Characterization of *CYP2B6* and *CYP2A6* Pharmacogenetic Variation in Sub-Saharan African Populations

David Twesigomwe[1,2,*], Britt I. Drögemöller[3], Galen E. B. Wright[4,5], Clement Adebamowo[6,7], Godfred Agongo[8,9], Palwendé R. Boua[1,10], Mogomotsi Matshaba[11,12], Maria Paximadis[13,14], Michèle Ramsay[1,2], Gustave Simo[15], Martin C. Simuunza[16], Caroline T. Tiemessen[13], Zané Lombard[2] and Scott Hazelhurst[1,17,*]

Genetic variation in *CYP2B6* and *CYP2A6* is known to impact interindividual response to antiretrovirals, nicotine, and bupropion, among other drugs. However, the full catalogue of clinically relevant pharmacogenetic variants in these genes is yet to be established, especially across African populations. This study therefore aimed to characterize the star allele (haplotype) distribution in *CYP2B6* and *CYP2A6* across diverse and understudied sub-Saharan African (SSA) populations. We called star alleles from 961 high-depth full genomes using StellarPGx, Aldy, and PyPGx. In addition, we performed *CYP2B6* and *CYP2A6* star allele frequency comparisons between SSA and other global biogeographical groups represented in the new 1000 Genomes Project high-coverage dataset ($n = 2{,}000$). This study presents frequency information for star alleles in *CYP2B6* (e.g., **6* and **18*; frequency of 21–47% and 2–19%, respectively) and *CYP2A6* (e.g., **4*, **9*, and **17*; frequency of 0–6%, 3–10%, and 6–20%, respectively), and predicted phenotypes (for CYP2B6), across various African populations. In addition, 50 potentially novel African-ancestry star alleles were computationally predicted by StellarPGx in *CYP2B6* and *CYP2A6* combined. For each of these genes, over 4% of the study participants had predicted novel star alleles. Three novel star alleles in *CYP2A6* (54*, **55*, and **56*) and *CYP2B6* apiece, and several suballeles were further validated via targeted Single-Molecule Real-Time resequencing. Our findings are important for informing the design of comprehensive pharmacogenetic testing platforms, and are highly relevant for personalized medicine strategies, especially relating to antiretroviral medication and smoking cessation treatment in Africa and the African diaspora. More broadly, this study highlights the importance of sampling diverse African ethnolinguistic groups for accurate characterization of the pharmacogene variation landscape across the continent.**

[1]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [2]Division of Human Genetics, National Health Laboratory Service, and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [3]Department of Biochemistry and Medical Genetics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada; [4]Neuroscience Research Program, Kleysen Institute for Advanced Medicine, Winnipeg Health Sciences Centre and Max Rady College of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada; [5]Department of Pharmacology and Therapeutics, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Manitoba, Canada; [6]Institute for Human Virology, Abuja, Nigeria; [7]Division of Cancer Epidemiology, Department of Epidemiology and Public Health, and the Marlene and Stewart Greenebaum Comprehensive Cancer Centre, University of Maryland School of Medicine, Baltimore, Maryland, USA; [8]Navrongo Health Research Centre, Ghana Health Service, Navrongo, Ghana; [9]Department of Biochemistry and Forensic Sciences, School of Chemical and Biochemical Sciences, C.K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana; [10]Clinical Research Unit of Nanoro, Institut de Recherche en Sciences de la Santé, Nanoro, Burkina Faso; [11]Botswana-Baylor Children's Clinical Centre of Excellence, Gaborone, Botswana; [12]Retrovirology, Department of Pediatrics, Baylor College of Medicine, Houston, Texas, USA; [13]Centre for HIV and STIs, National Institute for Communicable Diseases, National Health Laboratory Services and Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; [14]School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, South Africa; [15]Molecular Parasitology and Entomology Unit, Department of Biochemistry, Faculty of Science, University of Dschang, Dschang, Cameroon; [16]Department of Disease Control, School of Veterinary Medicine, University of Zambia, Lusaka, Zambia; [17]School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa. *Correspondence: David Twesigomwe (david.twesigomwe@wits.ac.za) and Scott Hazelhurst (scott.hazelhurst@wits.ac.za)

Received May 23, 2023; accepted November 16, 2023. doi:10.1002/cpt.3124

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ *CYP2B6* and *CYP2A6* genetic variation contributes to clinically relevant differences in response to antiretrovirals, nicotine, and bupropion (among other drugs) across individuals and populations. *CYP2B6* and *CYP2A6* are therefore important genes in clinical pharmacogenetic implementation initiatives globally. Current catalogues of *CYP2B6* and *CYP2A6* star alleles (haplotypes) are incomplete in part due to the high polymorphism in these genes and difficulty in interrogating their genomic loci.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ To date, the proportion of individuals with African ancestry has been relatively low across pharmacogenetic studies focused on *CYP2B6*, *CYP2A6*, and other key pharmacogenes. In particular, continental African populations have been underrepresented. This study addresses the paucity of information about the distribution of known star alleles in *CYP2B6* and *CYP2A6* across diverse African populations, and also highlights novel African-ancestry star alleles, with a view toward enabling relevant precision medicine strategies in Africa.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ This study highlights the varying distributions of known and novel *CYP2B6* and *CYP2A6* star alleles, and predicted metabolizer phenotypes (for CYP2B6) across previously under-represented African populations, and in comparison with global populations. For each of the two pharmacogenes, over 4% of the sub-Saharan African participants had predicted novel star alleles. Predicted novel *CYP2B6* and *CYP2A6* star alleles from our comparative analysis involving other global biogeographical groups are also provided. Furthermore, this study exemplifies the utility of high coverage whole genome sequence data and validated bioinformatics algorithms in catalyzing the investigation of haplotypes in hypervariable pharmacogenes, such as *CYP2B6* and *CYP2A6*. In addition, this is one of the first studies to demonstrate the use of targeted high-fidelity single-molecule real-time sequencing for characterizing novel *CYP2B6* and *CYP2A6* star alleles.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ This study highlights the need for clinical pharmacogenetics implementation strategies across Africa for substrates such as antiretrovirals, nicotine, and bupropion. Moreover, our findings (such as the relatively high number of novel star alleles) emphasize potential pitfalls in transferability of CYP2B6 and CYP2A6 phenotype prediction strategies based predominantly on European-ancestry populations. Therefore, pharmacogenomic studies and relevant variant functional impact assays involving more understudied populations—particularly in Africa—are warranted to inform effective drug efficacy and safety optimization across Africa, the African diaspora, and other global settings.

Genetic variation in the cytochrome P450 (CYP) supergene family is a major contributor to the drug response variability within and between populations. Of the 57 known functional CYP genes, 12 encode enzymes responsible for the metabolism and bioactivation of 70–80% of all clinically prescribed medications.[1,2] In particular, CYP2B6 and CYP2A6 combined are important for the metabolism of over 10% of drugs that have predominantly CYP-mediated pathways, including antiretrovirals (e.g., efavirenz and nevirapine), nicotine, and bupropion, among other substrates.[1,3]

The human *CYP2B6* and *CYP2A6* genes are located on chromosome 19q13 within the large *CYP2ABFGST* gene cluster.[4] Both of these genes are in close proximity with homologous pseudogenes, *CYP2B7* and *CYP2A7*, respectively. *CYP2B6* and *CYP2A6* are highly polymorphic, with over 37 and 45 star alleles (haplotypes) catalogued for these genes, respectively, by the Pharmacogene Variation Consortium (https://www.pharmvar.org). Star alleles comprise various combinations of single nucleotide variations (SNVs), small insertions and deletions (indels), and/or structural variants—which include copy number variations and other more complex re-arrangements.[5] A number of star alleles, such as *CYP2B6*6* (decreased function), *CYP2B6*22* (increased function), *CYP2A6*4* (gene deletion), and *CYP2A6*46* (58 bp 3′-UTR gene conversion), are known to contribute to variability in patient response to the aforementioned medications metabolized by CYP2B6 and CYP2A6, respectively.[6,7] However, the full catalogue of pharmacogenomically relevant star alleles is yet to be determined, especially in African populations, and across other under-represented biogeographical groups.[2,8]

African populations have higher genetic diversity compared with any other global superpopulations. Therefore, the paucity of information on *CYP2B6* and *CYP2A6* star allele distributions in under-represented African populations poses challenges for effective phenotype prediction[9,10] based on variants or diplotypes determined via various next-generation sequencing (NGS)-based platforms and bioinformatics pipelines. This effectively hampers efforts aimed at optimizing drug therapy adjustments in clinical settings, as there is often variability in measured drug response even within the same genotype-predicted metabolizer phenotype categories. The presence of the neighboring *CYP2B7* and *CYP2A7* pseudogenes and complex structural variant star alleles make diplotyping *CYP2B6* and *CYP2A6* challenging and oftentimes labor-intensive.[8,11] However, the recent availability of high coverage African genomes generated by various international[12] and Africa-based projects,[13] and development of star allele calling bioinformatics tools[14–16] provide the opportunity to study *CYP2B6* and *CYP2A6* pharmacogenetic variation across African populations at scale. Furthermore, recent NGS technologies, such as single-molecule real-time (SMRT) sequencing can facilitate validation of novel star alleles in these genes, as previously applied to other complex pharmacogenes, such as *CYP2D6*.[17,18] This study therefore

Table 1 Sources of the high-coverage whole genome sequence datasets used in the study

Populations/countries of origin	Project/institution	*n*
H3Africa Consortium data		
Fon from Benin (FNB)	H3Africa Baylor; University of Montréal	50
Berom of Nigeria (BRN)	H3Africa Baylor; Institute of Human Virology	49
Cameroon (CAM)	H3Africa Baylor; University of Dschang	26
Ghana (GHA)	H3Africa Baylor; AWI-Gen	26
Burkina Faso (BFA)	H3Africa Baylor; AWI-Gen	33
South Africa	AWI-Gen	100
Botswana (BOT)	H3Africa Baylor; CAfGEN	47
Bantu-speakers from Zambia (BSZ)	H3Africa Baylor; University of Zambia	41
Data from other Africa-based projects		
South Africa	SAHGP	15
South Africa	CBRL (NICD; Wits University)	40
African genomes in public repositories		
Botswana/Namibia	SGDP	3
Namibia	SGDP[a]	3
DRC	SGDP[b]	4
Gambia	SGDP	2
Kenya	SGDP[c]	5
Nigeria	SGDP	4
Senegal	SGDP	3
South Africa	SGDP[a]	3
South Sudan	SGDP[c]	3
Luhya in Webuye, Kenya (LWK)	1000 Genomes Project	99
Esan in Nigeria (ESN)	1000 Genomes Project	99
Yoruba in Ibadan (YRI)	1000 Genomes Project	108
Mende in Sierra Leone (MSL)	1000 Genomes Project	85
Gambian in Western Division, Mandinka (GWD)	1000 Genomes Project	113
Subtotal (Sub-Saharan Africa)		961
Public datasets with genomes from other global superpopulations (for comparative analysis)		
African Caribbean in Barbados (ACB)	1000 Genomes Project	96
People with African Ancestry in Southwest USA (ASW)	1000 Genomes Project	61
European (EUR)	1000 Genomes Project	503
Admixed American (AMR)	1000 Genomes Project	347
South Asian (SAS)	1000 Genomes Project	489
East Asian (EAS)	1000 Genomes Project	504
Subtotal (other global biogeographical groups)		2,000

AWI-Gen, Africa Wits-INDEPTH partnership for Genomics studies; CAfGen, The Collaborative African Genomics Network; CBRL, Cell Biology Research Laboratory; H3Africa, Human Heredity and Health in Africa; NICD, National Institute for Communicable Diseases; SAHGP, Southern African Human Genome Programme; SGDP, Simons Genome Diversity Project.
The genomes of all the study participants were sequenced to an average read depth of ~30× by the respective projects indicated in the table. Most of the continental African participants referred to in this table are from the Niger-Congo language family, except the following: [a]Khoe and San Hunter-gatherers; [b]Rain-Forest Hunter-gatherers; [c]Nilo-Saharan ($n=2$ for the Kenyan SGDP participants).

aims to extensively characterize the distribution of known and potential novel *CYP2B6* and *CYP2A6* star alleles, in particular across diverse sub-Saharan African (SSA) populations.

Our findings have significant implications for precision medicine strategies, particularly involving optimizing antiretroviral treatment, smoking cessation drug therapy, and treatment of major depressive disorders, across clinical settings in SSA, Africa at large, and those serving the African diaspora. Furthermore, our star allele distribution comparisons with other global biogeographical groups provide key insights into previously uncharacterized *CYP2B6* and

CYP2A6 pharmacogenetic variation globally. In addition, this study highlights the use of targeted SMRT sequencing for the validation of *CYP2B6* and *CYP2A6* star alleles.

METHODS

Study population and whole genome sequence data sources

SSA populations represented in this study and the primary data sources are summarized in **Table 1**. We analyzed 961 SSA whole genome sequence (WGS) samples in total. This included 272 genomes that were generated as part of the Human, Heredity, and Health in Africa (H3Africa)-Baylor dataset[13,19] (see **Table 1** for details), 100 genomes of south eastern Bantu Speakers (SEB) that are part of the Africa Wits-INDEPTH Partnership for Genomics Research (AWI-Gen) project,[20] 40 South African genomes generated by the Cell Biology Research Laboratory (CBRL; 39 African ancestry and one of mixed ancestry), 15 of the 24 genomes (7 excluded due to recent admixture) generated by the Southern African Human Genome Programme (SAHGP),[21] 31 African genomes generated by the Simons Genome Diversity Project,[22] and 504 continental African genomes generated by the 1000 Genomes Project.[12] In addition, we analyzed the rest ($n = 2{,}000$) of the 1000 Genomes Project high-coverage WGS samples, including African American/Afro-Caribbean participants ($n = 157$), and participants of European ($n = 503$), East Asian ($n = 504$), South Asian ($n = 489$), and Admixed American ($n = 347$) ancestry, in order to compare the *CYP2B6* and *CYP2A6* star allele distribution in Africa vs. that in global populations. All the genomes analyzed in this study were sequenced on Illumina platforms to a minimum depth of 30× by the primary research projects, and they were aligned to the GRCh38 reference genome.

DNA samples for star allele validation

Genomic DNA from 192 study participants was used during our long-read-based *CYP2B6* and *CYP2A6* star allele validation under ethics amendment terms in protocol M200993. This included DNA samples from the CBRL South African participants, the SEB AWI-Gen participants, and aliquots of samples (from Ghana and Burkina Faso participants) provided by AWI-Gen to H3Africa-Baylor for high coverage sequencing.

Star allele analysis

CYP2B6 and *CYP2A6* star alleles were called from WGS datasets using three separate tools: StellarPGx version 1.2.6,[14] Aldy version 4.4,[16] and PyPGx version 0.19.0[23] (successor to Stargazer[15]), in order to minimize the possibility of false-negative calls. Binary Alignment Map files were provided as input to StellarPGx and Aldy, which perform combinatorial-based diplotype assignment. For PyPGx, we supplied Variant Call Format files and depth of coverage files generated using the in-built *create-input-vcf* and *prepare-depth-of-coverage* scripts.[23] The 1000 Genomes Reference Panel was used for PyPGx's statistical phasing. Consensus diplotype calls were determined by considering star alleles called by at least two of the algorithms. However, for samples where complex structural variants (such as *CYP2A6*46* which is defined by a 58 bp 3′-UTR conversion to *CYP2A7*) were called by at least one tool, we performed manual visual inspections of the read coverage using the Integrative Genomics Viewer,[24] in addition to considering copy number and allele fraction profile plots output by PyPGx.

Metabolizer phenotype prediction

The *CYP2B6* consensus diplotype calls in this study were translated to CYP2B6 metabolizer phenotypes based on the Clinical Pharmacogenomics Implementation Consortium guidelines for efavirenz[9] as there were no corresponding guidelines for other CYP2B6-drug pairs at the time of this study. Participants with potentially novel star alleles were assigned an indeterminate metabolizer status. CYP2A6 phenotypes were not predicted in this study (see Discussion).

CYP2B6 and *CYP2A6* long-range PCR

As *CYP2B6* is a relatively large gene (~27 kb), multiple XL-PCR fragments were generated to cover various regions (see **Supplementary Material S1**). *CYP2B6* Frag1 (~9 kb) stretched from the *CYP2B6* upstream region to intron 1; Frag3 (~8.5 kb) covered exon 2 to exon 6; Frag4 (~10.3 kb) covered exon 4 to exon 9; and Frag5 (~7.1 kb) covered exon 8 to the *CYP2B6* downstream region. We ran multiple PCR optimisations to amplify Frag2 (~13 kb, exon 1 to exon 3) but none were successful. The XL-PCR primers and cycling conditions for each fragment are detailed in **Supplementary Material S1**.

For *CYP2A6*, 9.2 kb-long amplicons (FragA) that comprise the *CYP2A6* gene as well as upstream and downstream non-coding regions were generated following the long-range PCR (XL-PCR) protocols described by Wassenaar *et al.*[11] with some modifications. The XL-PCR forward and reverse primers used and PCR cycling conditions are detailed in **Supplementary Material S1**. We used previously published primers[25,26] to ascertain the presence of *CYP2A6* gene duplication(s). See **Supplementary Material S1** for depictions of the *CYP2A6* XL-PCR fragments targeted in this study.

The forward and reverse primers used to generate XL-PCR amplicons were tailed with universal sequences on the 5′ end (5′-GCAGTCGA ACATGTAGCTGACTCAGGTCAC-3′ and 5′-TGGATCACTTG TGCAAGCATCACATCGTAG-3′, respectively) to enable sample barcoding via a second PCR. Each 20 µL reaction mix contained 60–120 ng of genomic DNA, 10 µL of 2X LongAmp Taq ReadyMix (New England Biolabs, South Africa), 1 µL of 100% DMSO (Sigma-Aldrich/Merck, Johannesburg, South Africa), and 1 µL each of 10 µM forward and reverse primers (Inqaba Biotech, Pretoria, South Africa). The specific PCR cycling conditions are provided in **Supplementary Material S1**.

Amplicon pooling and barcoding

Amplicon pooling, barcoding, and the subsequent high fidelity (HiFi) sequencing were performed by Inqaba Biotech. Quality control of the PCR products from the first-round PCR was carried out using a 0.8% agarose gel for visual inspections and the Agilent 4200 TapeStation (Diagnostech, Johannesburg, South Africa) for quantification following the D5000 Screen tape kit. For each participant, *CYP2B6* amplicons were pooled equimolarly with *CYP2A6* amplicons, and also with *CYP2D6* amplicons from the related study.[17] All amplicons were in the size range of 5–12 kb. The amplicon pools were purified using the AMPure PB bead purification (Pacific Biosciences, California, USA). Thereafter, barcodes were added to the purified amplicons via a second round of PCR. The 25 µL reaction mix contained 5–10 ng/µL of initial pooled PCR product, 11 µL of 2X longAmp Taq ReadyMix (New England Biolabs, USA), 1 µL of DMSO (100%), and 5.0 µL of 2 µM barcoded universal primers (Inqaba Biotech). The PCR cycling conditions were as follows: 20 cycles of 95°C for 1 minute, 65°C for 30 seconds, and 72°C for 11 minutes.

Single-molecule real-time sequencing

SMRTBell libraries were constructed from ~500 ng of pooled barcoded fragments by following standard end-repair, adapter ligation, and purification strategies detailed in the Pacific Biosciences protocols (https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf). Annealing and binding of SMRTbell templates was performed using the Sequel II Binding kit 2.2 and sequencing primer version 5, and Circular Consensus Sequencing was performed for a movie time of 30 hours to generate HiFi reads via the SMRT Link software on the Sequel IIe instrument (Pacific Biosciences) at Inqaba Biotech.

Table 2 ***CYP2B6*** **and** ***CYP2A6*** **star allele frequencies (%) in sub-Saharan African populations compared with global populations**

		Allele frequencies (%)					
CYP2B6 star alleles	CPIC clinical function	SSA in this study (*n* = 935)	African American/ Afro-Caribbean (*n* = 153)	European (*n* = 499)	Admixed American (*n* = 344)	East Asian (*n* = 501)	South Asian (*n* = 479)
**1*	Normal	42.8	42.8	53.5	47.5	66.6	42.9
**2*	Normal	4.2	3.3	5.3	2.6	4.2	4.4
**5*	Normal	0.5	2	11.3	6.7	0.2	8.5
**17*	Normal	2.5	2.6	0	0.3	0	0
**32*	Normal	0.1	0	0	0	0	0
**6*	Decreased	32.6	34.3	22.2	34.7	19.8	36.3
**7*	Decreased	0.2	0	0	0	0	0.1
**9*	Decreased	3.2	0.3	0.6	1.2	0.3	1
**19*	Decreased	0.2	0	0	0	0	0
**20*	Decreased	0.1	0	0	0.1	0	0
**26*	Decreased	0	0	0	0	0.5	0
**29*	Decreased	0.4	0.7	0	0.1	0	0
**36*	Decreased	0.8	1	0	0.1	0	0.1
**12*	No function	0	0	0	0.1	0	0
**13*	No function	0	0.7	0.1	0	0	0.1
**18*	No function	9.5	7.5	0	1	0	0
**24*	No function	0	0	0	0	0.1	0
**4*	Increased	0.1	0.3	3	1	6.3	3.7
**22*	Increased	1.1	2	0.9	0.6	0.2	1.7
**3*	Uncertain	0	0.3	0.1	0	0	0.2
**10*	Uncertain	0	0	0.6	1.3	0.1	0.1
**11*	Uncertain	0.1	0	0.4	0	0	0
**15*	Uncertain	0	0	0.4	0.4	0	0
**23*	Unknown	0	0	0	0	0.2	0.1
**27*	Uncertain	0	0.3	0	0	0	0
**33*	Uncertain	0.1	0.3	0	0	0	0
CYP2A6 star alleles	Functional impact[a]	SSA in this study (*n* = 957)	African American/ Afro-Caribbean (*n* = 155)	European (*n* = 498)	Admixed American (*n* = 344)	East Asian (*n* = 503)	South Asian (*n* = 486)
**1*	Normal enzyme activity	55.7	54.8	51.3	38.8	32	46.5
**1x2*	Increased mRNA expression	0.8	2.6	0.9	0.7	0	0.5
**2*	Substantially decreased enzyme activity	0	0.3	3.2	0.7	0	0.6
**4*	No mRNA expression	3.1	2.3	0.9	1.5	11.8	3.8
**5*	Decreased enzyme activity	0	0	0	0	0.1	0.1
**7*	Decreased enzyme activity	0	0	0.2	0	8	0.2
**8*	Normal enzyme activity	0	0	0	0	0.1	0

(Continued)

Table 2 (Continued)

CYP2A6 star alleles	Functional impact[a]	SSA in this study (*n* = 957)	African American/ Afro-Caribbean (*n* = 155)	European (*n* = 498)	Admixed American (*n* = 344)	East Asian (*n* = 503)	South Asian (*n* = 486)
**9*	Decreased mRNA expression	7.5	7.7	7	9.4	19	12.9
**10*	Inactive enzyme	0	0	0	0	3	0.2
**11*	Decreased enzyme activity	0	0	0	0	0.8	0
**12*	Decreased enzyme activity	0.1	0	2	1.9	0	0.9
**13*	Unknown/uncertain	0	0	0	0	0.2	0
**14*	Unknown/uncertain	0	1	3.2	1.6	0	2.3
**15*	Unknown/uncertain	0	0	0	0	1.1	0
**17*	Substantially decreased enzyme activity	11	12.3	0	0.4	0	0
**18*	Decreased enzyme activity	0.3	0	1.4	1.3	0.6	1.7
**19*	Decreased enzyme activity	0	0	0	0	0.4	0
**20*	Substantially decreased protein levels	0.8	0.3	0	0	0	0
**21*	Decreased enzyme activity	0	0	1.2	0.6	0	0.8
**23*	Decreased enzyme activity	1.3	3.2	0	0	0	0
**24*	Decreased enzyme activity	1	0.6	0	0.1	0	0
**25*	Decreased enzyme activity	0.3	0.3	0	0	0	0
**26*	Decreased enzyme activity	0.3	0.6	0	0	0	0
**27*	Decreased enzyme activity	0.9	0.6	0	0	0	0
**28*	Decreased enzyme activity	2.4	0.6	0	0.3	0	0.1
**31*	Unknown/uncertain	2.3	1.3	0	0	0	0
**34*	Unknown/uncertain	0.1	0.3	0	0.1	0	0.1
**35*	Decreased enzyme activity	2.8	1.6	0.8	0.3	0.2	0
**36*	Unknown/uncertain	0	0	0	0	0.5	0
**37*	Unknown/uncertain	0	0	0	0	0.2	0
**39*	Decreased enzyme activity	0.1	0	0	0	0	0
**41*	Decreased enzyme activity	0.1	0	0	0	0	0
**46*	Increased mRNA stability	6.1	8.4	27.7	41.7	21.6	26.9

The star allele definitions followed in this study are according to PharmVar v5.2.14.1.

CPIC, Clinical Pharmacogenetics Implementation Consortium; *n*, individuals; PharmGKB, Pharmacogenomics Knowledge Base; PharmVar, Pharmacogene Variation Consortium; SSA, sub-Saharan African populations.

[a]The functional impacts of *CYP2A6* star alleles mentioned in this table were based on a review by Tanner and Tyndale,[2] but they have not yet been curated by the PharmVar *CYP2A6* expert panel and the PharmGKB team.

Table 3 ***CYP2B6*** **and** ***CYP2A6*** **star allele frequencies (%) across all sub-Saharan African populations included in the study**

CYP2B6 star alleles	CPIC clinical function	Allele frequencies (%)												
		FNB (*n* = 50)	BRN (*n* = 45)	BFA (*n* = 33)	GHA (*n* = 26)	CAM (*n* = 24)	BSZ (*n* = 41)	BOT (*n* = 46)	SEB (*n* = 151)	ESN (*n* = 97)	YRI (*n* = 106)	GWD (*n* = 111)	MSL (*n* = 82)	LWK (*n* = 95)
**1*	Normal	35	46.7	31.8	42.3	43.8	40.2	53.3	41.4	45.4	35.8	43.2	46.3	47.4
**2*	Normal	6	1.1	6.1	1.9	0	2.4	4.3	4.6	2.6	5.2	5	2.4	7.4
**5*	Normal	2	0	1.5	1.9	0	0	0	0.3	0	0.5	1.4	0.6	0
**17*	Normal	0	0	4.5	5.8	4.2	6.1	2.2	3	1.5	4.7	2.3	1.2	0.5
**32*	Normal	0	0	0	0	0	0	0	0.3	0	0	0	0	0
**6*	Decreased	42	37.8	47	38.5	27.1	28	21.7	32.5	39.7	34	26.1	28.7	31.1
**7*	Decreased	0	0	0	0	0	0	0	0	0.5	0.5	0.5	0.6	0
**9*	Decreased	0	4.4	3	3.8	0	4.9	1.1	2.6	1.5	4.7	4.5	4.3	3.2
**19*	Decreased	0	0	0	0	0	0	0	0	0	0.5	0.5	1.2	0
**20*	Decreased	0	0	0	0	0	1.2	0	0.3	0	0	0	0	0
**29*	Decreased	0	2.2	0	0	0	1.2	0	0	0.5	0.9	0.5	0	0
**36*	Decreased	2	3.3	0	0	2.1	1.2	1.1	0	0	0	0.9	0	2.1
**18*	No function	11	2.2	3	5.8	18.8	14.6	13	12.9	5.2	11.8	9.9	10.4	6.3
**4*	Increased	0	0	0	0	0	0	0	0	0	0	0.5	0	0
**22*	Increased	0	0	1.5	0	2.1	0	1.1	0.7	2.1	0.9	1.8	2.4	0.5
**11*	Uncertain	0	0	0	0	0	0	0	0	0	0	0	0	0.5
**33*	Uncertain	0	0	0	0	2.1	0	0	0	0	0	0	0	0

CYP2A6 star alleles	Functional impact[a]	FNB (*n* = 50)	BRN (*n* = 49)	BFA (*n* = 33)	GHA (*n* = 26)	CAM (*n* = 26)	BSZ (*n* = 41)	BOT (*n* = 47)	SEB (*n* = 155)	ESN (*n* = 98)	YRI (*n* = 108)	GWD (*n* = 113)	MSL (*n* = 84)	LWK (*n* = 98)
**1*	Normal enzyme activity	53	54.1	56.1	59.6	53.8	56.1	59.6	60.3	54.1	56.9	52.2	51.8	62.8
**1x2*	Increased mRNA expression	0	1	1.5	1.9	0	0	1.1	1	1	0.5	0.4	1.8	0.5
**4*	No mRNA expression	2	6.1	1.5	0	1.9	6.1	5.3	1	3.1	3.7	2.7	5.4	3.1
**9*	Decreased mRNA expression	10	4.1	3	9.6	9.6	9.8	5.3	8.4	5.1	7.9	8.8	8.3	8.2
**12*	Decreased enzyme activity	0	0	0	0	0	0	1.1	0	0	0	0	0	0
**17*	Substantially decreased enzyme activity	13	10.2	19.7	11.5	17.3	6.1	6.4	8.7	14.8	7.9	12.4	12.5	9.7
**17x2*	Unknown	1	0	0	0	0	0	0	0	0	0	0	0	0
**18*	Decreased enzyme activity	0	1	0	0	0	0	0	0	0.5	0.5	0.4	0	0.5
**20*	Substantially decreased protein levels	0	0	0	0	0	0	0	0.3	1.5	0.5	2.7	1.2	1
**23*	Decreased enzyme activity	4	0	1.5	1.9	0	1.2	0	0	1.5	0.9	1.8	1.8	2
**24*	Decreased enzyme activity	2	4.1	1.5	0	0	0	1.1	0.3	0	1.4	0.9	1.8	1

(Continued)

Table 3 (Continued)

CYP2A6 star alleles	Functional impact[a]	FNB (*n* = 50)	BRN (*n* = 49)	BFA (*n* = 33)	GHA (*n* = 26)	CAM (*n* = 26)	BSZ (*n* = 41)	BOT (*n* = 47)	SEB (*n* = 155)	ESN (*n* = 98)	YRI (*n* = 108)	GWD (*n* = 113)	MSL (*n* = 84)	LWK (*n* = 98)
**25*	Decreased enzyme activity	0	1	0	0	0	0	0	0	0.5	0.5	0.4	1.2	0
**26*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0.5	1.3	0	0.5
**27*	Decreased enzyme activity	1	1	0	0	0	4.9	2.1	1.3	0	0	0.9	1.2	0.5
**28*	Decreased enzyme activity	0	6.1	1.5	3.8	0	3.7	1.1	0.3	1	4.2	4.4	1.8	3.1
**28x2*	Unknown	0	0	0	0	0	0	0	0.3	0	0	0	0	0
**31*	Unknown	2	1	1.5	1.9	3.8	2.4	5.3	5.5	0.5	1.4	1.3	1.2	0
**34*	Unknown	0	0	0	0	0	0	0	0	0	0	0	0	0.5
**35*	Decreased enzyme activity	1	1	4.5	3.8	1.9	0	3.2	2.3	6.1	1.9	3.5	1.8	3.6
**39*	Decreased enzyme activity	0	0	0	0	0	0	0	0	0	0	0	0.6	0
**41*	Decreased enzyme activity	0	0	0	1.9	0	0	0	0	0	0	0	0	0
**46*	Increased mRNA stability	11	7.1	7.6	3.8	9.6	8.5	4.3	6.5	5.6	7.4	3.1	5.4	1.5

The star allele definitions followed in this study are according to PharmVar v5.2.14.1.

BFA, participants from Burkina Faso; BOT, participants from Botswana; BRN, Berom in Nigeria; BSZ, Bantu-speakers in Zambia; CAM, participants from Cameroon; ESN, Esan in Nigeria; FNB, Fon in Benin; GHA, participants from Ghana; GWD, Mandinka in Western Divisions of the Gambia; LWK, Luhya in Webuye, Kenya; MSL, Mende in Sierra Leone; PharmGKB, Pharmacogenomics Knowledge Base; PharmVar, Pharmacogene Variation Consortium; SEB, south-eastern Bantu in South Africa; YRI, Yoruba from Ibadan in Nigeria.

[a]The functional impacts of *CYP2A6* star alleles mentioned in this table were based on a review by Tanner and Tyndale,[2] but they have not yet been curated by the PharmVar *CYP2A6* expert panel and the PharmGKB team.

Raw HiFi sequence data were demultiplexed and processed into individual samples according to the corresponding barcode sequences using the NGSutils NGS data analysis software kit.[27] *CYP2B6* and *CYP2A6* HiFi reads were aligned to the corresponding regions in GRCh38, and also to the NG_007929.1 and NG_008377.1 reference sequences, respectively, using pbmm2 v1.7.0 (https://github.com/PacificBiosciences/pbmm2). Variant calling was done using DeepVariant.[28] Thereafter, variant phasing and read haplotagging was carried out for samples containing more than one heterozygous variant using WhatsHap.[29]

Variant functional prediction

CYP2B6 and *CYP2A6* variants were annotated using the Ensembl Variant Effect Predictor (VEP).[30] The functional effects (corresponding to the NM_000767.5 and NM_000762.6 transcripts) of potential novel star allele-defining variants on these genes were predicted using VEP plugins including SIFT,[31] Polyphen-2,[32] CADD,[33] LRT,[34] PROVEAN,[35] and VEST4,[36] taking into account the absorption, distribution, metabolism, and excretion (ADME)-optimized parameters suggested by Zhou *et al*.[37] SIFT Indel[38] was used to annotate frameshift variants while LOFTEE[39] was used to identify loss of function variation.

Statistical analysis

Star allele frequencies were summarized using percentages. Deviations from Hardy Weinberg equilibrium were investigated using the genetics package (https://www.rdocumentation.org/packages/genetics/versions/1.3.8.1.3) in R version 4.1.3 (https://www.r-project.org). The Fisher's exact test was used to determine significant differences in population *CYP2B6* and *CYP2A6* star allele frequencies. Any P values of < 0.05 were considered statistically significant.

Ethics statement

This study was approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand under protocol numbers M190631 and M200993. We performed secondary analysis of full genomes generated by contributing studies/centers based across Africa (each of which obtained local ethics approval), and further supplemented this with analysis of data from public repositories.

RESULTS

CYP2B6 star allele frequencies

Among the normal function *CYP2B6* star alleles, *CYP2B6*1* and **2* were the most frequent in SSA followed by **17* (**Table 2**). However, *CYP2B6*17* was not observed among the Berom in Nigeria and the Fon in Benin, whereas *CYP2B6*2* was not observed among the participants from Cameroon (**Table 3**). We found *CYP2B6*5* to be rare in SSA compared with frequencies among European, admixed American, and South Asian participants (**Table 2**).

Among the decreased function *CYP2B6* star alleles, *CYP2B6*6*—defined by rs3745274 (Q172H) and rs2279343 (K262R)—was by far the most frequent in SSA participants (Allele Frequency, AF = 32.6%) and also across African American/Afro-Caribbean participants (AF = 34.3%). The individuals of South Asian ancestry as well as the admixed American participants had comparable *CYP2B6*6* frequencies, that is, 36.3% and 34.7%, respectively. However, in comparison, *CYP2B6*6* was present at significantly lower frequencies among the European participants (AF = 22.2%, P = 4e-09) and East Asian participants (AF = 19.8%, P = 1.1e-13) diplotyped in this study (**Table 2**). The *CYP2B6*6* frequency was non-uniform across SSA as it ranged from 21.7% in the Botswana participants to 47% among individuals from Burkina Faso included in this study (**Table 3**). Among other key decreased function star alleles, we detected *CYP2B6*29* (hybrid deletion; SSA AF = 0.4%), **7*, **9*, **19*, **20*, and **36* across SSA (**Tables 2, 3**).

Among the known no function *CYP2B6* star alleles, only *CYP2B6*18*—which is defined by rs28399499 (I328T)—was present in SSA (AF = 9.5%). The frequency of **18* was lower in African American/Afro-Caribbean participants (AF = 7.5%) but this difference was not statistically significant (P = 0.3). In comparison, *CYP2B6*18* was found to be virtually absent among the European, East Asian, and South Asian participants represented in the 1000 Genomes Project dataset (**Table 2**). The frequency of **18* varied across the SSA populations, ranging from 2.2% among the Berom in Nigeria to 18.8% among the participants from Cameroon (**Table 3**).

With regard to the increased function *CYP2B6* alleles, *CYP2B6*4*—defined by rs2279343 (K262R) without rs3745274 (Q172H) in phase—was found in only one participant (among GWD) across SSA. Conversely, *CYP2B6*22* (defined by rs34223104) was more common and present in all SSA populations included in this study (combined AF = 1.1%), except for the Fon in Benin, Berom in Nigeria, Bantu speakers from Zambia, and Ghanaian participants.

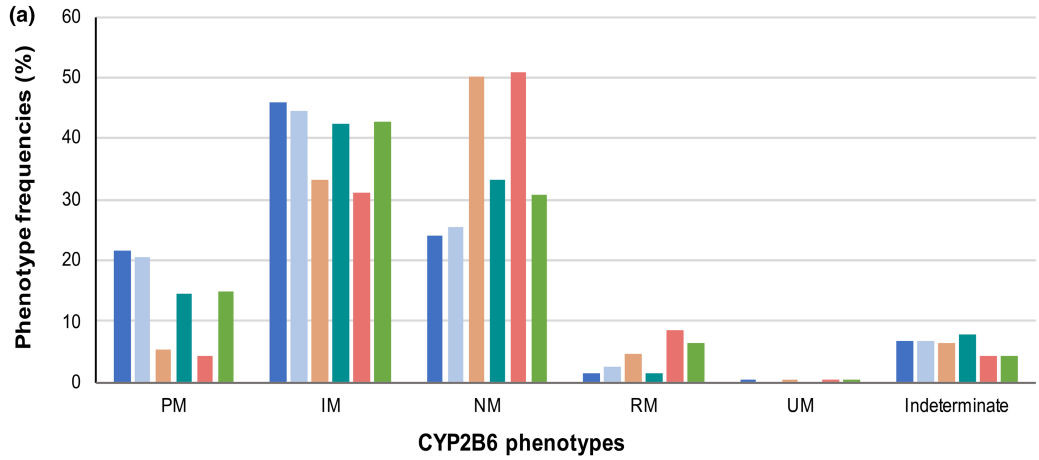
CYP2A6 star allele frequencies

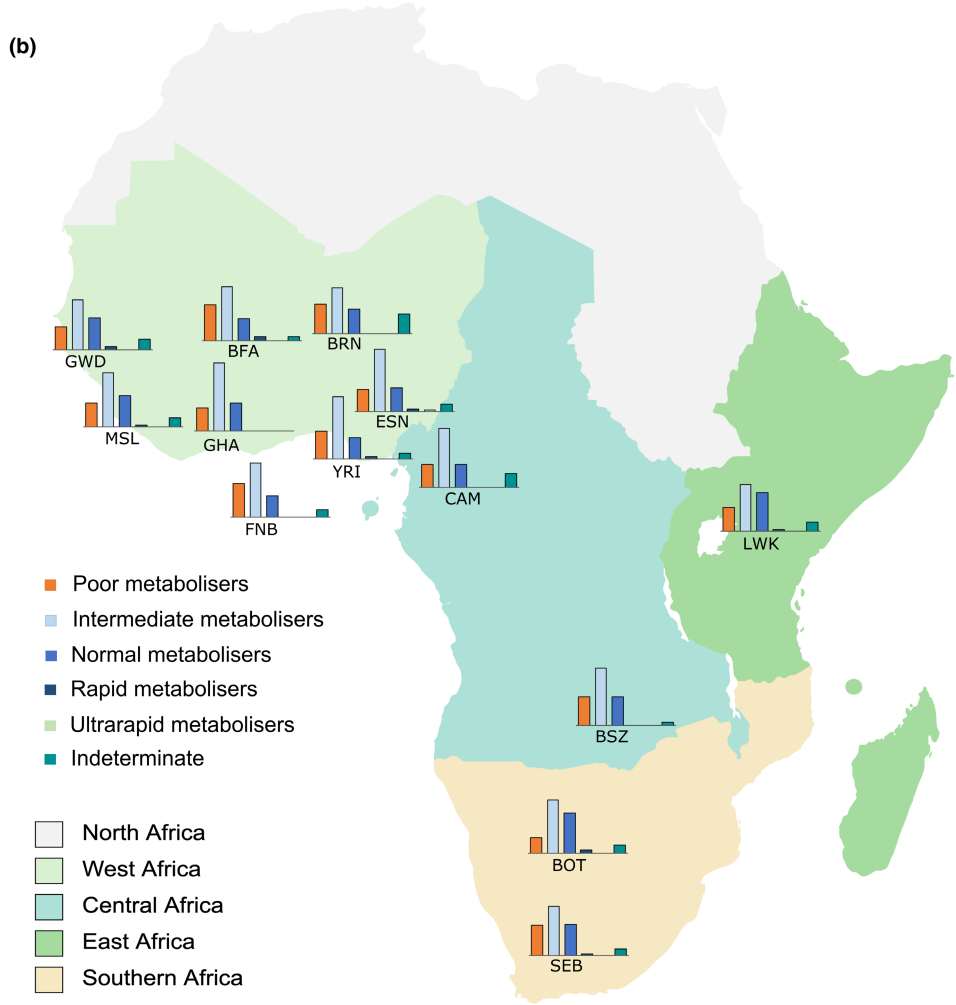
The frequencies of established *CYP2A6* star alleles are summarized in **Table 2** for SSA and other global populations (for comparison) represented in this study. *CYP2A6*1* (normal function star allele) was observed at the highest frequency across majority of the populations in this study.

*CYP2A6*46* (formerly *1B; defined by a 58 bp gene conversion to *CYP2A7* in the 3′-UTR) was observed at a frequency of 6.1% in SSA. Conversely, this star allele was observed at significantly higher frequencies in European, East Asian, and South Asian populations (**Table 2**). The frequency of *CYP2A6*46* varied across SSA populations in the study, ranging from 1.5% among the Luhya in Webuye (Kenya) to 11% among the Fon in Benin (**Table 3**).

Among the previously characterized duplication star alleles, we detected *CYP2A6*1x2* in this study. This star allele was present at a

Figure 1 Distribution of CYP2B6 phenotypes predicted in relation to efavirenz metabolism. (**a**) Comparison of CYP2B6 phenotypes across the global biogeographical groups included in this study. In general, SSA populations have a higher proportion of CYP2B6 poor and intermediate metabolizers compared with other biogeographical groups. This is in part accounted for by the high frequency of the *CYP2B6*6* (decreased function) and *CYP2B6*18* (no function) across Africa. (**b**) CYP2B6 phenotype distribution across the SSA populations in this study. BFA, participants from Burkina Faso; BOT, participants from Botswana; BRN, Berom in Nigeria; BSZ, Bantu speakers from Zambia; CAM, Cameroonian participants; ESN, Esan in Nigeria; FNB, Fon in Benin; GHA, Ghanaian participants; GWD, Gambian in Western Division (Mandinka); IM, intermediate metabolizer; LWK, Luhya in Webuye (Kenya); MSL, Mende in Sierra Leone; NM, normal metabolizer; PM, poor metabolizer; RM, rapid metabolizer; SEB, south-eastern Bantu in South Africa; UM, ultrarapid metabolizer; YRI, Yoruba in Ibadan, Nigeria.





frequency of 0.8% in SSA which was less than the frequency among African American/Afro-Caribbean participants (2.6%; $P = 0.01$), comparable to the frequency in European, admixed American, and South Asian participants, but absent from the East Asian populations in this study (Table 2). Among the SSA participants, *CYP2A6*1x2* was most frequent among the participants from

Table 4 Potentially novel *CYP2B6* and *CYP2A6* haplotypes inferred from African short-read whole genome sequence datasets in this study

Haplotype	Background star allele(s)	Additional core variant(s)	Allele count	Country/dataset [a]
CYP2B6				
1	*1	rs541486480~g.5164A>C (p.K53Q)	2	Nigeria (1000G)
2	*1	rs373926269~g.17867T>C (splice donor)	1	Gambia (1000G)
3	*1	rs370958436~g.18040C>T (stop-gained)	1	Nigeria
4	*1	rs766630605~g.20648G>A (p.A176T)	1	Nigeria
5	*1	rs537265436~g.20726T>C (p.F202L)	2	Botswana, South Africa
6	*1	rs1599849465~g.22995A>C (p.E240D)	1	South Africa
7	*1	rs150072531~g.26413A>G (p.H397R)	2	Gambia (1000G), Benin
8	*2	rs3211371~g.30512C>T (p.R487C)	2	Burkina Faso, Cameroon
9	*6	rs572134005~g.17785G>A (p.R85Q)	1	Sierra Leone
10	*6	rs183427203~17823C>T (p.R98W)	1	Kenya (1000G)
11	*6	rs142421637~g.17856C>T (p.R109W)	5	Botswana, South Africa, Kenya (1000G, SGDP)
12	*6	rs58871670~g.20669G>A (p.V183I)	4	Gambia (1000G), Nigeria (1000G)
13	*22	rs141666881~g.18137G>A (p.R158Q)	6	Gambia (1000G, SGDP), Nigeria (1000G), Sierra Leone (1000G)
14	*22	rs34698757~g.23790C>G (p.T306S)	1	Sierra Leone (1000G)
15	*22	rs147991149~g.30380C>T (p.R443C)	2	South Africa
CYP2A6				
1	*46	rs558145012~5128G>T (p.G36V)[b]	2	South Africa
2	*1	rs72549435~5615G>C (p.V110L)	1	Botswana
3	*1	rs780290198~6720C>T (p.L127F)	1	Cameroon
4	*1	rs554865113~6727G>C (p.R129P)	1	Nigeria (1000G)
5	*46	rs2545783~6798G>T (p.A153S)[c]	2	South Africa, Nigeria (1000G)
6	*1	rs557976670~8433C>T (p.P231S)	1	Kenya (1000G)
7	*1	rs138978736~8457C>A (p.Q239K)	2	Gambia (1000G), Nigeria (1000G)
8	*1	rs58720852~8515C>A (p.T258K)	1	Sierra Leone (1000G)
9	*1	rs111869995~8567G>T (p.M275I)	2	Sierra Leone (1000G), Zambia
10	*1	rs528089983~9445C>T (p.T309I)	2	Nigeria (1000G)
11	*1	rs58571639~9450C>T (p.R311C)	1	Kenya (1000G)
12	*1	rs145036049~10108G>A (p.R372H)	1	Nigeria (1000G)
13 (*CYP2A6*55*)	*1	rs114558780~10126C>T (p.T378I)	13	Botswana, South Africa, Nigeria (1000G) Congo (SGDP), Gambia (1000G)
14	*1	rs28399463~10766A>G (p.N418D)	4	Nigeria (1000G)
15	*1	rs8192730~10771G>C (p.E419D)	1	Nigeria (1000G)
16 (*CYP2A6*56*)	*46	rs113558392~9394T>G (p.V292G)	2	Namibia (SGDP), South Africa
17 (*CYP2A6*56x2*)[d]	*46	[(*46+rs113558392~9394T>G, p.V292G)×2]	10	Namibia (SGDP), Botswana, South Africa
18	*1	rs61605570~9990A>T (stop-gained)	1	Nigeria (1000G)
19	*1	rs533216061~8508C>A (Q256K)+rs771986786~7227C>G (Q218E)	3	South Africa

(Continued)

Table 4 (Continued)

Haplotype	Background star allele(s)	Additional core variant(s)	Allele count	Country/dataset [a]
20	**1*	[(rs533216061~8508C>A, p.Q256K+rs771986786~7227C>G, p.Q218E)×2]	1	South Africa
21	**5*	rs143731390~11479A>T (p.N438Y)+rs72549435~5615G>C (p.V110L)	1	Nigeria (1000G)
22	**9*	rs145308399~5576G>A (p.E97K)	2	Nigeria (1000G), Sierra Leone (1000G)
23	**9*	rs1809810~10689A>T (p.Y392F)	1	Nigeria
24	**17*	rs554920226~10778C>A (p.Q422K)	1	Gambia (1000G)
25	**17*	**17x2*	1	Benin
26	**18*	rs4997557~9400C>G (p.T294S)+rs2644906~9393G>A (p.V292M)	3	Gambia (1000G), Nigeria (1000G), South Sudan (SGDP)
27	**28*	rs28399454~10086G>A (p.V365M)	1	Nigeria
28	**28*	**28x2*	1	South Africa

Unresolved diplotypes with potentially novel star alleles

#	Background star alleles	Additional core variant(s)	Count	Country/Dataset[a]
CYP2B6				
1	**1/*18*	rs34698757~23790C>G (p.T306S)	1	Botswana
2	**1/*6*	rs33973337~5083A>T (p.T26S)	1	Cameroon
3	**2/*9*	rs28399499~26018T>C (p.I328T)	1	Botswana/Namibia (SGDP)
4	**6/*17*	rs45459594~20668C>G (p.I182M)+rs36079186~20715T>C (p.M198T)	1	Kenya (1000G)
5	**1* and **22*	Potential novel *CYP2B6/2B7* hybrid/ duplication (various breakpoints)	21	Various[a]
CYP2A6				
1	**9/*35*	rs137904044~9983G>T (p.E330D)	1	Nigeria (1000G)
2	**17/*35*	rs143067113~5546G>A (p.V87I)	1	Kenya (1000G)
3	**2/*27*	rs143731390~11479A>T (p.N438Y)	1	Sierra Leone (1000G)
4	**9/*17*	Unresolved *CYP2A6* duplication	1	South Sudan (SGDP)[a]

1000G, 1000 Genomes Project; HiFi, high fidelity; SGDP, Simons Genome Diversity Project; SMRT, single-molecule real-time.
[a]Where applicable, Coriell IDs for samples with these potential novel/complex *CYP2B6* and *CYP2A6* star alleles are provided in **Supplementary Material S2**.
[b]HiFi SMRT sequencing data later revealed an exon 3/intron3 conversion to *CYP2A7* in phase with this core variant (see **Figure 3**).
[c]This core variant may occur as part of the exon 3/intron3 conversion to *CYP2A7*.
[d]The duplicated form of *CYP2A6*56* is yet to be fully validated.

Ghana (AF = 1.9%) and the Mende in Sierra Leone (AF = 1.8%; **Table 3**).

Among the relatively large number of *CYP2A6* star alleles that cause decreased CYP2A6 expression and/or activity, **9* (defined by the rs28399433 variant in the TATA box) and **17* (rs28399454, V365M) were the most frequent across SSA (**Table 2**), but observed at non-uniform frequency distributions (**Table 3**). The frequency of the *CYP2A6*9* haplotype in SSA (7.5%) was comparable to that in other biogeographical groups, except in the East Asian populations (AF = 19%; **Table 2**). In contrast, *CYP2A6*17* was largely African-specific as it was observed at a frequency of 11% in SSA and 12.3% in African American/Afro-Caribbean participants, but absent among the European, East Asian, and South Asian participants (**Table 2**). Other largely African-specific *CYP2A6* star alleles that were observed in SSA in this study include **20*, **24*, **25*, **26*, **27*, **28*, **31*, **35*, **39*, and **41* (**Tables 2, 3**).

We observed the *CYP2A6*4* (*CYP2A6* gene deletion) at a frequency of 3.1% in SSA, which was similar to the **4* frequency in the African American/Afro-Caribbean participants, and South Asian participants. However, *CYP2A6*4* was less frequent among the European populations and highest among the East Asian populations (**Table 2**). Among the SSA participants in this study, *CYP2A6*4* was most frequent among the Berom in Nigeria, and Bantu-speakers from Zambia (AF = 6.1%), but it was not observed among the participants from Ghana.

Predicted *CYP2B6* phenotypes

The distributions of predicted CYP2B6 metabolizer phenotypes relative to efavirenz response are shown in **Figure 1**. Two-hundred seven participants (21.5%) were predicted to be CYP2B6 poor metabolizers (PMs) across the SSA populations in the study. This was similar to the proportion of CYP2B6 PMs in the African American/Afro-Caribbean participants (20.4%) but significantly higher than the proportion of PMs in European (5.4%; $P < 2.2e\text{-}16$), admixed American (14.7%; $P = 0.0058$), East Asian (4.4%; $P < 2.2e\text{-}16$), and South Asian (15.1%; $P = 0.004$) populations (**Figure 1a**). The proportion of predicted CYP2B6 PMs varied across SSA, with the lowest proportion being among the Botswana participants (12.8%), and the highest PM proportion observed among the participants from South Africa (25.8%), Benin (28%), and Burkina Faso (30.3%).

For the CYP2B6 intermediate metabolizer (IM) phenotype, participants from across SSA had a considerably higher frequency (46%) of predictive diplotypes compared with European participants (33.2%) and East Asian participants (31.2%; **Figure 1a**). The proportion of IMs in SSA populations was comparable to that in African American/Afro-Caribbean (44.6%), admixed American (42.4%), and South Asian (42.7%) participants. The distribution of IMs varied across SSA (**Figure 1b**) with relatively higher proportions among participants from Ghana (57%), Cameroon (50%), the Yoruba in Ibadan (Nigeria; 52.8%), and the Esan in Nigeria (52.5%), whereas the Berom in Nigeria (38.8%), Luhya in Webuye (Kenya) (39.4%), and SEB participants from South Africa (41.3%) had a relatively lower proportion of IMs.

The proportion of SSA participants (1.4%) with the CYP2B6 rapid metabolizer (RM) phenotype was significantly lower than that among the European (4.6%), South Asian (6.5%), and East Asian (8.7%) participants (**Figure 1a**). All the SSA populations in this study had ≤ 3 participants with the CYP2B6 RM phenotype and, in particular, it was not observed among the Fon in Benin, Berom in Nigeria, Bantu speakers from Zambia, and participants from Ghana and Cameroon (**Figure 1b**). The CYP2B6 ultrarapid metabolizer phenotype was only predicted in one participant across all SSA populations in the study. This participant (among the Esan in Nigeria) had the *CYP2B6*22/*22* diplotype.

Computationally predicted novel *CYP2B6* and *CYP2A6* haplotypes

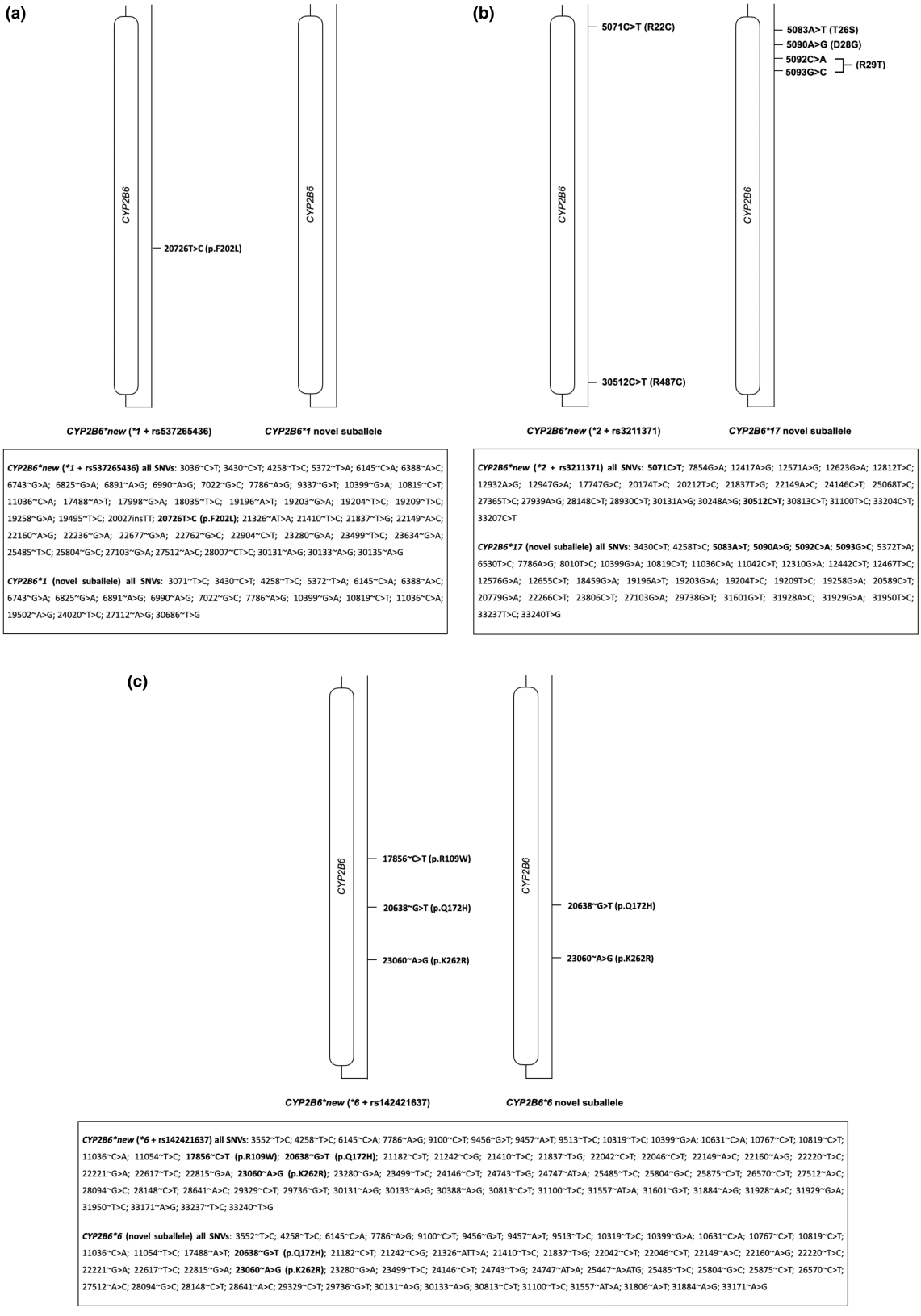
Four percent (41/961) of the SSA participants had potential novel CYP2B6 haplotypes—19 distinct haplotypes in total—based on the high coverage short-read WGS data used in the study. Fourteen of these alleles were fully phased computationally whereas five were observed in individuals whose diplotypes could not be resolved (i.e., the background allele on which the novel variants occurred could not be determined; see **Table 4**, **Supplementary Material S2**). For the phased potentially novel *CYP2B6* alleles, the phasing of novel core variants was made possible either by having homozygous background alleles in the same participant or observing multiple participants with the same novel core variant while also sharing one background allele. However, suballele definitions could not be determined by this approach. The core variants of these potentially novel *CYP2B6* star alleles were predicted to be deleterious by at least one VEP plugin used in the study, except for rs33973337 (T26S), rs541486480 (K53Q), rs537265436 (F202L), rs1599849465 (E240D), and rs3211371 (R487C; **Supplementary Material S2**). In addition to the aforementioned 19 potentially novel alleles, we found a potentially novel *CYP2B6-2B7* hybrid duplication in 21 individuals across various SSA populations. Using short-read data, we could not accurately estimate the breakpoints for this hybrid duplication in the genomes of these individuals or confirm whether it was similar to *CYP2B6*30*.[40]

For *CYP2A6* we identified 31 potentially novel haplotypes in 4% (40/961) of the SSA participants in the study (see **Table 4**, **Supplementary Material S2**). We determined the phase for 28 of these star alleles computationally, based on the same strategy used for inferring the aforementioned potentially novel *CYP2B6* star alleles. This included potential duplications of *CYP2A6*17* and *CYP2A6*28* observed in this study. The other three potentially novel star alleles occurred in unresolved diplotypes. From the *in silico* core variant effect predictions, rs58571639 (R311C), rs528089983 (T309I), rs554865113 (R129P), and rs558145012 (G36V) were predicted to be deleterious by all the VEP plugins used in this study, whereas rs143731390 (N438Y), rs8192730 (E419D), rs1809810 (N418D), rs2644906 (V292M), rs138978736 (Q239K), and rs72549435 (V110L) were predicted to be benign by all the tools (**Supplementary Material S2**).

Long-read-based characterization of novel *CYP2B6* and *CYP2A6* star alleles

Optimization for a *CYP2B6* Frag2 fragment (overlapping exon 1 to exons 2–3) was not successful. Therefore, it was challenging to fully characterize haplotypes with multiple novel core variants in both exon 1 and other exon(s). The targeted SMRT sequencing enabled further characterization of haplotypes 5, 8, and 11 (**Table 4**, **Figure 2**). The rs537265436 (NG_007929.1: g.20726T>C, F202L) variant which defines haplotype 5 (**Figure 2**) was replicated in targeted HiFi data from a South African participant with *CYP2B6*1/*1* background diplotype. Subvariants in the same phase-set as rs537265436 were ascertained from WhatsHap phased output and are summarized in **Figure 2**. For haplotype 8 (**Table 4**, **Figure 2**), HiFi data from a

Figure 2 Novel *CYP2B6* star alleles characterized via targeted single-molecule real-time (SMRT) sequencing. Panel (**a**) depicts a diplotype containing a novel *CYP2B6* star allele with rs537265436 (p.F202L) as the core single nucleotide variation (SNV) identified on a *CYP2B6*1* background. This haplotype was observed in a South African participant who also has a novel *CYP2B6*1* suballele as the second *CYP2B6* haplotype. Panel (**b**) shows a diplotype containing a novel *CYP2B6* star allele defined by rs3211371 (p.R487C) and the *CYP2B6*2*-defining variant rs8192709 (R22C). This haplotype was observed in a participant from Burkina Faso whose second *CYP2B6* haplotype was characterized as a novel *CYP2B6*17* suballele. Panel (**c**) depicts a diplotype containing a novel *CYP2B6* star allele defined by rs142421637 (p.R109W) on a *CYP2B6*6* background. This haplotype was observed in a South African participant whose second *CYP2B6* haplotype was characterized as a novel *CYP2B6*6* suballele.



Burkina Faso participant confirmed that rs3211371 (NG_007929.1: g.30512C>T, R487C) was in phase with the CYP2B6*2–defining variant, rs8192709 (NG_007929.1: g.5071C>T, R22C). For haplotype 11 (**Table 4**, **Figure 2**), HiFi data from a South African participant confirmed that rs142421637 (NG_007929.1: g.17856~C>T, R109W) was in phase with *CYP2B6*6*-defining

variants. For the other participants that have been successfully resequenced so far, the star alleles predicted from the short-read WGS data were concordant with the ones identified from the long-read data. Moreover, we also inferred phasing information for suballeles from the HiFi data mainly from exons 2–9.

From the targeted SMRT sequencing of the *CYP2A6* XL-PCR fragments, we further characterized *CYP2A6*54* which was identified in 2 SEB South African participants. This haplotype has rs558145012 (NG_008377.1: g.5128G>T, G36V) on a **46* background (see haplotype 1 in **Table 4**, **Figure 3**). In addition to rs558145012, we observed 4 other missense variants (**Figure 3**) on this haplotype arising due to a potential partial exon 3/intron 3 gene conversion (NG_008377.1:g.6798–6846 bp). The second *CYP2A6* haplotype for this participant was a novel **17* suballele (**Figure 3**). The second novel major *CYP2A6* star allele validated in this study is *CYP2A6*55* defined by rs114558780 (NG_008377.1: g.10126C>T, T378I) on a **1* background (see haplotype 13 in **Table 4**, **Figure 3**). The *CYP2A6*55* suballele depicted in **Figure 3** is from a South African participant. This haplotype (allele count = 13) was also identified in participants from Botswana, Nigeria, Congo, and the Gambia (**Table 4**), which adds another layer of validation in terms of its occurrence and definition. In addition, HiFi data from a South African participant enabled validation of *CYP2A6*56* defined by rs113558392 (NG_008377.1: g.9394T>G, V292G) on a **46* background (see haplotypes 16 and 17 in **Table 4**, **Figure 3**). The WGS data for this participant indicated the presence of a duplication of **56*, however, the duplicated allele was not successfully amplified for SMRT sequencing in this study. The potentially novel *CYP2A6*56* duplication had a frequency of 0.5% (allele count = 10) in SSA based on all the WGS datasets in this study.

DISCUSSION

CYP2B6 and *CYP2A6* are important pharmacogenes as genetic variation in these genes is known to impact the metabolism and response to medications, such as efavirenz and nevirapine (antiretrovirals), bupropion (antidepressant), and nicotine (major psychoactive component in cigarette smoke). These medications are important in the African context given the existing high HIV burden and the increasing prevalence of major depressive disorders, and smoking-related non-communicable diseases. In this study, we report the distribution of *CYP2B6* and *CYP2A6* star alleles across SSA based on the comprehensive analysis of 961 high coverage genomes representative of diverse populations from central, eastern, western, and southern Africa. These (short-read) data mainly include genomes generated by H3Africa projects (https://h3africa.org) and other collaborations within Africa,[13,20,21,41] and data from the 1000 Genomes Project Consortium.[12] For *CYP2B6*, we further present the distribution of the efavirenz-based predicted phenotype distribution. Our analysis also includes comparisons between the *CYP2B6* and *CYP2A6* allele distributions in SSA and other global populations. In addition, we infer 50 potentially novel African-ancestry alleles for *CYP2B6* and *CYP2A6* combined, and perform long-read-based characterization for some of these star alleles.

The known *CYP2B6* star allele distributions across SSA populations are mainly from studies among participants from Ghana,[42] Uganda,[43] Zimbabwe,[6] SEB in South Africa,[44] and the 5 SSA populations represented in the 1000 Genomes Project phase III dataset which largely comprises low coverage WGS data and whole exome sequence data.[45] For *CYP2A6*, the available frequency data for individuals of African ancestry is predominantly from African American populations, as reviewed by Tanner and Tyndale.[3] In comparison, this study presents *CYP2B6* and *CYP2A6* star allele distributions from a more diverse set of genomes from continental SSA populations, including previously understudied populations, for example, the Berom in Nigeria, Fon in Benin, Bantu speakers from Zambia, participants from Botswana, participants from Cameroon, and South African populations represented in the AWI-Gen and CBRL datasets (**Table 1**). Our findings highlight the varying allele distributions for both *CYP2B6* and *CYP2A6* (**Table 3**) across the SSA populations in this study. Some significant star allele frequency differences were also observed between populations in some neighboring African countries and/or ethnolinguistic groups within the same country which typifies the complex pharmacogenomic variation landscape observed across Africa by previous studies.[46]

As expected, *CYP2B6*6* was the most frequent decreased function star allele in SSA whereas the no function *CYP2B6*18* allele was also common but with varying frequencies. These two alleles are partly responsible for the high proportion of efavirenz PM phenotypes predicted in SSA in this study (**Figure 1**), which is consistent with previous research studies in SSA.[6,47,48] The pharmacogenetics analysis in these studies focused mainly on *CYP2B6*6* and **18*. However, based on the results in this study, genotyping only **6* and **18* among SSA populations, where *CYP2B6*9*, **20*, **22*, **29*, and **36* frequencies can be > 1%, could lead to inaccurate star allele calls and phenotype predictions. Furthermore, the *CYP2B6*6* frequency differences (**Table 3**) and presence of potential novel star alleles on a *CYP2B6*6* backbone (in addition to those on **1*, **2*, and **22* backbones; see **Table 4**) exemplify the caveats of blanket precision medicine implementation strategies. Among the *CYP2B6* star alleles with unknown/uncertain function currently catalogued by PharmVar, we only identified **11* and **33* in SSA, both of which were singletons. This was in contrast to the frequency for **11* (7.1%) inferred from previous studies across SSA in the PharmGKB *CYP2B6* reference materials (https://www.pharmgkb.org/page/cyp2b6RefMaterials), and emphasizes the importance of star allele assignment based on full haplotype information.

For *CYP2A6*, this study provides insights into the distribution of key star alleles, such as *CYP2A6*17* and **9* (associated with decreased CYP2A6 activity) across diverse continental African populations. Furthermore, the WGS data used in this study enabled detection of *CYP2A6* structural variants—including the *CYP2A6*1x2* and **46* star alleles, which are associated with greater *in vivo* nicotine metabolism.[3] *CYP2A6*46* (defined by a 58 bp gene conversion in 3′-UTR) is challenging to call from short-read WGS as the gene conversion causes read misalignments to *CYP2A7*, and it occurs in linkage disequilibrium with multiple other star alleles. StellarPGx was the only tool that enabled calling of *CYP2A6*46* in this study. The 3′-UTR gene conversion is associated with increased mRNA stability, thus contributing

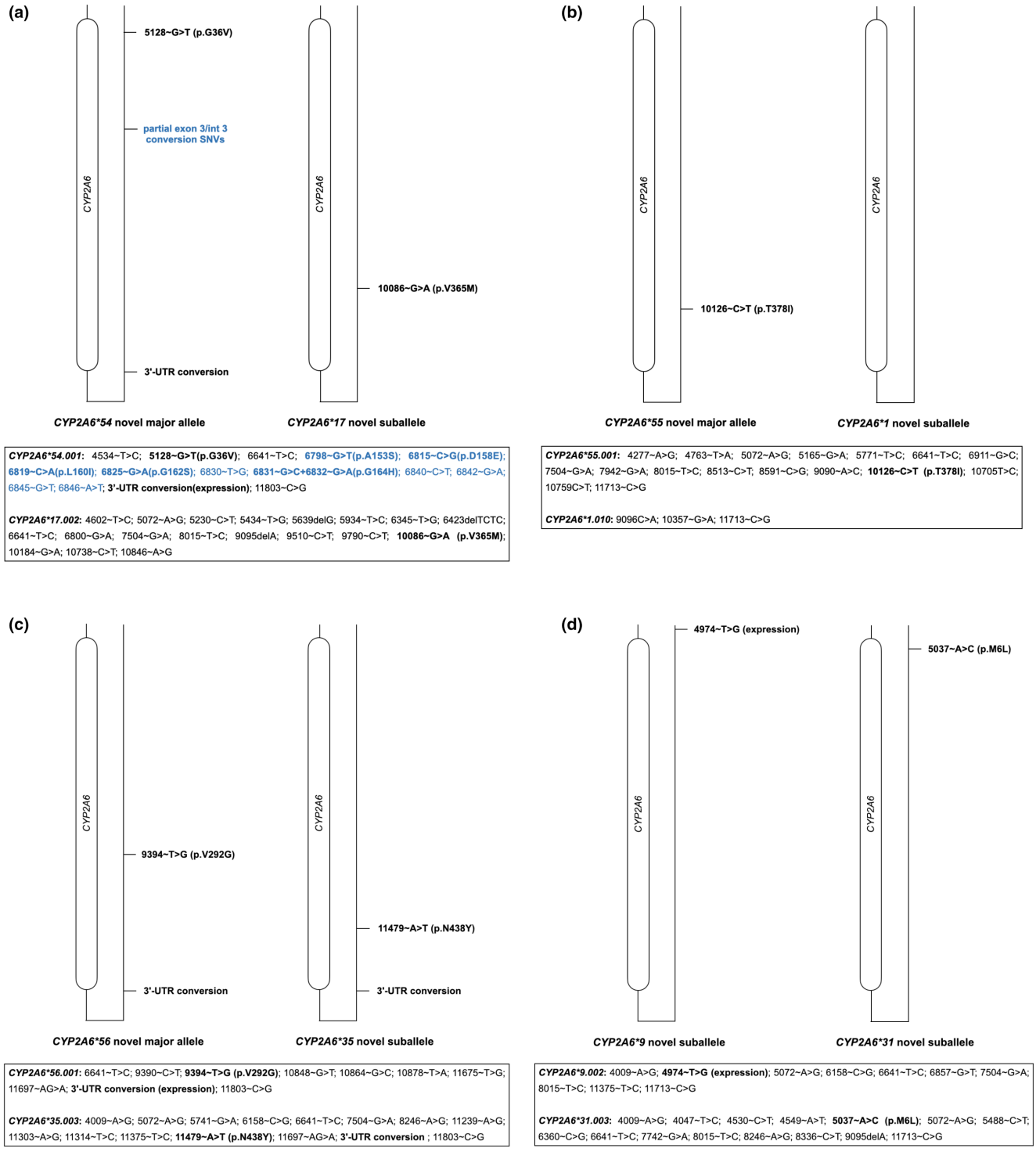


Figure 3 Novel *CYP2A6* star alleles characterized via targeted single-molecule real-time (SMRT) sequencing. Panel (**a**) depicts a *CYP2A6* diplotype observed in a South African participant with the novel *CYP2A6*54* star allele which is defined by rs558145012 (p.G36V) and core single nucleotide variations (SNVs) arising from a partial exon3/intron3 *CYP2A7* conversion, on a *CYP2A6*46* (58 bp 3′-UTR conversion to *CYP2A7*) backbone. High fidelity (HiFi) data facilitated the unambiguous read alignment spanning the entire *CYP2A6* region, including the 2 gene conversions. The second haplotype was characterized as a *CYP2A6*17* novel suballele (*CYP2A6*17.002*). Panel (**b**) depicts a *CYP2A6* diplotype observed in a South African participant with the novel *CYP2A6*55* star allele defined by rs114558780 (p.T378I) on a *CYP2A6*1* background. Panel (**c**) depicts a *CYP2A6* diplotype observed in a South African participant with the novel *CYP2A6*56* star allele defined by rs113558392 (p.V292G) on a *CYP2A6*46* background. The **56* allele appeared to be duplicated based on whole genome sequence data. However, XL-polymerase chain reaction (XL-PCR) for the duplicated gene copy was unsuccessful in this study. The second haplotype was characterized as a *CYP2A6*35* novel suballele (*CYP2A6*35.003*). Panel (**d**) depicts 2 novel *CYP2A6* suballeles (**9.002* and **31.003*) observed in a Ghanaian participant.

to increased CYP2A6 expression.[49] In general, **46* occurred at a lower frequency among SSA populations (6.1%) in comparison to other global populations (**Table 2**). For *CYP2A6*1x2* (frequency of 0.8% in SSA), we could not differentiate *CYP2A6*1x2A* from *CYP2A6*1x2B* via the tools and WGS data used in this study. *CYP2A6*1x2* results from unequal crossover involving *CYP2A6* and the neighboring *CYP2A7* pseudogene during recombination.[26] The reciprocal of this unequal crossover is the no function *CYP2A6*4* (gene deletion) allele, which we observed at frequencies as high as 6.1% in the Berom in Nigeria, and Bantu speakers from Zambia, and contrastingly lower frequencies among some of the other SSA populations (**Table 3**).

Regarding the CYP2B6 metabolizer phenotypes, we observed unique distributions for SSA compared with other global biogeographical groups (**Figure 1a**). This was mainly due to the differences in the diplotype frequencies and largely consistent with estimates in the PharmGKB *CYP2B6* reference materials.[2] Notably, the high proportions of participants with CYP2B6 poor and/or IM status observed across SSA (**Figure 1a,b**) emphasize the need for precision medicine implementation across Africa for medications that rely on CYP2B6-mediated metabolism. However, the predicted CYP2B6 phenotypes should be interpreted with caution as there are a number of non-genetic factors not investigated in this study (e.g., substrate specificity, phenoconversion, and environmental factors) that could influence the CYP2B6 phenotype. Therefore, pharmacokinetic studies in people with various CYP2B6 diplotypes in African populations are needed to determine appropriate drug dosage optimization algorithms.

This study inferred multiple potential novel African-ancestry star alleles for both *CYP2B6* and *CYP2A6* (**Table 4**). The core variants defining these star alleles were all rare and are not novel *per se*, but rather they are nonsynonymous variants that are either not currently catalogued as allele-defining by PharmVar or have been found in different combinations in our study. The functional impact of these novel star alleles is yet to be ascertained. However, *CYP2B6* novel haplotypes defined by rs373926269 (splice-donor) and rs370958436 (stop-gained), and the *CYP2A6* novel haplotype defined by rs61605570 (stop-gained; see **Table 4**) are likely to be nonfunctional as they have protein-truncating consequences. Although all the predicted novel star alleles in this study are independently relatively rare, collectively (frequency of 2% and 3% for *CYP2B6* and *CYP2A6* novel star alleles, respectively) they represent a significant challenge to pharmacogenetics strategies across SSA, if tests based only on common variants are implemented. Furthermore, the relatively high number of these previously uncharacterized haplotypes exemplify the considerable genetic diversity known to occur among African populations, including diversity in the pharmacogene variation landscape.[46,50] Regarding star allele validation, this is the first study to perform targeted SMRT sequencing to characterize novel *CYP2B6* and *CYP2A6* star alleles in an African setting. Three of the novel major *CYP2A6* star alleles (**54*, **55*, and **56*) inferred from the short-read WGS were further characterized via SMRT sequencing, as were multiple novel suballeles, and they also have been reviewed and designated by PharmVar. It is important to note that the partial exon 3/intron 3 conversion in *CYP2A6*54* can pose diplotype assignment challenges (similar to the *CYP2A6* 3′-UTR conversion) when using short-read WGS as some of the "conversion SNVs" may not be detected during variant calling—which is a result of read misalignments to *CYP2A7* (see **Supplementary Material S3**). *CYP2B6* targeted SMRT sequencing presented considerable challenges—discussed below in the limitations. However, it also provided resolution for 3 novel *CYP2B6* star alleles. The process of submitting these novel star alleles to PharmVar for naming is ongoing.

There were some limitations in this study. First, we used mostly short-read WGS data for our analysis. Therefore, computationally inferred novel star alleles for both *CYP2B6* and *CYP2A6* should be interpreted with caution given the difficulties associated with diplotyping these genes.[8,11] In the same vein, we were unable to computationally resolve *CYP2B6* diplotypes for 26 SSA participants and *CYP2A6* diplotypes for 4 SSA participants (**Supplementary Material S2**) either due to presence of novel core variants that could not be phased or due to potential uncharacterized structural variations that were novel to all the algorithms used in this study. We have reported the background star alleles for these participants and indicated the potential novel allele-defining variants (which require further experimental validation), and where possible provided sample IDs (Coriell and SGDP samples). Future NGS studies involving long-read platforms may be more informative in resolving *CYP2B6* and *CYP2A6* diplotypes for samples not further characterized in this study, and generally when analyzing variation in these complex pharmacogenes across understudied populations. Second, predicted *CYP2A6* phenotypes (relating to nicotine metabolism) could not be assigned for SSA participants in this study as the recently developed genetic risk score[10] has only been validated in African American/Afro-Caribbean individuals but not continental African populations. In addition, this genetic risk score only considers a few well-characterized *CYP2A6* star alleles and/or variants, which could be potentially misleading given the extent of novel star alleles predicted in this study. In the context of smoking initiation and cessation, we anticipate that *CYP2A6* PMs are less likely to become smokers, and if they do, they would smoke fewer cigarettes per day. For CYP2B6, the predicted phenotypes were based on efavirenz metabolism as there are no Clinical Pharmacogenetics Implementation Consortium guidelines for other drugs currently. About 7% of individuals had indeterminate phenotypes due to harboring novel/known star alleles with unknown function and/or ambiguous diplotypes. Last, in our laboratory validation, *CYP2B6* XL-PCR products for 27 participants failed to barcode during the HiFi sequencing library preparation due to technical challenges that arose from pooling the *CYP2B6* amplicons with those from *CYP2A6* (for which 28 samples failed barcoding). We mitigated this by optimizing barcoding for select individual amplicons from samples that had predicted novel star alleles or suballeles.

Future work entailing pharmacogenomics analysis based on WGS datasets from other under-represented African populations (e.g., Nilo-Saharan, Afroasiatic, and non-Bantu language families) would be critical for precision medicine across Africa and addressing disparities in comparison to global settings. Furthermore,

in-depth characterization of the *CYP2B6* and *CYP2A6* novel star alleles not validated in this study and also analyses to determine their clinical functional impact would be important in supporting clinical pharmacogenomics implementation strategies across Africa and the African diaspora. In the context of HIV treatment, it is important to assess how known and novel star alleles in *CYP2B6*, *CYP2A6*, and other pharmacogenes might affect response to new first-line drugs such as dolutegravir. Similarly, in the context of nicotine response, more studies are needed to assess the transferability of the CYP2A6 weighted genetic risk score[10] for phenotype prediction across various continental African populations, and further improve on its accuracy through inclusion of more African-ancestry star alleles.

In conclusion, this study presents an extensive characterization of the *CYP2B6* and *CYP2A6* pharmacogenetic variation in diverse SSA populations based on analysis of high-depth genomes generated by multiple projects. The differences in *CYP2B6* and *CYP2A6* star allele frequencies across SSA, and compared with other global populations, as well as the high number of potentially novel alleles in SSA emphasize the need for pharmacogenomic studies across under-represented populations for effective precision medicine implementation in Africa. Furthermore, these findings emphasize the advantage that sequencing-based strategies would present over targeted SNV genotyping tests for precision medicine purposes in genetically diverse populations. In addition, from our comparative star allele analysis we highlight potential novel *CYP2B6* and *CYP2A6* star alleles across other global populations, which is relevant in informing pharmacogenetic testing strategies worldwide.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The authors are grateful to colleagues in the various H3Africa groups who contributed to this work and to the launch of the Wits-H3Africa/GSK ADME collaboration in general. We thank Dr Philip Awadalla for providing permission to use the Benin portion of the H3Africa Baylor dataset in our study. We thank all the study participants for their generosity and invaluable contribution in making this research possible. We acknowledge the contributions of all the staff who contributed to the data and sample collections, processing, storage, and shipping in the respective primary studies that generated the high depth genomes used in this research. The high coverage 1000 Genomes Project datasets used in this study were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1. Special gratitude goes to: Gerrit Botha from the University of Cape Town for read alignment and joint calling of the H3Africa datasets; the biobank team (Busisiwe Mthembu and Natalie Smyth) at the Sydney Brenner Institute for Molecular Bioscience for their help with DNA sample processing and feedback during our XL-PCR work; Dr Andrea Gaedigk at the Children's Mercy Research Institute (Kansas City, Missouri, USA) for expert advice regarding generating XL-PCR fragments for validation of *CYP2B6* star alleles; Dr Rachel Tyndale and her team at the Department of Pharmacology and Toxicology, University of Toronto for their expert advice on generating XL-PCR fragments for validation of *CYP2A6* star alleles.

FUNDING

This study was funded by a grant from GlaxoSmithKline (GSK) to the Wits Health Consortium. GSK had no role in the study design, data collection and analysis. D.T. was partially supported by funding from the South African National Research Foundation (NRF grant number: 128895). The whole genome sequencing of the Human, Heredity and Health in Africa (H3Africa) Data was supported by a grant from the National Human Genome Research Institute, National Institutes of Health (NIH/NHGRI, Grant U54HG003273). The AWI-Gen Collaborative Center is funded by the NIH/NHGRI (Grant U54HG006938) as part of the H3Africa Consortium. M.R. is a South African Research Chair in Genomics and Bioinformatics of African Populations hosted by the University of the Witwatersrand, funded by the Department of Science and Technology, and administered by National Research Foundation of South Africa (NRF). The TrypanoGEN project was funded by the Wellcome Trust (study number 099310/Z/12/Z). The Collaborative African Genetics Network (CAfGEN) is funded by the NIH/NHGRI (Grant 1U54AI110398). The African Collaborative Center for Microbiome and Genomics Research is funded by the NIH/NHGRI (Grant U54HG006947). The primary work relating to DNA sample processing and whole genome sequencing at the Cell Biology Research Lab is based on research supported by grant awards from the Strategic Health Innovation Partnerships (SHIP) Unit of the South African Medical Research Council, a grantee of the Bill and Melinda Gates Foundation, and the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation of South Africa (84177). The opinions, findings and conclusions or recommendations expressed in this manuscript are solely the responsibility of the authors and not necessarily to be attributed to GSK, NRF, Wellcome Trust, or the NIH/NHGRI.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

D.T., B.I.D., G.E.B.W., M.P., G.A., P.R.B., C.A., M.M., G.S., M.C.S., C.T.T., M.R., Z.L., and S.H. wrote the manuscript. D.T., B.I.D., G.E.B.W., Z.L., and S.H. designed the research. D.T., Z.L., and S.H. performed the research. D.T. analyzed the data. C.A., M.M., G.S., M.C.S., C.T.T., and M.R. contributed new study datasets and materials.

DATA AVAILABILITY STATEMENT

The 1000 Genomes data (https://www.internationalgenome.org/) and the SGDP data (https://www.simonsfoundation.org/simons-genome-diversity-project/) are publicly available. The H3Africa-Baylor, CBRL, and SAHGP datasets used in this study are available from European Genome-Phenome Archive (https://ega-archive.org/) on application to the relevant Data Access Committees (EGADs: EGAD00001003791, EGAD00001006418, EGAD00001004220, EGAD00001004448, EGAD00001004505, EGAD00001004533, EGAD00001004557, EGAD00001004393, and EGAD00001007589).

1. Zanger, U.M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.* **138**, 103–141 (2013).
2. Whirl-Carrillo, M. *et al*. An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **110**, 563–572 (2021).
3. Tanner, J.-A. & Tyndale, R.F. Variation in CYP2A6 activity and personalized medicine. *J. Pers. Med.* **7**, 18 (2017).
4. Hoffman, S.M., Nelson, D.R. & Keeney, D.S. Organization, structure and evolution of the CYP2 gene cluster on human chromosome 19. *Pharmacogenetics* **11**, 687–698 (2001).
5. Gaedigk, A., Casey, S.T., Whirl-Carrillo, M., Miller, N.A. & Klein, T.E. Pharmacogene Variation Consortium: a global resource and repository for Pharmacogene variation. *Clin. Pharmacol. Ther.* **110**, 542–545 (2021).

6. Maimbo, M., Kiyotani, K., Mushiroda, T., Masimirembwa, C. & Nakamura, Y. CYP2B6 genotype is a strong predictor of systemic exposure to efavirenz in HIV-infected Zimbabweans. *Eur. J. Clin. Pharmacol.* **68**, 267–271 (2012).
7. Mwenifumbo, J.C. *et al*. Novel and established CYP2A6 alleles impair in vivo nicotine metabolism in a population of Black African descent. *Hum. Mutat.* **29**, 679–688 (2008).
8. Desta, Z. *et al*. PharmVar GeneFocus: CYP2B6. *Clin. Pharmacol. Ther.* **110**, 82–97 (2021).
9. Desta, Z. *et al*. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2B6 and efavirenz-containing antiretroviral therapy. *Clin. Pharmacol. Ther.* **106**, 726–733 (2019).
10. El-Boraie, A. *et al*. Transferability of ancestry-specific and cross-ancestry CYP2A6 activity genetic risk scores in African and European populations. *Clin. Pharmacol. Ther.* **110**, 975–985 (2021).
11. Wassenaar, C.A., Zhou, Q. & Tyndale, R.F. CYP2A6 genotyping methods and strategies using real-time and end point PCR platforms. *Pharmacogenomics* **17**, 147–162 (2016).
12. 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).
13. Choudhury, A. *et al*. High-depth African genomes inform human migration and health. *Nature* **586**, 741–748 (2020).
14. Twesigomwe, D. *et al*. StellarPGx: a Nextflow pipeline for calling star alleles in cytochrome P450 genes. *Clin. Pharmacol. Ther.* **110**, 741–749 (2021).
15. Lee, S.-B., Wheeler, M.M., Thummel, K.E. & Nickerson, D.A. Calling star alleles with stargazer in 28 pharmacogenes with whole genome sequences. *Clin. Pharmacol. Ther.* **106**, 1328–1337 (2019).
16. Hari, A. *et al*. An efficient genotyper and star-allele caller for pharmacogenomics. *Genome Res.* **33**, 61–70 (2023).
17. Twesigomwe, D. *et al*. Characterization of CYP2D6 pharmacogenetic variation in sub-Saharan African populations. *Clin. Pharmacol. Ther.* **113**, 643–659 (2022).
18. Buermans, H.P.J. *et al*. Flexible and scalable full-length CYP2D6 long amplicon PacBio sequencing. *Hum. Mutat.* **38**, 310–316 (2017).
19. The H3Africa Consortium *et al*. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).
20. Ramsay, M. *et al*. H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob Health Epidemiol Genom* **1**, e20 (2016).
21. Choudhury, A. *et al*. Whole-genome sequencing for an enhanced understanding of genetic variation among south Africans. *Nat. Commun.* **8**, 2062 (2017).
22. Mallick, S. *et al*. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
23. Lee, S., Shin, J.-Y., Kwon, N.-J., Kim, C. & Seo, J.-S. ClinPharmSeq: a targeted sequencing panel for clinical pharmacogenetics implementation. *PLoS One* **17**, e0272129 (2022).
24. Robinson, J.T., Thorvaldsdóttir, H., Wenger, A.M., Zehir, A. & Mesirov, J.P. Variant review with the integrative genomics viewer. *Cancer Res.* **77**, e31–e34 (2017).
25. Mwenifumbo, J.C., Zhou, Q., Benowitz, N.L., Sellers, E.M. & Tyndale, R.F. New CYP2A6 gene deletion and conversion variants in a population of Black African descent. *Pharmacogenomics* **11**, 189–198 (2010).
26. Rao, Y. *et al*. Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. *Mol. Pharmacol.* **58**, 747–755 (2000).
27. Breese, M.R. & Liu, Y. NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* **29**, 494–496 (2013).
28. Poplin, R. *et al*. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
29. Patterson, M. *et al*. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol.* **22**, 498–509 (2015).
30. McLaren, W. *et al*. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
31. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
32. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit7.20 (2013).
33. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
34. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).
35. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R. & Chan, A.P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688 (2012).
36. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).
37. Zhou, Y., Mkrtchian, S., Kumondai, M., Hiratsuka, M. & Lauschke, V.M. An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *Pharmacogenomics J.* **19**, 115–126 (2019).
38. Hu, J. & Ng, P.C. SIFT indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One* **8**, e77940 (2013).
39. Karczewski, K.J. *et al*. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
40. Martis, S., Mei, H., Vijzelaar, R., Edelmann, L., Desnick, R.J. & Scott, S.A. Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. *Pharmacogenomics J.* **13**, 558–566 (2013).
41. Mboowa, G. *et al*. The Collaborative African Genomics Network (CAfGEN): applying genomic technologies to probe host factors important to the progression of HIV and HIV-tuberculosis infection in sub-Saharan Africa. *AAS Open Res.* **1**, 3 (2018).
42. Klein, K. *et al*. Genetic variability of CYP2B6 in populations of African and Asian origin: allele frequencies, novel functional variants, and possible implications for anti-HIV therapy with efavirenz. *Pharmacogenet. Genomics* **15**, 861–873 (2005).
43. Mukonzo, J.K. *et al*. Pharmacogenetic-based efavirenz dose modification: suggestions for an African population and the different CYP2B6 genotypes. *PLoS One* **9**, e86919 (2014).
44. Swart, M., Skelton, M., Ren, Y., Smith, P., Takuva, S. & Dandara, C. High predictive value of CYP2B6 SNPs for steady-state plasma efavirenz levels in south African HIV/AIDS patients. *Pharmacogenet. Genomics* **23**, 415–427 (2013).
45. Auton, A. *et al*. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
46. da Rocha, J.E.B. *et al*. The extent and impact of variation in ADME genes in sub-Saharan African populations. *Front. Pharmacol.* **12**, 634016 (2021).
47. Dhoro, M. *et al*. CYP2B6*6, CYP2B6*18, body weight and sex are predictors of efavirenz pharmacokinetics and treatment response: population pharmacokinetic modeling in an HIV/AIDS and TB cohort in Zimbabwe. *BMC Pharmacol. Toxicol.* **16**, 4 (2015).
48. Nyakutira, C. *et al*. High prevalence of the CYP2B6 516G→T(*6) variant and effect on the population pharmacokinetics of efavirenz in HIV/AIDS outpatients in Zimbabwe. *Eur. J. Clin. Pharmacol.* **64**, 357–365 (2008).
49. Wang, J., Pitarque, M. & Ingelman-Sundberg, M. 3′-UTR polymorphism in the human CYP2A6 gene affects mRNA stability and enzyme expression. *Biochem. Biophys. Res. Commun.* **340**, 491–497 (2006).
50. Rajman, I., Knapp, L., Morgan, T. & Masimirembwa, C. African genetic diversity: implications for cytochrome P450-mediated drug metabolism and drug development. *EBioMedicine* **17**, 67–74 (2017).