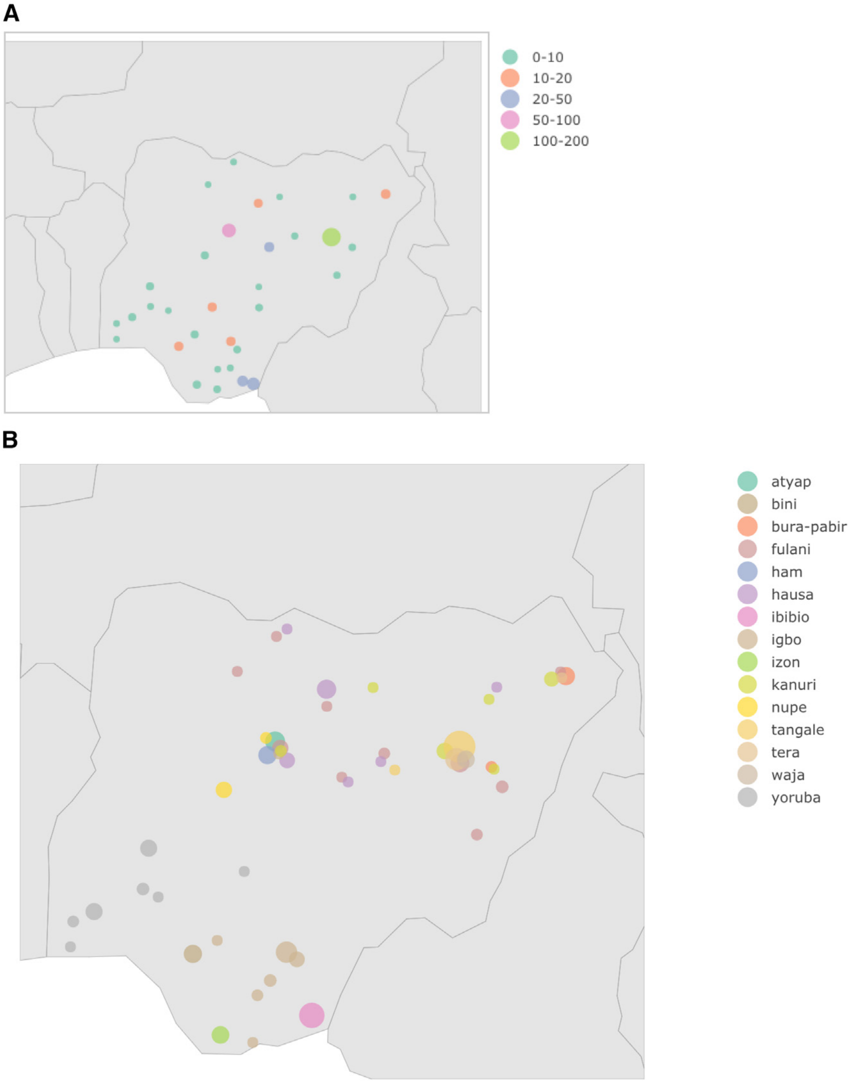


Figure 2. Collection sites in Nigeria where individuals of the 54gene dataset were sampled

(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.

(B) Reported ethnolinguistic group and state of origin for top 15 most prevalent groups. Marker size is in proportion to the number of individuals sampled.

Several recent large-scale genomic efforts focused on African populations have improved our understanding of the extensive genetic variation present on the continent and have expanded our knowledge of human demographic history.[11–13] Our work represents an effort to add more granularity to sample collections in Africa, specifically through 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 had 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 (Tables 1 and S7).

Given that current variant-calling approaches have been largely benchmarked using populations of European descent, we incorporated a non-Genome in a Bottle (GIAB) Yoruba sample (NA19238) as a control among our sequencing cohorts 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 precision/recall/F1 scores of 97.9%/91.4%/94.5% for SNPs and 79.6%/57.6%/66.9% for insertions or deletions (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 African (AFR) dataset, our results demonstrate that our variant-calling pipeline and post-processing filtering strategies are well suited for variant discovery in this dataset.

Patterns of variation across ELGs in Nigeria

We compared the properties of observed genetic diversity in the our dataset of 449 individuals (hereafter referred to as the "54gene dataset") with 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 the 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 also observed an increase in counts for unknown variation (variants not present in dbsnp154) across the 54gene

Table 1. Demographics and clinical characteristics of the 54gene dataset

	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

Samples were ascertained across 7 phenotypic groups and 47 unique self-reported ELGs (mean: 9, median: 5 samples per group, interquartile range [IQR]: 7.5). The top 15 groups are shown.

dataset with 3,748,259 unobserved SNPs and indels relative to the subset of the 1000 Genomes Project with 1,446,210. 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.

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/Pabir from Borno, Kanuri from Gombe, Tera from Gombe), who tend to have higher numbers of novel variants (Figures 1 and 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.[14,15] 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 4), which we hypothesize may play a role in this observation of a decrease in rare variation overall.[16] 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 an increased degree of population differentiation relative to other populations within the same continental grouping.[10]

Population structure across ELGs across Nigeria

We applied principal-component analysis (PCA) to investigate patterns of population structure across the ELGs in Nigeria. For example, we noted three distinct groups of genetically similar ELGs (Figures 5 and S3). 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 putatively North African ancestry.[16–18] We found specifically that the Hausa, Fulani, and Kanuri groups share a higher degree of genetic similarity with Mozabite-ancestry individuals, suggesting higher rates of North African ancestry within these populations from Northern Nigeria. 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 (Figures S4 and S5).

Admixture clustering had the lowest cross-validation error between K = 1 and K = 3 (Figure S6). We found similar patterns of ancestry between the Yoruba and Esan ELGs within our dataset and between the YRI (Yoruba in Ibadan, Nigeria) and ESN (Esan in Nigeria) populations from 1000 Genomes, respectively (Figure 4). Individuals reported as Yoruba, Esan, Igbo, Ibibio, Bini, and Izon showed evidence of similar ancestral composition (Figure 4). 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 North African or European groups (using Mozabite as a proxy for this ancestry and also incorporating European populations from the 1000 Genomes Project in the admixture analysis), corroborating results from PCA (Figure S3).

Table 2. Counts of variants in high-level classes of functional impact for 54gene and NYGC datasets

Dataset	Variant subset	Variant counts	T_i/T_v	Median variants per subject	IQR
Pre-filtering 54gene cohort (n = 543)	all	53,360,383	1.81	5,675,617	82,912.5
	SNPs	45,709,746	N/A	4,886,715	64,853.5
	indels	7,650,637	N/A	787,822	17,059.0
Post-filtering 54gene cohort (n = 451)	all	36,822,733	2.09	4,580,531	62,566.0
	SNPs	32,496,712	N/A	4,002,707	52,381.5
	indels	4,326,021	N/A	576,997	12,900.0
Pre-filtering NYGC AFR cohort (n = 661)	all	63,816,296	1.72	6,053,976	82,416.0
	SNPs	55,258,388	N/A	4,937,900	66,728.0
	indels	8,557,908	N/A	1,116,645	16,421.0
Post-filtering NYGC AFR cohort (n = 650)	all	43,845,824	2.08	4,628,977	71,259.3
	SNPs	40,451,499	N/A	3,978,564	56,881.0
	indels	3,394,325	N/A	650,738	14,111.3

T_i/T_v is defined as the ratio of transition (T_i) to transversion (T_v) SNPs with their interquartile range (IQR) provided.

Variation of clinical importance

To get a broad understanding of the relative frequencies of genetic variation that may be of clinical relevance to our cohort, we subsetted our dataset to annotated variants classified as "pathogenic" and having established evidence as being disease causing in the ClinVar Database. Additionally, we stratified variants by whether they belonged to genes from the American College of Medical Genetics and Genomics (ACMG)'s recommended list of 73 genes with reportable variants.[19] We identified a total of 134 variants classified as "pathogenic" in our cohort (Table S3). Fourteen individuals from our cohort carried at least one potential reportable ACMG variant, three carried a variant in *BRCA2* (associated with breast and ovarian cancers), four carried a variant in *BTD* (associated with biotinidase deficiency), and two carried a variant in *GAA* (associated with lysosome-associated glycogen storage disease) (Table S4).

Of the 134 variants identified as "pathogenic," eight were found to have a minor allele frequency (MAF) >5% in at least one of the self-reported ELGs in our cohort (Table S5). These eight variants were further compared to observed allele frequencies available for global populations and African population subsets in GnomAD[14] and the 1000 Genomes Project.[10] Similar to previous comparisons performed,[11] we observed several of these variants with disease associations to rare disorders with an MAF <5% across all populations in GnomAD and the 1000 Genomes Project. Larger sample sizes across these ELGs would be helpful to better understand differences in allele frequencies of these variants across multiple regions in Nigeria. These data could inform more precise classifications of "pathogenic" as well as "likely pathogenic" variants and could increase confidence when making disease associations across global populations. These results fall within a larger effort to re-examine alleles associated with rare diseases in more comprehensive population reference datasets.

Table 3. Counts of variants observed across all individuals by type

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

Impact categories are defined as indicated here: https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html. Counts are shown for both 54gene (449 subjects and 2 controls) and NYGC AFR datasets.

Allele frequencies of known variants associated with response to indicated drugs

Understanding how genetic variation impacts drug efficacy and safety across diverse population groups can improve individualized clinical utility of pharmacogenomic profiling. Variants in pharmacogenes such as *CYP2C9, CYP4F2*, and *VKORC1* have been implicated in the efficacy of warfarin, a commonly used anticoagulant for prevention of venous thrombosis, and have been included in pharmacogenomic screens to assess interindividual variability and dosing criteria for warfarin. Common variants in these genes have been found to differ in allele frequency between African- and European-ancestry individuals.[20] To assess the value of studying underrepresented ancestries in pharmacogenomics, we surveyed the frequencies of variants in key pharmacogenes across the ELGs from the 54gene dataset. We then compared the frequencies of variants in these key pharmacogenes across ELGs to selected ancestry groups from the 1000 Genomes Project (Table S6).[21]

Several polymorphisms within the *CYP4F2* gene encoding for the cytochrome P450 4F2 enzyme have been implicated in altered warfarin sensitivity and metabolism. We note elevated frequencies of pharmacogenomic variants within this gene 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, designated as *CYP4F2*2*, has a frequency of approximately 40% in the Fulani and Hausa ELGs but is closer to 30% frequency in the Yoruba.[22]

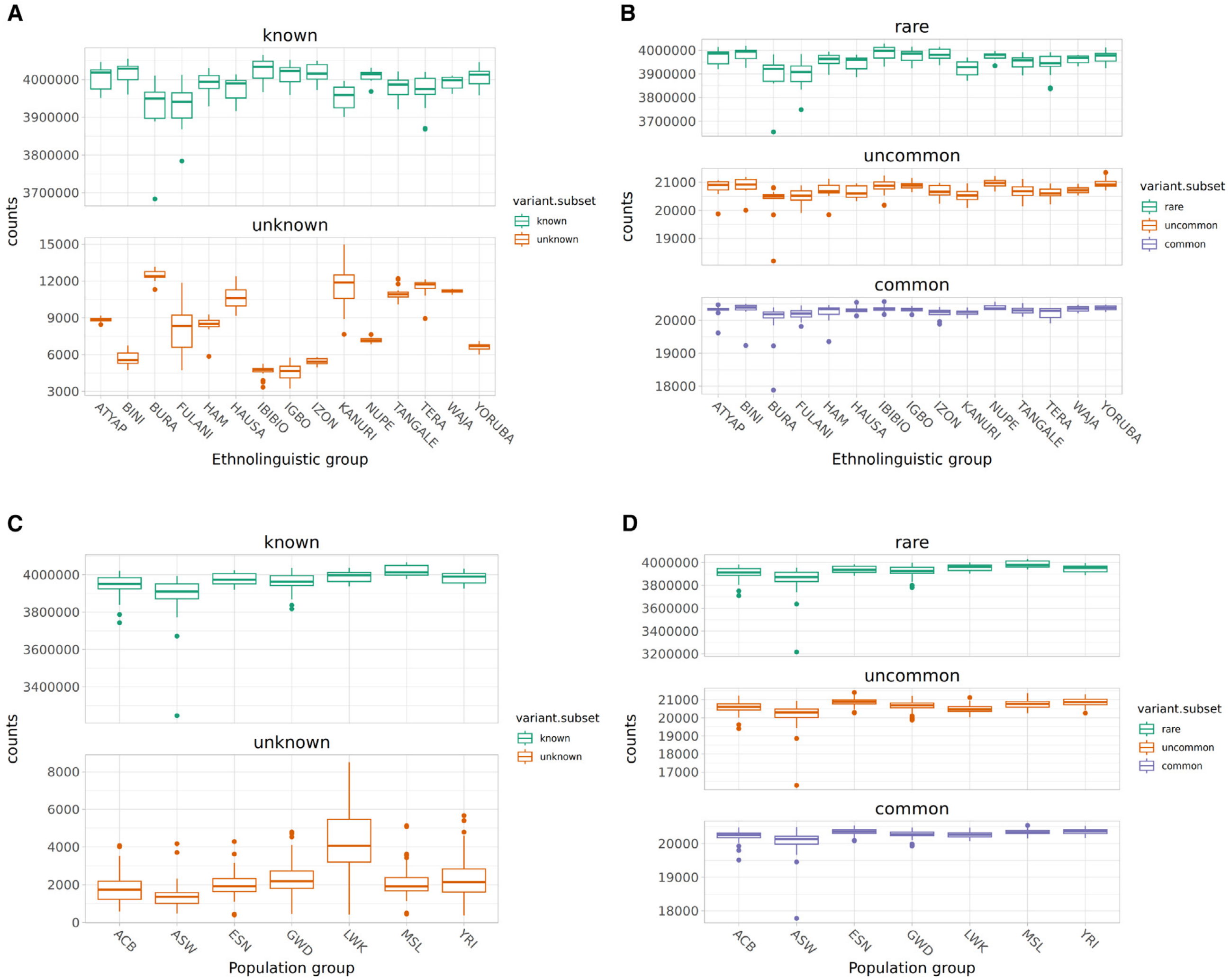


Figure 3. Variant counts across ELGs in the 54gene dataset and population groups in the NYGC AFR cohort

(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 < 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 are the same as in B).

We also observed an elevated frequency of rs2108622 (*CYP4F2*3*) in the Fulani and Hausa ELGs (15%–17% vs. ∼5% in YRI) (Table S6). This polymorphism has been described as reducing enzymatic levels of cytochrome P450 4F2 required for metabolism of vitamin K and is typically included in pharmacogenomic screens with evidence of association to warfarin response in European and Han-Chinese populations.[23–25] In African populations specifically, there have been little-to-no associations made between the *CYP4F2*3* allele and the warfarin dosage response because of the typically low frequency of this allele observed in the available, but limited, data in admixed and sub-Saharan African groups.[26,27] These findings highlight the necessity for added representation of allele frequencies from diverse ELGs, which can improve our understanding of how genetic variability contributes to drug efficacy and how population-specific data may be applied to improve the predictive power of dosing algorithms for commonly indicated drugs. 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 in medications.

DISCUSSION

This report represents an initial assessment using WGS to understand variation within, and the population structure of, some of the predominant ELGs in Nigeria. This resource also

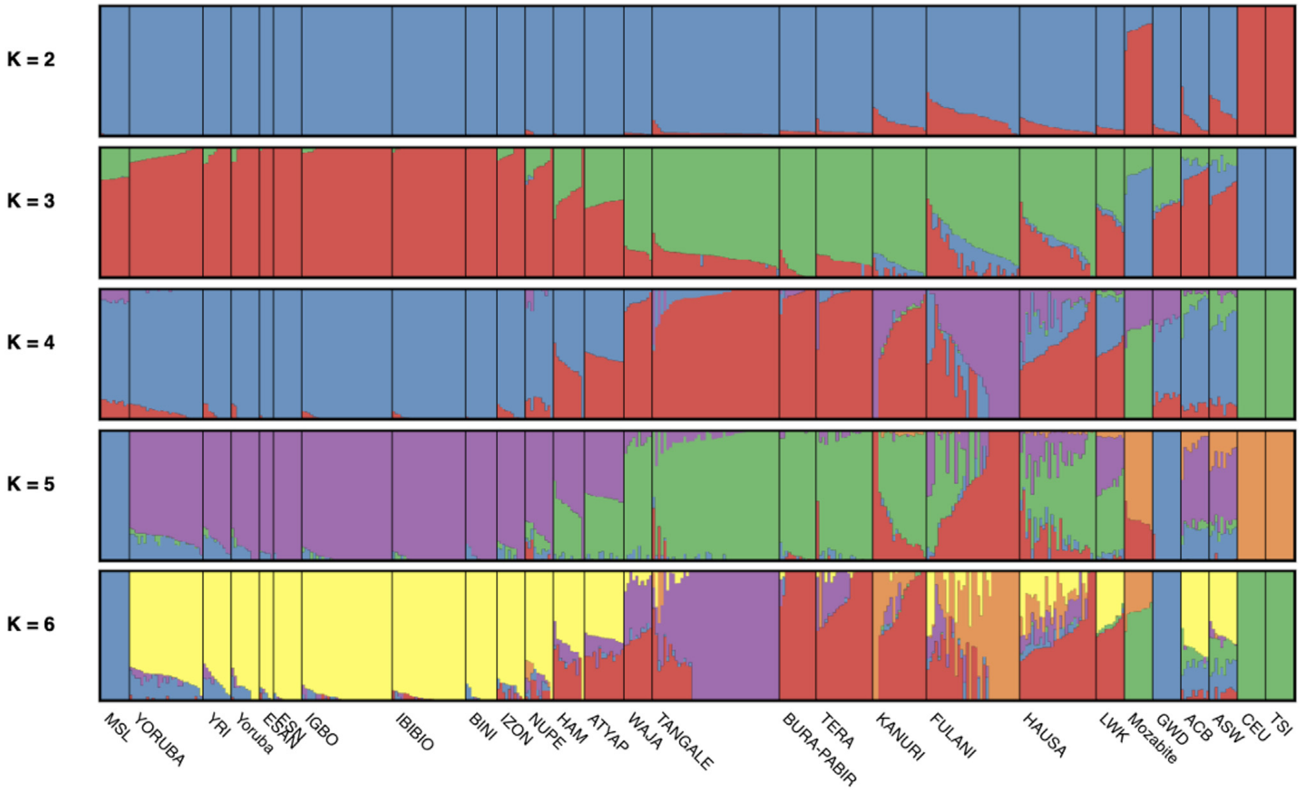


Figure 4. Population structure analysis using ADMIXTURE of ethnolinguistic groups listed in Table 1, alongside select populations from 1000 Genomes Project (10 random samples from African Caribbean in Barbados [ACB]; African Ancestry in Southwest USA [ASW]; Utah residents [CEPH] with Northern and Western European ancestry [CEU]; Esan in Nigeria [ESN]; Gambian in Western Division, The Gambia - Mandinka [GWD]; Luhya in Webuye, Kenya [LWK]; Mende in Sierra Leone [MSL]; Toscani in Italy [TSI]; Yoruba in Ibadan, Nigeria [YRI])

HDGP populations included were Yoruba in Nigeria (Yoruba) and Mozabite in Mzab, Algeria (Mozabite). Total sample size was n = 422.

demonstrates the capacity for conducting large-scale genome analyses in the region, speaking to the promise of building research capacity on the African continent.[11,28] We present results for several ELGs 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 among 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 the 1000 Genomes Project. We have added to this by sampling from a wider set of ELGs across Nigeria. By using the NA19238 African control in addition to the gold-standard NA12878 to perform benchmarking using field standard strategies,[29] 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 AFR sample data.

Using a broader representation of genetic diversity within Nigeria, we find several features of population structure within ELGs within Nigeria. We observe specific groups that are more genetically similar to one another within Nigeria (e.g., Yoruba, Igbo, and Izon). A specific and notable example of population structure is the gradient of North African-related ancestry (approximated using Mozabite individuals) across multiple groups in Nigeria. Previous literature has shown high North African-related ancestry in Fulani individuals, but our analysis here considers this across a much wider range of groups within Nigeria.[16] For example, we find elevation of this ancestry within the Hausa and Kanuri groups from Northern Nigeria as well. A finer-scale resolution of population structure could benefit from more detailed sampling with respect to the ELG of an individual's grandparents to highlight. We anticipate that further studies within these groups may shed light on potential trait-associated variants at higher frequencies in specific ELGs relative to the entire Nigerian population (e.g., elevated in Hausa), where each ELG consists of a sizable number of people, highlighting the importance of understanding fine-scale population structure within this region.

In order to derive tangible benefits from genomics research for global populations, making the resulting genomic data and metadata available is essential. However, the accessibility and availability of genomics data remains a persistent challenge for the field.[30] There are notable exceptions to this. The public availability of data from the 1000 Genomes Project, HGDP, and the UK Biobank—to name a few—has removed major barriers to conducting human genetics research, particularly for researchers with limited funding.[10,31,32] There are additional efforts for which a subset of the data are public (e.g., TOPMed,

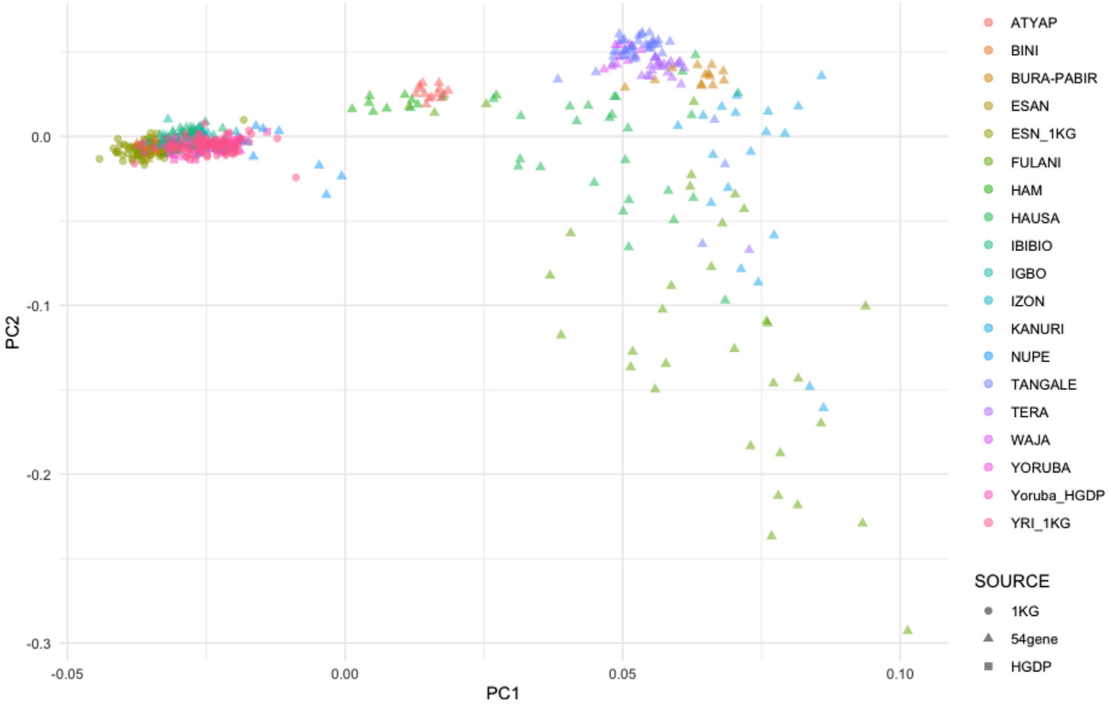


Figure 5. Principal-component plot of ethnolinguistic groups listed in Table 1 in addition to Esan from 54gene and 1000 Genomes Project and Yoruba from 1000 Genomes Project and HGDP

An additional version of this plot with all ethnolinguistic groups is shown in Figures S3 and S4.

imputation server but limited direct access to phased data) and others that, though publicly funded, remain difficult to access (e.g., H3Africa whole-genome sequences). We note that the data presented here were not funded by major public grants or other non-profit support, unlike some of the datasets highlighted above. While the data are not completely publicly available, and some level of access control is enforced, we are hopeful that this is a step in the right direction, where both public and private initiatives make every effort to release and share data with the broader research community.

While there is a critical need to facilitate open-access sharing of high-quality genomic data, there is also a need to balance the interests of the researchers generating the data and the ethical and privacy obligations to the participants. Specifically, ensuring the data are used for non-commercial purposes and that the data producers fully benefit from their contributions in the form of formal credit and/or acknowledgment drives progress and capacity building in genomics research in regions such as Nigeria. Ethical use of genomic data requires that there are safeguards for protecting patient privacy, confidentiality, and prevention of data misuse or unauthorized access. Implementing controlled and/or restricted access to genomic data with robust but transparent governance mechanisms allows researchers to find a balance between these challenges. Repositories such as the European Genome-phenome Archive (EGA) and dbGAP can facilitate secure and structured methods of data sharing. While this framework may create barriers in the form of application procedures, documentation, and longer turn-around times from assessment committees, it remains the best current solution to address security concerns. However, the burden of enabling data sharing highlights a larger need to re-evaluate international guidelines and best practices in genomics for effective data sharing to maximize scientific discoveries and health equity.

This resource provides an approach for conducting further population genomic studies in Nigeria using WGS with larger sample sizes to provide more definitive insights into novel or rare variation in certain ELGs and to provide a high-level summary of population structure. Our results also emphasize the utility of publicly available WGS 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.

Limitations of the study

The sample sizes across the self-reported ELGs in our cohort and their depths of coverage limit the interpretations that can be made from the discovery of clinically relevant variants and potential conclusions that can be made about the distribution of pathogenic disease-associated variants in Nigeria. This also limits our ability to make conclusions on the relative frequencies of novel or known pharmacogenetic variants that exist within the population. Nevertheless, our findings of the relative counts of ACMG-reportable variants and broad comparisons of pathogenic and pharmacogene variant frequency can serve as a template for cataloging variation at the level of ELGs. An additional limitation of the currently generated data is that the lower depth of coverage limits our ability to draw demographic insights from patterns of rare-variant sharing across ELGs. Data of higher depth and quality and increased sample sizes across lesser-represented ELGs will allow for more robust conclusions about complex genomic regions and mutations that could have significant impacts on health or disease outcomes. As more complete demographic and health data emerge for these understudied population groups, we foresee significant opportunities for health interventions that will improve the health and well-being of patients, particularly in areas such as pharmacogenomics.

CONSORTIA

The members of the 54gene team consortium are 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, 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, and Teresia L. Bost.

The members of the NCD-GHS Consortium are Segun Fatumo, Aminu Yakubu, Abdullahi Musa, Abdulrasheed M. Mujtaba, Abiodun Popoola, Abubakar M. Bello, Anthony Anyanwu, Ashiru Yusuf, Gesiye E.L. 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, Kenneth Chima Nwankwo, Marcus Inyama Asuquo, Timothy Ekwere, Ezechukwu Aniekwensi, Chidi Ezeude, Olayemi Awopeju, Tolutope Kolawole, Olubiyi Adesina, Vandi Ghyi, Olaolu Oni, Zumnan Gimba, and Abasi Ene-Obong.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANTS
- METHOD DETAILS
  - ○ Sample processing and whole Genome sequencing
  - ○ FASTQ generation
  - ○ Variant calling and QC
  - ○ Post-calling variant filtering
  - ○ Analysis of population structure
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100378.

ACKNOWLEDGMENTS

We thank the members of NCD-GHS Scientific Advisory Board, 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, and Nnamdi Ilodiuba. Funding for this study was provided by the African Centre for Translational Genomics (ACTG) with support from 54gene, Inc. Above all, we thank the participants for their support, participation, and engagement as we seek to promote genomics at scale on the continent of Africa.

AUTHOR CONTRIBUTIONS

Conceptualization, E.J., A.B., J.P., S.F., and C.O'D.; software, E.J. and A.B.; formal analysis, E.J., A.B., 54gene team, and C.O'D.; investigation and project administration, J.P., A.Y., O. Osakwe, D.A., 54gene team, O. Oyedele, and G.E.-E.; data curation, E.J., A.B., J.P., and O. Oyedele; writing – original draft, E.J., A.B., and C.O'D.; writing – review & editing, E.J., A.B., B.S., O.N., O.S., S.F., A.E.-O., and C.O'D.; supervision, J.P., D.A., NCD-GHS Consortium, A.E.-O., and C.O'D.

DECLARATION OF INTERESTS

E.J., A.B., J.P., A.Y., O.O., D.A., E.D., O.O., G.E.-E., A.E.-O., and C.O'D. were employed by 54gene, Inc., at the time this research was conducted. Funding for this study was provided by 54gene, Inc. C.O'D. is currently employed at insitro, San Francisco, CA 94080, USA. insitro had no involvement in the design or implementation of the work presented here. E.D., J.P., A.Y., and A.E.-O. are current employees of Syndicate Bio.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research. We worked to ensure gender balance in the recruitment of human subjects. We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. We worked to ensure that the study questionnaires were prepared in an inclusive way. One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a gender minority in their field of research. We avoided "helicopter science" practices by including the participating local contributors from the region where we conducted the research as authors on the paper.

Received: January 12, 2023
Revised: February 23, 2023
Accepted: July 13, 2023
Published: August 8, 2023

REFERENCES

1. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.
2. International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789–796. https://doi.org/10.1038/nature02168.
3. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299. https://doi.org/10.1038/s41586-021-03205-y.
4. 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. https://doi.org/10.1126/science.1172257.
5. United Nations Department of Economic and Social Affairs, Population Division (2022). World Population Prospects.
6. United States; Department of State; Bureau of Public Affairs (1987). Nigeria. Backgr. Notes Ser. *1–8*.
7. Gouda, H.N., Charlson, F., Sorsdahl, K., Ahmadzada, S., Ferrari, A.J., Erskine, H., Leung, J., Santamauro, D., Lund, C., Aminde, L.N., et al. (2019). Burden of non-communicable diseases in sub-Saharan Africa, 1990-2017: results from the Global Burden of Disease Study 2017. Lancet. Glob. Health *7*, e1375–e1387. https://doi.org/10.1016/S2214-109X(19)30374-2.
8. Fatumo, S., Yakubu, A., Oyedele, O., Popoola, J., Attipoe, D.A., Eze-Echesi, G., Modibbo, F.Z., Ado-Wanka, N., et al.; 54gene Team; NCD-GHS Consortium (2022). Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. Nat. Genet. *54*, 531–536. https://doi.org/10.1038/s41588-022-01071-6.
9. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498. https://doi.org/10.1038/ng.806.
10. 1000 Genomes Project Consortium; 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., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
11. Choudhury, A., Aron, S., Botigué, L.R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y.J., et al. (2020). High-depth African genomes inform human migration and health. Nature *586*, 741–748. https://doi.org/10.1038/s41586-020-2859-7.
12. Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C.S., Prado-Martinez, J., Bouman, H., Abascal, F., Haber, M., Tachmazidou, I., et al. (2019). Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. Cell *179*, 984–1002.e36. https://doi.org/10.1016/j.cell.2019.10.004.
13. 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. https://doi.org/10.1038/nature13997.
14. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.
15. Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104. https://doi.org/10.1126/science.1217876.
16. Vicente, M., Priehodová, E., Diallo, I., Podgorná, E., Poloni, E.S., Černý, V., and Schlebusch, C.M. (2019). Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. BMC Genom. *20*, 915. https://doi.org/10.1186/s12864-019-6296-7.
17. Cuadros-Espinoza, S., Laval, G., Quintana-Murci, L., and Patin, E. (2022). The genomic signatures of natural selection in admixed human populations. Am. J. Hum. Genet. *109*, 710–726. https://doi.org/10.1016/j.ajhg.2022.02.011.
18. 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. (2016). Admixture into and within sub-Saharan Africa. Elife *5*, e15266. https://doi.org/10.7554/eLife.15266.
19. Miller, D.T., Lee, K., Chung, W.K., Gordon, A.S., Herman, G.E., Klein, T.E., Stewart, D.R., Amendola, L.M., Adelman, K., Bale, S.J., et al. (2021). 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. https://doi.org/10.1038/s41436-021-01172-3.
20. Flockhart, D.A., O'Kane, D., Williams, M.S., Watson, M.S., Flockhart, D.A., Gage, B., Gandolfi, R., King, R., Lyon, E., Nussbaum, R., et al. (2008). Pharmacogenetic testing of CYP2C9 and VKORC1 alleles for warfarin. Genet. Med. *10*, 139–150. https://doi.org/10.1097/GIM.0b013e318163c35f.
21. GenomeAsia100K Consortium (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature *576*, 106–111. https://doi.org/10.1038/s41586-019-1793-z.
22. Caldwell, M.D., Awad, T., Johnson, J.A., Gage, B.F., Falkowski, M., Gardina, P., Hubbard, J., Turpaz, Y., Langaee, T.Y., Eby, C., et al. (2008). CYP4F2 genetic variant alters required warfarin dose. Blood *111*, 4106–4112. https://doi.org/10.1182/blood-2007-11-122010.
23. Alvarellos, M.L., Sangkuhl, K., Daneshjou, R., Whirl-Carrillo, M., Altman, R.B., and Klein, T.E. (2015). PharmGKB summary: very important pharmacogene information for CYP4F2. Pharmacogenet. Genomics *25*, 41–47. https://doi.org/10.1097/FPC.0000000000000100.
24. Zhang, J.E., Klein, K., Jorgensen, A.L., Francis, B., Alfirevic, A., Bourgeois, S., Deloukas, P., Zanger, U.M., and Pirmohamed, M. (2017). Effect of Genetic Variability in the CYP4F2, CYP4F11, and CYP4F12 Genes on Liver mRNA Levels and Warfarin Response. Front. Pharmacol. *8*, 323. https://doi.org/10.3389/fphar.2017.00323.
25. Singh, O., Sandanaraj, E., Subramanian, K., Lee, L.H., and Chowbay, B. (2011). Influence of CYP4F2 rs2108622 (V433M) on warfarin dose requirement in Asian patients. Drug Metab. Pharmacokinet. *26*, 130–136. https://doi.org/10.2133/dmpk.dmpk-10-rg-080.
26. Cavallari, L.H., Langaee, T.Y., Momary, K.M., Shapiro, N.L., Nutescu, E.A., Coty, W.A., Viana, M.A.G., Patel, S.R., and Johnson, J.A. (2010). Genetic and clinical predictors of warfarin dose requirements in African Americans. Clin. Pharmacol. Ther. *87*, 459–464. https://doi.org/10.1038/clpt.2009.223.
27. Perini, J.A., Struchiner, C.J., Silva-Assunção, E., and Suarez-Kurtz, G. (2010). Impact of CYP4F2 rs2108622 on the stable warfarin dose in an admixed patient cohort. Clin. Pharmacol. Ther. *87*, 417–420. https://doi.org/10.1038/clpt.2009.307.
28. Adoga, M.P., Fatumo, S.A., and Agwale, S.M. (2014). H3Africa: a tipping point for a revolution in bioinformatics, genomics and health research in Africa. Source Code Biol. Med. *9*, 10. https://doi.org/10.1186/1751-0473-9-10.

29. Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., De La Vega, F.M., Moore, B.L., Gonzalez-Porta, M., Eberle, M.A., Tezak, Z., Lababidi, S., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. Nat. Biotechnol. *37*, 555–560. https://doi.org/10.1038/s41587-019-0054-x.

30. Ramsay, M. (2022). African genomic data sharing and the struggle for equitable benefit. Patterns (N Y) *3*, 100412. https://doi.org/10.1016/j.patter.2021.100412.

31. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012. https://doi.org/10.1126/science.aay5012.

32. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.

33. Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., et al. (2021). A complete reference genome improves analysis of human genetic variation. Preprint at bioRxiv. https://doi.org/10.1101/2021.07.12.452063.

34. Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., et al. (2015). The European Genome-phenome Archive of human data consented for biomedical research. Nat. Genet. *47*, 692–695. https://doi.org/10.1038/ng.3312.

35. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data.

36. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

37. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arXiv.1303.3997.

38. 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. https://doi.org/10.1101/gr.107524.110.

39. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at bioRxiv. https://doi.org/10.1101/201178.

40. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

41. Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doheny, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. Am. J. Hum. Genet. *91*, 839–848. https://doi.org/10.1016/j.ajhg.2012.09.004.

42. Pedersen, B.S., Bhetariya, P.J., Brown, J., Kravitz, S.N., Marth, G., Jensen, R.L., Bronner, M.P., Underhill, H.R., and Quinlan, A.R. (2020). Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. Genome Med. *12*, 62. https://doi.org/10.1186/s13073-020-00761-2.

43. 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. https://doi.org/10.1186/s13059-016-0974-4.

44. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354. https://doi.org/10.1038/s41598-019-45839-z.

45. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067. https://doi.org/10.1093/nar/gkx1153.

46. Sherry, S.T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. *9*, 677–679.

47. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data *3*, 160025. https://doi.org/10.1038/sdata.2016.25.

48. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*, eabf7117. https://doi.org/10.1126/science.abf7117.

49. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., 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. https://doi.org/10.1086/519795.

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

51. Behr, A.A., Liu, K.Z., Liu-Fang, G., Nakka, P., and Ramachandran, S. (2016). pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics *32*, 2817–2823. https://doi.org/10.1093/bioinformatics/btw327.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw and analyzed data	This paper	EGA: EGAS00001007036
NYGC 1000 Genomes Project WGS dataset	Byrska-Bishop et al.[1]	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: known sites VCFs and scattered calling intervals	BROAD Institute	https://s3.amazonaws.com/broad-references/broad-references-readme.html
Human reference genome NCBI build 38, GRCh38	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Software and algorithms		
Illumina bcl2fastq2	Illumina	https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html
54gene-wgs-germline	This paper	https://54gene-wgs-germline.readthedocs.io/en/latest/
FastQC	Andrew et al.[35]	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Fastp	Chen et al.[36]	https://github.com/OpenGene/fastp
BWA	Li et al.[37]	N/A
GATK	McKenna et al.[38] Poplin et al.[39]	http://www.genome.org/cgi/doi/10.1101/gr.107524.110 https://www.biorxiv.org/content/10.1101/201178v3
Samtools and bcftools	Danecek et al.[40]	https://github.com/samtools/samtools
VerifyBamID	Jun et al.[41]	https://github.com/statgen/verifyBamID/
Somalier	Pederson et al.[42]	https://github.com/brentp/somalier
Ensembl Variant Effect Predictor (VEP)	McLaren et al.[43]	https://useast.ensembl.org/info/docs/tools/vep/script/index.html
hap.py	Krusche et al.[29]	https://github.com/Illumina/hap.py
PLINK	Purcell et al.[49]	https://www.cog-genomics.org/plink/
ADMIXTURE	Alexander et al.[50]	https://dalexander.github.io/admixture/
Code for figure generation	This paper	https://gitlab.com/data-analysis5/wgs_449_figure_generation
Pong	Behr et al.[51]	https://github.com/ramachandran-lab/pong
Other		
ENCODE Blacklist regions	Amemiya et al.[44]	https://github.com/Boyle-Lab/Blacklist/tree/master/lists
ClinVar Database of clinically relevant variation	Landrum et al.[45]	https://www.ncbi.nlm.nih.gov/clinvar/
ACMG v3.0 List of genes from secondary findings	Miller et al.[19]	N/A
dbSNP Database for SNPs	Sherry et al.[46]	https://www.ncbi.nlm.nih.gov/snp/
Genome-in-a-bottle: genome stratification regions	Krusche et al.[29]	https://github.com/genome-in-a-bottle/genome-stratifications
NIST high-confidence regions	Krusche et al.[29] Zook et al.[47]	https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and data should be directed to and will be fulfilled by the lead contact, Colm O'Dushlaine (codushlaine@gmail.com), or by Segun Fatumo (segun.fatumo@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The whole-genome raw sequence data reported in this study cannot be deposited in a public repository because of ethical and subject/patient privacy restrictions. Sequence data in the form of CRAMs and aggregate, per-chromosome VCF files have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under the study accession number EGAS00001007036. They are available through controlled access. Further information about EGA can be found on https://ega-archive.org.[34] Access can be requested by contacting the NCD-GHS consortium through the lead contact, Colm O'Dushlaine (codushlaine@gmail.com) or Segun Fatumo (segunfatumo@gmail.com), and providing information about the intended non-commercial use of the requested data.
- All original code is available in this paper's supplemental information.
- Any additional information required to reanalyze the data reported in this paper can be made available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANTS

Individuals for this study were recruited from numerous study sites across Nigeria (Figure 2).[8] 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.

METHOD DETAILS

Sample processing 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 (v2.20) 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 key resources table) was used to process the raw sequencing data. FastQC (v0.11.9) reports were generated for all raw FASTQs,[35] followed by read trimming and adapter removal with fastp.[36] 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,[37] followed by deduplication. GATK (v4.2.5.0)[38,39] BaseRecalibrator was used to generate a recalibration table for base quality scores using the VCFs for known SNP and INDELs sites in dbSNP (build 138) from the Broad's genome references on Amazon Web Services, then applied to each BAM file. Samtools stats (v1.15) were then generated for all BAMs.[40] 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 . 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'.[40] 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 (v2.0.1)[41]; subject relatedness was estimated using Somalier (0.2.14) in order to identify unexpected genetic duplicates[42]; sex discordances were detected using two orthogonal techniques, Somalier and the ‘bcftools guess-ploidy‘ plugin (v1.10).[40,42] 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.

Variant annotation

We used Ensembl’s Variant Effect Predictor (VEP)[43] 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.[44] Annotations for variant classifications of clinical importance using the ClinVar Database (v.20221113),[45] ACMG v3.0 list of genes from secondary findings, and variant labels (rs IDs) from dbSNP (v.154) were also included.[19,46]

Post-calling variant filtering

We explored different variant filtering strategies to apply to our dataset and to evaluate sensitivity and specificity (Table S2, Supplemental Notes). Several of these filtering strategies included masking for the aforementioned regions that are prone to yield missing or unreliable data, applied as custom annotations that encompassed simple repeats, centromeric and telomeric regions, segmental duplications, microsatellites, and custom regions defined in the ENCODE blacklist. 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[47]) and NA19238, a Yoruba reference sample from the 1000 Genomes Project (benchmarked against a publicly available HiFi call set).[48] 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.

To assess performance of variant calling, we used BED files provided by the Genome in a Bottle Consortium and T2T Consortium to stratify true positive, false positive, and false negative variant calls over difficult regions of the genome, 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).[29,33] We used hap.py (v0.3.15) to assess performance and applied various variant and region filtering strategies; see Supplemental Notes for details on all filters applied.[29]

Preparing publicly available data

We acquired publicly available data produced by the New York Genome Center (NYGC).[1] We subsetted the data to 661 subjects from the populations in the African-ancestry regional grouping (ACB, ASW, ESN, GWD, LWK, MSL, and YRI) (Table 2) 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. Similarly, we also acquired publicly available data produced by the Human Genome Diversity Project (HGDP),[31] subsetting the data to 420 subjects from population groups spanning Africa, Europe and the Middle-East; (Adygei, Bantu Kenya, Bantu South Africa, Basque, Bedouin, Bergamo Italian, Biaka, Druze, French, Mandenka, Mbuti, Mozabite, Orcadian, Palestinian, Russian, San, Sardinian, Tuscan, Yoruba).

Analysis of population structure

We merged the subsetted VCFs containing population groups of interest from the NYGC (n=650) and HGDP (n=420) datasets, with our 54gene dataset (n=449) (Table 2). We used BCFtools[40] (v1.10) with the ‘-m none‘ to output no new multiallelics, but multiple records instead, and the ‘–force-samples‘ parameter. Using PLINK[49] (v1.9), a starting call set of 53,649,772 variants were filtered to 0.5% minor allele frequency within the cohort (18,721,257 variants), 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 5,726,648 variants 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 611,322 variants for population genetic analyses. Principal components were estimated using the default settings of ‘plink –pca‘.[49] We used ADMIXTURE[50] (v1.3.0) to characterize the population structure of 54gene samples, alongside African and European samples from the 1000 Genomes Project.[1] We applied the clustering approach in ADMIXTURE across a range of cluster counts (K), from K=1 to K=10. The admixture plots were generated using Pong.[51]

QUANTIFICATION AND STATISTICAL ANALYSIS

Included in Method Details.