scientific reports

OPEN

On whole-genome demography of world's ethnic groups and individual genomic identity

Byung-Ju Kim[1,2,5], JaeJin Choi[1,3] & Sung-Hou Kim[1,2,4]✉

All current categorizations of human population, such as ethnicity, ancestry and race, are based on various selections and combinations of complex and dynamic common characteristics, that are mostly societal and cultural in nature, perceived by the members within or from outside of the categorized group. During the last decade, a massive amount of a new type of characteristics, that are exclusively genomic in nature, became available that allows us to analyze the inherited whole-genome demographics of extant human, especially in the fields such as human genetics, health sciences and medical practices (e.g., 1,2,3), where such health-related characteristics can be related to whole-genome-based categorization. Here we show the feasibility of deriving such whole-genome-based categorization. We observe that, within the available genomic data at present, (a) the study populations form about 14 genomic groups, each consisting of multiple ethnic groups; and (b), at an individual level, approximately 99.8%, on average, of the whole autosomal-genome contents are identical between any *two individuals* regardless of their genomic or ethnic groups.

Background. Classification or categorization of human population, such as race, ethnicity, and ancestry, has been commonly made based on mostly physical, cultural and societal characteristics, combined with other non-genomic characteristics such as presumed ancestry, language, cultural history, religion, socioeconomic status and others. Such categorizations have been very useful, or debatable sometimes. During the last decade, a massive amount of a new type of characteristics, inherited genomic characteristics, became available that is revolutionizing our understanding of biological and genomic characteristics of human diversity, especially in the fields such as human genetics, health sciences and medical practices. Yet, we do not have a whole-genome-based categorization of extant human population that can be correlated between the categories and the characteristics of whole-genome data gathered objectively and quantitatively in such health-related fields. Thus, there is an urgent need for such genome-based categorization of extant human population[1–3].

Several large-scale germ-line genomic studies have been published during the last decade to address the extent and types of genomic diversity of the extant human species, e.g., for 2504 individuals from 26 large "population groups (PGs)" from diverse geographic locations in the 1000 Genomes Project (1KGP)[4], for 300 individuals from 142 ethnic groups (EGs) based on ancestry, linguistic, faith and cultural differences in the Simons Genome Diversity Project (SGDP)[5], and for 44 African ethnic populations in 4 linguistic groups[6]. As a result of these studies as well as other similar or related studies (e.g.,[7,8]), a large body of whole-genome Single Nucleotide Polymorphism (SNP) data became publicly available for a wide range of individuals representing different ethnic groups or "population" groups throughout the world.

Therefore, we revisit these data and chose (a) the SGDP data for the purpose of finding a whole-genome-based categorization, regardless of ethnicity, using a text-comparison method of Information Theory[9], which has not been applied in any of the earlier studies, and (b) both the 1KGP and the SGDP for the purpose of quantifying the fraction of whole genome that enables such genome-based categorization.

Objectives. The itemized objectives of this study are twofold: first, (a) using the concept of the "contextually-linked Single-Nucleotide-Variation (c-SNV) genotypes", identify whole-genome-based genomic groups (GGs)

[1]Department of Chemistry and Center for Computational Biology, University of California, Berkeley, CA 94720, USA. [2]Human Genome Research Center, Incheon National University, Incheon 22012, Republic of 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. [5]Convergence Research Center for Insect Vectors, Incheon National University, Incheon 22012, Republic of Korea. ✉email: sunghou@berkeley.edu

of the population in the SGDP database[5], where c-SNVs are defined by overlapping short strings of ordered SNV genotypes, (b) identify, for each GG, a set of different EGs that have very similar whole-genome content, thus, belong to the same GG, (c) suggest the order of emergence of the GGs; second, (d) quantify the magnitudes of autosomal genomic identity between two individuals within as well as between two different GGs of the study population to estimate the fraction of whole autosomal-genome that contributes toward an individual's GG; and (e) perform a similar study as (d) above for the 1KGP population, which has 26 geographically defined "Population Groups (PGs)", but with larger sample size for each PG.

Approach. Our approach is to adapt a computational method of comparing and categorizing linear information such as texts or books in the field of Information Theory[9] by "word frequency profiles"[10]. We generalized the method to be applicable to other linear information, 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)"[11], where each "Feature" (which corresponds to a unique "word" in a book) 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 Data source and Methods" of Supplementary Materials). 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 ordered SNV genotypes. Then, for a given length of the c-SNVs, all pair-wise divergence of the FFPs of a 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 pattern (as manifested by clustering) of the study population (see FFP-based PCA in data source and Methods of Supplementary Materials) and to find the genome-based clading pattern from a rooted neighbor-joining tree ([13] see FFP-based rooted NJ tree and Tree rooting in Data Source and Methods of Supplementary Materials). Among many such trees, each corresponding to a given length of the c-SNVs, the one with the most topologically stable tree is selected as the final tree. From the final tree the order of the emergence of the founder nodes of GGs are also predicted on an evolutionary progression scale, which corresponds to the normalized cumulative branch-lengths from the tree root to each of the founder nodes of the GGs.

The uniqueness of our approach is we compare the entirety of each individual's whole autosomal-genome variations *in context,* and with that of each of all other individuals, not to that of one Eurocentric "human reference genome".

Results

Our results are divided into two parts. The first part focuses on, at the group level, the whole-genome-based grouping pattern of the study population in the recent SGDP database, and on the emergence order of the founders of the Genomic Groups (GGs). These provide the results for the first three specific objectives ((a)–(c)) listed under *Objectives* in the Introduction. 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 two databases: the SGDP sample, which contains the largest number of ethnic groups, and the 1KGP populations, which has the largest number of samples per "population group". Then, we extrapolate the percent identity of all SNP loci to that for the respective whole-genome length to get an estimate of intuitively understandable magnitude of the whole autosomal-genome identity between two individuals within a given GG or PG and between two different GGs or PGs. These provide the results for the remaining two specific objectives ((d) and (e)) listed under *Objectives* in the Introduction.

Part I: Genomic demography: 14 "genomic groups (GGs)" and the order of their emergence. At a group level, we examined the genome-based grouping patterns and their relationships among all GGs by two different methods, both based on the individual's genomic variation expressed by the FFPs of c-SNVs: Principal Component Analysis (PCA)[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 recent SGDP database[5], where the germline 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 (see FFP based PCA in Data source and Methods of Supplementary Materials). The figure shows that:

(a) There are about 14 clusters in genomic variation space (i.e., c-SNV space) which we named "Genome-based Groups" or "Genomic Groups (GGs)" (one of which, GG3, consists of a collection of several not-well-resolved sub-clusters of various sizes) represented in the SGDP data;
(b) All GGs are divided into two interconnected super groups: one defined by a long linear arm in Fig. 1 containing all five African GGs (GG0–GG4), most of which are linearly linked but not well clustered (due to sparse availability of the whole-genome sequences representing the vast diversity of African EGs) and account for all 45 African EGs available in the SGDP database. The other is a more fanned-out arm containing all non-African GGs (GG5–GG13), most of which are well clustered, accounting for all of the 119 non-African EGs in the data;
(c) Most of African GGs are not well resolved and linearly connected, with GG0 (consisting of the EGs of Khomani_San and Jo_hoan) at one end, and GG4 (consisting of the EGs of Mozabite and Saharawi) at the other end, but all non-African GGs appear to have originated from the Middle East GG (GG5). Most of non-African GGs are tightly-clustered and well-resolved, but, all remaining un-clustered and isolated samples are found in this super group;

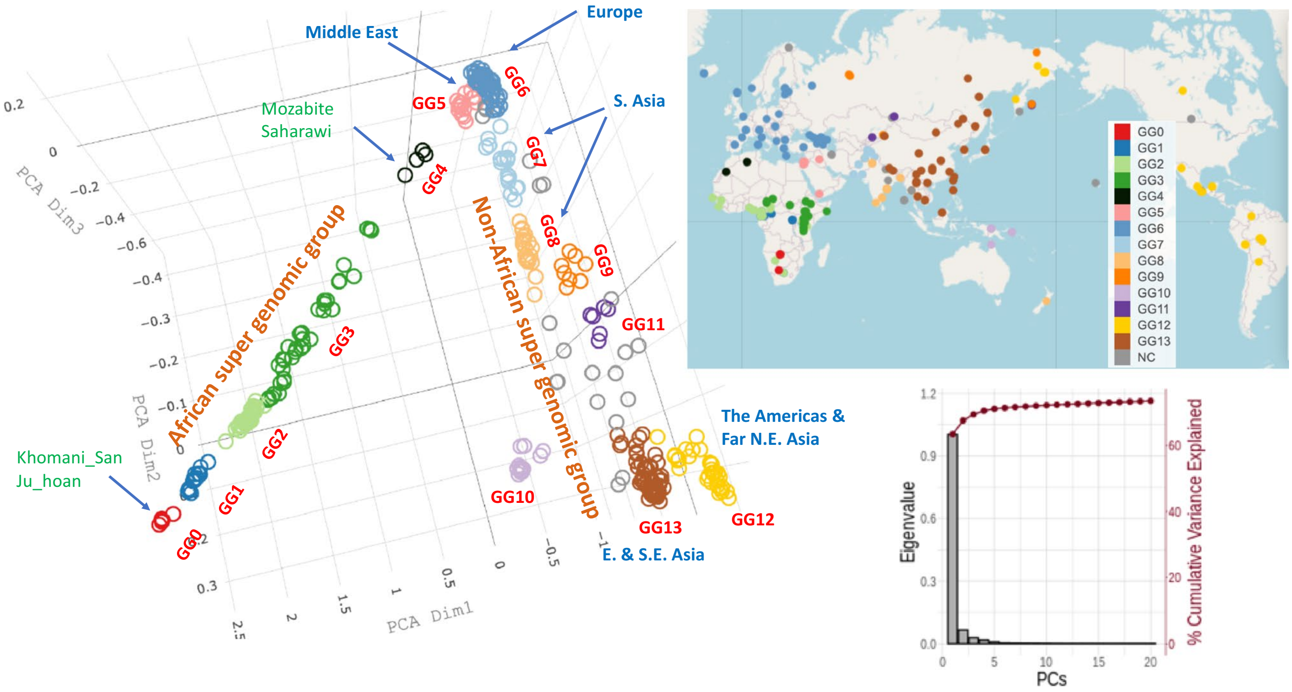


Figure 1. Clustering pattern by genomic-divergence-based PCA plotted for the three major principal component axes. PCA was performed using the "distance matrix", where each "distance" is calculated by JS divergence between two FFPs of c-SNVs. Most of 345 individuals from 164 ethnic groups from the updated SGDP form about 14 clusters shown in different colors and labeled as genomic groups (GGs) GG0–GG13, of which most are tight clusters, but some are not, e.g., GG3 consists of multiple un-resolved clusters of African ethnic groups. All un-clustered individuals are shown as gray circles and located outside of Africa. Although the EG members of each GG are tightly clustered in genomic divergence space (PCA space), they are widely spread out within a large region bound by great geological barriers, as shown at the geographical locations of the sample collection sites, suggesting broad migration of the EGs within the region (see the top right inset). Sorted eigenvalues and variance explained in % for the first 20 principal component axes are shown in the bottom right inset. Grouping of GGs is based on a combination of the clustering pattern of individuals in PCA (in genomic divergence space), the clading pattern in NJ tree (in genomic divergence space), and the clustering pattern in geological/geographical map. A video of the rotating PCA plot is shown in Supplementary Fig. S1A.

(d) Each GG consists of multiple EGs (see Supplementary Table S1), thus there is no one-to-one correspondence between GGs and EGs in the data;
(e) Many of the 14 GGs can be assigned to various geographical or geological regions (see the color coding in the upper inset of Fig. 1);
(f) The extant members of some large GGs (GG6, GG12 and GG13) in the data 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, Fig. S1A in Supplementary Materials).

Some of the features described above (such as (b), (c) and (f) mentioned above) have been observed in the first SGDP 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 in Supplementary Materials).

Clading in neighbor-joining (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 Data source and Methods of Supplementary Materials) by a neighbor-joining method[13,14] (for a corresponding topological tree with the names of EGs and clading details is shown in Supplementary Fig. S1B, which can be expanded for easier viewing). 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 method, two additional conditions are imposed for the evolutionary model, which constrains the tree topology and clading: (a) bifurcating branches emerging from each internal node and (b) maximum parsimony (minimal evolution) when choosing the neighbors to join as a sister pair. For comparison, all the GGs identified by the PCA clustering in Fig. 1 are also shown in the middle circular band of Fig. 2. Notable features from the tree are:

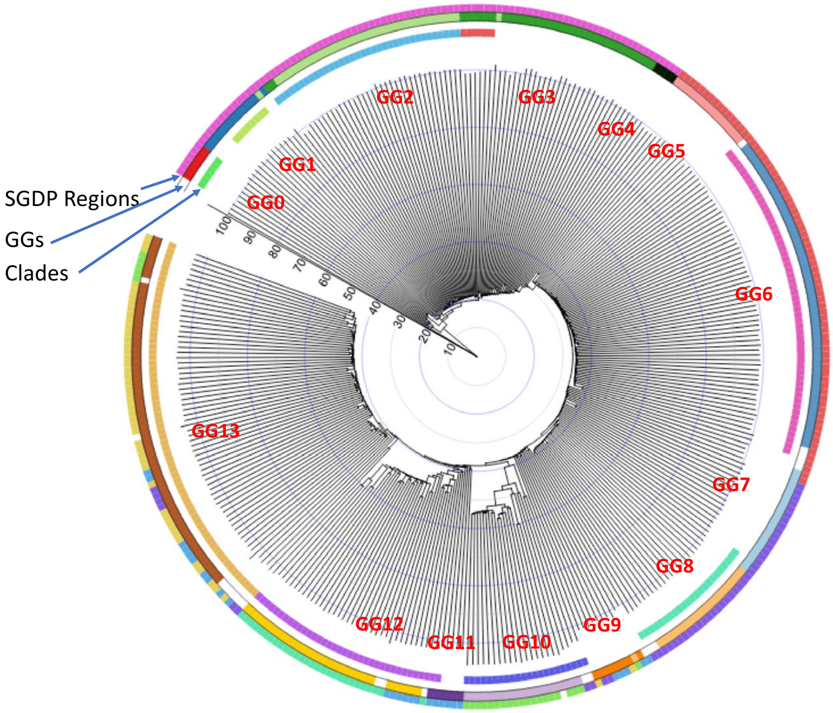


Figure 2. Rooted c-SNV-based NJ tree in circular form[15] with the "evolutionary progression scale (EPS)". The same "distance matrix" as was used in Fig. 1 is used to construct the tree (see Rooting of c-SNV-based tree in Data Source and Methods in Supplementary Materials). "Evolutionary progression scale" is the scaled cumulative branch-length bounded such that EPS = 0.0 at the root of the tree and EPS = around 100 at the leaf nodes. Nine clades in the tree are shown in the inner circular band. Fourteen 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 no SGDP regional assignment (outer band), not clustered among GG0-GG13 (middle band), or not in any of the 9 clades (inner band). A corresponding linear topological tree with the names of EGs is shown in Supplementary Fig. S1.

(a) As was observed in Fig. 1, all African GGs (GG0–GG4) emerged sequentially in multiple steps, but all non-African GGs (GG5–GG13) 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 14 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.

The NJ tree of the first SGDP publication[5], when compared to ours, has several differences in clading pattern, relative branch lengths and branching order within each region (Extended Data Fig. 4B of reference[5]). These differences as well as those of clustering differences mentioned in previous section ("Clustering by PCA") are ultimately due to the different ways the two methods describe the individual genomic variations, i.e., isolated SNVs in reference[5] vs. contextual SNVs in this study.

Combining the clustering pattern in PCA, the clading pattern of NJ tree, and the clustering pattern in geographical map. Our GG classification is not based on PCA clustering alone. It is based on the cross-checking of the clustering pattern of individuals in the PCA (in genomic divergence space), the clading pattern in the NJ tree (in branch-length and topology space), and the clustering pattern in the geographical map. Figure 2 shows that 9 out of 14 GGs agree approximately between the PCA clustering and the clading in NJ tree (see the middle color band and inner color band, respectively). Among the remaining five GGs, four are based on the degree of agreement between the PCA clusters and much loose clusters in the current geographical locations of individuals (shown in the upper right inset of Fig. 1). For GG3, all three disagree in various extent. So, we arbitrarily lumped together into one and called GG3 with the hope that, once more data become available in future, we may be able to subdivide into GG3a, GG3b, etc.

Order of emergence of the founders of GGs on "Evolutionary Progression Scale (EPS)". The order of emergence of all the extant GGs in the SGDP data can be inferred from Fig. 1 based on the nearest-neighbor relationship among the centers of each GGs (see also Supplementary Fig. S2A; Table S3). Furthermore, the point on EPS (see

"Evolutionary Progression Scale" in Supplementary Materials) at which the "founder(s)" of each GG emerged can be derived from Fig. 2 under the following processes: we start with two assumptions: (a) from the point of view of genomic information, the progression of evolution can be considered as the process of increasing divergence of genomic sequences; (b) the "founder(s)" of an extant GG can be considered as a selected subpopulation from the population of an internal node at a specific point on EPS (see the radial line scaled zero to 100 in Fig. 2), and, then, the founders diversify and migrate to generate all the extant EG members of the GG in the data. The genomic divergence, which corresponds to the cumulative branch-length, is set to EPS = 0 at the origin of our tree, and EPS = around 100 for the leaf-nodes of all extant individuals (see Fig. 2 and Evolutionary Progression Scale in Supplementary Materials). For example, the founder(s) of the first GG (GG0) emerges from the first internal node located at EPS = about 15.2 in Fig. 2.

Finally, we take the identification of each GG from Fig. 1 and its nearest-neighbor relationship from Supplementary Table S3, and combine with the EPS value for the corresponding internal node from our tree to estimate the order of the emergence of each GG *and* the point on the EPS of the emergence of the founder(s) of the GG in two spaces: one on PCA space (Supplementary Fig. S2A) and the other on a world map (large circles in Supplementary Fig. S2B). Such combination suggests that:

(a) All extant African GGs (GG0 – GG4) in the SGDP data emerged *sequentially*, and their founders emerged between EPS of 15.2 and 29.4, but any presumably extinct earlier GGs must have emerged during the period corresponding to EPS between 0.0 and 15;
(b) The first extant GG (GG0) consists of two EGs of Khomani_San and Jo_hoan (see Fig. 1 for the numbering of the GGs and Supplementary Table S1 for the names of EGs in each GG) currently residing in the southern tip of Africa, and the last African GG (GG4) is composed of the EGs of Mozabite and Saharawi in Northern Africa at the end of the African supergroup in Fig. 1;
(c) The founders of the non-African GGs in the SGDP emerged in a "burst" into several lineages of the GGs during a period corresponding to a narrow range of EPS of 33.5–39.5 (see Fig. 2, Supplementary Fig. S2A,B). They all emerged from GG5 (the first non-African GG nearest to the last African GG (GG4) of Northern Africa). GG5 consists of the EGs of Bedouin B, Druze, Iraqi Jew, Jordanian, Palestinian, Samaritan, and Yemenite Jew in the Middle East;
(d) The founders of GG12 and GG13 emerged most recently at EPS of 39.5.
(e) The EGs of a newly emerged GG are often found in a large region bound by great geological barriers different from those of previous GG, suggesting that some members of the previous GG may have migrated through the great geological barriers.

Part II: Average genotype identity between two individuals. The genome-based grouping discussed above has been derived based on the "genomic-divergence distance" between two FFPs of *contextually-linked* SNVs in each of all pairs of individuals in the SGDP data. Although the distance is precisely defined from the Information-Theoretic viewpoint it is not obvious what the distance between the two FFPs corresponds to physically in terms of two respective whole genome contents. Furthermore, although 1KGP study[4] showed that the average SNV difference between the reference genome of European ancestry and each individual's genome of different population groups vary from 3.54 million (M) to 4.32 M SNP loci, corresponding to 4.2–5.1% of total SNP loci, it is not certain whether similar variations will be observed between two individuals among *all* study populations (not comparing each to the one Eurocentric reference genome). To get a more intuitive understanding, we ask alternative questions: (a) What is the average "percent identity" in terms of whole genome content between two individuals within one GG vs. from two different GGs? (b) Are the percent identity unique to each GG? and (c) What are the implications of the answers to the above questions?

For this portion of our study, genomic identity has been quantified in two steps: first, we quantify the genotype identity of all SNP loci within each pair of two individuals in each group as well as between two different groups, using the genomic data of the larger population size (the 1KGP data) and of the broader 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 a percent of the most recent *whole* human genome sequence with no gaps (about 3.06 billion (B) base-pairs[16]). These steps are taken under a few approximating and simplifying assumptions described below:

There are many types of mutational or variational events that result in individual genomic variations, such as single nucleotide substitutions, short Indels, large deletions/insertions, inversions, recombination, interbreeding, admixture and others. Of these, about 99.7% are due to SNPs, and the rest of variational events are one or more order-of-magnitude rarer 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 quantify and compare. Thus, we estimate the extent of the individual SNV *identity* by counting identical genotypes between the genomes of two individuals at all available SNP loci, then we re-scale the SNV identity among all SNP loci to that for a whole-genome length. This process was repeated for both data sets of the 1KGP[4] and the SGDP[5]. We estimate the average percent-identity in four different ways, the results of which are summarized below. The detailed calculations are described in Supplementary Materials under *Four ways of estimating the average identity of whole genomes between two individuals among the world's ethnic and population groups.*

(a) The average genomic identity between the human reference haploid genome, GRCh37[17], and each of all individual genomes from 26 "population groups (PGs)" of the 1KGP was calculated. Each genotypes of 95.6%, on average, of all SNP loci (84.7 M) between the reference genome and each individual is identical. Re-scaling this number to whole-genome length, 99.86% of the whole genome of each individual have

Figure 3. "Violin plots" of the genotype identity for all SNP loci. (**A1**) Violin plot of the % genotype identity for 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 categorization by the population grouping of the American continental group in the 1KGP samples. The extent of the "streaking" is correlated to the standard deviation (SD) of each average value, which, in turn, is related to the "looseness" of each cluster (see Fig. 4A). The average % identities within African PGs (about 94.4%) is slightly lower than those of non-African PGs (about 95.9%), revealing that (a) African PGs and non-African PGs form two separate super groups, and (b) the African PGs have slightly more diverse genomes than those among the members of non-African PGs. (**A2**) Violin plot of the % genotype identity of pair-wise SNVs among all SNP loci *between* the European group (consisting of the PGs of CEU, FIN, GBR, IBS, and TSI) and the rest. The % identity between two different PGs are slightly lower than those within the same PG. The 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) as in (**A1**) with the 1KGP samples above, African GGs and non-African GGs of the SGDP samples 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 GG6 and the 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 GG6 and GG3 is because GG3 is not a tight cluster but a diffuse distribution of the individuals from several loose sub-clusters.

identical genotypes as those of the reference genome under the simplifying approximations mentioned above.

(b) Furthermore, the average genomic identity between two individuals (*excluding* the human reference genome) among *all* members of the 1KGP population is also found to be about 95.08% of the total SNP loci (Figs. 3A1,A2,4A), which corresponds to 99.87% of the whole genome of the recently updated human reference genome length[18,19].

(c) Similar to the results with 1KGP population above, the average genotype identity between two individuals among all members of 164 ethnic groups of the SGDP is 90.39% for a total of 34.4 M SNP loci (Figs. 3B1,B2,4B), which corresponds to about 99.84%, on average, of whole-genome length.

(d) Extrapolating these consistently high estimates for the genome identity between two individuals in the 1KGP and the SGDP samples to the latest values for the whole genome SNP loci[19] and for the complete human genome length[16], we arrive at a "generalized" and conservative estimation of 99.8% as the genotype identity between two individuals regardless of the types of categorization by the "population groups" in the 1KGP or GGs of the ethnic groups of the SGDP.

Summaries and prospects. At the level of the population diversity represented in the SGDP, we show the feasibility of categorizing the study population into approximately 14 groups based exclusively on whole-genome characteristics, different from all other characteristics used so far for human categorization, such as cultural, societal, physical, ancestral, language, cultural history, religions, socioeconomic status and other characteristics. Such germ-line genome-based categorization, which is expected to be improved as more diverse whole-genome data become available in future, should provide a quantitative footing for genome-based demographic studies of various health-related fields in:

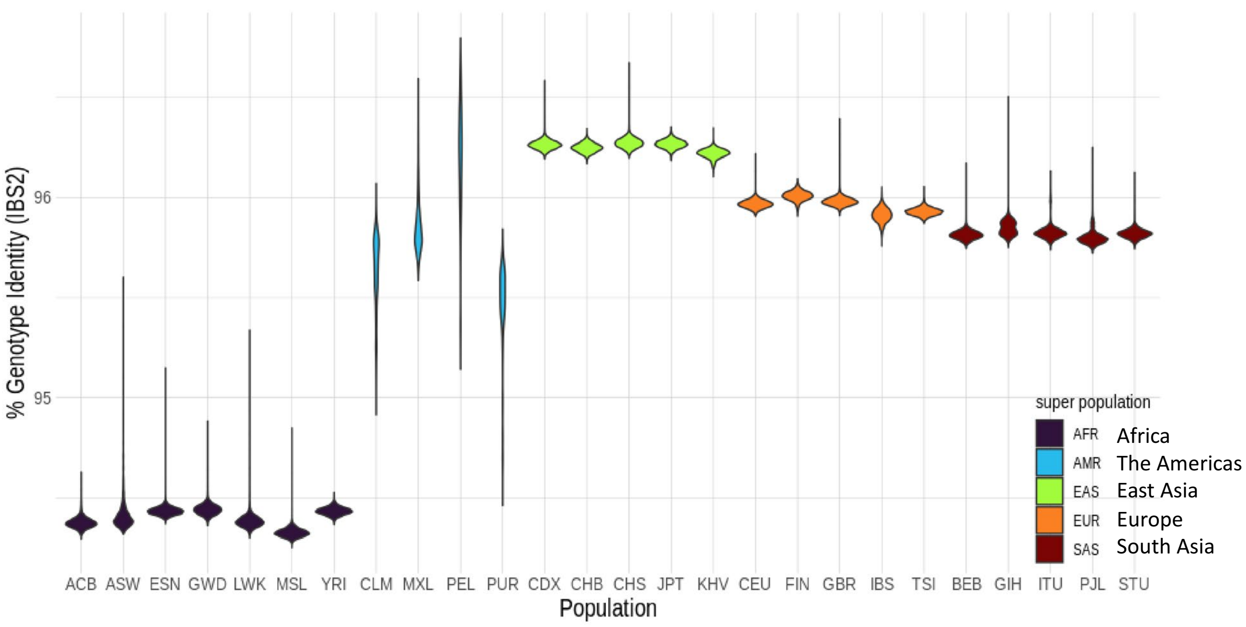
(a) Estimating or predicting the role of the inherited whole-genome components that contribute to certain phenotypes specific to a given GG in the health-related fields, such as epidemiology, disease diagnosis, disease susceptibility, and clinical practice in predicting, for example, drug or therapy responses;

(b) Training and testing of various computational algorithms using Information Technology, e.g., Machine Learning and Artificial Intelligence in the health-related fields mentioned above, where the development of such algorithms depend on having objectively definable demographic categorizations (called "labels" in this field) of the genomic input data.

(c) Suggesting the need for additional genomic diversity data for the ethnic groups especially in Northern Africa, Nordic Europe, North, Central and South East Asia, Oceania and the Americas, for which currently existing data are very sparse. Such data may lead to discovering additional GGs and resolving loosely clustered GGs, especially in Africa.

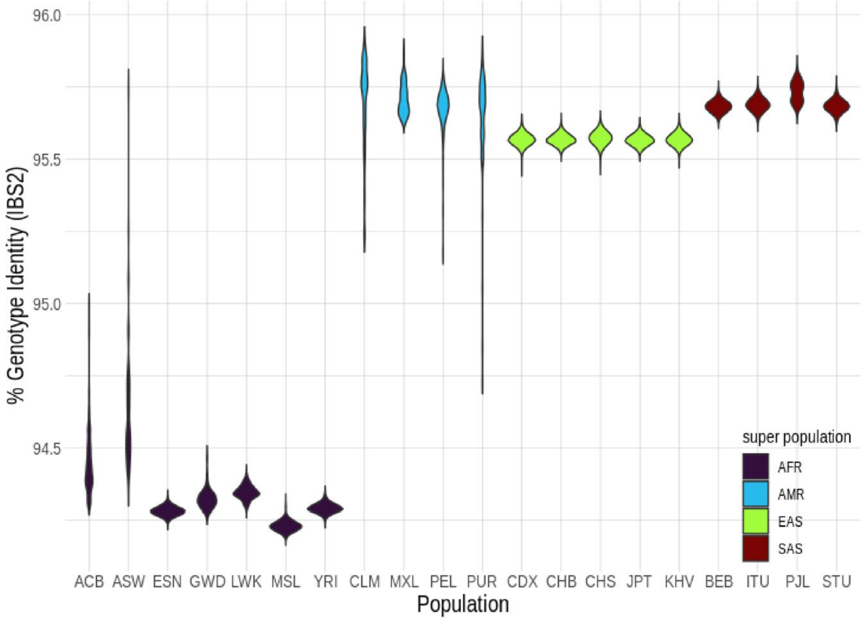
At an individual level, taking all four estimates derived above together we make a simplified and conservative approximation that the genomic identity is, on average, about 99.8% of whole autosomal-genome(ranges of the four estimates above: 99.82–99.87%) between two individuals in the study population regardless of categorization, i.e., the remaining 0.2% accounts for non-identical genotypes, which amount to about 6 M genomic loci. 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. Possible implications

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

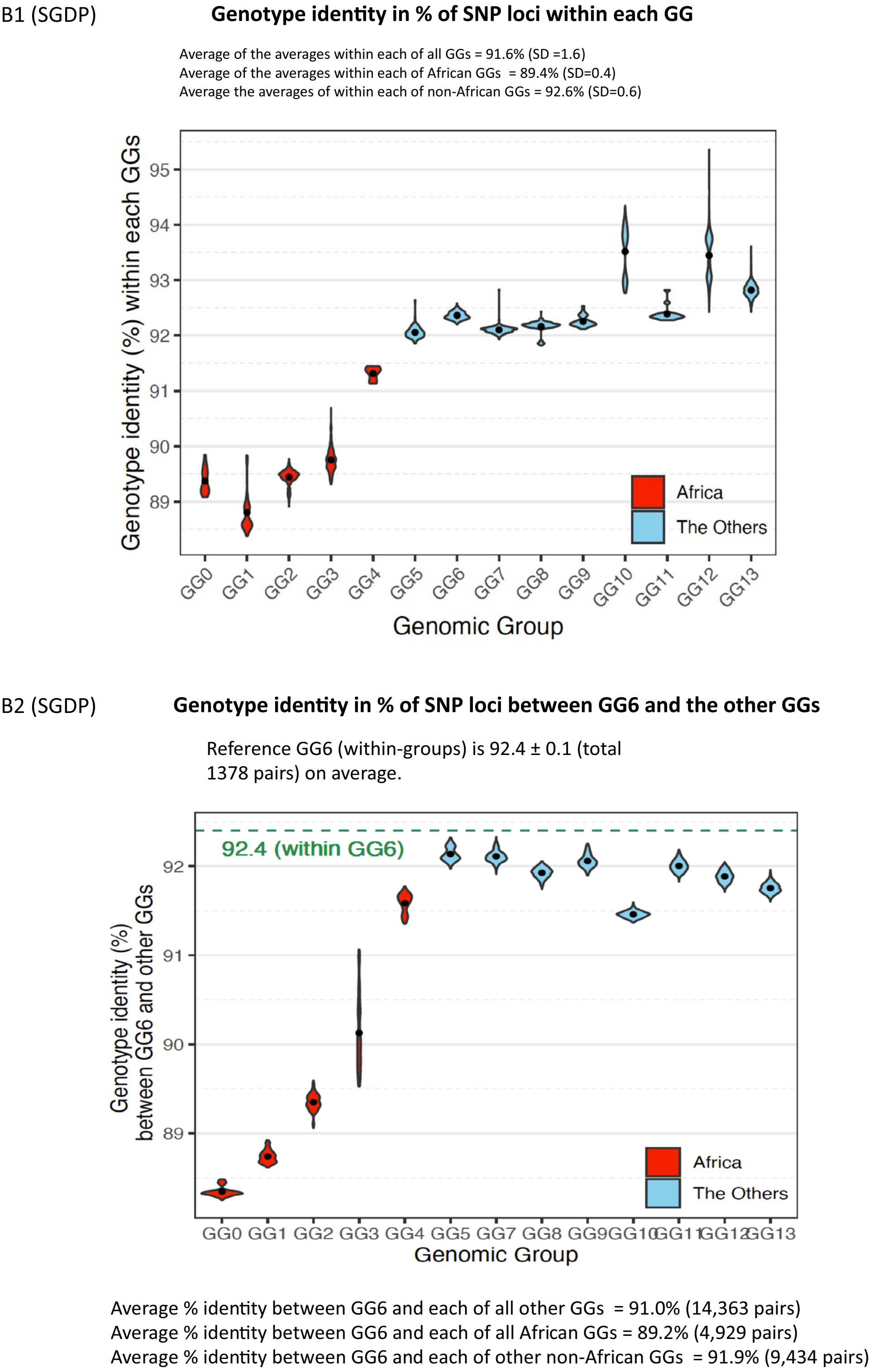


For the three-alphabet names of "population" groups see Table 1A

A2 (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



B1 (SGDP) Genotype identity in % of SNP loci within each GG

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)

B2 (SGDP) Genotype identity in % of SNP loci between GG6 and the other GGs

Reference GG6 (within-groups) is 92.4 ± 0.1 (total 1378 pairs) on average.

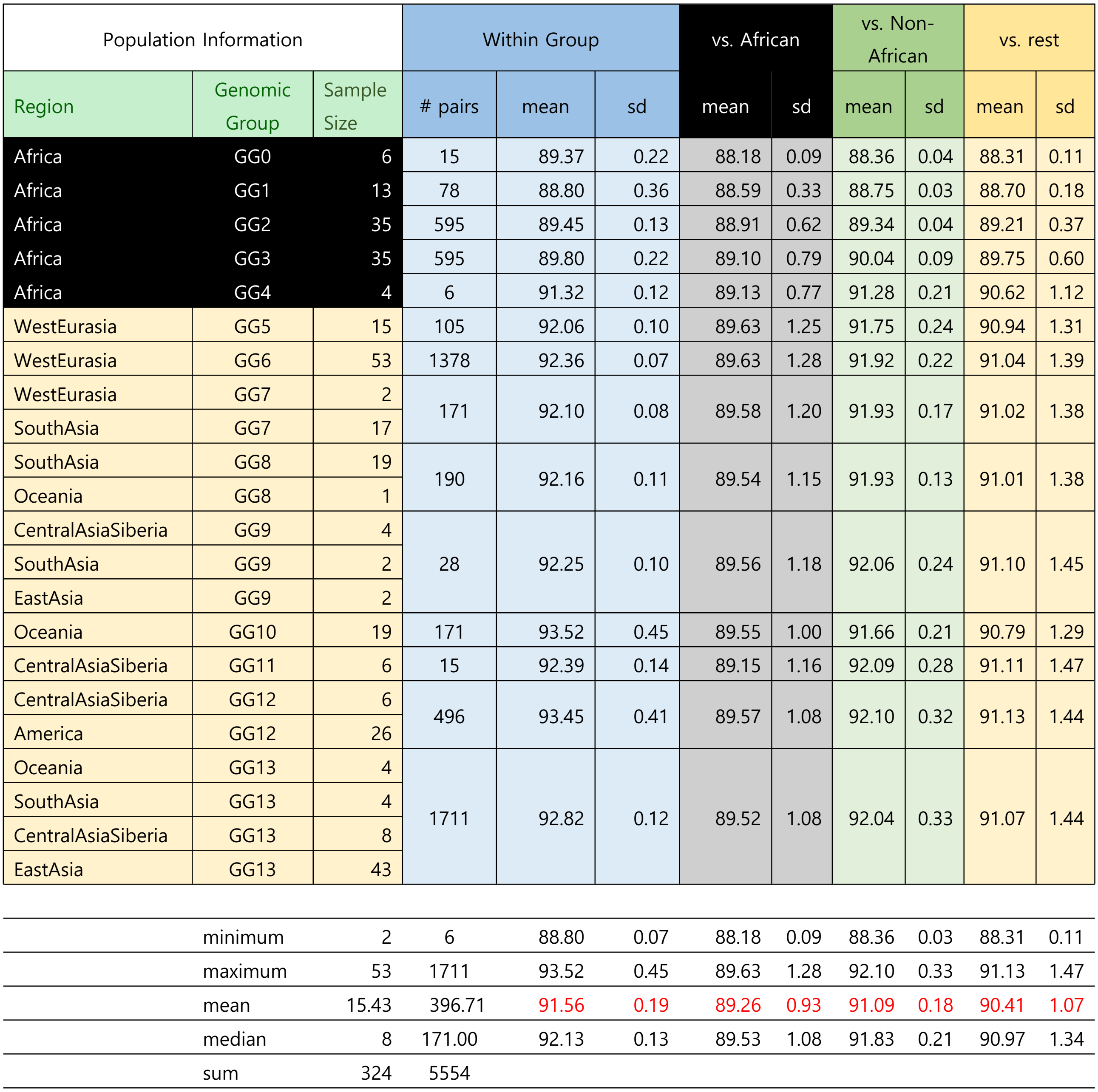
Average % identity between GG6 and each of all other GGs = 91.0% (14,363 pairs)
Average % identity between GG6 and each of all African GGs = 89.2% (4,929 pairs)
Average % identity between GG6 and each of other non-African GGs = 91.9% (9,434 pairs)

Figure 3. (continued)

derived from the observation of the uniformly very high genomic identity between two individuals regardless of categorization are discussed below in *Limitations, Implications and Discussions*.

Limitations, implications and discussions

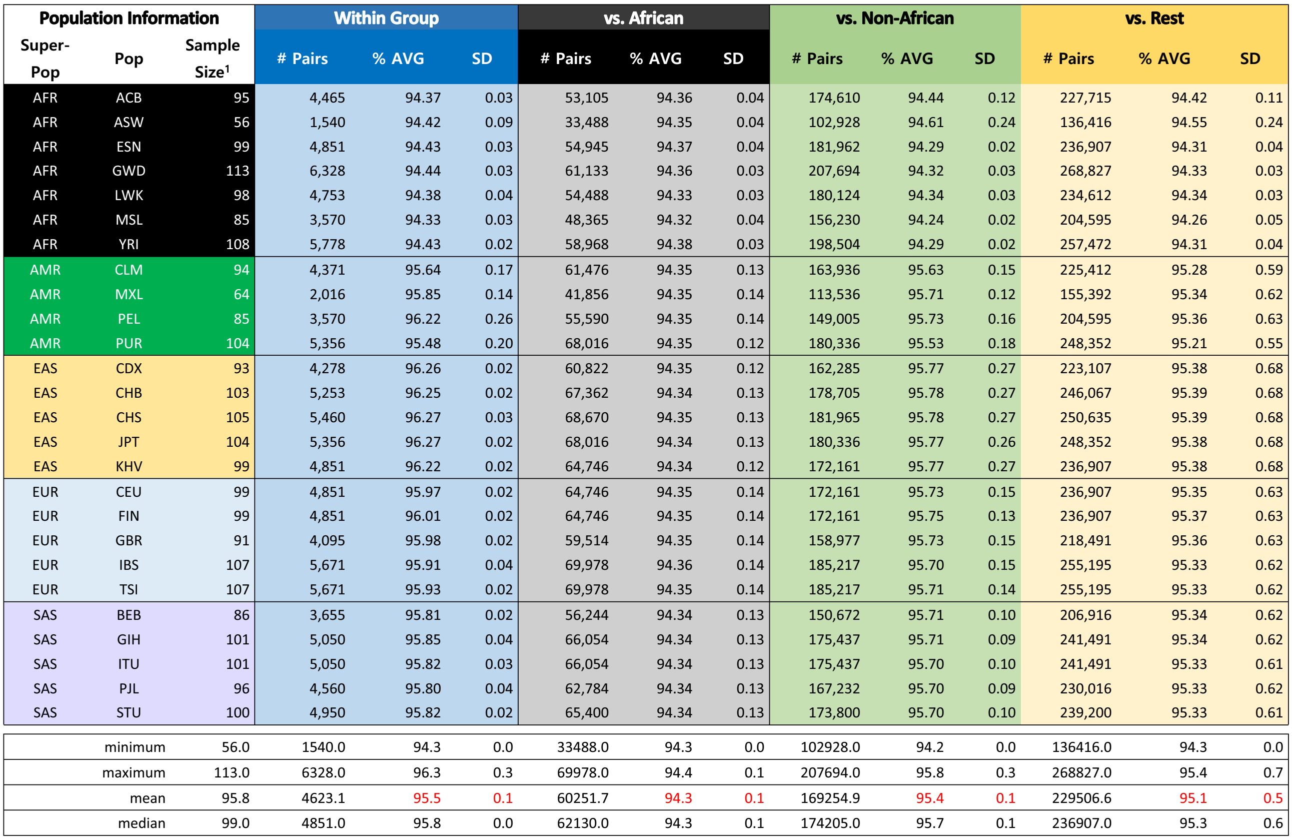
On the question of whether the SGDP data, which is sampled based on *ethnic diversity*, has sufficient *genomic diversity* to discover all the genome-based groups, we can only say that the SGDP data is the most genomic-diverse among all available genomic data at present. The ideal data for our genomic grouping studies should be



Population Information			Within Group			vs. African		vs. Non-African		vs. rest	
Region	Genomic Group	Sample Size	# pairs	mean	sd	mean	sd	mean	sd	mean	sd
Africa	GG0	6	15	89.37	0.22	88.18	0.09	88.36	0.04	88.31	0.11
Africa	GG1	13	78	88.80	0.36	88.59	0.33	88.75	0.03	88.70	0.18
Africa	GG2	35	595	89.45	0.13	88.91	0.62	89.34	0.04	89.21	0.37
Africa	GG3	35	595	89.80	0.22	89.10	0.79	90.04	0.09	89.75	0.60
Africa	GG4	4	6	91.32	0.12	89.13	0.77	91.28	0.21	90.62	1.12
WestEurasia	GG5	15	105	92.06	0.10	89.63	1.25	91.75	0.24	90.94	1.31
WestEurasia	GG6	53	1378	92.36	0.07	89.63	1.28	91.92	0.22	91.04	1.39
WestEurasia	GG7	2	171	92.10	0.08	89.58	1.20	91.93	0.17	91.02	1.38
SouthAsia	GG7	17									
SouthAsia	GG8	19	190	92.16	0.11	89.54	1.15	91.93	0.13	91.01	1.38
Oceania	GG8	1									
CentralAsiaSiberia	GG9	4	28	92.25	0.10	89.56	1.18	92.06	0.24	91.10	1.45
SouthAsia	GG9	2									
EastAsia	GG9	2									
Oceania	GG10	19	171	93.52	0.45	89.55	1.00	91.66	0.21	90.79	1.29
CentralAsiaSiberia	GG11	6	15	92.39	0.14	89.15	1.16	92.09	0.28	91.11	1.47
CentralAsiaSiberia	GG12	6	496	93.45	0.41	89.57	1.08	92.10	0.32	91.13	1.44
America	GG12	26									
Oceania	GG13	4	1711	92.82	0.12	89.52	1.08	92.04	0.33	91.07	1.44
SouthAsia	GG13	4									
CentralAsiaSiberia	GG13	8									
EastAsia	GG13	43									

	minimum	2	6	88.80	0.07	88.18	0.09	88.36	0.03	88.31	0.11
	maximum	53	1711	93.52	0.45	89.63	1.28	92.10	0.33	91.13	1.47
	mean	15.43	396.71	91.56	0.19	89.26	0.93	91.09	0.18	90.41	1.07
	median	8	171.00	92.13	0.13	89.53	1.08	91.83	0.21	90.97	1.34
	sum	324	5554								

Figure 4. Average genotype identity in % of whole-genome 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 sample sizes. (**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. Also listed are the minima, maxima, means, medians, and sums of sample sizes.



Population Information			Within Group			vs. African			vs. Non-African			vs. Rest		
Super-Pop	Pop	Sample Size[1]	# Pairs	% AVG	SD	# Pairs	% AVG	SD	# Pairs	% AVG	SD	# Pairs	% AVG	SD
AFR	ACB	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	56.0	1540.0	94.3	0.0	33488.0	94.3	0.0	102928.0	94.2	0.0	136416.0	94.3	0.0
	maximum	113.0	6328.0	96.3	0.3	69978.0	94.4	0.1	207694.0	95.8	0.3	268827.0	95.4	0.7
	mean	95.8	4623.1	95.5	0.1	60251.7	94.3	0.1	169254.9	95.4	0.1	229506.6	95.1	0.5
	median	99.0	4851.0	95.8	0.0	62130.0	94.3	0.1	174205.0	95.7	0.1	236907.0	95.3	0.6

Figure 4. (continued)

the one with the maximum whole-genome sequence diversity from all genome-altering events of germline cells, such as large-scale events of interbreeding, intermixing, introgression, etc., and small-scale events of point mutation, indels, inversion, repetition, etc. Since such ideal data is not available at present we used the SGDP data as the "proxy", albeit limited, because it contains the broadest genomic diversity represented by 164 ethnic groups, in contrast to most other available whole-genome data that are highly biased to the European "white" population.

The genome-based grouping pattern and individual genomic identity described in Results are the *average* features of the grouping and quantifications of inherited genomic characteristics for the *majority* of the study population of the SGDP and the 1KGP, two of the most diverse whole-genome variation data publicly available at the time of this study. Thus, although some interpretation about "outliers" may change, the overall average features of the results will stand, at least at the level of order-of-magnitude.

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 long "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[20], for which no comprehensive data are available, at present, to make quantitative estimates of their contributions to a genome-based categorization. With these caveats and limitations, we describe the implications of the Results with discussions below.

"Burst" of all extant non-African ethnic groups from the Middle Eastern GG.

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 (GG0), and (b) all founders of the non-African ethnic groups emerged in a "burst" from the Middle Eastern genomic group (GG5), which emerged after the two Northern African ethnic groups of Mozabite and Saharawi (in GG4) in Northern Africa emerged. What event or events may have caused such burst?

About 99.8% genomic identity for individuals and 96.4% for all populations. Under the simplifying assumptions mentioned in part II of the Results above and at the level of order-of-magnitude, we find that the contents of the whole autosomal-genomes of any two individuals in this study populations are very similar (about 99.8% identical on average (with a narrow range of 99.82–99.87%) between two individuals not only within a given GG or PG but also from 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 (see section (d) of Four ways of estimating the average identity of whole genomes between two individuals among the world's ethnic groups in Supplementary Materials). 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, although the magnitude of the difference is about the same for both pairs. 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 members of the study population as a single group has 96.4% identical genome, but, at an individual level, 99.8% of whole genome is identical between two individuals, in general.

Inherited "Passive" genomic information. Our observation of almost uniform degree of very high genomic identity indicates that any two extant individuals have inherited an *almost complete* set of identical genomic information, despite apparently complex phenotypic differences. This apparent "inconsistency" implies that the information in the 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 an individual's inherited genome (see next paragraph, and Indirect implication on a molecular scenario for converting environmental diversity to genomic divergence in Supplementary Note 2 in Supplementary Materials) 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 and essential proteins and other chemicals to start life, until certain environmental signals, such as heat and essential nutrients, are received 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 a part of critical "activating" agents for the "passive" genomic information.

Environmentally-selected genomic variants. Although most of each GG form tight genomic clusters, the members of each GG are found spread broadly within one of about 11 geological/geographical regions (S. Africa, Mid Africa, N. Africa, Middle East, Europe, Central Asia, S. Asia, Oceania, North Asia, the Americas plus NE Asia, and E. plus SE. Asia) as shown in Fig. 1 and Supplementary Table S2. Many of the regions are definable by various major geological barriers such as high mountain ranges, large body of water or desert, etc. For example, the members in GG6 are widely spread in Europe, but segregated from those of other non-African GGs by the Ural mountain range and Caspian sea; GG7 and GG8 (both in S. Asia) are separated by a large geological barrier, Thar desert; most members of GG12 are in the North and South America; and the members of GG13 (in E. and South East Asia) are segregated from those of the rest of Asia by Himalaya mountains on the South, Tianshan mountains, Alta mountains and the Gobi desert on the North. Thus, interestingly, the categorization of GGs, which are exclusively genome-based, appears to correlate with non-biological environment bordered by major geological barriers. This observation implies or suggests that even the genome-based categorization of a GG may have been strongly influenced by the environmental selection of a unique set of genomic sub-variants (from a diverse genomic variant population accumulated during a long environmental exposure specific to the survival advantage of the GG) on an evolutionary time scale.

Environment-based vs. genome-based categorization. The fact that there is no one-to-one correspondence between the 14 *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 biologically inherited genomic diversity does not play significant role in the categorization by the current ethnicity or race. From the genomic view point, the overwhelming portion of the inherited genomes, which is about 99.8% identical between two individual's autosomal-genomes, do not play a very influential role in the categorization of ancestry, ethnicity or race. Furthermore, even the genome-based categorization of GGs is not only derived from a very small fraction (about 0.2%) of the whole genome, but also may have originated from the environmental selection of the genomic variations of the 0.2%. Thus, our results imply that both ethnicity and race are non-genomic, i.e. *environmental*, categorization, be they biological and/or non-biological.

Data availability

All genomic data used in this study have been released and are publicly available from: (a) the Simons Genome Diversity Project (https://www.simonsfoundation.org/simons-genome-diversity-project/; https://reichdata.hms.harvard.edu/pub/datasets/sgdp/) and (b) the 1000 Genomes Project (https://www.internationalgenome.org/data; http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502). For this study, The SNP data from The Simons Genome Diversity Project database[5,6] were last accessed in July of 2020, and those from The One Thousand Genome Project database[4,19] were last accessed in June of 2020. No additional new genomic data have been generated from this study.

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.

Received: 12 September 2022; Accepted: 25 March 2023
Published online: 18 April 2023

References

1. Yudell, M., Roberts, D., DeSalle, R. & Tishkoff, S. Taking race out of human genetics. *Science* **351**(6273), 564–565 (2016).
2. Sirugo, G., Tishkoff, S. A. & Williams, S. M. The quagmire of race, genetic ancestry, and health disparities. *J. Clin. Invest.* **131**(11), e150255. https://doi.org/10.1172/JCI150255 (2021).
3. Borrell, L. N. *et al.* Race and genetic ancestry in medicine—a time for reckoning with racism. *N. Engl. J. Med.* **384**, 474–480. https://doi.org/10.1056/NEGMms2029562 (2021).
4. The 1000 Genomes Project, A global reference for human genetic variation. *Nature* 52, 68–73 (2015).
5. Mallick, S. *et al.*, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206 (2016). Updated April 12, 2017. https://www.simonsfoundation.org/simons-genome-diversity-project/. Accessed 17 July 2020.
6. Fan, S. *et al.* African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biol.* **20**(82), 1–14. https://doi.org/10.1186/s13059-019-1679-2) (2019).
7. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302–310. https://doi.org/10.1038/nature21347; pmid:28102248 (2017).
8. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, 6484 (2020).
9. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
10. Deerwester, S., Dumais, S. T., Furnas, G. W. & Landauer, T. K. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**, 391–407 (1990).
11. Sims, G. E., Jun, S., Wu, G. A. & Kim, S. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. USA* **106**, 2677–2682 (2009).
12. R Core Team, R: A language and environment for statistical computing (R Foundation for Statistical Computing , 2017). https://www.R-project.org/.
13. Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**(4), 406–425 (1987).
14. Gascuel, O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**(7), 685–695 (1997).
15. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**(W1), W256–W259 (2019).
16. Nurk, S., *et al.* The complete sequence of a human genome. *bioRxiv* preprint (2021). https://doi.org/10.1101/2021.05.26.445798.
17. GRCh37. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/. Accessed 15 Mar 2021.
18. GRCh38. https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26. Accessed 15 Mar 2021.
19. Byrska-Bishop, M., *et.al.*, High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv preprint (2021). https://doi.org/10.1101/2021.02.06.430068.
20. Horsthemke, B. A critical view on transgenerational epigenetic inheritance in humans. *Nat. Commun.* **9**(2973), 1–4 (2018).

Acknowledgements

We are grateful to the Simons Genome Diversity Project 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[16] and Dr. Michael Zody of New York Genome Center, NY, USA on updated information about human reference genome GRCh38[19]. 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 beginning part of manuscript preparation.

Author contributions

Conceptual design of the study and speculative interpretations and implications of the results by S.H.K.; all computational work including algorithm design, programming and execution for this study was done mostly by B.J.K. and Neighbor Joining tree construction by J.J.C.; discussion and interpretation of computational results by B.J.K., J.J.C. and S.H.K.; manuscript preparation by S.H.K. with extensive discussions with B.J.K. and J.J.C.; figures designed by B.J.K., J.J.C. and S.H.K.

Funding

World Class University Project, Ministry of Education, Science and Technology, Republic of Korea; a gift grant to the University of California (41349-10805-44-CCSHK), Berkeley, CA. USA; The Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A03041954).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-32325-w.

Correspondence and requests for materials should be addressed to S.-H.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.