Title: Genomic demography of world's ethnic groups and their genomic identity between two individuals

Authors: Byung-Ju Kim[1,2], JaeJin Choi[1,3], and Sung-Hou Kim[1,2,4*]

Affiliations:

[1]Department of Chemistry and Center for Computational Biology,
University of California, Berkeley, CA, 94720, USA;
[2]Human Genome Research Center and Convergence Research Center for Insect Vectors,
Incheon National University, Incheon 22012, Republic of Korea (South Korea);
[3]Division of Biological Systems and Engineering, Lawrence Berkeley National Laboratory,
Berkeley, CA, 94720, USA;
[4]Department of Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley
National Laboratory, Berkeley, CA, 94720, USA.

*Corresponding author. Email: Sung-Hou Kim; sunghou@berkeley.edu

Abstract: A genome-based grouping of world's ethnic groups (EGs) is presented using a computational method developed for semantic analysis and categorization of books. For simplicity, we represent each individual as a "book" of all overlapping short strings of contextually-linked genotypes at all loci of Single Nucleotide Polymorphism, the most common type of individual's genomic variation. The study suggests that: (a) the EGs can be grouped into about 13 "genomic groups (GGs)"; (b) each GG consists of multiple EGs; (c) all African GGs emerged sequentially, followed by all non-African GGs, which emerged in a "burst" from the Middle East GG; and (d), at an individual level, approximately 99.8%, on average, of the whole genome contents are identical between any *two individuals* regardless of their GGs or EGs.

One Sentence Summary: About 99.8% of genome content is identical between two individuals among world's ethnic populations in 13 genomic groups.

Main Text:

Introduction

Background: "Ethnicity" or "race" is a categorization difficult to define objectively and quantitatively, yet they have been widely used to describe a group of people who have a "perceived" notion that its members share a unique set of inherited (mostly genomic) and acquired (mostly non-genomic) traits with unknown quantitative ratio between the two. Nevertheless, the presumed ethnic or race group identity has a strong psychological and emotional component that often divides people into opposing categories of "us" and "them" (*1*), one of the primary causes for human conflict and suffering. However, some objective categorizations, if properly and quantitatively defined and labeled, may have practical utilities for understanding the diverse issues in healthcare and healthpolicy, education, societal issues, demography, or others, as well as academic utility for understanding the genomic and phenotypic diversity of the modern human species and their evolutionary relationship. Therefore, the

validity and practical utility of existing categorization, such as ethnicity and race, have been revisited and debated since large numbers of human genome sequences from diverse ethnicities have become available (e.g., *2, 3*). A narrowly defined *genome-based* categorization may have a practical utility for addressing some of the issues with objectivity and quantification.

Although the non-genomic traits or attributes remain difficult to describe objectively and measure quantitatively, the genomic attributes may now be objectively quantifiable due to the rapid increase of public availability of whole genome sequence information from a large number of ethnic groups over the last 6 years. Thus, it is now possible to enquire about questions that address academic issues such as whether a genome-based categorization or grouping of extant human species is possible. Then, one can ask the evolutionary relationship among genome-based groups as well as practical issues such as genome-based health research, clinical practice and others.

There have been several studies on the first enquiry: The 1000 Genomes Project (1KGP) (*4*) revealed that there are many types of genomic variations between individual's genomes and a "reference human genome". The project showed that Single Nucleotide Polymorphism (SNP) accounts for the overwhelmingly common genomic variation, and the rest of the variations are rare or very rare at the level of one or more order-of-magnitude. Thus, as a first approximation, individual's Single Nucleotide Variations (SNVs) have been used commonly to study inherited genomic component of different ethnic individuals. Several such studies have been published for a world-wide extant population groups (e.g., 2,504 individuals from 26 large "population groups (PGs)" from diverse geographic locations in the 1KGP, for 345 individuals from 142 ethnic groups (EGs) based on genetic, linguistic, faith and cultural variation in the Simons Genome Diversity Project (SGDP) (*5*), and for 44 African ethnic populations in 4 linguistic groups (*6*). Common in these studies as well as most of other similar or related studies (e.g., *7, 8*), it has been assumed, for computational simplicity, that each SNV genotype is independent of those of its neighbors, despite experimental observations that it is, in general, at "linkage disequilibrium" with those of its neighbors in varying degrees depending on the SNP loci and population of study (e.g., see *9*).

Objectives: The purpose of this study is two-fold: (a) Using the concept of the "contextually-linked SNV genotypes (c-SNVs)" to identify "genome-based groups (GGs)" of world's EGs, where c-SNVs are represented by overlapping short strings of ordered SNV genotypes, to correlate the relationship between GGs and EGs, and to suggest the order of emergence of GGs. Such GGs will be identified for the publicly available SNP database of the SGDP, which has the genomic information for the largest number of narrowly defined EGs (142 EGs) at present; (b) to quantify the magnitudes of genomic identity/difference between two individuals from all GGs, and to interpret them to estimate, qualitatively, the extent of the relative contribution of genomic factors toward individual's GG for the SGDP samples; (c) to do the same as (b) for the 1KGP samples, which has fewer EGs divided into 26 broadly defined "Population Groups(PGs)". .

Approach: Our approach is to adapt a computational method of comparing and categorizing natural language texts or books in the field of Information Theory (*10*) by "word frequency profiles" (*11*). We generalized the method to be applicable to other linear informational "Features", such as genomic sequence, proteomic sequence, or SNV sequence. In this study we convert the whole-genome SNV genotypes of an individual into a "book" of

"Feature Frequency Profile (FFP)", where each "Feature" (which corresponds to "word" in books) consist of a *unique* c-SNV, as a "character" of the genomic feature, and its frequency in a genome, as the feature's "character state", for a given length of c-SNVs (see **Feature Frequency Profile** in Materials and Methods). Thus, an FFP of an individual's c-SNV genotypes and their frequencies contains all the information necessary to reconstruct the original sequence of the SNV genotypes. Then, for a given length of the c-SNVs, all pair-wise divergence of the FFPs for all study population can be calculated to assemble a "genomic divergence matrix" for the given c-SNV length. This matrix is used both for Principal Component Analysis (PCA) (*12*) to discover the genome-based demographic grouping (clustering) pattern of the study population (see **FFP-based PCA** in Materials and Methods) and for constructing a rooted neighbor-joining tree (see **FFP-based rooted NJ tree** and **Tree rooting** in Materials and Methods). Among many such trees, each corresponding to a given length of the c-SNVs, one with the most topologically stable tree is selected as the final tree, and the order of the emergence of GGs is estimated on an evolutionary progression scale, which corresponds to the normalized cumulative branch-lengths, from the final tree.

Results

Our results are divided into two parts: the first one focuses on, at the group level, the genome-based grouping of the population of 142 ethnicities of the SGDP and on the order of emergence of the groups based on inherited genomic information of individual's SNVs. The second part focuses, at the individual level, on the extent of genomic identity at all SNP loci between two individuals among all members of the SGDP population with the largest number of ethnic groups and, separately, of the 1KGP population, which has the largest number of samples. Then, we extrapolate the percent identity for all SNP loci to that for the respective whole genome length to get an intuitively understandable magnitude of the genomic identity/difference between two human individuals.

I. Genomic demography (grouping and the order of their emergence):

At group level, we examined the genome-based grouping patterns and their relationships among all GGs and EGs by two different methods, both based on the genomic variation expressed by the FFPs of c-SNVs: Principal Component Analysis (*12*) and neighbor-joining (NJ) phylogeny methods (*13*).

Clustering by PCA: Figure 1 shows the grouping pattern, based on clustering, revealed by PCA for all members (345 individuals) of the study population from 164 EGs in the SGDP database, where the genomes were sequenced to an average coverage of 43-fold. In the PCA, the clustering is strictly based on the genomic divergence matrix calculated using FFPs of c-SNVs without any pre-labeled group statistics. The figure shows that:

(a) There are about 13 clusters in genomic variation space (i.e., c-SNV space) which we named "Genome-based Groups" or "Genomic Groups (GGs)";
(b) All GGs are divided into two interconnected super groups: one defined by a long linear arm in Fig. 1 containing all four African GGs (GG0 – GG3*), all of which are linearly linked but

not well clustered and account for all 45 African EGs in the SGDP. The other is a more fanned-out arm containing all non-African GGs (GG4 – GG12), most of which are well separated, accounting for all of the 119 non-African EGs;

(c) All African GGs are linearly connected, with the EGs of Khomani_San and Jo_hoan at one end, and the EGs of Mozabite and Saharawi at the other, but all non-African GGs appear to originated from the Middle East GG (GG4). Most of non-African GGs are tightly-clustered and well-resolved groups;

(d) Each GG consists of multiple EGs (see Supplementary Table S1), thus there is no one-to-one correspondence between any GGs and EGs;

(e) Most of the 13 GGs can be assigned to various geographical or geological regions (see the upper inset of Fig. 1);

(f) The members of some large GGs (GG5, GG11 and GG12) are widely scattered in geographical space (see the upper inset of Fig. 1), but their genomes are closely clustered in genomic variation space (see the PCA plot of Fig. 1);

Some of the features described above (such as (b), (c) and (f) mentioned above) have been shown in the first SDGP publication (see Extended Data Fig. 4A of reference *5*). We also noticed a few unexpected observations in our PCA plot where some members of a given ethnic group are found in two geographic locations very far apart (see Supplementary Note 1).

Clading in NJ tree: Figure 2 shows the clading pattern revealed in our rooted FFP-based tree with cumulative branch-lengths (see **FFP-based rooted NJ tree** in Materials and Methods**)** by a neighbor-joining (NJ) method (*13, 14*) (for a corresponding linear topological tree with the names of EGs is shown in Supplementary Fig. S1). The genomic divergence matrix used as the input of the tree is the same as that used in the PCA clustering for Fig. 1 above, but, in the NJ tree building, two additional constraints are hypothesized for the evolutional model which influences the tree topology and clading: bifurcately branching from each internal node, and maximum parsimony (minimal evolution) when choosing the nearest neighbors to join. For comparison, all the GGs identified by the PCA clustering in Fig. 1 are also shown. Noticeable from the tree are:

(a) As was observed in Fig. 1, all African GGs (GG0 – GG3*) emerged sequentially in multiple steps, but all non-African GGs (GG4 – GG12) emerged in a burst within one giant clade nested by 4 large clades and a few small clades plus isolated branches;

(b) We can identify 9 clades (inner circular bands), 8 of which are closely related to 8 out of the 13 GGs defined by the PCA clustering (middle circular band), and;

(c) Most of the 8 clades can be assigned to various geographical or geological regions; and

(d) Most of the EGs of GG9 and GG11 appear to have emerged most recently.

In the first SDGP publication (5) they showed that the non-contextual SNP-based NJ tree using Fst values for all EGs with at least two samples (Extended Data Fig. 4B of reference *5*) revealed overall similarity in PCA plot at supergroup level and some similarity in grouping at the level of 6 continental regions. However, there are differences in tree topology, clading pattern, relative branch lengths and branching order within each region.

Order of emergence of the founders of GGs on "Evolutionary Progression Scale (EPS)": From the point of view of genomic information, the progression of evolution can be

178 considered as the process of increasing genomic divergence. To find the order of emergence of
179 GGs, we define a founder population of a GG as the selected sub-population from an internal
180 node (consisting of multiple founder populations) of our FFP-based tree, from which all extant
181 members of the GG emerged at a point during the progression of the evolution. Then, we can
182 represent the point of emergence of the founders of each GG on a "evolutionary progression
183 scale (EPS)", where EPS corresponds to a scaled cumulative branch-length of the corresponding
184 internal node of the tree, which is proportional to the genomic divergence, expressed in unitless
185 values, where EPS = 0.0 for the root of the tree and EPS = around 100 for the leaf-nodes of the
186 individuals (see Fig. 2 and **Evolutionary Progression Scale** in Materials and Methods).
187 The order of emergence can also be inferred from Fig. 1 based on the nearest-neighbor
188 relationship among the GGs (see also Supplementary Fig. S2 and Table S3), which agrees, in
189 general, with that of clades in Fig. 2. We combine the identification of GGs and their nearest-
190 neighbor relationship from the FFP-based PCA plot (Fig. 1) with EPS values for the founders of
191 all GGs from the FFP-based tree (Fig. 2) to estimate the order of the emergence of the founders
192 of all GGs (see the combined information in Supplementary Fig. S2). Such combination
193 suggests that:
194 (a). All extant African GGs (GG0 – GG3*) emerged *sequentially* between EPS of 15.2 and 29.4,
195 but any presumed GGs emerged from the root of the tree during a period corresponding to EPS
196 between 0.0 and 15.2 are extinct;
197 (b). The first extant GG is GG0 consisting of two EGs of Khomani_San and Jo_hoan (see Fig. 1
198 for the numbering of the GGs and Supplementary Table S1 for the names of EGs in each GG)
199 currently residing in the southern tip of Africa, and the last portion of African GG3* is composed
200 of the EGs of Mozabite and Saharawi in northern Africa at the end of African supergroup in Fig.
201 1;
202 (c). The founders of the non-African GGs emerged in a *"burst"* into several lineages of the GGs
203 during a period corresponding to a narrow range of EPS of 33.5 – 39.5 (see Fig. 2 and
204 Supplementary Fig. S2) from GG4 (the nearest non-African GG from the last African EGs of
205 Mozabite and Saharawi in Northern Africa), which consists of the EGs of Bedouin B, Druze,
206 Iraqi Jew, Jordanian, Palestinian, Samaritan, and Yemenite Jew in the Middle East (Levant);
207 (d). The founders of two of the largest genomic groups, GG12 and GG11, emerged most recently
208 at EPS of 39.5; and
209 (e). These observations generally agree with those observed in the PCA plot Fig. 1, where the
210 joining of the consecutive nearest neighbors (see and Supplementary Table S3) recapitulates
211 approximately the order of the emergence of GGs predicted from the branching order of the tree
212 in Fig. 2 (see also Supplementary Fig. S2).
213
214
215 **II. Average genotype identity/difference between two individuals:**
216 The genome-based grouping discussed above has been derived based on the "distance"
217 between two FFPs of *contextually-linked* SNVs in each of all pairs of individuals. But it is not
218 obvious what the "distance" between the two FFPs corresponds to in terms of two respective
219 whole genome contents. Furthermore, although 1KGP study showed (4) that the average SNV
220 difference between the reference genome of European ancestry and each individual's genome of
221 different population groups vary from 3.54M to 4.32M SNP loci, corresponding to 4.2% to 5.1%
222 of total SNP loci, it is not certain whether similar variations will be observed between two
223 individuals among *all* study populations (not comparing to the one reference genome). To get

more intuitive understanding, we ask alternative questions: (a) What is the average percent identity/differences in terms of whole genome content between two individuals from one GG vs. from two different GGs?; (b) Are the percent identity/difference unique to each GG? ; and (c) What are the implications of the answers to the above questions?

In this study genomic identity has been quantified in two steps: First, we quantify the genotype identity at all SNP loci within each pair of two individuals in each group, using the publicly available genomic data of the largest population size (the 1KGP data) and of the broadest diversity in ethnicity (the SGDP data). Then, we combine the results to generalize and extrapolate to estimate an approximate average magnitude for the genomic identity between two individuals as percent of the most recent *whole* human genome sequence with no gaps (about 3.06 billion (B) base-pairs (*15*). These steps are taken under a few approximating and simplifying assumptions (see below).

Quantifying genotype *differences* precisely is very challenging at present because of many types of differences. But we can quantify genotype *identity* more straight-forwardly to make an approximate (at the level of order-of-magnitude) estimate under a few reasonable approximations: There are many types of mutational events that result in individual genomic variations, such as single nucleotide substitutions (which result in SNP for a population), short Indels, large deletions/insertions, inversions and others. Of these, about 90% are due to SNPs, and the rest of variational events are an order-of-magnitude or more rare than SNPs (*4*). Therefore, under the first approximation that SNPs account for the overwhelming portion of all mutational events, we ignore all other mutational events, which are very rare and difficult to quantitate and compare. Thus, we estimate the extent of the individual SNV *identity* by counting identical genotypes between two individuals' genomes at all available SNP loci, then we extrapolate the SNV identity among all SNP loci to a whole genome length. This process was repeated for two data sets: the 1KG (4) and the SGDP (5). We estimate the average percent-identity in four different ways:

(a). Average genomic identity between one human reference genome and each of all individual genomes from 26 "population groups (PGs)" of the 1KGP: The first hint of the extent of the genotype difference between two individuals came from the 1KGP (*4*), where the genotypes of whole genome SNP loci of each of 2,504 individuals from 26 PGs of the 1KGP were compared with those of a single reference haploid human genome, GRCh37 (*16*), which is not of the genome of one individual, but a synthetic hybrid of multiple individuals' genomes of mostly European ancestries with only the ABO blood type of "O". Comparing each of all individual genomes to this single reference haploid genome, the study found that, on average, only about 3.58 million (M) (for non-African individuals) to 4.31M (for African individuals) SNP loci out of a total of 84.7M SNP loci (for whole "callable" genome that excludes the regions for which genomic variant calls are difficult) have genotypes different from those of the reference genome (from Table 1 of reference 4). When averaged over the 5 regional populations (Africa, Americas, East Asia, Europe, and South Asia), 3.73M SNP loci (4.4% of total 84.7M SNP loci) show individual genotypes (SNVs) different from those of the reference genome, thus, 95.6% of SNP loci have identical genotypes, on average, between the reference genome and each of all study population. At this point we make the second approximation that the "ungapped" reference genome GRCh37, which has about 2.87 B base-pairs (16), may be a reasonable proxy for the "genotype-callable" genome. Extrapolating the 95.6% of SNP loci identity within all SNP loci to the whole "ungapped"genome of the reference genome (which is about 92.5% of the

total GRCh37 sequence length, but excludes the regions such as pericentromeric and subtelomeric regions, ampliconic gene arrays, and ribosomal DNA arrays (*15, 16*)), 99.86% of the ”ungapped” whole genome of each individual, on average, have identical genotypes as those of the reference genome under the two simplifying approximations.

(b). Average genomic identity between two individuals among 26 PGs of the 1KGP: However, the above result is biased toward European genomes, because the reference genome to which each of all other individual genomes are compared is constructed from multiple European-descendant’s genomes. To avoid potential effects of the bias we calculated, for the second estimation, the genotype identity of all-to-all pairs (3,103,786 pairs) from 2492 individuals of the 1KGP **(**after excluding the reference genome and filtering out 12 kinship related samples from 2504 samples) from the 26 PGs. Table 1A shows a summary of the results averaged for each of the 26 PGs of 5 continental groups in four different ways: The average of the SNV identity between the pairs of all individuals within each of all 26 PGs equals 95.54% (with a standard deviation (SD) of 0.05%); that between one from each of African PGs and the other from the rest of PGs is 94.35% (SD 0.11%); that between one from non-African PGs and the other from the rest, is 95.35% (SD 0.14%); and that between one from all groups and the other from the rest of all groups is 95.08% (SD 0.48%). Thus, the overall SNV genotype identity between two individuals is found to be about 95.08% (range: 94.26% - 95.39% per PG), on average, of total SNP loci of 84.7 M. The narrow range of the % SNV identities indicates a very high genomic identity between two individuals regardless of whether they are from the same or different PGs. Extrapolating the % average SNV identity among all SNP loci to 2.87 B base-pairs of the whole length of the “ungapped” reference genome of GRCh37 (*16*), we get 99.87% (range: 99.83% - 99.86% per PG) as an overall genotype identity, on average, for the “ungapped” GRCh37. Almost identical % identity is found even when we use the recently updated human reference genome of GRCh38.p13, which shows a longer “total ungapped length” (*17*) and an increased number of total SNP loci (*18*)). For easy visualization of Table 1A, two examples are given: Fig. 3A1 shows the SNV % identities of the genotypes among all SNP loci between two individuals within each of the 26 PGs (blue column of Table 1A), and Fig. 3A2, as one example among many between-group comparisons, shows that between each of five EUR (European) populations and the rest of all other population groups. It is noticeable that the % genotype identity within as well as among seven African PGs is slightly lower, i.e., slightly more diverse, than those of non-African groups (Fig. 3A1), consistent with the segregation of all African GGs and all non-African GGs observed with PCA clustering in Fig. 1.

Thus, under the approximating assumptions, all pairs of two individuals in the study population have identical genotypes to each other at 99.87%, on average, of their “ungapped” genomes.

(c). Average genomic identity between two individuals among all members of 164 ethnic groups of the SGDP: So far, the above two estimates are based on the 1KGP data that covers 26 PGs, which are much broadly defined geographical/ethnic population groups than the ethnic groups of the SGDP based on genetic, linguistic and cultural variations. Thus, for our third estimation, we examined the degree of genotype identity between two individuals of the study population of the SGDP database containing 164 ethnicity-based EGs, where the genomes were sequenced to an average coverage of 43-fold, which is much deeper than that of 1KGP. However, we cannot directly compare the SGDP study (*5*) with that of the 1KGP (*4*), because in

the SGDP study the authors "retained" an average of 2.13B base-pairs per genome after filtering at "filter level 1", and identified 34.4M SNP loci (in contrast to 84.7M SNP loci out of 2.87B "ungapped length" of the reference genome (*16*) in the 1KGP) using also the GRCh37 reference human genome. Thus, we compare all-to-all pairs (59,340 pairs) of the whole genome SNP genotypes from 345 individuals in the SGDP database within the "filtered" genomes. The results of comparison of all pairs of the SGDP individuals are summarized in Table 1B. On average, 90.39% (range: 88.31% – 91.13%, in the last yellow column of Table 1B) of 34.4M SNP loci have identical genotypes. Extrapolating the average SNV identity among all SNP loci of SGDP to 2.13B base-pairs, on average, of the "filtered" whole genome represented in the SGDP data indicates that 99.84% (range: 99.81% - 99.86% per GG) of the "filtered" genome has identical genotypes, on average, between two individuals. For easy visualization of Table 1B, two examples are given: Fig, 3B1 shows the % SNV identities of the genotypes among all SNP loci between two individuals within each given GG, and Fig. 3B2, as one example among many between-group comparisons, shows that between each of GG5 (European) populations and each of the rest of all other GGs groups. It is also noticeable, as with 1KGP population, that the % SNV identity within as well as among African GGs is slightly lower, i.e., slightly more diverse, than those of non-African GGs, consistent to the segregation of two super groups of all African GGs and all non-African GDs observed with PCA clustering in Fig. 1.

Thus, under the approximating assumptions, all pairs of two individuals in this study population have, on average, 99.84% identical genotypes to each other in their "filtered" genomes.

(d). "Generalized" average genomic identity of 99.8% between two individuals of world ethnic population: Continuing advances in quality and quantity of whole genome sequencing techniques suggest that we can expect many more complete genome sequences with no gaps and no "unplaced bases". Such advance has already been apparent with the recent *complete* telomere-to-telomere sequence of about 3.06B base-pairs (without Y chromosome) of CHM13 **(***15***),** which also identified 110M SNP loci for the population of the 1KGP. Similarly, an updated GRCh38.p13 sequence shows an "ungapped" length of about 2.95B base-pairs out of an estimated 3.10B total sequence length (*17*), which also identified 111M SNP loci (*18*) for the population of the 1KGP with respect to GRCh38.p13. These observations suggest that the total whole genome sequence lengths of CHM13 and GRCh38.p13 as well as the numbers of the total SNP loci for over 3,000 individuals of the study population of the1KGP may be reaching "asymptotic" values: The total genome length of human genome with no gaps of about 3.1B base-pairs and the total size of SNP loci of about 111M. These recent values also allow us to make a *conservative* estimation that the genomic identity at the level of all populations of the 1KGP is about 96.4% of whole complete genome, i.e., (3.1B - 111M) / 3.1B.

Thus, for our fourth estimate, we use the % SNP loci identity of 95.08% between two individuals (in section (b) above for the 1KGP sample) combine with the asymptotic values of 111M for the whole genome SNP loci and of 3.1B base-pairs for whole complete genome length, and arrive at the estimation of 99.82% as the % genotype identity between two individuals.

III. Summaries and Prospects:

At the level of the population of all 164 ethnic groups in the SGDP, we show the feasibility of categorizing the population into approximately 13 genome-based groups (GGs): By using the SNP genotypes of an individual, it is now possible to classify objectively the GG to which the individual's genome is most likely to belong, rather than relying on self-declared or subjectively perceived categorization, such as race or ethnic group. Such genome-based categorization should help improve estimating or predicting the role of the inherited genomic component that may contribute to certain phenotypes specific to a given GG in the health-related fields, such as epidemiology, disease diagnosis, disease susceptibility, and clinical practice in predicting, for example, drug or therapy responses. With increasing genomic sequences with more diverse ethnicity and increased sample size, especially those in Northern Africa, Nordic Europe, North and Central Asia, and the Americas, the quality of categorization in the future is expected to improve. Furthermore, in this period of dramatic increase of massive genomic data becoming available, the critical roles being played by Information Technology, e.g., Machine Learning and Artificial Intelligence, depend on having objectively definable categorizations (called "labels" in this field) of the genomic input data for training and testing during the development of various algorithms using the technology. Such labeling also should help interpretation of genomic information to some other non-medical fields as well.

At an individual level, taking all four estimates derived above together we make a simplified and conservative approximation that the genomic identity is about 99.8% of whole genome, on average, (ranges of the four estimates above: 99.82% - 99.87%) between two individuals, i.e., the remaining 0.2% accounts for about 6M genomic loci with non-identical genotypes. This approximation is also under the final assumption that the degree of the genomic diversity of the ethnic groups in this study represents, approximately, the genomic diversity of the extant human species. The genomic identity of 99.8% between two individuals is more intuitively understandable than the genomic identity of 96.4% for the whole population of the extant human species. Possible implications derived from the observation of the uniformly very high genomic identity between two individuals regardless of categorization are discussed below.

Direct Implications and Discussions

The genome-based grouping pattern and individual genomic identity/difference described in Results are the *average* features of the grouping and quantifications for the *majority* of the ethnic and population groups at the time of this study, when the field is continuously moving toward obtaining higher quality (*18*) and completeness (*15*) of the reference human genome sequence. Not discussed in the Results are: (a) those features contributed by the "outliers" with rare degrees of genomic identity/difference between two individuals located in the "tails" of the main bodies of the "violins" in Fig. 3, and (b) those contributed by other inherited but non-genomic elements, such as epigenomic chemical modifications of genome, non-genomic but inherited nucleic acids, microbiome, and others (*19*), for which no comprehensive data are available, at present, to make quantitative estimates of their contributions to a genome-based categorization. With these caveats we describe the direct implications of the Results with discussions below and some indirect implications in Supplementary notes.

"Burst" of all extant non-African ethnic groups from the Middle East: Under the assumption that the overall divergence of the whole genome sequence of an individual in all ethnic groups increases as they evolve, our observations in part I of the Results above imply that**:** (a) the most recent ancestors of the extant ethnic groups in this study emerged sequentially from Southern Africa, and (b) all founders of the non-African ethnic groups emerged in a "burst" from the Middle Eastern genomic group, which emerged after the two northern African ethnic groups of Mozabite and Saharawi in Northern Africa emerged. What event or events may have caused such burst is unknown.

About 99.8% genomic identity for individuals and 96.4% for all population: Under the simplifying assumptions mentioned in part II of the Results above and at the level of order-of-magnitude estimation, we find that the contents of the whole genomes of any two individuals in this study population are strikingly similar (about 99.8 % identical on average (with a narrow range of 99.82% - 99.87%) not only within a given GG or PG but also between two different GGs or PGs. On the other hand, the genomic identity at all SNP loci among *all* PGs in 1KGP is about 96.4% of the whole genome of 3.1B base-pairs (see section (d) in Result **II. Average genotype identity/difference between two individuals**). This observation infers that the loci with different genotypes between two individuals in one pair is different, in general, from those in a different pair. Thus, the union of the loci with identical genotypes among *all* pairs is significantly lower than the number of loci with identical genotypes *between* two individuals, i.e., all study population as a single group has 96.4% identical genome, but 99.8% identical genome between two individuals, in general.

"Passive" genomic information: The observation of almost uniform degree of very high genomic identity (99.8%) indicates that any two extant individuals have inherited an *almost complete* set of identical genomic information, despite the multiple phenotypic differences. This apparent "inconsistency" implies that the information in inherited whole genome may be a near-complete set of *"passive" information (or potentials)* that are differentially activated by a combination of the diverse non-genomic (environmental) factors and a very small fraction (0.2% of whole genome) of individual's inherited genomic variations (see next paragraph, and **A molecular scenario for converting environmental diversity to genomic divergence** in Supplementary notes) for basic survival and reproduction of the individuals. For example, a frozen embryo does not have life, although it has the complete set of genomic information for life as well as essential proteins and other chemicals, until certain environmental signals, such as heat and essential nutrients to start its life. In this scenario the environmental signals, be they from the "biological environment" (in-utero, microbiome, ecology, food, family, society, culture, faith, ideology, lifestyles, etc.) or "non-biological environment" (heat, radiation, geology, climate, atmosphere, etc.), can be considered as critical "activating" agents for the "passive" genomic information.

Interconnection between environment and individual's genomic variations: Although most of each GG form tight genomic clusters, the members of each GG are found spread broadly within one of about 10 geological/geographical regions (S. Africa, Mid Africa, N. Africa, Middle East (Levant), Europe, S. Asia, Central Asia, Oceania, The Americas plus NE Asia, and E. plus SE. Asia) as shown in Fig. 1 and Supplementary Table S2, many of which are

definable by various geological features such as high mountain ranges, large body of water or desert, etc. For example, the members in GG5 are widely spread in Europe, but segregated from those of other non-African GGs by the Ural mountain range and Caspian sea; GG6 and GG7 (both in S. Asia) are separated by a large geological barrier, Thar desert; the members of GG12. (in E. and SE Asia) is segregated from those of the rest of Asia by Himalaya mountains on South, Tianshan mountains, Alta mountains and the Gobi desert on North. Thus, interestingly, the categorization of GGs, which are genome-based exclusively, appears to correlate with non-biological environment bordered by geological barriers. This observation implies or suggests that even the genome-based categorization of a GG may have been strongly influenced by the environmental factors due to the selection of a unique set of genomic variants from a diverse genomic population accumulated during a long environmental exposure specific to the GG.

"Environmental construct": The fact that there is no one-to-one correspondence between the 13 *explicitly* genome-based GGs and the 164 EGs (as designated based on "ethnicity" in the SGDP) or the 5 "racial groups" (Black or African Americans, White, Asian, Native Hawaiian or Other Pacific Islander, and American Indian or Alaska Native, as defined in US 2020 census based on skin color and geographical regions) (see Supplementary Table S3) indicates that the overwhelming portion of the inherited genomes, which is about 99.8% identical between two individuals, do not play a very influential role in the categorization of ethnicity or race. Furthermore, even the genome-based categorization of GGs is not only derived from a very small fraction (0.2%) of the whole genome, but also may have originated from the environmental selection of the genomic variations of the 0.2% (see previous paragraph). Thus, our results imply that current categorizations, such as ethnicity and race, can be considered as the constructs due mostly to the non-genomic, i.e. *environmental,* factors, be they biological (in-utero, microbiome, ecology, family, society, culture, ideology, lifestyles, etc.) and/or non-biological (heat, radiation, geology, climate, etc.) environmental factors.

Figure and Table legends

Figure 1. Distribution of world's ethnic groups in PCA space and in geographic space. Clustering pattern by genomic variation-based PCA plotted for the three major principal component axes. PCA was performed using the "distance matrix", where each "distance" is calculated by JS divergence between two FFPs of c-SNVs. Most of 345 individuals from 164 ethnic groups from the SGDP form about 13 clusters shown in different colors and labeled as genomic groups (GGs) GG0 – GG12, of which 12 are tight clusters, but GG3* consists of more than 5 un-resolved clusters of African ethnic groups. All un-clustered individuals are shown as gray circles. Although the members of each cluster are tightly clustered in genomic variation space, they are broadly spread out within a given geological region as shown at the geographical locations of the sample collections (see the top inset). Sorted eigenvalues and variance explained in % for the first 20 Principal Component axes are shown in the bottom inset.

Figure 2. Rooted c-SNV-based NJ tree in circular form (*21*) with the "evolutionary progression scale (EPS)". The same "distance matrix" as in Fig. 1 is used to construct the tree (see **Rooting of c-SNV-based tree** in Materials and Methods). "Evolutionary progression scale" is the scaled cumulative branch-length bounded such that EPS = 0.0 at the root of the tree and EPS = about

100, on average, at the leaf nodes. Nine clades in the tree are shown in the inner circular band. Thirteen clusters (with GG labels in red) from Fig. 1 are shown in the middle band for comparison with the 9 clades of the tree. The outer band shows the regional classification used in the SGDP. The blanks on the circular bands are for those individuals with SDGP regional assignment (outer bend), but not clustered among GG0-GG12 (middle band), or not in any of the 9 clades. A corresponding linear topological tree with the names of EGs is shown in Supplementary **Fig. S1.**

Figure 3. Violin plots of the genotype identity for all SNP loci. (A1) Violin plot of the % genotype identity of pair-wise SNVs among all SNP loci within each of 26 "population groups" of the 1KGP samples. The long streak for the distribution of % identity within each of 4 groups of the Americans reflect loose clustering of the American continental group in the 1KGP samples. The extent of the "streaking" is correlated to the SDs of each average value, which, in turn, is related to the "looseness" of each cluster (see the blue column in Table 1A). The average % identities within African PGs (about 94.4%) is slightly lower, i.e., slightly more diverse genomes, than those of non-African PGs (about 95.9%), revealing that (a) African Continental Groups (CGs) and non-African CGs form two separate super groups, and (b) the African CGs have slightly lower % genotype identity, i.e., slightly more diverse genomes than those among the members of non-African CGs within each CG. **(A2)** Violin plot of the % genotype identity of pair-wise SNVs among all SNP loci between each of 5 European groups (CEU, FIN, GBR, IBS, and TSI) and the rest. The % identity between two different CGs are slightly lower than those within the same CG. Plots for those between each of non-European groups and the rest look similar, and not shown.
(**B1**) "Violin plot" of the % identity of pair-wise SNVs among all SNP loci within each GG of the SGDP samples. The average of the averages of % identity within each of all GGs equals 91.6%, that within African GGs is 89.4%, and that within non-African GGs is 92.6%, revealing that (a) African GGs and non-African GGs form two separate super groups, and (b) the African GGs have slightly lower % genotype identity, i.e., slightly more diverse genomes than those among the members of non-African GGs within each GG. (**B2**) The % identity of pair-wise SNVs between GG5 and other GGs. Plots for those between each of non-European GGs and the rest look similar, and not shown. The % identity between two different GGs are slightly lower than those within the same GG. The long streak for the distribution of % identity between GG5 and GG3* is because GG3* is not a tight cluster but a diffuse distribution of the individuals from several loose clusters.

Table 1. Average genotype identity in % at all SNP loci between two individuals. **(A).** Average % SNV genotype identity between two individuals within each "population group" (PG) of the 1KGP population in column 1 and each of four categories on row 1. The numbers of pairs compared, average % SNV genotype identity, and standard deviations (SDs) are listed. Also listed are the minima, maxima, means, medians, and sums of each column. **(B).** Average % SNV genotype identity of the SGDP samples between two individuals within each GG in column 1 and each GG category on row 1. The averages of each column are shown in the last row.

References and Notes:

1. J. Peoples, G. A. Bailey, Humanity: An Introduction to Cultural Anthropology (11th ed.). p. 389 (2018), Cengage learning. Boston, MA
2. G. Sirugo, S. A. Tishkoff, S. M. Williams, The quagmire of race, genetic ancestry, and health disparities. *J Clin Invest.* 2021;131(11):e150255, p. 1-3 (2021). https://doi.org/10.1172/JCI150255.
3. L. N. Borrell, J. R. Elhawary, E. Fuentes-Afflick, J. Witonsky, N. Bhakta, A.H.B. Wu, K. Bibbins-Domingo, J. R. Rodriguez-Santana, M. A. Lenoir, J. R. Gavin III, R. A. Kittles, N. A. Zaitlen, D. S. Wilkes, N. R. Powe, E. Ziv, and E. G. Burchard, Race and Genetic Ancestry in Medicine— A Time for Reckoning with Racism. *N Engl J Med* 384: p.474-480 (2021). DOI: 10.1056/NEGMms2029562
4. The 1000 Genomes Project, A global reference for human genetic variation. *Nature* Vol. 52, p.68-73 (2015).
5. S. Mallick *et al*., The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*. Vol. 538, p.201-206 (2016). Updated April 12, 2017: https://www.simonsfoundation.org/simons-genome-diversity-project/ accessed July 17, 2020.
6. S. Fan, D. E. Kelly, M. H. Beltrame, M. E. B. Hansen, S. Mallick, A. Ranciaro, J. Hirbo, S. Thompson, W. Beggs, T. Nyambo, S. A. Omar, D. W. Meskel, G. Belay, A. Froment, N. Patterson, D. Reich and S. A. Tishkoff, African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology* 20:82, p.1-14 (2019), https://doi.org/10.1186/s13059-019-1679-2)
7. R. Nielsen *et al*., Tracing the peopling of the world through genomics. *Nature* 541, p.302–310 (2017), doi: 10.1038/nature21347; pmid: 28102248;
8. A. Bergström , S. A. McCarthy, R. Hui , *et al*., Insights into human genetic variation and population history from 929 diverse genomes. *Science*; 367(6484): eaay5012, p.1-10 (2020).
9. Y. Y. Teo, X. Sim, Patterns of linkage disequilibrium in different populations: implications and opportunities for lipid-associated loci identified from genome-wide association studies. *Current Opinion in Lipidology*. **21** (2): p.104–15 (2010), doi:10.1097/MOL.0b013e3283369e5b. PMID 20125009.
10. C. Shannon, A Mathematical Theory of Communication. *Bell Syst. Tech. J.* Vol.27, p.379–423 (1948).
11. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* 41, p.391–407 (1990).
12. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2017). URL https://www.R-project.org/.
13. N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), p.406–425 (1987),.
14. O. Gascuel, BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), p.685–695 (1997).
15. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, *et al.*, The complete sequence of a human genome, *bioRxiv* preprint (2021), doi: https://doi.org/10.1101/2021.05.26.445798;
16. GRCh37. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/ (retrieved March 15, 2021)

17. GRCh38. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26. (retrieved March 15, 2021)
18. M. Byrska-Bishop , U. S. Evani , X. Zhao , A. O. Basile , *et.al.*, High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv preprint (2021), doi: https://doi.org/10.1101/2021.02.06.430068
19. B. Horsthemke, A critical view on transgenerational epigenetic inheritance in humans. *Nat Commun* 9, 2973 p.1-4 (2018).
20. https://doi.org/10.1038/s41467-018-05445-5I. US census, https://www.census.gov/topics/population/race/about.html
21. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*. 47(W1), p.W256–W259 (2019),
22. G. E. Sims, S. Jun, G. A. Wu, S. Kim, Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U. S. A.* **106**, p.2677– 2682 (2009).
23. J. Lin, Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*. **37**, p.145–151 (1991).
24. D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1), p.131–147 (1981).

Acknowledgements:
We are grateful to the Simons Genome Diversity Project (https://www.simonsfoundation.org/simons-genome-diversity-project/ and https://reichdata.hms.harvard.edu/pub/datasets/sgdp/) and the 1000 Genomes Project for making their SNP data available publicly. We thank Prof. Rajiv McCoy of Johns Hopkins University for their updated information on the first complete human genome sequence of CHM13 (*15*) and Dr. Michael Zody of New York Genome Center, NY, USA on updated information about human reference genome GRCh38 (*18*). SHK acknowledges having appointments as Visiting Professorships at Yonsei University, Korea Advanced Institute of Science and Technology, and Incheon National University in South Korea during the manuscript preparation.

Funding
World Class University Project, Ministry of Education, Science and Technology, Republic of Korea.
A gift grant to University of California (41349-10805-44-CCSHK), Berkeley, CA. USA.

Author contributions

Conceptual design of the study and speculative interpretations and implications of the results by SHK; all computational work including algorithm design, programming and execution for the study by BJK and JJC; discussion and interpretation of computational

results by BJK, JJC and SHK; manuscript preparation by SHK with extensive discussions with BJK and JJC; figures designed by BJK, JJC and SHK.

Competing financial interests

Authors declare that there are no competing financial interests in connection with this paper.

Computer code availability

The codes to parse VCF files, determine/visualize PCA and display geological locations are available at https://github.com/UCB-KIMLAB/peopling, and the programs to generate JSD distance matrix are available at https://github.com/jaejinchoi/FFP, using the version 2v.3.1.

Supplementary Materials:

Lists:

Materials and Methods

Supplementary Notes

Legends for Supplementary Figures and Tables

Supplementary Figures and Tables

Materials and Methods.

SNV data source: Publicly available SNP data from the Simons Genome Diversity Project database (SGDP (*5, 6*), last accessed August of 2020) and from the One Thousand Genome Project phase 3 data (1KGP (*4*), released on May 02, 2013), last accessed June of 2020).

Feature Frequency Profile (FFP). For the purpose of introducing the concept of the contextual SNV (c-SNV), each individual's whole genome SNVs are represented by a vector, FFP (*22*), consisting of multiple millions of unique c-SNVs, as genomic characters, and their respective frequencies, as character states. Each c-SNV is a short string of linked SNVs of a given length (ℓ), which has ℓ -1 overlapping sequence of its neighbor c-SNVs. Since FFP is a collection of all overlapping c-SNVs and respective frequencies, it contains all information needed to reconstruct the original ordered SNVs, thus, an FFP is a convenient vector to represent all SNVs of an individual's whole genome for mathematical comparisons and manipulations.

Optimal feature length of c-SNVs. One of the critical parameters in FFP method is to predict the optimal length of c-SNVs, but there is no *a priori* method to find it. We therefore take an empirical approach to find the optimal Feature length by constructing a series of FFP-based trees

using c-SNVs of incrementally longer lengths and choosing the length that produces the most topologically stable tree when compared with its neighbor trees. The results for this study are shown in Supplemental Fig. S3.

FFP-based PCA. The input for the classical PCA used in this study is the pairwise "distance" matrix, where the distance is calculated by pairwise Jensen-Shannon divergence between two leaf nodes represented by two FFPs (see **Jensen-Shannon divergence and "Evolutionary distance matrix"** below)

FFP-based rooted Neighbor-Joining (NJ) tree. The neighbor-joining trees (13) in this study were built using BIONJ (*14*). Two unique differences in this study are: (a) as in PCA above, the starting point is the pair-wise "distance" matrix, where the distance is calculated by pair-wise JS divergence between two leaf nodes represented by two FFPs (see **Jensen-Shannon divergence and "Evolutionary distance matrix"** below); and (b) the tree is rooted by an artificial out-group (see **Out-group** below). Since each branch-length is proportional to the genomic divergence as a result of evolution, the cumulative branch-length between two nodes represents the extent of evolutionary progression toward a higher genomic divergence.

Jensen-Shannon (JS) divergence. The "genomic distance" between two FFPs for a given ℓ, represented by two frequency vectors P and Q for all of the equivalent c-SNVs, are calculated by the JS divergence (*23*):

$$JS_l(\mathbf{P}_l,\ \mathbf{Q}_l) = \frac{1}{2} KL(\mathbf{P}_l,\ \mathbf{M}_l) + \frac{1}{2} KL(\mathbf{Q}_l,\ \mathbf{M}_l)$$

Where M = (P + Q)/2 and *KL* is the Kullback-Leidler (KL) divergence:

$$KL(\mathbf{P}_l,\ \mathbf{M}_l) = \sum_{i=1}^{K} p_{l,i} \log_2 \frac{p_{l,i}}{m_{l,i}}$$

For vector Q, replace P and p by Q and q in the above equation, where p and q are the components of P and Q vectors.
K is the union of all unique c-SNVs in P and Q, and runs usually in millions. JS divergence has useful properties for our study: They are symmetric and bound between 0 and 1 for two identical FFPs and two completely different FFPs, respectively; they reflect the extent of genomic divergence as evolution progresses. The collection of all JS divergence provides a matrix, which we take as the "evolutionary distance matrix", and use as the starting input for PCA as well as for the construction of neighbor-joining trees.

Out-group. Since our FFP method does not require multiple sequence alignment of the linked SNVs, we construct two SNV sequences of artificial outgroup members by taking two actual SNV sequences and shuffling their sequences. Such synthetic and artificial SNV sequences would have the same length and composition of the original SNV sequences without their

contexts of the original SNVs. The rooting based on such synthetic SNVs as the outgroup allows us to identify the first emerging GG as two EGs of Khomani_San and Jo_hoan of Africa.

"Evolutionary progression scale". In Information Theory, the Jensen-Shannon (JS) divergence, bound between zero and one, is commonly used as a measure of the dissimilarity between two probability distribution of linear informational features. The FFP as the distribution of frequencies (character states) of unique c-SNVs (genomic characters) is such a probability distribution. Thus, a JS divergence of two FFPs of a given Feature length, used as a measure of the genomic-divergence distance between the two FFPs of c-SNVs, is also bound between 0 and 1, corresponding to the JS divergence value between two identical FFPs, representing two organisms with identical genome sequences, and two completely different FFPs, representing two organisms with completely different genome sequences, respectively. In addition, since the genomic divergence is defined mathematically similar to physical entropy, which increase on time progression, the JS divergence will have the same direction and magnitude of evolutionary progression, thus "evolutionary progression scale". Any whole c-SNV "dissimilarity" between two individuals accumulated during the evolution from their common ancestor can be considered as the sum of the two divergences, each expressed as cumulative branch-length from the common ancestor to respective leaf node as caused by changes of, ultimately, genomic sequences due to all types of mutations, such as point substitutions, indels, inversion, recombination, loss/gain of genes, etc. as well as other unknown mechanisms, and they will bring JS divergence somewhere between 0.0 and 1.0 depending on the degree of the sequence divergences from the most recent common ancestor.
In this study the collection of the JS divergences for all pairs of the study organisms plus 2 out-group members (see **Outgroup** in Materials and Methods) constitutes the "distance matrix" for BIONJ (*13, 14*). Since all the branch-lengths are derived from the JS divergence values, the cumulative branch-length of an internal node corresponds to "cumulative genomic divergence (CGD)" of the node along the presumed evolutionary lineage. Thus, it can be considered as the point of evolutionary stage reached by the node on an "evolutionary progression scale". For convenience of assigning the nodes on the progression scale, CGDs are scaled such that the CGD value at the root node of the tree is set to zero and the leaf nodes of the individuals to 100, on average, corresponding to the fully evolved genomic states of the individuals at sampling time, which we define as the end point of the "evolutionary progression scale" for the extant individual. (see Fig. 2, where the cumulative branch-lengths are multiplied by 200, so that the average cumulative distance between the root and the leaf nodes become about 100 of CGD, which are used to represent "Evolutionary Progression Scale").

Supplementary Note 1. Some notable unexpected observations in PCA plot:

1. There are two Bantu Tswanas, but one belongs to GG1 and the other to GG2
2. There are 4 Punjabis, one of which is in GG6 and the other three in GG7
3. There are two Daurs: Daur2 in GG12 and Daur1, which has the highest missing SNVs, as an un-clustered individual near to Hawaiian singleton.

4. Maori belongs to GG7 in India although its current geographic location is in New Zealand.
5. Tlingit2 belongs to G8, but Tlingit1 is next to unclustered Saami pair sampled in N. Finnland.
6. Most of ethnic groups in Southeast Asia including many islands cluster with G12 or GG7.

Supplementary Note 2.

Indirect implications: ***A molecular scenario for converting environmental diversity to genomic divergence***. Such a high genomic identity (99.8%) leaves only about 0.2% of the whole genome content to account for any genotype and phenotype differences between two individuals from within a given PG/GG or between two different PGs/GGs. In theory, it is possible that this 0.2% may be just random genotypes uniformly spread through all PGs/GGs. But that possibility is highly unlikely because such uniform noise would not produce 13+ genome-based distinct clusters of the study population as we have shown in Fig. 1. A more likely possibility is that, although some of this 0.2% (which corresponds to about 6M genomic loci located mostly outside of the coding regions) have little functional roles and distributed randomly, the others may be non-randomly distributed genotypes, and they participate in regulating, directly or indirectly, the transcription of a set of key genes in response to environmental signals, thus generating environment-dependent phenotype/trait variations. This possibility implies that the functionally important portion of the 0.2% of the whole genome genotype differences may play a role in connecting the environmental (non-genomic) signals, be they biological or non-biological, to the portion of the individual's genomic variations to produce phenotypic differences between the two individuals, thus, phenotypic diversity for the population. Then, some of the phenotypes are selected due to better adaptability to the given environment, and the germline genotypes of the selected individuals are inherited for the subsequent generations, and, thus, allowing genome-based categorization such as GGs based on genomic variations alone. Furthermore, this bottlenecking step of the environmentally selected subpopulation with a set of non-random genotypes from the 0.2% results in becoming the genome of the founder group of a specific GG, which gives the appearance of correlating with geological regions (see "**Interconnection between environment and individual's genomic variations"** under **Direct Implications and Discussions)**

Above scenario assumes three types of events:
(a) "**Random genomic diversification**" of the founder population of each of the all extant groups, resulting in the genomic variation of 0.2% between two individuals as the "base-line genomic noise" (for example, due to errors caused by the biological environment such as the errors during DNA replication or/and repair, or non-biological environment such as radiation and environmental mutagens), thus generating genomic diversity
(b) "**Environmental selection"** by one or more cycles of environmental selections of the phenotype variants, which carry non-random set of genotypes from the 0.2%, better suited for the environment of each GG; and
(c) "**Genotypic selection**" by the sexual reproduction among the population of the environment-selected variants (thus, non-randomly selected genotype variants) in each GG.

Multiple cycles of the three types of events would result in, ultimately, a formation of a particular GG with a set of distinct genomic variations at the level of about 0.2% of their whole genomes with apparent correlation to different geological regions.

Legends for Supplementary Figures and Tables

Supplementary Figure S1. Rooted topological tree based on c-SNVs (For rooting see **FFP-based rooted NJ tree** in Materials and Methods).

Supplementary Figure S2. Proposed sequence of emergence of GGs shown on PCA plot (**A**) and world map (**B**). (**A**) Each number in blue is the "Evolutionary progression scale" value, which is proportional to the scaled cumulative branch-length in the FFP tree (Fig. 2) at which the founders of corresponding GG emerged. The dotted arrows show speculated directions toward the next emerging GGs (based on the geological locations of the un-clustered samples between the both ends of the arrows). (**B**) Each colored polygon represents the area within which the samples from all members of a given GG were collected.

Supplementary Figure S3. Topologically stable trees. To find the range of the length of c-SNVs, ℓ, (on X-axis) that gives topologically stable FFP trees. Robinson-Foulds (RF) metrics (*24*) for two trees, one for the length of c-SNVs equals ℓ and the other for ℓ +1, are plotted on Y-axis. The RF matrix become zero at ℓ = 18 – 21, where tree topology remains stable.

Supplementary Table S1. Correspondence between genomic groups and ethnic groups. African GGs in Brown and non-African GGs in red.

Supplementary Table S2. Correspondence among 13 GGs, 164 EGs and 5 RGs.

Supplementary Table S3. Euclidean distances among the center position of each GGs in the PCA plot. The distances from GG3* are not useful, because GG3* consists of many unresolved small clusters; The distances in red identify the pairs of GGs that are closest to each other in both our PCA plot and the FFP tree.

Fig. 1

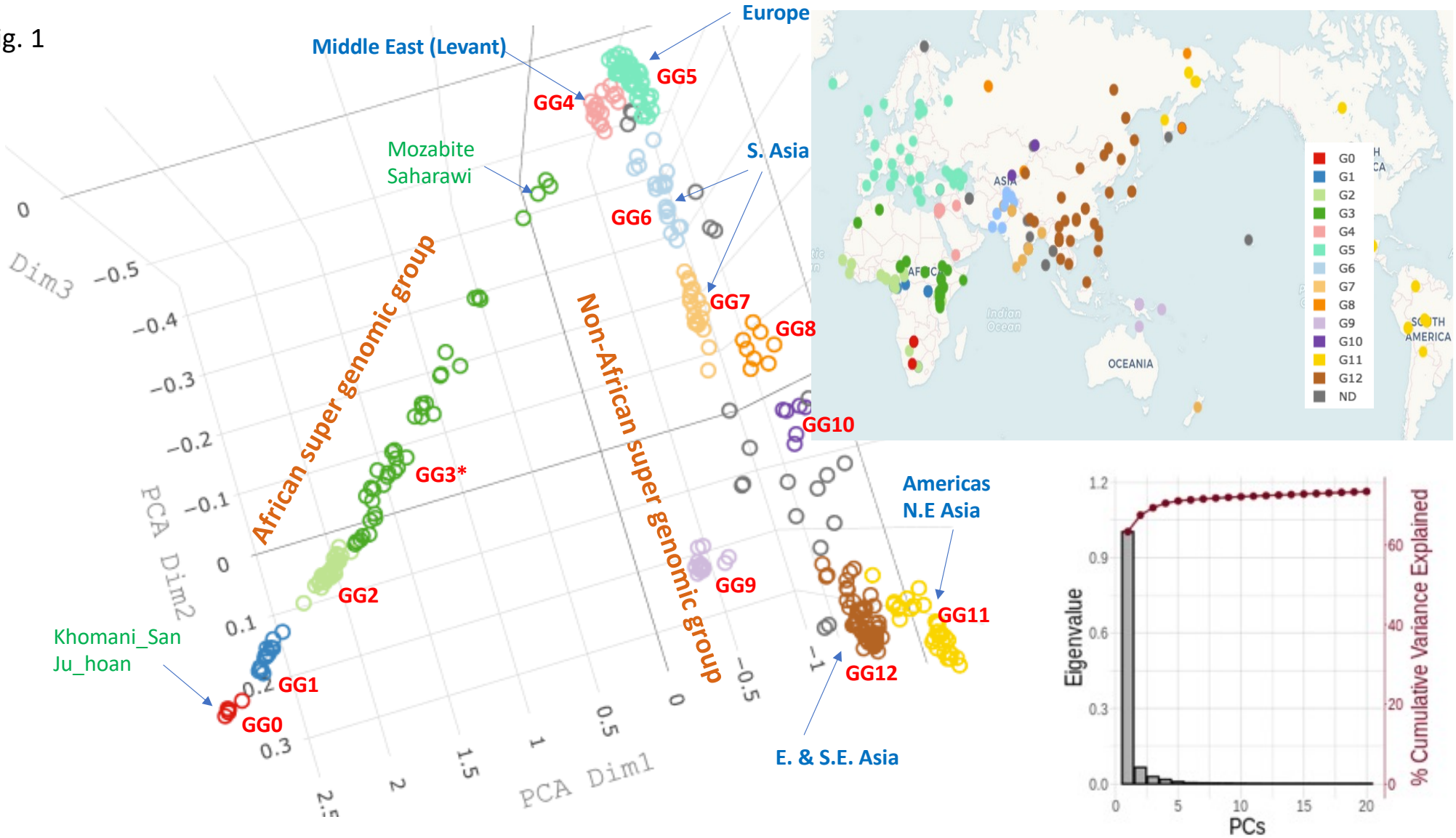

Europe
Middle East (Levant)
GG5
GG4
Mozabite
Saharawi
S. Asia
GG6
GG7
GG8
GG10
African super genomic group
Non-African super genomic group
GG3*
GG2
GG9
Americas
N.E Asia
GG11
GG12
Khomani_San
Ju_hoan
GG1
GG0
E. & S.E. Asia
PCA Dim1
PCA Dim2
Dim3
G0
G1
G2
G3
G4
G5
G6
G7
G8
G9
G10
G11
G12
ND
ASIA
OCEANIA
Indian Ocean
Eigenvalue
% Cumulative Variance Explained
PCs

Fig. 2

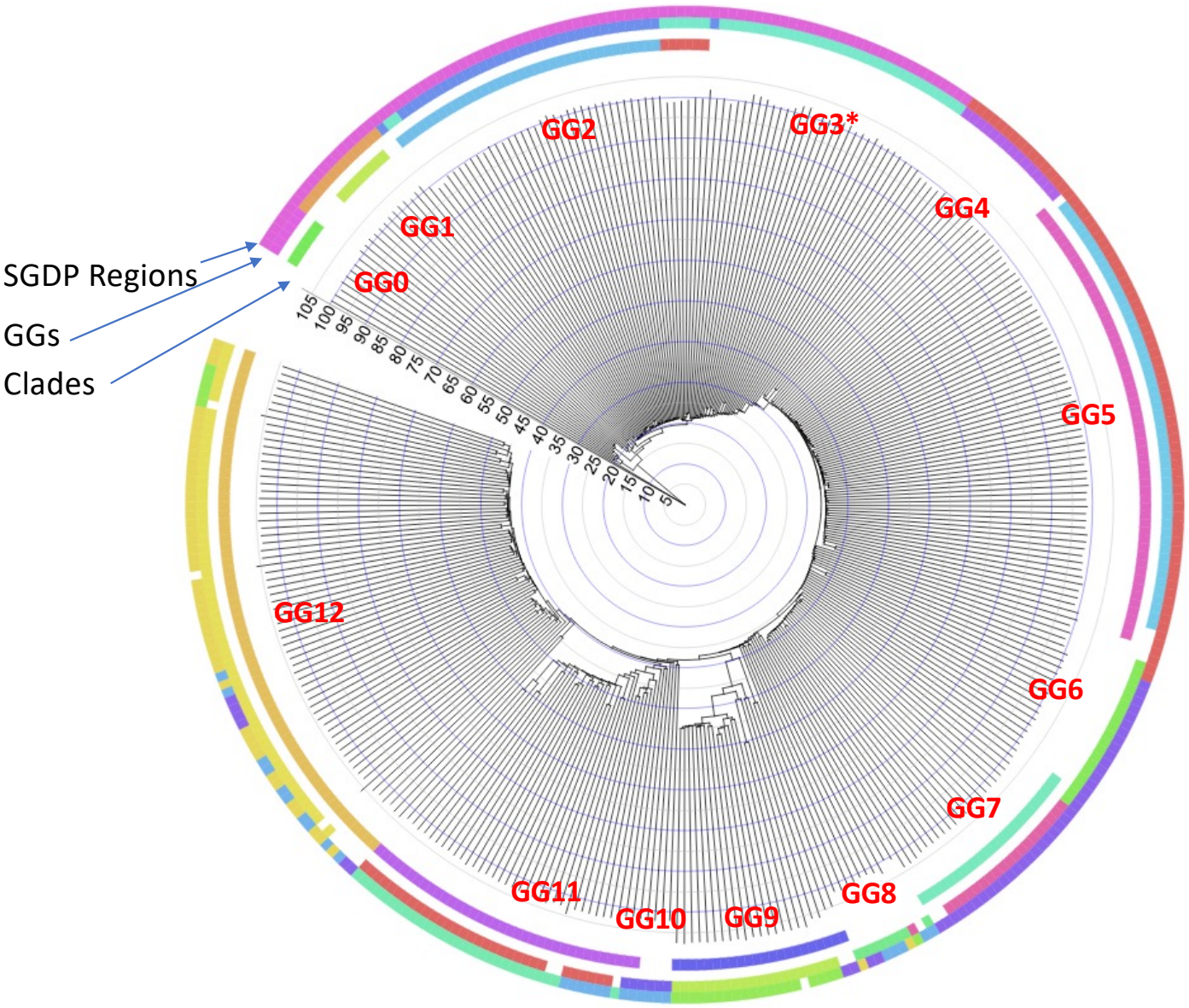
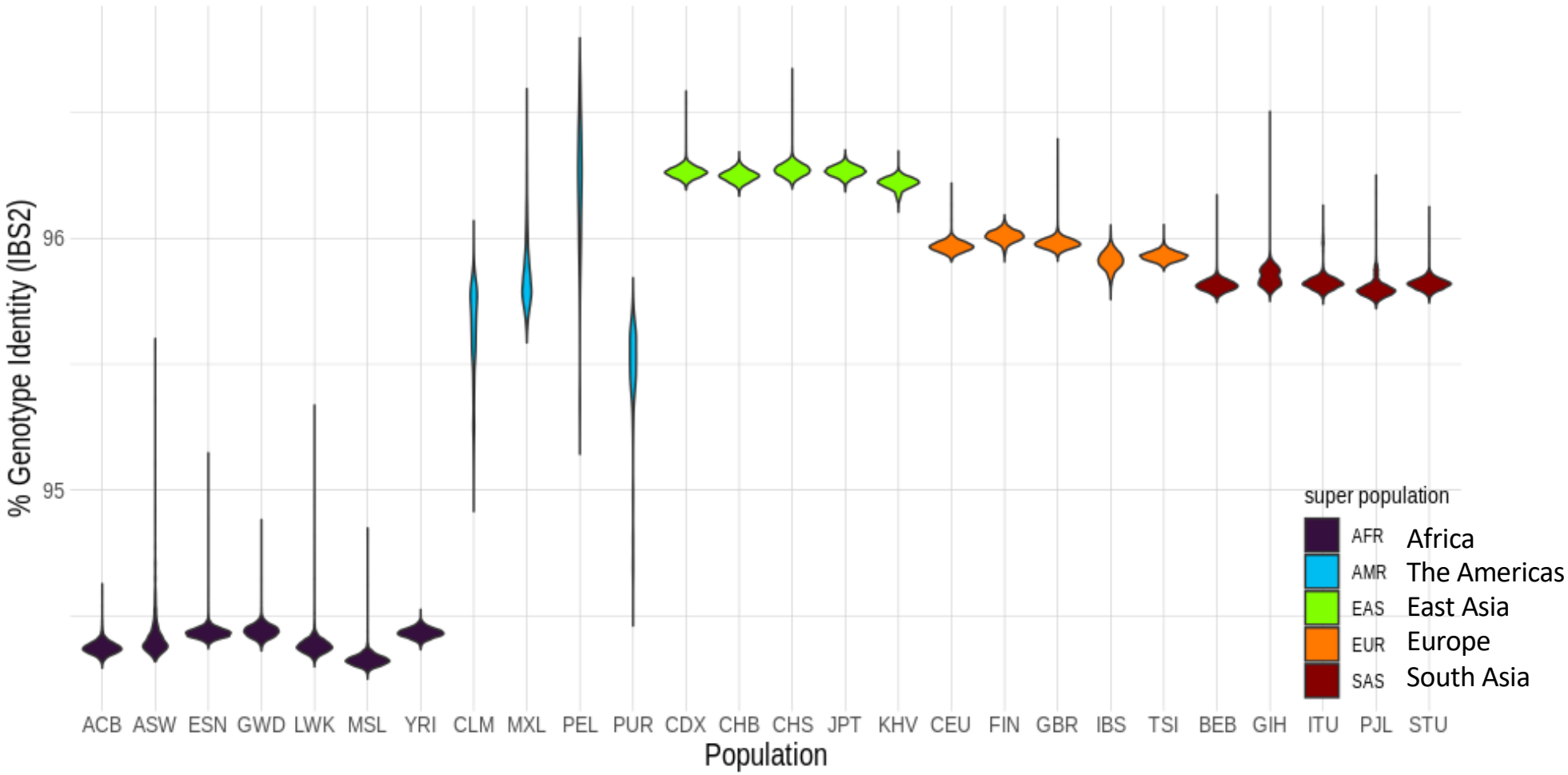
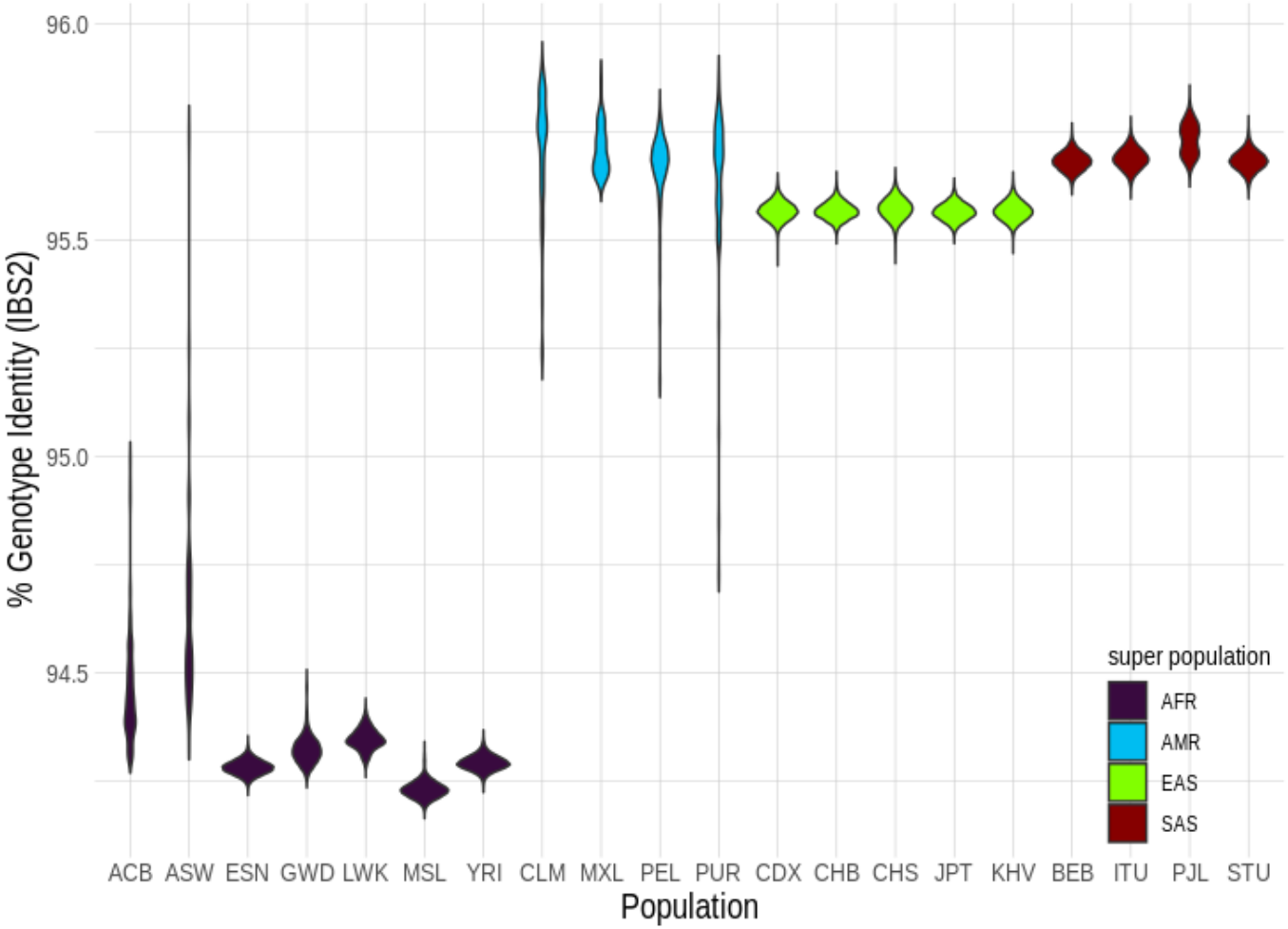

SGDP Regions
GGs
Clades
GG0
GG1
GG2
GG3*
GG4
GG5
GG6
GG7
GG8
GG9
GG10
GG11
GG12
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5

Fig. 3A1 (1KGP) Genotype identity in % of SNP loci between two individuals within each population group



For the three-alphabet names of “population” groups see Table 1A

Fig. 3A2 (1KGP) Genotype identity in % of SNP loci between Europe (EUR) and the rest of population groups

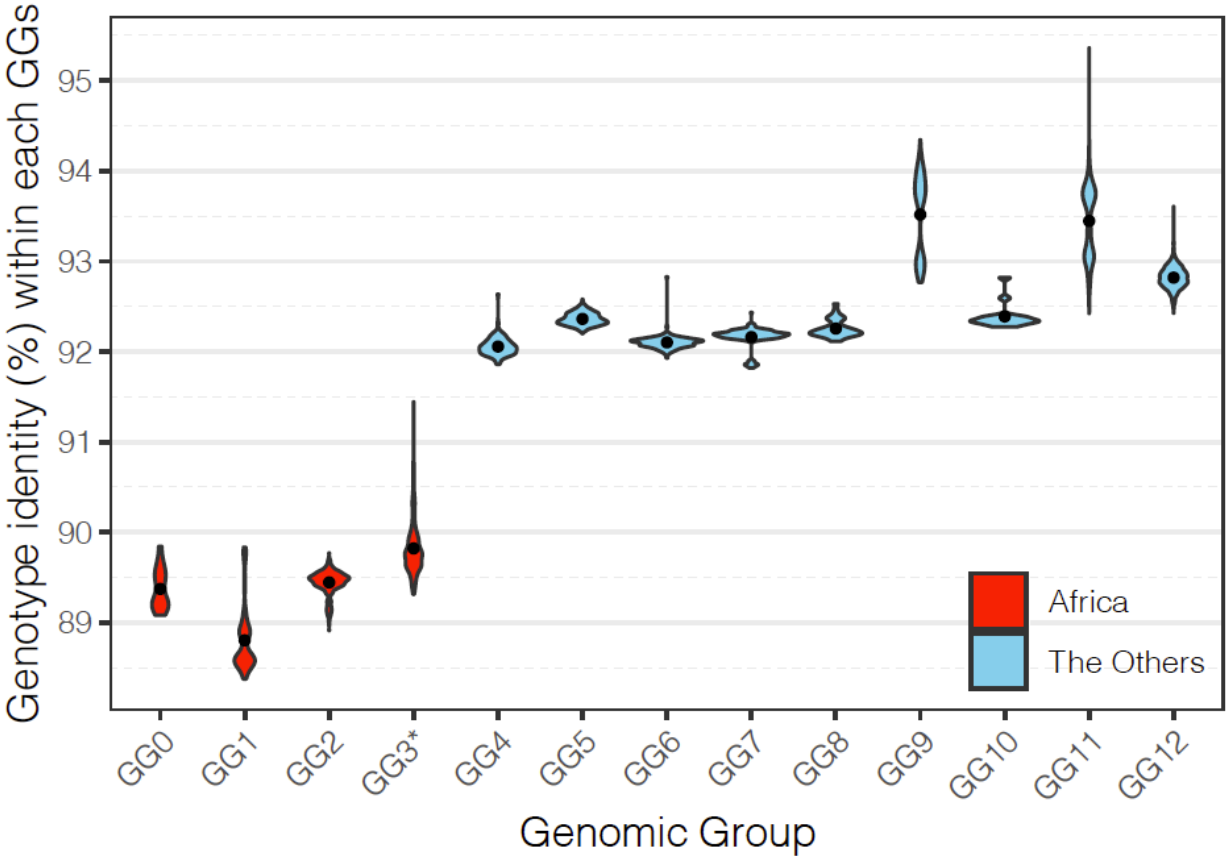


For the three-alphabet names of “population” groups see Table 1A

Genotype identity in % of SNP loci within each GG

Fig. 3B1 (SGDP)

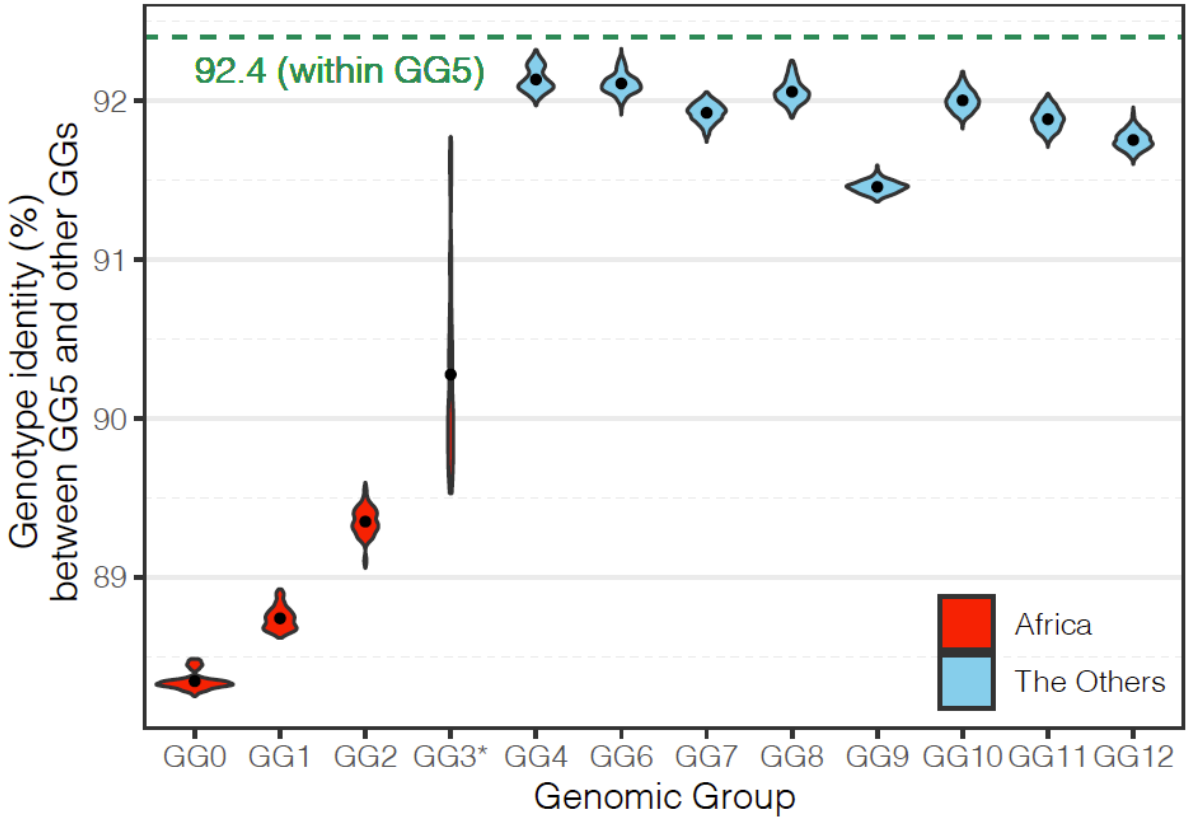
Average of the averages within each of all GGs = 91.6% (SD =1.6)
Average of the averages within each of African GGs = 89.4% (SD=0.4)
Average the averages of within each of non-African GGs = 92.6% (SD=0.6)



Genotype identity in % of SNP loci between GG5 and other GGs

Fig. 3B2 (SGDP)

Reference GG5 (within-groups) is 92.4 ± 0.1 (total 1378 pairs) on average.



Average % identity between GG5 and each of all other GGs = 91.0% (14,363 pairs)
Average % identity between GG5 and each of all African GGs = 89.2% (4,929 pairs)
Average % identity between GG5 and each of other non-African GGs = 91.9% (9,434 pairs)

Table 1A (1KGP). Average genotype identity in percent of SNP loci between two individuals

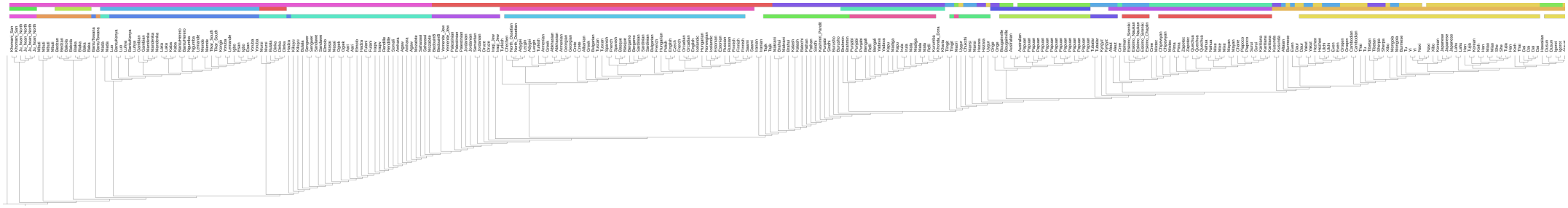
Population Information				Within Group			vs. African			vs. Non-African			vs. Rest		
Super-Population	Population	Description	Sample_Size[1]	# Pairs	% AVG	SD	# Pairs	% AVG	SD	# Pairs	% AVG	SD	# pair	% AVG	SD
AFR	ACB	African Caribbean	95	4,465	94.37	0.03	53,105	94.36	0.04	174,610	94.44	0.12	227,715	94.42	0.11
AFR	ASW	African Ancestry in Southwest US	56	1,540	94.42	0.09	33,488	94.35	0.04	102,928	94.61	0.24	136,416	94.55	0.24
AFR	ESN	Esan in Nigeria	99	4,851	94.43	0.03	54,945	94.37	0.04	181,962	94.29	0.02	236,907	94.31	0.04
AFR	GWD	Gambian in Western Division, The Gambia - Mandinka	113	6,328	94.44	0.03	61,133	94.36	0.03	207,694	94.32	0.03	268,827	94.33	0.03
AFR	LWK	Luhya in Webuye, Kenya	98	4,753	94.38	0.04	54,488	94.33	0.03	180,124	94.34	0.03	234,612	94.34	0.03
AFR	MSL	Mende in Sierra Leone	85	3,570	94.33	0.03	48,365	94.32	0.04	156,230	94.24	0.02	204,595	94.26	0.05
AFR	YRI	Yoruba in Ibadan, Nigeria	108	5,778	94.43	0.02	58,968	94.38	0.03	198,504	94.29	0.02	257,472	94.31	0.04
AMR	CLM	Colombian in Medellin, Colombia	94	4,371	95.64	0.17	61,476	94.35	0.13	163,936	95.63	0.15	225,412	95.28	0.59
AMR	MXL	Mexican Ancestry in Los Angeles, California	64	2,016	95.85	0.14	41,856	94.35	0.14	113,536	95.71	0.12	155,392	95.34	0.62
AMR	PEL	Peruvian in Lima, Peru	85	3,570	96.22	0.26	55,590	94.35	0.14	149,005	95.73	0.16	204,595	95.36	0.63
AMR	PUR	Puerto Rican in Puerto Rico	104	5,356	95.48	0.20	68,016	94.35	0.12	180,336	95.53	0.18	248,352	95.21	0.55
EAS	CDX	Chinese Dai in Xishuangbanna, China	93	4,278	96.26	0.02	60,822	94.35	0.12	162,285	95.77	0.27	223,107	95.38	0.68
EAS	CHB	Han Chinese in Beijing, China	103	5,253	96.25	0.02	67,362	94.34	0.13	178,705	95.78	0.27	246,067	95.39	0.68
EAS	CHS	Han Chinese South	105	5,460	96.27	0.03	68,670	94.35	0.13	181,965	95.78	0.27	250,635	95.39	0.68
EAS	JPT	Japanese in Tokyo, Japan	104	5,356	96.27	0.02	68,016	94.34	0.13	180,336	95.77	0.26	248,352	95.38	0.68
EAS	KHV	Kinh in Ho Chi Minh City, Vietnam	99	4,851	96.22	0.02	64,746	94.34	0.12	172,161	95.77	0.27	236,907	95.38	0.68
EUR	CEU	Utah residents (CEPH) with Northern and Western European ar	99	4,851	95.97	0.02	64,746	94.35	0.14	172,161	95.73	0.15	236,907	95.35	0.63
EUR	FIN	Finnish in Finland	99	4,851	96.01	0.02	64,746	94.35	0.14	172,161	95.75	0.13	236,907	95.37	0.63
EUR	GBR	British in England and Scotland	91	4,095	95.98	0.02	59,514	94.35	0.14	158,977	95.73	0.15	218,491	95.36	0.63
EUR	IBS	Iberian population in Spain	107	5,671	95.91	0.04	69,978	94.36	0.14	185,217	95.70	0.15	255,195	95.33	0.62
EUR	TSI	Toscani in Italy	107	5,671	95.93	0.02	69,978	94.35	0.14	185,217	95.71	0.14	255,195	95.33	0.62
SAS	BEB	Bengali in Bangladesh	86	3,655	95.81	0.02	56,244	94.34	0.13	150,672	95.71	0.10	206,916	95.34	0.62
SAS	GIH	Gujarati Indians in Houston, TX	101	5,050	95.85	0.04	66,054	94.34	0.13	175,437	95.71	0.09	241,491	95.34	0.62
SAS	ITU	Indian Telugu in the UK	101	5,050	95.82	0.03	66,054	94.34	0.13	175,437	95.70	0.10	241,491	95.33	0.61
SAS	PJL	Punjabi in Lahore, Pakistan	96	4,560	95.80	0.04	62,784	94.34	0.13	167,232	95.70	0.09	230,016	95.33	0.62
SAS	STU	Sri Lankan Tamil in the UK	100	4,950	95.82	0.02	65,400	94.34	0.13	173,800	95.70	0.10	239,200	95.33	0.61
		minimum	56.00	1540.00	94.33	0.02	33488.00	94.32	0.03	102928.00	94.24	0.02	136416.00	94.26	0.03
		maximum	113.00	6328.00	96.27	0.26	69978.00	94.38	0.14	207694.00	95.78	0.27	268827.00	95.39	0.68
		mean	95.85	4623.08	95.54	0.05	60251.69	94.35	0.11	169254.92	95.35	0.14	229506.62	95.08	0.48
		median	99.00	4851.00	95.84	0.03	62130.00	94.35	0.13	174205.00	95.71	0.13	236907.00	95.33	0.62
		sum	2492	120200			1566544			4400628			5967172		

1 some samples in family relationship (<= 2nd kinship) were filtered out in the analysis.

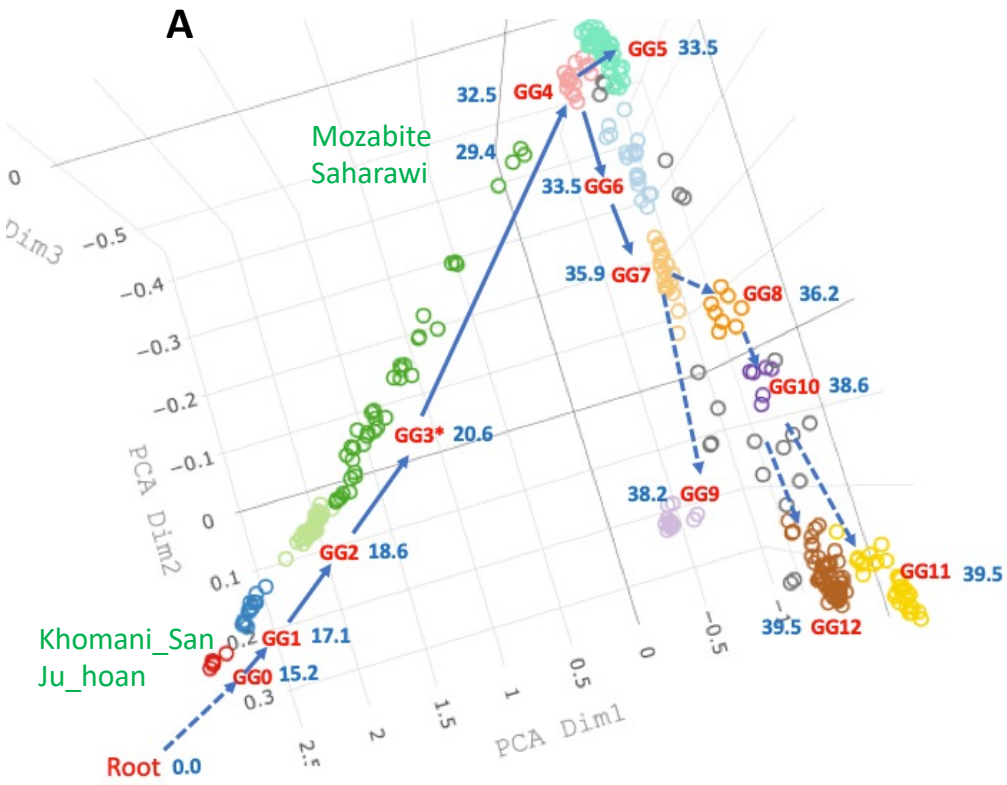
Table 1B (SGDP). Average genotype identity in % of SNP loci between each GG in column 1 and GGs on row 1 (continental groups)

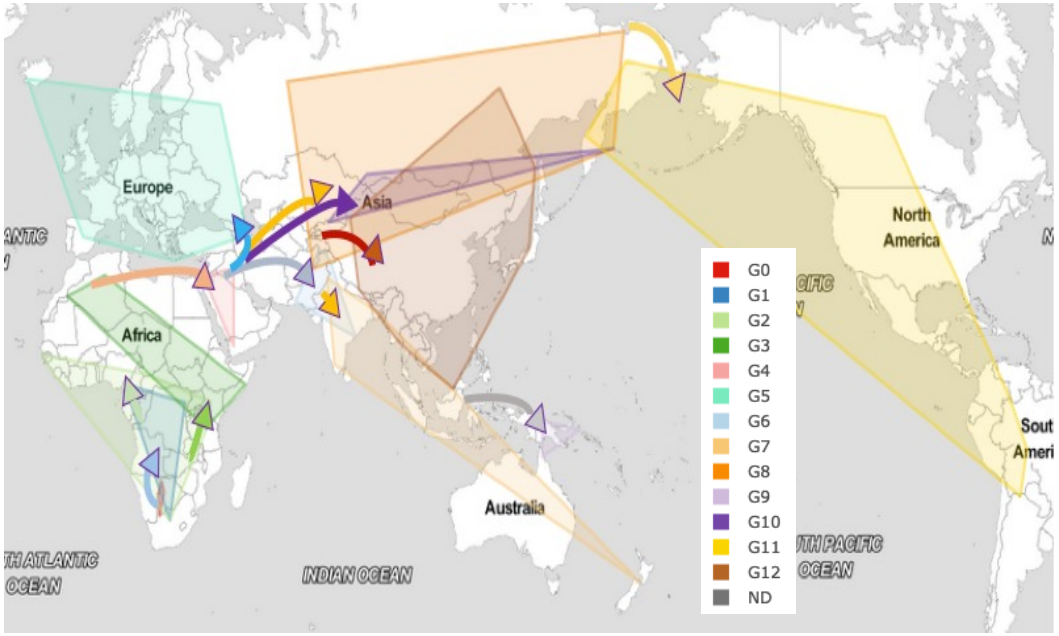
Population Information			Within Group			vs. African		vs. Non-African		vs. rest	
Region	Genomic Grp	Sample Size	# pairs	mean	sd	mean	sd	mean	sd	mean	sd
Africa	GG0	6	15	89.37	0.22	88.14	0.07	88.36	0.04	88.31	0.11
Africa	GG1	13	78	88.80	0.36	88.54	0.38	88.75	0.03	88.7	0.19
Africa	GG2	35	595	89.45	0.13	88.75	0.64	89.34	0.04	89.19	0.38
Africa	GG3	39	741	89.82	0.30	88.78	0.57	90.16	0.1	89.82	0.38
WestEurasia	GG4	15	105	92.06	0.10	89.20	0.84	91.75	0.24	90.9	0.68
WestEurasia	GG5	53	1378	92.36	0.07	89.18	0.84	91.92	0.22	91	1.35
WestEurasia	GG6	2	171	92.10	0.08	89.16	0.81	91.93	0.17	91.01	1.43
SouthAsia	GG6	17									
SouthAsia	GG7	19	190	92.16	0.11	89.14	0.79	91.93	0.13	91	1.43
Oceania	GG7	1									
CentralAsiaSiberia	GG8	4	28	92.25	0.10	89.15	0.8	92.06	0.24	91.09	1.44
SouthAsia	GG8	2									
EastAsia	GG8	2									
Oceania	GG9	19	171	93.52	0.45	89.06	0.69	91.66	0.21	90.79	1.51
CentralAsiaSiberia	GG10	6	15	92.39	0.14	89.15	0.79	92.1	0.28	91.11	1.34
CentralAsiaSiberia	GG11	6	496	93.45	0.41	89.20	0.73	92.04	0.32	91.13	1.5
America	GG11	26									
Oceania	GG12	4	1711	92.82	0.12	89.16	0.75	91.08	0.33	91.08	1.5
SouthAsia	GG12	4									
CentralAsiaSiberia	GG12	8									
EastAsia	GG12	43									
	minimum	2	15	88.804201	0.06884747	88.14	0.07	88.36	0.03	88.31	0.11
	maximum	53	1711	93.5161634	0.45348395	89.2	0.84	92.1	0.33	91.13	1.51
	mean	15.43	438.00	91.58	0.20	88.97	0.67	91.01	0.18	90.39	1.02
	maedian	8	171.00	92.16	0.13	89.15	0.75	91.75	0.21	91.00	1.35
	sum	324	5694								

Supplementary Fig. S1. Topological rooted NJ tree based on c-SNV (png)

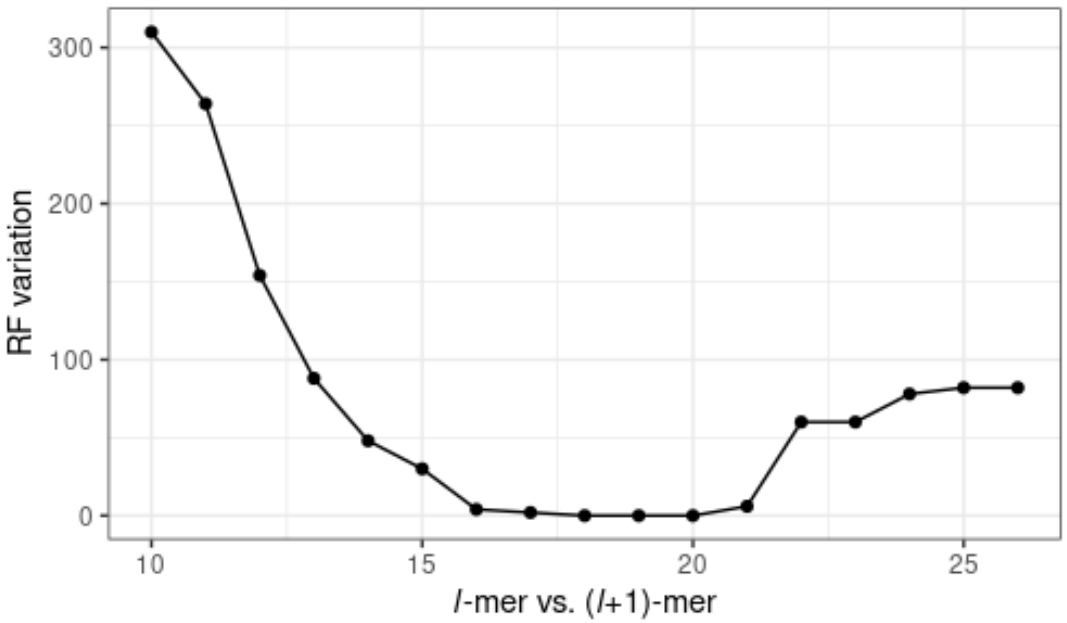


Supplementary Fig. S2





Supplementary Fig. S3



l vs. l+1	RF variation
10 vs 11	310
11 vs 12	264
12 vs 13	154
13 vs 14	88
14 vs 15	48
15 vs 16	30
16 vs 17	4
17 vs 18	2
18 vs 19	0
19 vs 20	0
20 vs 21	0
21 vs 22	6
22 vs 23	60
23 vs 24	60
24 vs 25	78
25 vs 26	82
26 vs 27	82

Suppl. Table S1

GG0: Ju_hoan, Khomani_San (2 EGs)

GG1: Baka, Bakola, BantuTswana, Bedzan, Biaka, Mbuti (6 EGs)

GG2: BantuKenya, BantuTswana, BentuHerero, Esan, Gambian, Hazda, Igbo, Kaba, Kongo, Laka, Lemande, Luhya, Luo, Mandenka, Mende, Ngumba, TikarSouth, Yaruba (18 EGs)

GG3*: Aari, Agaw, Amhara, Bulala, Dinka, Elmolo, Fulani, Iraqw, Kikuyu, Mada, Masai, Mozabite, Mursi, Ogiek, Rendille, Saharawi, Sengwer, Somali (18 EGs—many loosely clustered)

GG4: BedouinB, Druze, IraqiJew, Jordanian, Palestinian, Samaritan, YemeniteJew (7 EGs)

GG5: Abkhsian, Adygei, Albanian, Armenian, Basque, Bergamo, Bulgarian, Chechen, Crete, Czech, English, Estonian, Finnish, French, Georgian, Greek, Hungaria, Icelandic, Leznin, NorthOssetian, Norwegian, Orcadian, Polish, Russian, Sardinian, Spanish, Turkish, Tuscan (28 EGs)

GG6: Balochi, Brahmin, Brahui, Brusho, Kalash, KashmiriPandit, Makrani, Pathan, Punjabi, Sindhi, Tajik (11 EGs)

GG7: Bengali, Irula, Kapu, Kharia, KhondaDora, Kurumba, Madiga, Mala, Maori, Punjabi, Relli, Yadava (12 EGs)

GG8: Chuckchi, Hazara, Mansi, Tlingit2, Uygur (5 EGs)

GG9: Australian, Bougainville, Papuan (3 EGs)

GG10: Aleut, Kyrgyz, Tubalar (3 EGs)

GG11: Chane, Chipewayan, Cree2, Eskimo, Itelman, Karitiana, Mayan, Mixe, Mixtec, Nahua, Piapoco, Pima, Quechua, Surui, Zapotec (15 EGs)

GG12: Ami, Atayal, Burmese, Cambodian, Dai, Daur2, Dusun, Even, Han, Hezhen, Igorot, Japanese, Kinh, Korean, Lahu, Miao, Mongolia, Naxi, Orogen, She, Sherpa, Thai, Tibetan, Tu, Tujia, Ulchi, Xibo, Yakut, Yi

(29 EGs)

Unclustered : Altaian, Bermese, Cree1, Daur1, Even, Hawaiian, Iranian1,2, Kusunda, Mixtec2, Onge1+2, Saami1+2, Tlingit1, Turkish1 (13 EGs)

Blue: Singletons; Underline: Members found in more than one GGs

Supplementary Table S2

13 GGs), 164 EGs, 5 RGs, 10 Geological regions

GGs.	EGs.	"Race" in 2020 US census	Geological Regions
GG0:	2 EGs	"Black or African American"	S. Africa
GG1:	6 EGs	"Black or African American"	(S. Africa), Mid Africa
GG2:	19 EGs	"Black or African American"	Mid Africa
GG3*:	16+ EGs	"Black or African American"	Mid and N. Africa
2 EGs (Mozabite and Saharawi)		"White"	N. Africa
GG4:	7 EGs	"White"	Middle East (Levant)
GG5:	28 EGs	"White"	Europe
GG6:	11 EGs	"Asian"	S. Asia
GG7:	12 EGs.	"Asian"	S. Asia
GG8:	5 EGs	"Asian"	Central Asia
GG9:	3 EGs	"Native Hawaiian or Other Pacific Islander".	Oceania
GG10:	3 EGs	"Asian"	Central Asia
GG11:	15 EGs	"American Indian or Alaska Native"	the Americas, NE Asia
GG12:	29 EGs	"Asian"	E. and SE. Asia

Supplementary Table S3

Euclidean distances among the center positions of each GGs (from PCA plot)

	GG0	GG1	GG2	GG3*	GG4	GG5	GG6	GG7	GG8	GG9	GG10	GG11	GG12
GG0	0.00	0.31	0.76	1.36	2.75	2.93	2.89	2.89	3.04	2.81	3.09	3.26	3.15
GG1	0.31	0.00	0.46	1.07	2.48	2.66	2.62	2.62	2.78	2.54	2.83	3.00	2.89
GG2	0.76	0.46	0.00	0.64	2.06	2.24	2.20	2.20	2.36	2.13	2.41	2.59	2.48
GG3*	1.36	1.07	0.64	0.00	1.42	1.60	1.56	1.57	1.73	1.52	1.79	1.99	1.88
GG4	2.75	2.48	2.06	1.42	0.00	0.18	0.20	0.32	0.48	0.60	0.58	0.88	0.81
GG5	2.93	2.66	2.24	1.60	0.18	0.00	0.14	0.28	0.39	0.61	0.49	0.79	0.74
GG6	2.89	2.62	2.20	1.56	0.20	0.14	0.00	0.14	0.28	0.47	0.38	0.68	0.62
GG7	2.89	2.62	2.20	1.57	0.32	0.28	0.14	0.00	0.18	0.33	0.27	0.56	0.49
GG8	3.04	2.78	2.36	1.73	0.48	0.39	0.28	0.18	0.00	0.35	0.11	0.41	0.35
GG9	2.81	2.54	2.13	1.52	0.60	0.61	0.47	0.33	0.35	0.00	0.33	0.47	0.36
GG10	3.09	2.83	2.41	1.79	0.58	0.49	0.38	0.27	0.11	0.33	0.00	0.30	0.25
GG11	3.26	3.00	2.59	1.99	0.88	0.79	0.68	0.56	0.41	0.47	0.30	0.00	0.11
GG12	3.15	2.89	2.48	1.88	0.81	0.74	0.62	0.49	0.35	0.36	0.25	0.11	0.00