Title: Standardized biogeographic grouping system for annotating populations in pharmacogenetic research

Authors:

Rachel Huddart[1,8], Alison E. Fohner[1,2,8], Michelle Whirl-Carrillo[1,8], Genevieve L. Wojcik[1], Christopher R. Gignoux[3], Alice B. Popejoy[1,4], Carlos D. Bustamante[1,5], Russ B. Altman[1,5,6,7] , Teri E. Klein[1,7]*

Affiliations:

[1] Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA

[2] Department of Epidemiology, University of Washington, Seattle, WA 98195, USA

[3] Division of Bioinformatics and Personalized Medicine, and Department of Biostatistics, University of Colorado, Anschutz Medical Campus, Aurora, CO 80045, USA

[4] Stanford Center for Integration of Research on Genetics and Ethics, Stanford, CA 94305, USA

[5] Department of Genetics, Stanford University, Stanford, CA 94305, USA

[6] Department of Biomedical Engineering, Stanford University, Stanford, CA 94305, USA

[7] Department of Medicine, Stanford University, Stanford, CA 94305, USA

[8] These authors contributed equally to this work

* **Correspondence:** Dr. Teri E. Klein, Shriram Center for BioE & ChemE, 443 Via Ortega, Stanford, CA 94305-4125, Phone: (650) 736-0156, feedback@pharmgkb.org

Abstract:

The frequencies of pharmacogenetic alleles vary considerably between populations which has important implications for the impact of these alleles in different populations. However, current population grouping methods to communicate these patterns are insufficient as they are inconsistent and fail to reflect the distribution of genetic variability around the world. To facilitate and standardize the reporting of global variability in pharmacogenetic allele frequencies, we present nine broad biogeographical groups, defined by global autosomal genetic structure. These groups are based on data from large-scale initiatives, including the 1000 Genomes Project and the Human Genome Diversity Project, and reflect population genetic history, genetic distances between populations, and geographical proximity, with borders falling predominantly along national boundaries to simplify application of the grouping system. The geographically-defined groups are American, Central/South Asian, East Asian, European, Near Eastern, Oceanian, and Sub-Saharan African. Because of their distinct genetic patterns and frequent inclusion in published data, we also present two admixed groups: African American/Afro-Caribbean and Latino. Here, we characterize this population grouping system in the context of broad genomic data and present its utility for annotating pharmacogenetic studies and alleles. We recognize that broadly grouping global populations is an oversimplification of human diversity, does not capture complex social and cultural identity, and uses arbitrary geographic borders. However, this grouping method is consistent with robust population genetic patterns and meets a need of the pharmacogenetics field by enabling consistent communication of the scale of variability in global allele frequencies.

44 **Report Text:**

45 Interindividual variability in pharmacogenes has important consequences for drug
46 efficacy and toxicity.[1; 2] Unlike the low frequencies of alleles that are considered actionable with
47 respect to disease risk, pharmacogenetic variants with clinical relevance are common and, in fact,
48 both presence and absence of variants provide valuable dosing information.[3; 4] The frequencies
49 of many pharmacogenetic alleles vary greatly by global population, meaning that people with
50 different ancestries can have considerably different likelihoods of carrying an allele that is
51 associated with a particular drug response. For example, the CYP3A5*3 allele has been found at
52 a frequency of 98% in an Iranian population but at 11% in a Ngoni population from Malawi. [5; 6]
53 A single value for global allele frequency would fail to reflect this pattern. Presenting the
54 differences in frequencies of pharmacogenetic alleles is important for communicating the scale
55 of their expected impact on drug response and the degree of variation between populations. This
56 information is invaluable for furthering pharmacogenetic research and implementation.

57 Many pharmacogenetic studies present allelic data for very specific populations, such as
58 from a single country or ethnic group, which are difficult to incorporate into broader research or
59 implementation. Literature curation and gene summaries, such as those from the
60 Pharmacogenomics Knowledgebase (PharmGKB: www.pharmgkb.org), must group these
61 specific populations when annotating pharmacogenetic studies to allow users to easily compare
62 information from multiple studies. As such, tagging studies with population group identifiers is
63 an important component of knowledge extraction from curated literature. These population group
64 labels then are used in aggregating and evaluating overall evidence for gene-drug associations,
65 which eventually inform clinical implementation guidelines, such as those of the Clinical
66 Pharmacogenetics Implementation Consortium (CPIC: www.cpicpgx.org).

67 Current methods for grouping global populations in pharmacogenetics are based on
68 subjective, vague, and inconsistent geographical boundaries, or on populations that are
69 geographically straightforward to cluster and reflect little admixture.[7-11] As an example of the
70 issues with current grouping methods, some studies cluster participants of Egyptian descent with
71 African populations, while others cluster them with Middle Eastern populations.[12; 13] While this
72 discrepancy illustrates inconsistencies of geographic borders, the clustering of African-descent
73 populations of the Americas with populations from Africa, as seen in the 1000 Genomes African
74 (AFR) superpopulation, provides another example of the challenges of employing a small
75 number of categories. The genetic patterns seen in American populations with African ancestry
76 differs dramatically from populations in Africa due to admixture primarily with European and
77 American Indian populations. [14-16] While sharing common ancestry, the recent admixture
78 typically observed in the Americas can complicate average allele frequency estimation or, at a
79 minimum, make these combined groupings less homogeneous.[15] These insufficient grouping
80 systems, often ad-hoc and not fully representative of what we have learned through population
81 genomic studies, create a barrier to understanding and interpreting pharmacogenetic allele
82 frequencies in a globally representative fashion.

83 Up until July 2018, PharmGKB annotated studies using the five race categories defined
84 by the US Office of Management and Budget (OMB): White, Black or African American,
85 American Indian or Alaska Native, Asian, and Native Hawaiian or Pacific Islander, with an
86 additional ethnicity OMB category of Hispanic/Latino. While PharmGKB serves as a global
87 resource, these OMB groups are US-centric and lack the capacity to truly capture the scale of
88 global human diversity. We also investigated the utility of the biogeographic categories
89 employed by the Human Genome Diversity Panel - Centre d'Etude du Polymophisme Humain

90 (HGDP - CEPH), which groups its 52 populations into Africa, Europe, Middle East, South and
91 Central Asia, East Asia, Oceania and the Americas.[7; 17; 18] These population labels work well for
92 the populations included in the HGDP data set, which are not located in ambiguous regions
93 between group borders and which mostly contain populations with little admixture. However,
94 papers curated at PharmGKB can include populations located all over the world, including in the
95 transitional zones between HGDP geographical regions and admixed populations. This leads to
96 ambiguity in how such populations would be grouped using HGDP categories. In conclusion,
97 existing systems are insufficient for capturing the diversity of study populations in a replicable
98 manner that is consistent with patterns of human genetic variation.

99 Therefore, we sought to define a grouping system of global populations that could be
100 used consistently to annotate pharmacogenetic studies and relevant alleles, and could capture
101 global human population genetic patterns. Using population genetics data sources, including the
102 1000 Genomes Phase 3 data release and the HGDP, we propose a simple and robust grouping
103 pattern based on nine broad biogeographic regions that represent major geographic regions of the
104 world (**Figure 1**). It is important to note that classifying individuals and communities into a few
105 distinct groups with defined boundaries conflicts with our understanding of human variation,
106 history, and social/cultural identities. *As a result, we respectfully present these groups as a tool*
107 *to represent broad differences in frequencies of genetic variation rather than as a classification*
108 *of human diversity*.

109 We chose this geographic clustering pattern because geography has historically been the
110 greatest predictor of genetic variation between human populations, with genetic distance
111 increasing as geographic distance increases.[19] This geographic pattern aids consistency in
112 population groupings by setting boundaries along national borders. We have also included two

113 admixed groups due to their distinct genetic patterns compared to the geographically-defined
114 groups and because they are relatively frequently described in pharmacogenetic cohorts. Nine
115 groups were selected to be enough to illustrate the broad diversity in global allele frequencies,
116 yet small enough to apply easily and to be tractable in grouping specific populations.[20-23] We
117 then used autosomal genotype data to validate the ability of these groups to capture broad genetic
118 variation and to optimally reflect global patterns. The groups are given below with their
119 abbreviations.
120
121 **Geographical populations:**
122 To simplify utility, geographic boundaries between groupings are drawn predominantly
123 along country borders, with only Russia divided into east and west along the Ural Mountains
124 boundary due to the large size and genetic heterogeneity of the country. We intend these groups
125 to represent peoples with a predominance of ancestors who were in the region pre-Diaspora and
126 pre-colonization.
127 *American (AME)*: The American genetic ancestry group includes all Native, Indigenous, and
128 Aboriginal people from both North and South America, including American Indian, Alaska
129 Native, First Nations, Inuit, and Métis in Canada, and Indigenous peoples of Central and South
130 America.
131 *Central/South Asian (SAS)*: The Central and South Asian genetic ancestry group includes
132 populations from Pakistan, Sri Lanka, Bangladesh, India, and ranges from Afghanistan to the
133 western border of China.

East Asian (EAS): The East Asian genetic ancestry group includes Japan, Korea, and China, and stretches from mainland Southeast Asia through the islands of Southeast Asia. In addition, it includes portions of central Asia and Russia east of the Ural Mountains.

European (EUR): The European genetic ancestry group includes populations of European descent, including European Americans. We define the European region as extending west from the Ural Mountains and south to the Turkish and Bulgarian border.

Near Eastern (NEA): The Near Eastern genetic ancestry group encompasses northern Africa, the Middle East, and the Caucasus. It includes Turkey and African nations north of the Saharan Desert.

Oceanian (OCE): The Oceanian genetic ancestry group includes the indigenous populations of the Pacific Islands, including Hawaii, Australia, and Papua New Guinea.

Sub-Saharan African (SSA): The Sub-Saharan African genetic ancestry group includes individuals from all nations of Sub-Saharan Africa, including Madagascar.[24]

Admixed populations:

We include two admixed groups representing populations with recent mixing of geographically-based populations and therefore, have distinct genetic patterns which are not adequately reflected by any single geographically-based group. While many populations reflect a degree of admixture, we selected these two populations because they are frequently reported in pharmacogenetic studies.

African American/Afro-Caribbean (AAC): Individuals in the African American/Afro-Caribbean genetic ancestry group reflect the extensive admixture between African, European, and Indigenous ancestries[25] and, as such, display a unique genetic profile compared to individuals

157 from each of those lineages alone. Examples within this cluster include the Coriell Institute's
158 African Caribbean in Barbados (ACB) population and the African Americans from the
159 Southwest US (ASW) population, [26] and individuals from Jamaica and the US Virgin Islands.
160 *Latino (LAT)*: The Latino genetic ancestry group is not defined by an exclusive geographic
161 region, but includes individuals of Mestizo descent, individuals from Latin America, and self-
162 identified Latino individuals in the United States. Like the African American/Afro-Caribbean
163 group, the admixture in this population creates a unique genetic pattern compared to any of the
164 discrete geographic regions, with individuals reflecting mixed Native and Indigenous American,
165 European, and African ancestry.
166
167 The Central/South Asian, East Asian and European groups presented here are equivalent
168 to the 1000 Genomes South Asian (SAS), East Asian (EAS) and European (EUR) super
169 populations, respectively. As such, we have adopted the relevant 1000 Genomes super
170 population codes as abbreviations for each of these groups to maintain consistency. While the
171 1000 Genomes Ad Mixed American (AMR) super population shows complete overlap with the
172 Latino group, we have opted to use the abbreviation LAT for this group. This removes the
173 potential for confusion between the Latino group and the other admixed group of African
174 American/Afro-Caribbean.
175 **Figure 1** illustrates the countries included in each group and removes any ambiguity of
176 the group boundaries. We intend this map to be used as a guide for grouping populations. Study
177 subjects of an ancestry that is not within the geographic cluster in which they currently live will
178 be included in the geographic cluster reflecting their ancestry. For example, South Africans of
179 Dutch descent would be included in the European cluster rather than the Sub-Saharan African

180 cluster. However, when lacking a clear description otherwise, the population will be included in
181 the group that includes its home country. This approach highlights the importance of
182 understanding and recording detailed self-identity and self-reported ethnicity in the context of
183 genetic studies.

184 To validate the genetic variability distinguished by these population groups, we
185 conducted Principal Components Analysis (PCA) of unrelated individuals from 1000 Genomes.
186 As seen in **Figure 2A**, the first two principal components (PCs) separate populations by
187 geographic region, especially along continental boundaries, and illustrate the increasing genetic
188 distance between populations of increasing geographic distance. As can be seen in the
189 overlapping PC distribution of individuals of different population groups, human genetic
190 diversity is a spectrum,[18] and therefore the geographic boundaries of these groups should be
191 understood as an obligatory divide to create relevant groupings, with the acknowledgement that
192 these borders are constrained by modern country borders and therefore are inherently arbitrary in
193 geographic space.[18] However, as shown in **Figure 2B**, only a few PCs are needed to accurately
194 predict these population clusters. Even with only 4 PCs, the minimum area under the curve
195 (AUC) for correct cluster prediction is 97.9% for most populations using multiple logistic
196 regression. The only outlier is the African American/Afro-Caribbean cluster, consistent with
197 ancestral similarity to the African cluster.[14; 27] Here still, with a larger number of PCs, the AUC
198 is above 93%, even with the observed ancestry outliers present in the 1000 Genomes African
199 Americans in the Southwest US (ASW) population.[28] While no categorization will result in
200 perfect prediction, given the spectrum of human diversity, the statistical validation of this
201 clustering from broad autosomal data makes these clusters both relevant and useful for
202 PharmGKB.

203 In **Figure 3**, we demonstrate that the groups we have selected are effective for
204 representing the diversity of global allele frequencies in pharmacogenes. We present here the
205 frequency of four single nucleotide polymorphisms (SNPs) with important pharmacogenetic
206 implications. The 'A' allele of rs1065852 is the defining SNP of the *cytochrome P450 2D6*
207 *(CYP2D6) *10* haplotype and is also found in combination with other variants in multiple
208 CYP2D6 haplotypes. Haplotypes containing this SNP are associated with decreased CYP2D6
209 activity, which has important implications for drugs that are CYP2D6 substrates, including
210 codeine, selective serotonin reuptake inhibitors, ondansetron, and tricyclic antidepressants.[29-31]
211 The *CYP2C9* alleles **2* (defined by rs1799853), **3* (defined by rs1057910), and **8* (defined by
212 rs7900194) are associated with reduced enzyme function and therefore are associated with
213 recommended changes to the dosing of warfarin and phenytoin, which are substrates of
214 CYP2C9.[32; 33] Using data from the 1000 Genomes, we show the frequency of the four SNPs in
215 these biogeographic groups. The range of frequencies globally illustrates the importance of
216 showing allele frequency by group in order to convey its impact on drug response globally.
217 The SNP rs1065852 shows stark continental patterns (**Figure 3A**). The 'A' allele is found
218 at high frequencies within East Asian populations, ranging from 66.2% in Vietnam (KHV) to
219 36.1% in Japan (JPT). This allele is less frequent in other continental populations, such as Africa
220 (3.5-16.5%), Europe (14.6-24.7%), and South Asia (10.4-25.6%). As can be seen from the range
221 of frequencies of the three *CYP2C9* alleles, the most common reduced function allele varies
222 globally, with the **8* allele much more common in African populations (1.8-7.6%) than the **2*
223 (<1%) or **3* (monomorphic in Africa) (**Figure 3B-D**). Conversely, the **8* allele is rare in
224 European populations (<1%), while **2* (8.1-15.2%) and **3* (5.6-8.4%) are more common.
225 Patterns such as this one can result in bias in the utility of dosing algorithms, such as the

226 International Warfarin Pharmacogenetics Consortium (IWPC) dosing algorithm for warfarin,
227 which adjusts dose based on the presence of the **2* and **3* alleles but does not include the **8*
228 allele.[34]

229 While individual pharmacogenetic testing (either pre-emptive or at point-of-care) remains
230 the most effective and appropriate way to implement pharmacogenetic knowledge for the care of
231 an individual,[35; 36] we recognize the need in clinical and genetic research for a standardized
232 method to broadly group populations based on biogeographic region. For example, identifying
233 populations with high frequencies of certain pharmacogenetic alleles can help to direct targeted
234 screening when resources are constrained and inform priorities for future pharmacogenetic
235 research.[19] However, the groups we present are large and the summary information presented
236 should be understood as an approximation dependent on existing studies in that region, which
237 may be limited to a few locations. As such, these groups are not suitable for use in guiding
238 specific implementation programs; rather, they should be seen as a tool for research purposes.

239 It should be noted that this grouping system does have some limitations. Classifying
240 individuals into these population groups can be complicated by social and cultural identities,
241 especially for admixed individuals,[7; 9; 37-39] and membership of an individual within one of these
242 population groups is inherently an imperfect surrogate for predicting the likelihood that the
243 individual carries a particular genetic variant.[36; 40] As can be seen in the analysis of rs1065852
244 above, the frequency of the 'A' allele can vary by up to 30% between populations which are all
245 included in the East Asian group. Furthermore, while the grouping system is based on overall
246 genome-wide average patterns, which typically follow a clinal variation pattern correlated with
247 geographic proximity,[7; 22; 23; 41; 42] variation in individual genes or individual populations do not
248 always follow these gradual patterns.[8-11; 36]

249 Because aggregate annotations of pharmacogenetic research and summary allele
250 frequencies are based only on available studies, additional studies are needed that include a
251 greater diversity of populations to make pharmacogenetic research and allele frequency
252 summaries more representative.[43] For example, the Sub-Saharan Africa (SSA) grouping
253 represents a large swath of human genomic diversity, and increased representation of these
254 populations in pharmacogenetics studies may lead to the discovery of clinical differences within
255 the larger grouping. Furthermore, large, reference genetic studies with targeted allele
256 information, like that emerging from the Population Architecture using Genomics and
257 Epidemiology (PAGE) study (www.pagestudy.org), may provide compelling evidence to adjust
258 these group boundaries based on frequency patterns specific to pharmacogenetic alleles.
259 Despite these limitations, broad population groups are needed for illustrating global
260 diversity with respect to pharmacogenetic variation and the average predicted phenotypes in
261 populations. These nine proposed biogeographic groups provide a consistent way to present
262 these data based on a system that is grounded in robust data on population genetic patterns.
263 PharmGKB has begun to use these population groups in curation activities, and we recommend
264 that these groups and accompanying map be considered the standard grouping mechanism for
265 population pharmacogenetics. Ultimately, individual pharmacogenetic testing of all patients,
266 regardless of ancestry, is needed to deliver truly personalized medicine. However, the population
267 groups we present are useful for the standardized presentation of pharmacogenetic studies, global
268 allele frequency summaries in pharmacogenetic research and broad clinical screening.
269

270 **Acknowledgments:**
271 This work was funded by NIH/NIGMS R24 GM61374 and NIH/NHGRI U01 HG007419-04.

Conflicts of Interest:

CRG owns stock in 23andMe, Inc and is a founder of and advisor to Encompass Bioscience, Inc. CDB is a member of the scientific advisory boards for Liberty Biosecurity, Personalis, 23andMe Roots into the Future, Ancestry.com, IdentifyGenomics, and Etalon and is a founder of CDB Consulting. RBA is a stockholder in Personalis Inc. and 23andMe, and a paid advisor for Youscript. Remaining authors have no conflicts of interest.

References:

1. Roden, D.M. (2006). PHarmacogenomics: Challenges and Opportunities. Annals of Internal Medicine 145.
2. Dunnenberger, H.M., Crews, K.R., Hoffman, J.M., Caudle, K.E., Broeckel, U., Howard, S.C., Hunkler, R.J., Klein, T.E., Evans, W.E., and Relling, M.V. (2015). Preemptive clinical pharmacogenetics implementation: current programs in five US medical centers. Annu Rev Pharmacol Toxicol 55, 89-106.
3. Tabor, H.K., Auer, P.L., Jamal, S.M., Chong, J.X., Yu, J.H., Gordon, A.S., Graubert, T.A., O'Donnell, C.J., Rich, S.S., Nickerson, D.A., et al. (2014). Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: implications for the return of incidental results. Am J Hum Genet 95, 183-193.
4. Wright, G.E.B., Carleton, B., Hayden, M.R., and Ross, C.J.D. (2018). The global spectrum of protein-coding pharmacogenomic diversity. The pharmacogenomics journal 18, 187-195.
5. Rahsaz, M., Azarpira, N., Nikeghbalian, S., Aghdaie, M.H., Geramizadeh, B., Moini, M., Banihashemi, M., Darai, M., Malekpour, Z., and Malekhosseini, S.A. (2012). Association between tacrolimus concentration and genetic polymorphisms of CYP3A5 and ABCB1 during the early stage after liver transplant in an Iranian population. Experimental and clinical transplantation : official journal of the Middle East Society for Organ Transplantation 10, 24-29.
6. Bains, R.K., Kovacevic, M., Plaster, C.A., Tarekegn, A., Bekele, E., Bradman, N.N., and Thomas, M.G. (2013). Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa. BMC genetics 14, 34.
7. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science (New York, NY) 298, 2381-2385.
8. Rajagopalan, R., and Fujimura, J.H. (2012). Will personalized medicine challenge or reify categories of race and ethnicity? The virtual mentor : VM 14, 657-663.
9. Gannett, L. (2005). Group Categories in Pharmacogenetics Research. Philosophy of Science 72, 1232-1247.
10. Wilson, J.F., Weale, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bradman, N., and Goldstein, D.B. (2001). Population genetic structure of variable drug response. Nat Genet 29, 265-269.

11. Race, E., and Genetics Working Group. (2005). The Use of Racial, Ethnic, and Ancestral Categories in Human Genetics Research. Am J Hum Genet 77, 519-532.
12. Relling, M.V., Gardner, E.E., Sandborn, W.J., Schmiegelow, K., Pui, C.H., Yee, S.W., Stein, C.M., Carrillo, M., Evans, W.E., and Klein, T.E. (2011). Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. Clin Pharmacol Ther 89, 387-391.
13. Scott, S.A., Sangkuhl, K., Stein, C.M., Hulot, J.S., Mega, J.L., Roden, D.M., Klein, T.E., Sabatine, M.S., Johnson, J.A., and Shuldiner, A.R. (2013). Clinical Pharmacogenetics Implementation Consortium guidelines for CYP2C19 genotype and clopidogrel therapy: 2013 update. Clin Pharmacol Ther 94, 317-323.
14. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.M., Wambebe, C., Tishkoff, S.A., et al. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proceedings of the National Academy of Sciences of the United States of America 107, 786-791.
15. Mathias, R.A., Taub, M.A., Gignoux, C.R., Fu, W., Musharoff, S., O'Connor, T.D., Vergara, C., Torgerson, D.G., Pino-Yanes, M., Shringarpure, S.S., et al. (2016). A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nature communications 7, 12522.
16. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 100, 635-649.
17. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science (New York, NY) 296, 261-262.
18. Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., and Feldman, M.W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1, e70.
19. Burchard, E.G., Ziv, E., Coyle, N., Gomez, S.L., Tang, H., Karter, A.J., Mountain, J.L., Perez-Stable, E.J., Sheppard, D., and Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. The New England journal of medicine 348, 1170-1175.
20. 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.
21. Elhaik, E., Tatarinova, T., Chebotarev, D., Piras, I.S., Maria Calo, C., De Montis, A., Atzori, M., Marini, M., Tofanelli, S., Francalacci, P., et al. (2014). Geographic population structure analysis of worldwide human populations infers their biogeographical origins. Nature communications 5, 3513.
22. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451, 998-1003.
23. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science (New York, NY) 319, 1100-1104.
24. Hurles, M.E., Sykes, B.C., Jobling, M.A., and Forster, P. (2005). The dual origin of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and paternal lineages. Am J Hum Genet 76, 894-901.

361 25. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a
362 discriminative modeling approach for rapid and robust local-ancestry inference. Am J
363 Hum Genet 93, 278-288.
364 26. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M.,
365 Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map
366 of genetic variation from 1,092 human genomes. Nature 491, 56-65.
367 27. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J.,
368 Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., et al. (2016). The Great
369 Migration and African-American Genomic Diversity. PLoS Genet 12, e1006059.
370 28. Mimno, D., Blei, D.M., and Engelhardt, B.E. (2015). Posterior predictive checks to quantify
371 lack-of-fit in admixture models of latent population structure. Proceedings of the National
372 Academy of Sciences of the United States of America 112, E3441-3450.
373 29. Bell, G.C., Caudle, K.E., Whirl-Carrillo, M., Gordon, R.J., Hikino, H., Prows, C.A., Gaedigk,
374 A., Agundez, J., Sadhasivam, S., Klein, T.E., et al. (2016). Clinical Pharmacogenetics
375 Implementation Consortium (CPIC) guideline for CYP2D6 genotype and use of
376 ondansetron and tropisetron. Clin Pharmacol Ther.
377 30. Hicks, J.K., Sangkuhl, K., Swen, J.J., Ellingrod, V.L., Muller, D.J., Shimoda, K., Bishop, J.R.,
378 Kharasch, E.D., Skaar, T.C., Gaedigk, A., et al. (2016). Clinical pharmacogenetics
379 implementation consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and
380 dosing of tricyclic antidepressants: 2016 update. Clin Pharmacol Ther.
381 31. Hicks, J.K., Bishop, J.R., Sangkuhl, K., Muller, D.J., Ji, Y., Leckband, S.G., Leeder, J.S.,
382 Graham, R.L., Chiulli, D.L., A, L.L., et al. (2015). Clinical Pharmacogenetics
383 Implementation Consortium (CPIC) Guideline for CYP2D6 and CYP2C19 Genotypes
384 and Dosing of Selective Serotonin Reuptake Inhibitors. Clin Pharmacol Ther 98, 127-
385 134.
386 32. Caudle, K.E., Rettie, A.E., Whirl-Carrillo, M., Smith, L.H., Mintzer, S., Lee, M.T., Klein, T.E.,
387 Callaghan, J.T., and Clinical Pharmacogenetics Implementation, C. (2014). Clinical
388 pharmacogenetics implementation consortium guidelines for CYP2C9 and HLA-B
389 genotypes and phenytoin dosing. Clin Pharmacol Ther 96, 542-548.
390 33. Johnson, J.A., Caudle, K.E., Gong, L., Whirl-Carrillo, M., Stein, C.M., Scott, S.A., Lee, M.T.,
391 Gage, B.F., Kimmel, S.E., Perera, M.A., et al. (2017). Clinical Pharmacogenetics
392 Implementation Consortium (CPIC) Guideline for Pharmacogenetics-Guided Warfarin
393 Dosing: 2017 Update. Clin Pharmacol Ther.
394 34. International Warfarin Pharmacogenetics, C., Klein, T.E., Altman, R.B., Eriksson, N., Gage,
395 B.F., Kimmel, S.E., Lee, M.T., Limdi, N.A., Page, D., Roden, D.M., et al. (2009).
396 Estimation of the warfarin dose with clinical and pharmacogenetic data. The New
397 England journal of medicine 360, 753-764.
398 35. Foster, M.W., Sharp, R.R., and Mulvihill, J.J. (2001). Pharmacogenetics, Race, and
399 Ethnicity: Social Identities and Individualized Medical Care. Therapeutic Drug Monitoring
400 23, 232-238.
401 36. Yen-Revollo, J.L., Auman, J.T., and McLeod, H.L. (2008). Race does not explain genetic
402 heterogeneity in pharmacogenomic pathways. Pharmacogenomics 9, 1639-1645.
403 37. Braun, L., Fausto-Sterling, A., Fullwiley, D., Hammonds, E.M., Nelson, A., Quivers, W.,
404 Reverby, S.M., and Shields, A.E. (2007). Racial categories in medical practice: how
405 useful are they? PLoS medicine 4, e271.
406 38. Ortega, V.E., and Meyers, D.A. (2014). Pharmacogenetics: implications of race and ethnicity
407 on defining genetic profiles for personalized medicine. J Allergy Clin Immunol 133, 16-
408 26.
409 39. Bamshad, M., Wooding, S., Salisbury, B.A., and Stephens, J.C. (2004). Deconstructing the
410 relationship between genetics and race. Nat Rev Genet 5, 598-609.

411 40. Urban, T.J. (2010). Race, ethnicity, ancestry, and pharmacogenetics. Mt Sinai J Med 77,
412 133-139.
413 41. Risch, N., Burchard, E., Ziv, E., and Tang, H. (2002). Categorization of humans in
414 biomedical research: genes, race and disease. Genome Biology 3.
415 42. Bamshad, M.J., Wooding, S., Watkins, W.S., Ostler, C.T., Batzer, M.A., and Jorde, L.B.
416 (2003). Human population genetic structure and inference of group membership. Am J
417 Hum Genet 72, 578-589.
418 43. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world.
419 Nature 475, 163-165.
420

421

422

Figures:

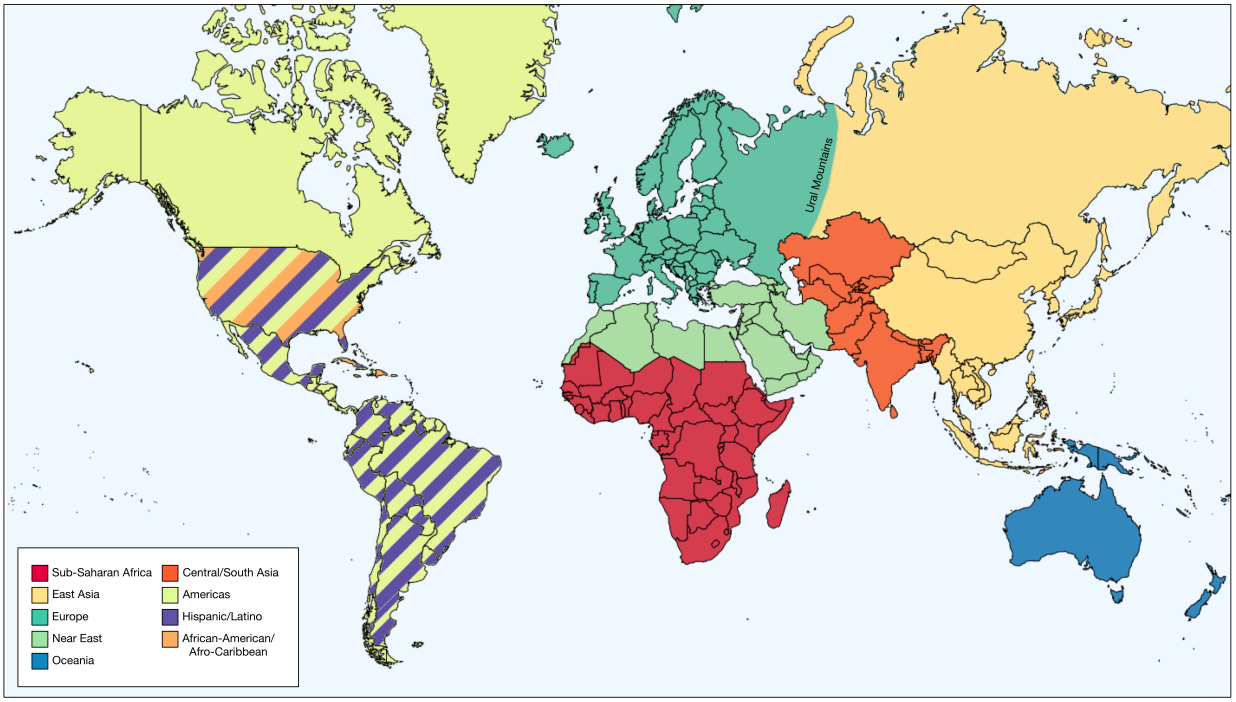

Figure 1: Map of geographical boundaries included in each genetic ancestry group. The biogeographical groups are color-coded and the group boundaries fall predominantly along national boundaries to aid the assignment of group membership.

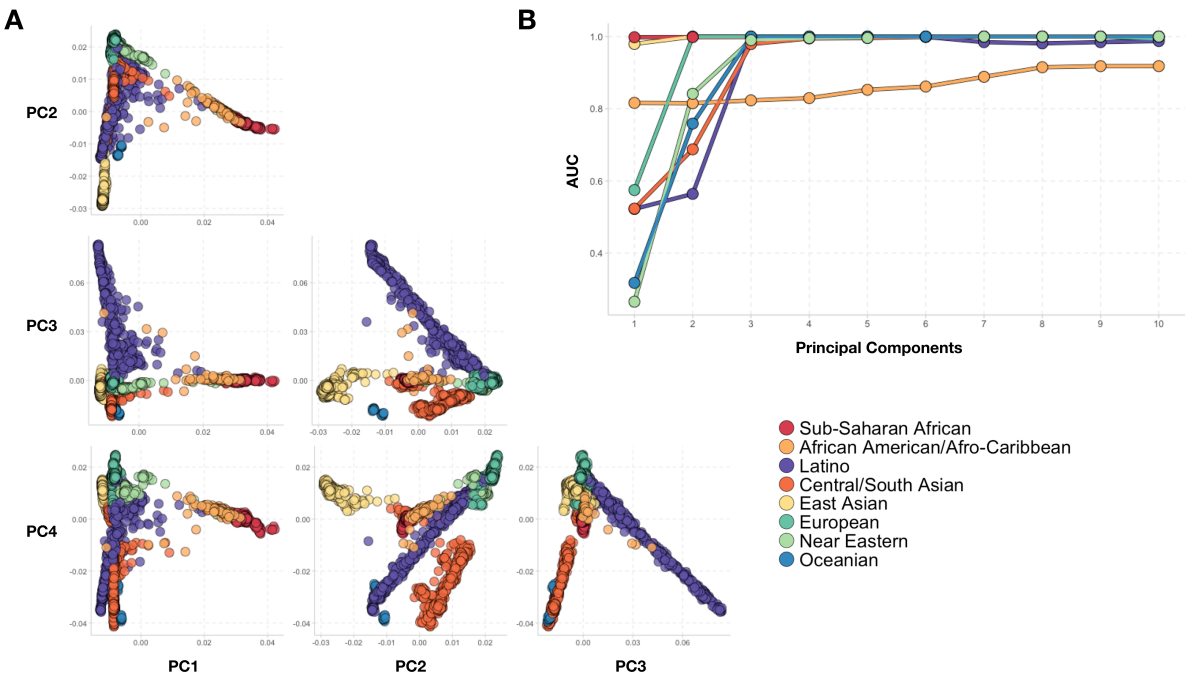

Figure 2: Principal component analysis comparing genetic distances of populations with close geographic proximity using 1000 Genomes participants. (A) The genetic gradient between populations is illustrated along PCs 1 vs 2 and PCs 3 vs 4, showing that, while completely discrete population boundaries are challenging, the groupings proposed here provide a statistically robust grouping. (B) AUCs of logistic regression to predict cluster membership, showing high degree of population structure. Note that, because none of the 1000 Genomes populations fall into the American (AME) group, no reference data were available to include this group in the analysis.

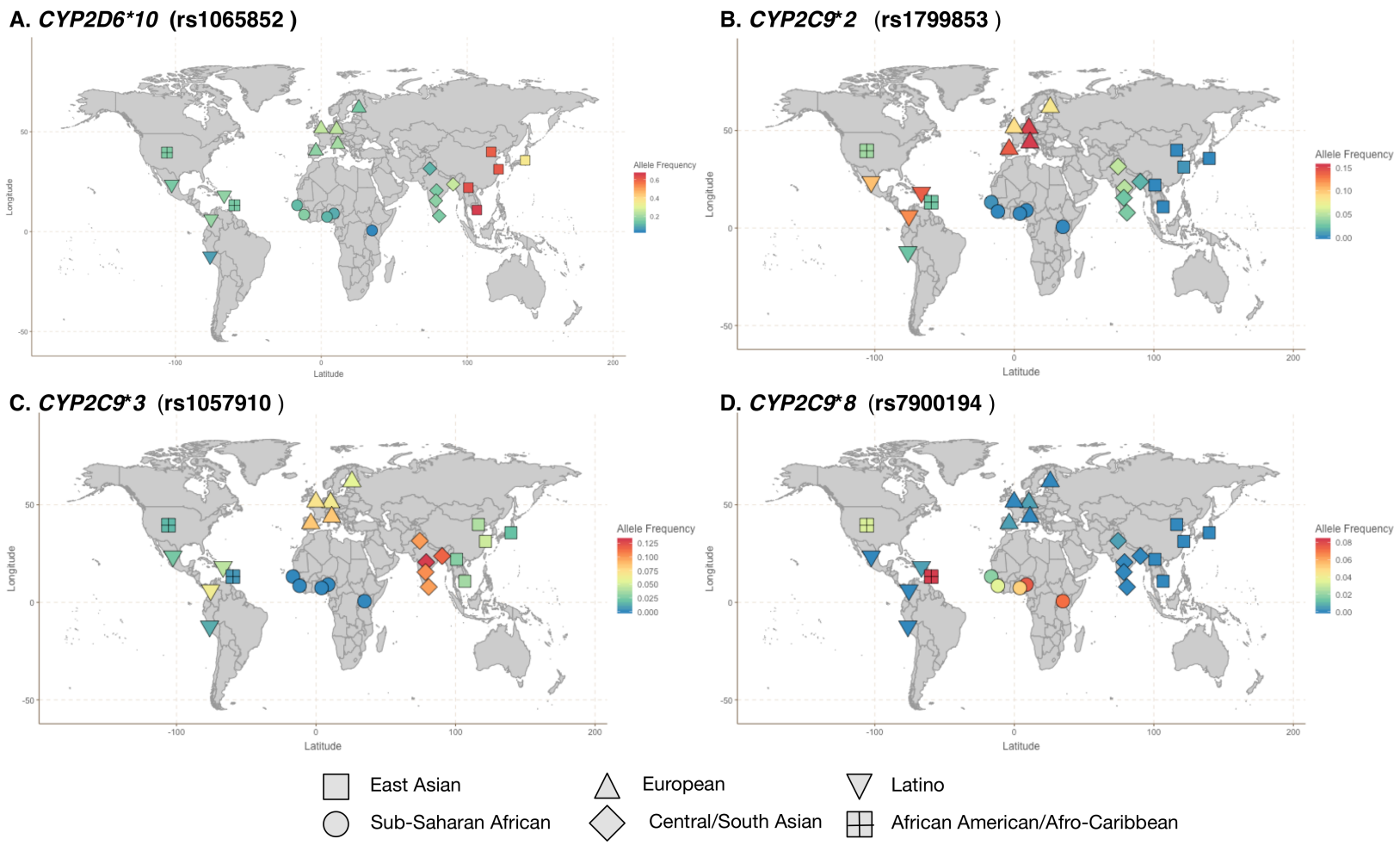

Figure 3: **Maps illustrating how the proposed biogeographical grouping system can be used to illustrate the variability in global allele frequencies.**