Genome Biology and Evolution Advance Access published January 7, 2015 doi:10.1093/gbe/evv003

Genome Biology and Evolution

SMBE

Inference Of Distant Genetic Relations In Humans Using "1000 Genomes"

Ahmed Al-Khudhair[1], Shuhao Qiu[2,3], Meghan Wyse[2], Shilpi Chowdhury[2], Xi Cheng[2], Dulat Bekbolsynov[2], Arnab Saha-Mandal[1], Rajib Dutta[2], Larisa Fedorova[4], Alexei Fedorov[1,3*]

[1]Program in Bioinformatics and Proteomics/Genomics, University of Toledo, Health Science Campus, OH 43614, USA.

[2]Program in Biomedical Sciences, University of Toledo, Health Science Campus, OH 43614, USA.

[3]Department of Medicine, University of Toledo, Health Science Campus, OH 43614, USA.

[4]GEMA-biomics, Ottawa Hills, OH 43606, USA

To whom correspondence should be addressed. Tel +419-383-5270; Fax +419-383-3102; e-mail: Alexei.fedorov@utoledo.edu

Abstract

Nucleotide sequence differences on the whole-genome scale have been computed for 1092 people from 14 populations publicly available by the 1000 Genomes Project. Total number of differences in genetic variants between 96,464 human pairs has been calculated. The distributions of these differences for individuals within European, Asian or African origin were characterized by narrow unimodal peaks with mean values of 3.8, 3.5, and 5.1 million respectively and standard deviations of 0.1-0.03 million. The total numbers of genomic differences between pairs of all known relatives were found to be significantly lower than their respective population means and in reverse proportion to the distance of their consanguinity. By counting the total number of genomic differences it is possible to infer familial relations for people that share down to 6% of common loci identical-by-descent. Detection of familial relations can be radically improved when only very rare genetic variants are taken into account. Counting of total number of shared very rare SNPs from whole-genome sequences allows establishing distant familial relations for persons with 8th and 9th degree of relationship. Using this analysis we predicted 271 distant familial pair-wise relations among 1092 individuals that have not been declared by 1000 Genomes Project. Particularly, among 89 British and 97 Chinese individuals we found three British-Chinese pairs with distant genetic relationships. Individuals from these pairs share identical by descent DNA fragments that represent 0.001%, 0.004%, and 0.01% of their genomes. With affordable whole-genome sequencing techniques, very rare SNPs should become important genetic markers for familial relationships and population stratification.

Introduction

Accomplishment of "1000 Genome Project" revealed immense amount of information about variation, mutation dynamics, and evolution of the human DNA sequences. The obtained critical data were originally reported by the Project Consortium (Abecasis *et al.* 2010; Abecasis *et al.* 2012). These genomes have been already used in a number of studies, which added essential information about human populations, allele frequencies, local haplotype structures, distribution of common and rare genetic variants, and determination of human ancestry and familial relationships (see, for example, articles most relevant to this paper (Fagny *et al.* 2014; Gravel *et al.* 2013; Harris and Nielsen 2013; Hochreiter 2013; Moore *et al.* 2013)).

Knowledge of population stratification is important for medicine, specifically, in case-control association and cohort studies since unknown distant familial relationships could potentially compromise interpretation of collected data. Proper genetic identification of familial relationships is also critical for forensic identification, in criminal investigations, inheritance claims, and in other areas of human life.

Widely used haplotype data such as Y chromosome or mitochondrial DNA for identification of distant genetic relationships have limited applications due to the consideration of male or female lines of descent (Parson and Bandelt 2007; Willuweit *et al.* 2011). Estimation of genetic relatedness on autosomal genomic sequences is mainly based on genome-wide averages of the estimated number of alleles shared identically by descent (IBD) (Browning and Browning 2013; Huff *et al.* 2011; Weir *et al.* 2006). Various methods have been used to detect IBD familial relationships (Boehnke and Cox 1997; Li *et al.* 2014; Thompson 1975). The most commonly used

GEMLINE, fastBD, ISCA and ERSA. A most sophisticated approach, ERSA2.0, for IBD identification depends on the complicated statistical methods. Yet, only with confidence (97%) it can identify up to 5th degree relatives while deeper relations with confidence of less than 80% in simulated or mixed populations using genome-wide genotyping arrays and whole genome sequencing (Huff *et al.* 2011; Li *et al.* 2014). Recent analysis by Durand and co-authors demonstrated that GEMLINE method when applied for analysis of nearly three thousand real, non-simulated, father-mother-child trios had over 67% of false positive rate (Durand *et al.* 2014). The same authors introduced non-probabilistic additional computationally effective metric to score IBD fragments, HaploScore, to improve accuracy of IBD detection methods. However the efficiency and reliability of such approach to testing of familial relationship in generations deeper than first was not tested.

Aiming to advance identification of distant familial relationships, we undertook computational examination of publicly available 1092 genomes. Genomic differences across all autosomes (total number of different genetic variants) have been computationally assessed for all possible 45,747 human pairs from the same populations and also for 50,717 pairs of individuals taken from different populations, which represent 9% of all possible inter-population pairs and chosen randomly. We found that in-line with previous publications most genetic variations are found within human populations (Barbujani *et al.* 1997; Jobling and Gill 2004). We also observed that pairs with declared familial genetic relations have the least genomic differences compared to other non-related pairs from the same population. By simply counting the total number of genomic differences it is possible to infer familial relations for people that share down to

6% of common IBD genetic materials. Here we demonstrated that the detection of familial relations would be drastically improved (by the order of magnitude) when only very rare genetic variants (vrGVs, with frequencies less than 0.2%) are taken into account. This paper demonstrates that simple counting of total number of shared vrGVs from whole-genome sequences allows establishing with high certainty (p-value < 0.001) distant familial relations for persons with 8th and 9th degree of relationship (people that have merely a fraction of a percentage of a coefficient of relationship (***r***) as defined by Sewall Wright (Wright 1922)). This is a very simple and powerful method for estimation of familial relationship based on vrGVs comparison, which requires whole-genome sequencing. With the availability of Illumina's new HiSeq X Ten device, the price of human genome sequencing this year was reduced three times to $1000 per genome. After accomplishment of the technology race to $100 per genome in the nearest future, vrGVs should become affordable important genetic markers for familial relationships and a broad range of population genetics studies.

Materials and Methods

Assessing the Total Number of the Genomic Variants Differences

We used data from the 1000 Genomes Project that are available through public ftp site ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/ (Abecasis *et al.* 2012). Specifically, Variant Call Format (VCF) files version 4.1 that contained a total of 38.2 million SNPs, 3.9 million short insertions/deletions and 14 thousand deletions for all the human chromosomes have been used. Information about genotype for each sequenced

individual was extracted from the GT-field of VCF files "as is" in the 1000 Genomes dataset. The genotype likelihood information (GL field) has not been considered.

A large-scale computational analysis using a combination of Perl programs was carried out to process and assess the total genetic differences between each pair of individuals. The programs were run on the Oakley supercomputer (https://www.osc.edu/supercomputing/computing/oakley) in the Ohio Supercomputer Center or their optimized versions on our local Linux workstation. All Perl Programs utilized in this project are available at our web page "http://bpg.utoledo.edu/~afedorov/lab/prog.html". These programs include the following: 1) *Intra_PopGenomeDif.pl* and *Inter_PopGenomeDif.pl* that computes the total number of genetic variant differences between pairs of individuals from the same and different populations respectively; 2) shell script *Batch_Populations.sh* for batch-distributing the program to multiple cores in Oakley; 3) *2individualsGenomeDif_vrGVs.pl;* 4) *IDs_seperator_rareSNPs.pl*; 5) *Intra_PopGenomeDif_vrGVs.pl* and *Inter_PopGenomeDif_vrGVs.pl* that computes the total number of shared vrGVs between individuals from the same and different populations respectively. The step-by-step description how to use these programs is presented in the Supplementary file S1.

Computer modeling of genomic differences has been performed with the program *GenomeDiffSimulation.pl.* The explicit instructions to this program are inserted as the comments into this script.

Each insertion or deletion has been counted as a single disparity not taking into account the length. Both parents' alleles have been considered. As an example, for a polymorphic site containing alleles A_1 and A_2, we counted as two differences between

persons homozygous with A_1 and homozygous with A_2 and as a single difference between a heterozygous person and a homozygous one. For a population of size *N*, all possible pairs ($N^2/2$) have been computationally processed for their intra-population genomic differences.

Supplementary Table S1 shows a summary of the samples the 1000 Genomes Project has sequenced and been used in this project. Our analysis included the entire set of human autosomes while X and Y-chromosomes have been omitted to allow a proper comparison between males and females.

Statistics

A non-parametric statistical method (Kruskal-Wallis Test (KRUSKAL 1952) for testing equality of population medians among groups is used to assess for significant differences among populations and among continental ancestors.

Kruskal-Wallis Test is identical to an ANOVA (5.1.4) with the data replaced by their ranks. The data analysis is performed using R commander package.

Number of Very Rare Genetic Variants Shared Between Relatives

We set our frequency threshold for vrGVs as less than 0.2% based on the number of studied individuals (1092) that provide data for the 2184 haploid genomes. With this threshold, the genetic variants with less than 5 minor allele counts (in other words singletons, doubletons, tripletons, and quadrupletons) among 2184 studied haploid genomes were considered as vrGVs.

A subset table of the autosomal vrGVs information for the 1092 individuals is created using a Perl program (*IDs_seperator_rareSNPs.pl*). The table included solely variants (very rare genomic variants) with frequency as less as 0.2%. Using the rare variants table, a second Perl program (*Intra_PopGenomeDif_vrGVs.pl*) used to assess the number of rare variants shared between each pair of individuals within the same population. In order to assess the rare variants shared between individuals from different population, a Perl program named (*Inter_PopGenomeDif_vrGVs.pl*) was developed.

We referred to familial relations following Sewall Wright (Wright 1922) in degree of relationship and coefficient of relationship (***r***). However, 1000 Genomes Project uses another term – first, second, and third order of relations, which is not well defined. Since we examined 1000 Genomes datasets we also used "order of relations" referring to the 1000 Genomes Project data.

Results

Genomic Differences Among Humans

We have computed the total number of genomic differences between pairs of individuals whose DNA sequences are available from the "1000 Genomes" project. Our analysis included the entire set of human autosomes while X and Y-chromosomes have been omitted to allow a proper comparison between males and females. Figure 1 illustrates the intra-population results for 14 populations from Africa, America, Asia, and Europe. All pairs of individuals with declared family relationships are marked by stars in Figure 1B. These pairs have significantly fewer genomic differences than the remaining non-related pairs from the same population. Statistical examination of the intra-

population distributions using Kruskal-Wallis test showed that, with the 0.05 significance level, the distributions are different from each other except for CHB and JPT populations (see statistical details in Supplementary file S2). The inter-population genomic differences are presented in the Supplementary Figure S1.

Computer Modeling Of Genomic Differences

Intriguingly, the number of genomic differences within Asian, European and African populations are shaped as narrow peaks with mean values of 3.5, 3.8, and 5.1 million respectively, and standard deviations in the range of 0.03-0.1 million (Figure 1A). Since a majority of human genes have several major mutually-exclusive haplotypes, comprising dozens to hundreds of frequent SNPs (Consortium 2003), the number of genomic differences for a particular gene between pairs of human individuals should range from 0 (when compared individuals carry the same gene haplotypes) to dozens or hundreds of differences (when compared individuals carry different haplotypes of the gene under analysis). In order to understand the reason why the genomic differences for African, Asian, and European populations on the Figure 1 are distributed as single narrow peaks, a computer program *GenomeDiffSimulation.pl* has been created. This program models the genomes of virtual individuals that, on an average, contain 3,800,000 different SNPs between them. In addition, these SNPs are grouped into several (four by default) mutually exclusive haplotypes for each genomic locus of the virtual individuals. The variable parameter for this program is the total number of loci that are in linkage equilibrium with each other.

The computational results for the distribution of the total differences in SNPs between pairs of virtual individuals are shown in Figure 2. The width of the peaks in the Figure 2A essentially depends on the number of genomic loci, in which SNPs are in linkage equilibrium with each other. In the model where the number of loci with linkage equilibrium is 5,000, the peak for the total genomic differences between virtual individuals (shown in blue) closely matches the shape of the peak computed for the actual Great Britain population (which, for comparison, is also present in Figure 2A and shown as a red bold line). This number (5,000) of chromosomal loci with linkage equilibrium with each other roughly corresponds to that in the human genome. There is an ambiguity in the estimation of the exact number of such loci in humans because of the fact that linkage disequilibrium between SNPs in humans decays continuously with increasing physical distance between SNPs, and also depends on the local recombination rate, which is highly variable along chromosomes (Arnheim *et al.* 2003). If the human genome consisted of 5,000 loci with mutual linkage equilibrium, the average size of the locus would have to be 600 Kb. This nucleotide length in the human genome corresponds to 0.6 centimorgan for genetic distance, which seems reasonable for modeling of the locus size. Hence, 5,000 loci with mutual linkage equilibrium give a rough approximation of the human genome. This estimation is congruent to common view in Hartl and Clark textbook (page 543) (Hartl 2007). However, for more precise estimation, the population history and demography should be taken into account. All in all, we attribute the narrow width of the peaks for the genomic differences in long-established African, Asian, and European populations to the presence of several thousand chromosomal loci in mutual linkage equilibrium. In each of these relatively old populations, the haplotypes of the loci

have been well shuffled and all individuals have equal chances of carrying a particular haplotype. Figure 1, also reveals much wider peaks for the American populations. We attribute this increased width to the recent admixture in populations of the New World, where European, African, and Native American genomic ancestry may be observed in various proportions in different people.

Our *GenomeDiffSimulation.pl* program has an option to mimic close genetic relations for several pairs of virtual individuals. A user may assign specific genetic relations for these pairs such as siblings (which share 50% of common genetic material IBD), second order of genetic relations (for example aunt/niece with 25% of common genetic material IBD), third order of relations (cousins with 12.5% of common IBD loci), or other more distant relatives with any user-defined percentage for common genetic loci. The genetically related pairs of virtual individuals have been simulated and five of these computational experiments are presented on the Figure 2B, where positions of pairs with genetic relations are marked by stars. Positions of virtual individual pairs with first and second order of genetic relationships (50% and 25% of common IBD loci respectively) correspond well to the positions of the actual human pairs having declared family relationships from the 1000 Genomes. For example, genetically-related pairs of virtual individuals are compared with pairs from Great Britain populations in Figure 2B. We observed that positions of siblings and parent/child pairs are always located in the extreme left of their corresponding population peak, followed by pairs with the second order of relations, which are closer to the corresponding peaks, and so on.

In the Figure 1B, the positions of several pairs within Luhya in Webuye, Kenya (LWK), Southern Han Chinese, China (CHS), British in England and Scotland (GBR) populations that are located close to the left slopes of their respective population peaks should correspond to the fourth or fifth order of genetic relations (6.2%-3.1% of shared IBD genetic materials). The genetic relations for these pairs have not been declared, yet with this analysis we can infer their putative genetic relations (which also has been confirmed by the distributions of very rare SNPs, see next paragraph). However, according to our computer simulations, the pairs with the fifth and higher order of relations (3.1% and less percentage of common genetic materials) may frequently be located within the left slopes of the corresponding peaks together with genetically non-related pairs (see Figure 2B). Thus, prediction of fifth and higher orders of genetic relations based on the total number of genomic differences appears to be unreliable. This limitation in identifying genetic relationships exists because a majority of genomic differences between pairs of individuals is contributed by frequent SNPs that form several (usually from two to five) major haplotypes in each loci (CONSORTIUM 2003). These major haplotypes have a high probability of being the same between genetically non-related individuals. This obstacle can be overcome if we consider only the very rare SNPs, for which probabilities of being shared by chance in non-related individuals drop dramatically (in the direct reverse proportion to the frequency of the considered SNPs).

Distributions Of Shared Very Rare Genetic Variants In Humans

In order to explore this possible method for predicting distant genetic relations in humans, we computationally filtered a complete subset of very rare genetic variants (vrGVs) from the “1000 Genomes” database having frequencies of less than 0.2% in the

2184 chromosomes from 1092 sequenced individuals. The distributions of positions of vrGVs along chromosomes are uniform and cover a vast majority of genomic regions, as exemplified in Figure 3 and detailed in the Supplementary Table S2. About 99% of these vrGVs are inside introns or intergenic regions. The number of shared vrGVs between each pair of individuals from the same population has been calculated (Figure 4). The graph reveals that a vast majority of examined pairs from American, Asian, and European populations shared from 50 to 300 vrGVs and form unimodal peaks for each population (Figure 4A). A majority of pairs from three African populations (African Ancestry in Southwest US (ASW), Luhya in Webuye, Kenya (LWK), and Yoruba in Ibadan, Nigeria (YRI)) share from 200 to 800 vrGVs, and also form unimodal peaks for each population. However, among all 14 populations, 311 pairs shared much higher number of vrGVs, (more than a thousand per pair) with the highest number of shared vrGVs being 46,745. Such extra-long tails in the distributions of shared vrGVs were even problematic to illustrate in the same figure together with the main peaks. Therefore, we presented these tails separately in Figure 4B, which has a 50 fold different scale compared to the peaks in Figure 4A. All 40 pairs with declared genetic relationships from 1000 Genomes are marked by stars in Figure 4B. These declared relatives share 6,252 to 46,745 vrGVs and represent the right-most points in the tails of distributions in Figure 4B. Besides these 40 pairs of known relatives, there are 271 pairs on Figure 4B that shared more than a thousand vrGVs (see Supplementary Table S3) and also dozens of pairs in Figure 4A that share several hundreds of vrGVs, which are on the right side of corresponding peaks and clearly separated from the peaks.

Interestingly, these right tails of distributions of vrGVs have population-specific patterns. For example, one of the African populations, LWK, has the highest number of pairs (260), each with more than a thousand of shared vrGVs. At the same time another African population (YRI) has only two of such pairs that share 1193 and 1841 vrGVs. Since the information about the individuals and strategies of their sampling for 1000 Genome Project is publicly unavailable, it is impossible to investigate this issue further. We hypothesize that pairs of individuals that share more than a thousand of vrGVs should have family relationships. Even those pairs, that share hundreds of vrGVs and are clearly separated from the main peaks, are likely formed by distant relatives.

This hypothesis is strongly supported by the calculations of the number of shared vrGVs between populations, shown in Supplementary Table S4. All studied 44,278 pairs formed by individuals from two different continents have less than 118 shared vrGVs. (For example, the highest number of shared vrGVs between LWK and JPT is 37; LWK-FIN is 80; and GBR-CHB is 117.) The number of shared vrGVs between populations from the same Asian or European continent is also low (for instance, maximal number between GBR and FIN is 159 and between CHB and JPT is 78). This means that a pair of European and/or Asian individuals that shares more than 300 vrGVs very likely has a familial relationship. The distributions of shared vrGVs between African populations (LWK vs. YRI and LWK vs. ASW) are demonstrated in Figure 5. With three exceptions, all studied 14,453 pairs formed by individuals from two different African populations have less than 623 shared vrGVs (these three exception pairs are discussed in the next section). Detailed examination of the inter-population distribution of shared vrGVs was

performed on the entire set of 8633 British-Chinese pairs formed by one individual from GBR and another individual from CHB population (see Table 1). This table demonstrates that a vast majority (8547) of these pairs have only single digit numbers of shared vrGVs. Only 3 out of 8633 pairs have 30 or more shared vrGVs. The distribution of shared vrGVs along chromosomes for these three pairs has been analyzed with a Perl program – *2individualsGenomeDif_vrGVs.pl*. The results for the HG00255-NA18614 pair, which has 30 shared vrGVs, are shown in the Table 2, while the data for other two pairs with 59 and 117 shared vrGVs are shown in the Supplementary Table S5. Table 2 demonstrates that 27 out of 30 shared vrGVs are located inside a 71 Kb genomic segment (positions form 90,787,654 to 90,858,949 nts) within chromosome 11. All clustered vrGVs do not show correlations with structural variants in this region. In addition, supplementary Table S6 demonstrates that shared vrGVs for HG00255 and NA18614 individuals are present on the same haplotype background. Similar clustering of shared vrGVs was observed for two other British-Chinese pairs (see Table S5). The pair HG01334-NA18627 has all 59 shared vrGVs located within 284 Kb locus on chromosome 1, while another HG00263-NA18541 pair has 115 shared vrGVs within a 806 Kb region inside chromosome 6. Given the enormous size of the human genome (3,300 Mb), the probability (*P*) of occurrence by chance for the case presented in the Table 2 that corresponds to 27 out of 30 shared vrGVs located inside 71 Kb region is less than 10^{-117}, according to the formula (1).

$$P = C_{30}^{3} * (71000/33000000000)^{26} \qquad (1)$$

This formula (1) assumes that all 30 vrGVs are independent and in equilibrium with each other. Therefore, undoubtedly, 27 out of 30 independent vrGVs cannot be located within

the same short locus by chance. This means that these three British-Chinese pairs represent very distant genetic relatives and their shared vrGVs located in the same locus are identical by descent and are in linkage disequilibrium with each other. Our observations of the chromosomal distributions of shared vrGVs are in a complete accordance with the population genetics theory that genetic inheritance occurs through chromosomal IBD segments, which are likely to become smaller and smaller with generations due to meiotic recombination events (Browning and Browning 2010; Huff *et al.* 2011). In agreement with this theory, these three British-Chinese pairs with very distant genetic relations should likely have only one short IBD per pair. The percentage of common genetic materials (C%) identical by descent for the British-Chinese pairs under consideration may be calculated by the formula:

$$C\% = (\Delta l/2L)*100\% \qquad (2)$$

Where Δl is the size of the IBD segment and L is the size of haploid genome. According to (2), these three pairs with 30, 59, and 117 shared vrGVs should have 0.0011%, 0.0043% and 0.012% of common genetic materials respectively.

If we consider relatively old population that existed for many hundreds of years (like GBR, FIN, or CHS), a majority of its individuals are likely to be in extremely distant genetic relations to each other (let's say 20 generations apart). Hence, they should share multiple and very short IBD chromosomal segments (a few thousands of nucleotides) because these IBD segments have been divided by recombinations in multiple generations. All these short IBD segments should contain only a few vrGVs due to their small size. In this respect, let's consider for example the Chinese (CHS) population in which the distribution of shared vrGVs has a peak of 94 (see Fig 4A). A

NA18548-NA18567 pair from this population has 303 shared vrGVs and is clearly separated from the corresponding peak on the Fig 4A. The distribution of shared vrGVs for this pair is demonstrated in the Table S5. This pair also has a single 36.9 Mb IBD segment on chromosome 2 that contains 199 shared vrGVs. The rest 104 vrGVs have a relatively random distribution across all chromosomes. Several of these 104 shared vrGVs may occasionally be grouped within a short chromosomal region (see Table S5). On the contrary, if we consider a pair from CHS that has a number of shared vrGVs around the peak value of 94 (for instance pair HG00557-HG00610 with 80 shared vrGVs) the distribution of shared vrGVs along chromosomes for this pair does not have any prominent IBD that contains more than 9 shared vrGVs (see Table S5). Supplementary Table S5 also contains examples of two intra-population pairs for GBR individuals (HG00109-HG00117 and HG00101-HG00099) containing 276 and 324 shared vrGVs respectively. The number of shared vrGVs corresponding to the two pairs are significantly higher than the peak value of 42 for this population. These pairs have several IBD segments on different chromosomes each containing dozens of shared vrGVs, so these individuals should be in distant genetic relation to each other.

Finally, we examined the inter-population distribution of shared vrGVs for three populations with African origin (see Figure 5). There are three pairs that have the highest numbers of shared vrGVs and they are clearly separated from the rest of the pairs illustrated in Figure 5. They are the following: (NA19443– NA18508) pair for LWK-YRI populations with 1121 shared vrGVs and two pairs for LWK-ASW populations, NA19350 - NA20348 and NA19397- NA20348 with 903 and 939 shared vrGVs

respectively. Distributions of shared vrGVs along chromosomes for these three pairs are also presented in the Table S5. The LWK-YRI pair has a prominent 8.5 Mb IBD region on chromosome 8 that contains a vast majority (1037) of all shared vrGVs. Therefore this pair has 0.13% of common genetic material according to formula (2). The other two pairs from LWK-ASW share the same person NA20348 from the ASW population. These two pairs also have a single prominent IBD spanning 14 Mb genomic segment on chromosome 11, which contains more than half (709) of all shared vrGVs for these two pairs. Therefore these individuals share 0.21% of common IBD genetic materials and should be distantly related to one another.

We did not perform the exhaustive inter-population comparisons of shared vrGVs because of the enormous amount of computational space required for computation of 549,842 pairs in total, which is beyond the scope of our resources. However, we expect that many more cases for inter-and intra-population distant genetic relationships will be revealed for the 1092 sequenced individuals. All in all, our approach is able to detect distant genetic relations that may share as small as 0.001% of genomic DNA.

Discussion

We demonstrated that human populations are distinct from one another by distribution of genomic differences among their individuals (see Figure 1) and also distribution of shared vrGVs (see Figure 4). Those populations that were formed thousands years ago -- African (LWK and YRI), Asian (CHS, CHB, and JPT), and European (GBR, FIN, TSI, CEU) have sharp and narrow peaks in the corresponding distributions of genomic differences, while populations from America that experienced

admixture a few hundred years ago, via inclusion of people from different continents, have much wider distributions of genomic differences (see Figure 1A).

Some human populations differ from others by distribution of shared vrGVs. For example, in the LWK population we observed the largest number of pairs (156) that shared more than 800 vrGVs. However, another African population, YRI, has only 7 of such pairs shared >800 vrGVs (see Fig 4). LWK population has the widest peak of the distribution of shared vrGVs with the mean-to-SD ratio of 1.2, whereas this ratio in European populations is about 0.3. One of the possible interpretations of this observation is that LWK might have experienced a high level of inbreeding, or it has a distinct subpopulation structure and the sample of LWK individuals were collected disproportionately from a few subpopulations.

Here we showed that genetic relationships can be effectively determined by the analysis of distribution of shared vrGVs between individuals. This analysis should take into account population structure. For example, number of vrGVs per individual varies among different geographic regions, being the highest in Africa (average vrGVs per individual in LWK is 67,200 and standard deviation, σ, is 7,500) and dropping to 16,200 in Europe (GBR population; σ=2,650) and 24,100 in Asia (CHB population; σ=4,100). In these calculations the threshold (0.2%) for vrGVs determination has been established based on the entire set of 1092 people from 14 populations. It makes sense to put such a threshold for each population discretely. This has not been done in this paper since we have not got enough statistics (the number of people in each population is less than 100).

Due to the differences in population structures, we observed significant variations in the number of shared vrGVs between the first and the second order relatives in different populations (see Figure 4 and Supplementary Table S3). First order relatives (shared 50% common genetic materials) have 28,000-46,000 shared vrGVs in Africa and only about 14,000-20,000 in Asia. This number is proportionally decreased for the second order relatives and further on.

There is a constant and intense influx of novel mutations in humans and other species. On average, each person has from 40 to 100 novel mutations that are absent in the genome his/her parents (Conrad *et al.* 2011; Kondrashov and Shabalina 2002; Li and Durbin 2011). A majority of these novel mutations are eliminated soon after their arrival by genetic drift and selection. Yet the remaining portion of novel mutations is an important endless source for vrGVs, which pool continuously renovates and maintains at a very high level (14-40 thousand vrGVs per individual in European and Asian populations). Recent computational analysis of the 1000 Genome database by Moore and coauthors (Moore *et al.* 2013) also demonstrated the highest abundance of rare GVs, yet they used slightly higher threshold (0.3%) for their frequencies. In the review by Keinan and Clark the authors summarized the common viewpoint that an excess of rare genetic variants has resulted from the recent explosive growth of human population (Keinan and Clark 2012). Whole-genome dynamics of millions of genetic variants is a very intricate issue that only recently has been touched in computer simulations (Qiu *et al.* 2014) and also in large-scale computations of 1000 Genomes Project data (Moore *et al.* 2013).

Impact of sequencing errors on the analysis of shared vrGVs.

As demonstrated on the Figure 3, the distribution of vrGVs along chromosomes is relatively even. A majority of vrGVs occurs inside largest genomics regions with the longest spans, namely the intergenic regions and introns. According to the publication of 1000 Genome consortium, these non-exome regions have the lowest sequencing coverage (on average x5 times), and thus they have the highest level of sequencing errors. On page 1065 of the 1000 Genome publication, the authors estimated that "in low-coverage project, the overall genotype error rate was 1-3%" (Abecasis *et al.* 2010). According to the same publication (page 1067), in some cases the error rate maybe ~4% (for CEU population) and ~10% for YRI depending on the sequence coverage for a genomic region. Misinterpretation of heterozygous sites with homozygous sites is the main cause of errors in interpreting genomic regions with low depth of sequencing coverage. For example, for a heterozygous person with a (G/A) SNP, when a sequence coverage is x6, there is a 1/32 chance that only G or only A nucleotides will be detected in all of the six reads (3% error). It means that, on average, 3% (and in some occasions up to 10%) of vrGVs are randomly missed in 1000 Genomes database. This effect partially explains the large intra-population variations in total number of vrGVs between individuals (see the Results section and Supplementary table S2). Another type of sequencing error is the misinterpretation of one nucleotide instead of another. The frequency of such type of errors has not been explicitly discussed in the reports of 1000 Genome. However, such errors should occur pretty randomly across the genome and in a majority of cases should be interpreted as an arrival of a novel mutation – a singleton. Such singletons should be

sparsely distributed across the genome and should increase the number of vrGVs in individuals. Since the length of the human genome is huge (3 billion nucleotides), one vrGV occurs, on an average, per 100 Kb region. Hence the probability that non-related individuals share the same vrGV is very low (less than one shared vrGVs) per pair. Taking into account that mutations did not occur randomly, but rather at particular hot-spots, this estimation may be raised to a handful of randomly shared vrGVs between non-related individuals. Indeed, when we compared number of shared vrGVs between continents (see Supplementary Table S4) the median number of shared vrGVs was 2 (for CHB-GBR populations), 6 (LWK-FIN), and 8 (for LWK-JPT). Therefore, sequencing errors due to nucleotide misinterpretation should be at most accountable for a handful of shared vrGVs between pairs of individuals and their impact on the overall vrGV analysis should be negligibly small.

In some populations marriage between relatives is a common practice. [http://www.consang.net/index.php/Global_prevalence). For example, we detected a pair from Colombian in Medellin, Colombia (CLM) (HG01277 and HG01278) that has the highest number (2863) of shared vrGVs for this population. According to "1000 Genomes" annotation table, this pair represent a husband and wife, and we project that they share about 6% of common IBD genetic materials. Therefore, we expect that their child (HG01279, not sequenced yet) should have more than 50% of common genetic materials with each of his/her parents. Presumably, due to this reason, the observed variation of numbers of shared vrGVs among the first order relatives is very wide compared to our modeling. For instance, in LWK population, this variation is from

31,000 up to 46,500. We conjecture that the highest numbers may correspond to the families where marriage occurred between genetic relatives. It is also worth mentioning that actual relationship between siblings or parent/offspring pairs may fluctuate noticeably from 50% (Odegard and Meuwissen 2012). Finally, even within the same population, the number of vrGVs among individuals significantly varies. For example, in Chinese population CHB the average number of vrGVs per individual is 24,100 while σ=4,100. In this population the lowest number of vrGVs (16,745) was detected in HG00403 person, while the highest 40,444 in HG00702 individual. All these facts together may explain the large variations in the numbers of shared vrGVs between the pairs of relatives with the same degree of relationship.

In summary, if two individuals share less than a dozen of vrGVs they should descend from different ethnic and geographically diverse populations. In case persons share several dozens of vrGVs located in the same chromosomal region they should have some degree of genetic relationship to each other. Finally, a pair may have dozens to hundreds of shared vrGVs that have a uniform spread over all chromosomes without a strong signal for preferential association or clustering within a particular locus. This means that some predecessors of these individuals belonged to the same population.

All in all, in addition to well-established DNA fingerprinting, application of vrGVs analysis for obtaining distant genetic relations could be a valuable molecular genetic technique in criminal investigations, in civil familial searching as well as for population, clinical and association studies.

Acknowledgements

We are grateful to Dr. Robert Blumenthal, University of Toledo Health Science Campus, for his insightful discussion of the project. The computations were performed in Oakley supercomputer with support from Ohio Supercomputer Center. We also appreciate the financial support from the Department of Medicine to conduct our research.

DISCLOSURE: The patent of our approach for detection of distant genetic relationships is pending.

# Shared vrGVs	# Human Pairs
0	903
1	1828
2	2045
3	1584
4	1009
5	605
6	298
7	149
8	68
9	58
10	22
11	20
12	13
13	3
14	3
15	5
16	4
17	2
18	2
19	0
20	0
21	3
22	1
23	1
24	0
25	0
26	0
27	0
28	1
29	1
30	1 *
……	0
59	1 *
……	0
117	1*

Table 1. Distribution of numbers of shared vrGVs for 8633 human pairs, where one person of a pair represents British population (GBR), while another person – Chinese population (CHB). *NOTES: detail characterization of shared vrGVs for the pair, which has 30 shared vrGVs, is shown in the Table 2. Detail characterization of

shared vrGVs for three pairs at the bottom of this table (marked by *) is shown in the Supplementary Table S5.

Chromo-some	Position of vrGVs	Identifiers of vrGVs	Reference allele	Alternative allele
CHR3	163910979	rs147633047	C	T
CHR9	42323192	rs184959358	G	A
CHR11	90787654	rs138781903	A	G
CHR11	90788511	rs141690807	C	T
CHR11	90788759	rs187621230	T	C
CHR11	90798281	rs144138129	G	A
CHR11	90806962	rs183908202	A	G
CHR11	90808684	rs147862657	C	A
CHR11	90812601	rs147197102	G	A
CHR11	90815124	rs190205439	C	T
CHR11	90816996	rs147226573	A	G
CHR11	90817266	rs185201515	G	A
CHR11	90826778	rs139867381	T	A
CHR11	90828732	rs149763439	G	A
CHR11	90835123	rs140038072	G	A
CHR11	90835556	rs140255793	A	G
CHR11	90840943	rs147799849	A	G
CHR11	90842479	rs142999510	G	T
CHR11	90843070	rs141928306	T	C
CHR11	90844258	rs139273514	A	G
CHR11	90847782	rs150575842	C	T
CHR11	90848531	rs147400508	G	A
CHR11	90848728	rs149904020	G	C
CHR11	90850157	rs138217375	G	T
CHR11	90852915	rs187692214	A	G
CHR11	90856705	rs144056495	G	A
CHR11	90858178	rs189208470	C	T
CHR11	90858721	rs139417643	G	A
CHR11	90858949	rs150070179	A	G
CHR20	42290810	rs146883107	C	T

Table 2. Characterization of 30 shared vrGVs for the British-Chinese pair composed by HG00255 and NA18614 individuals. Those vrGVs that are located in the same locus on chromosome 11 are shaded. The detailed description of shared vrGVs for this pair and also for eight other pairs described in the Results section, is provided in the Supplementary Table S5.

FIGURES

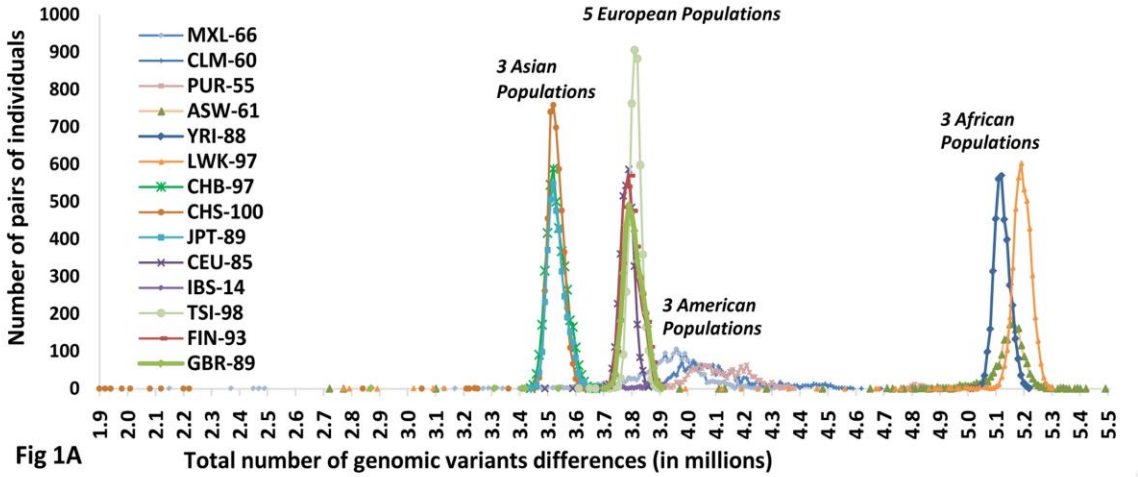

Fig 1A

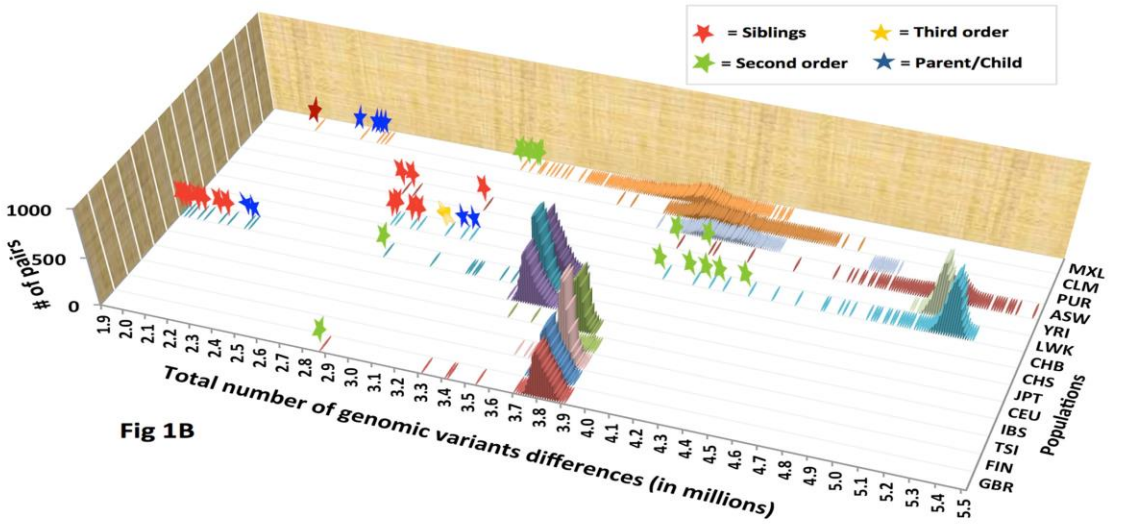

Fig 1B

FIGURE 1. Distribution of number of genetic variants (GVs) between all possible pairs of individuals within the same population. Three populations from Africa (ASW, LWK, YRI), three populations from America (CLM, MXL, PUR), three from Asia (CHB, CHS, JPT), and five from Europe (CEU, FIN, GBR, IBS, TSI) have been examined. Numbers of individuals in the populations are shown on the graph behind the population identifier (e.g. 66 people for MXL-66). The number of pairs has been calculated for bins (X; X+10,000), where number of genetic variants X is plotted on the graph and the bin size was 10,000 genetic

variations. **A** – Two-dimensional view of the distribution. **B** -three dimensional view of the distribution where all pairs with declared genetic relations are marked by stars. The color of a star reflects a specific genetic relationship: red stands for siblings, blue – parent/child pair, green – second order relations, and yellow – third order.

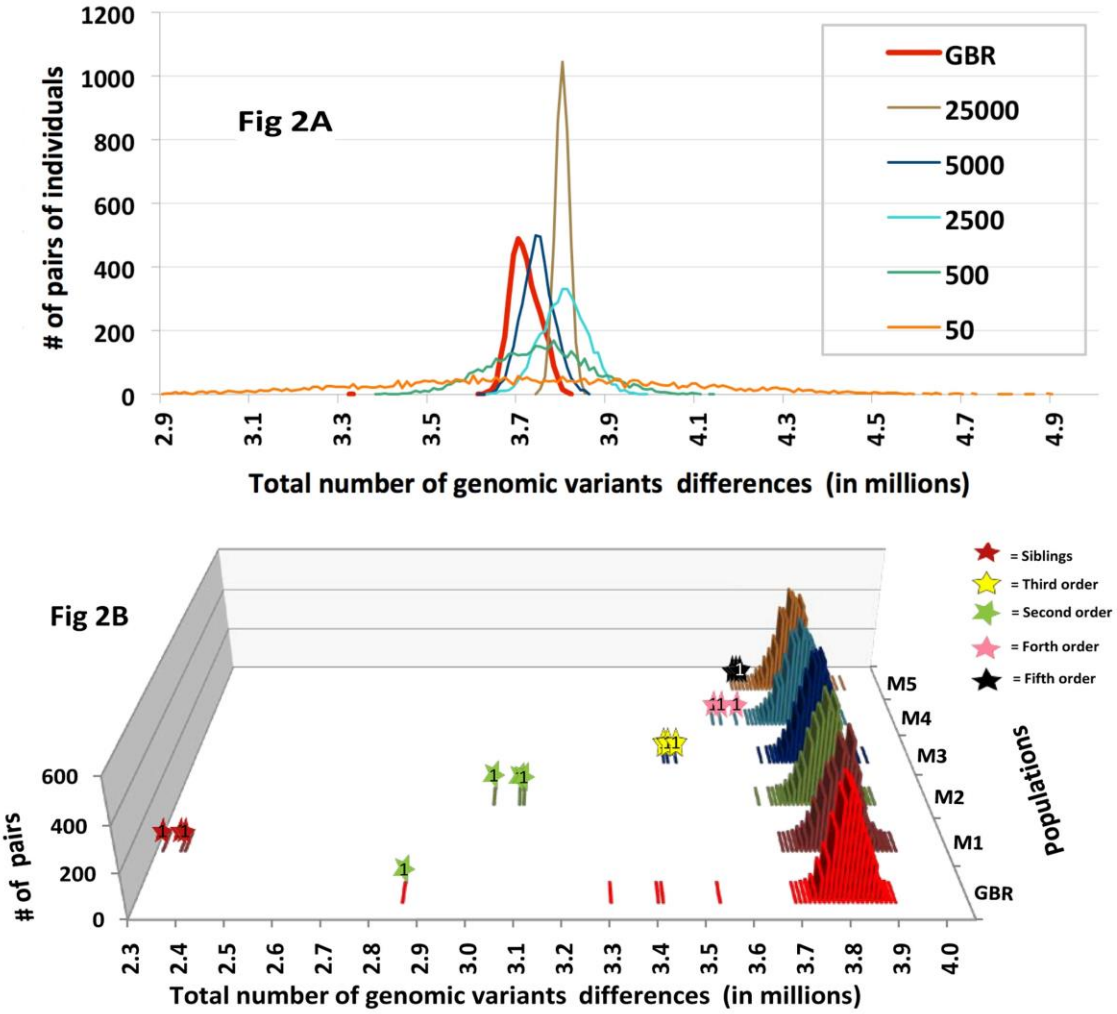

FIGURE 2. Distribution of number of genetic variants (GVs) between pairs of individuals from the same real and modeling populations.

A – Two-dimensional expanded view of the distribution. Real population from Great Britain (GBR) is shown as a red bold line, while the five other curves represent model populations of virtual individuals. Virtual individuals in all models have on average 3.8 million differences of genetic variants between them. Various models have different number of genomic loci that are in linkage equilibrium with each other. The model with the lowest number (50) of loci with equilibrium is shown by orange line and has the widest span. The model with the

highest number of loci in linkage equilibrium, 25,000, has the narrowest peak (brown line). When the number of loci with equilibrium in the modeling genome is 5,000 (navy blue line) the modeling distribution is most similar to the real one from GBR population (red line).

B – Three-dimensional view of the distribution where pairs with known genetic relations are marked with stars. The color of a star reflects a specific genetic relationship: red stands for siblings, green – second order relations, and yellow – third order, pink - fourth order, and black - fifth order of genetic relations. The front most distribution (red) represents the real population from Great Britain (GBR). The next five curves represent distributions for five model populations of virtual individuals (M1 to M5). In each of these five models the number of loci in linkage equilibrium with each other is the same - 5,000. Three pairs of virtual individuals mimic genetic relationships in every model. In M1 these three pairs are represented by siblings (that share 50% of common genetic materials from the most recent common ancestor). M2 represents three pairs with the second order of relations that share 25% of common genetic materials (e.g. aunt/niece). M3 represents three pairs with third order of relations that share 12.5% of common genetic materials (e.g. cousins). M4 –fourth order with 6.25%; and M5 – fifth order with 3.12% of common genetic materials. All three pairs with fifth order of genetic relations from M5 model are located in the same left-most bin together with one pair of virtual individuals that does not have genetic relations.

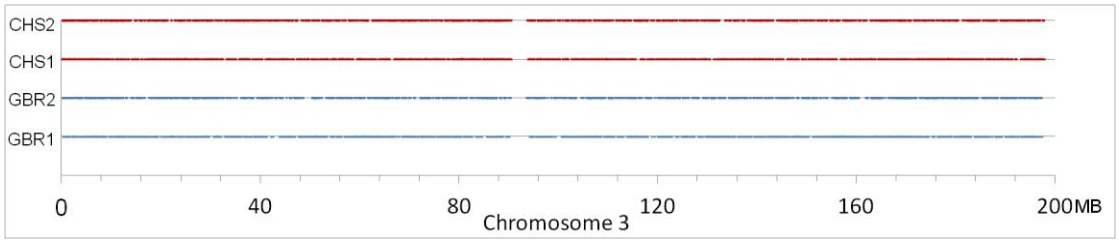

FIGURE 3. Distribution of vrGVs along chromosome 3 for four randomly picked individuals: two from Chinese (CHS) population (HG00404 and HG00407 individuals) and two from British (GBR) population (HG00097 and HG00099). Every vrGV is represented by a dot. The detail information about distribution of vrGVs along all chromosomes for these individuals is available from Supplementary Table S2.

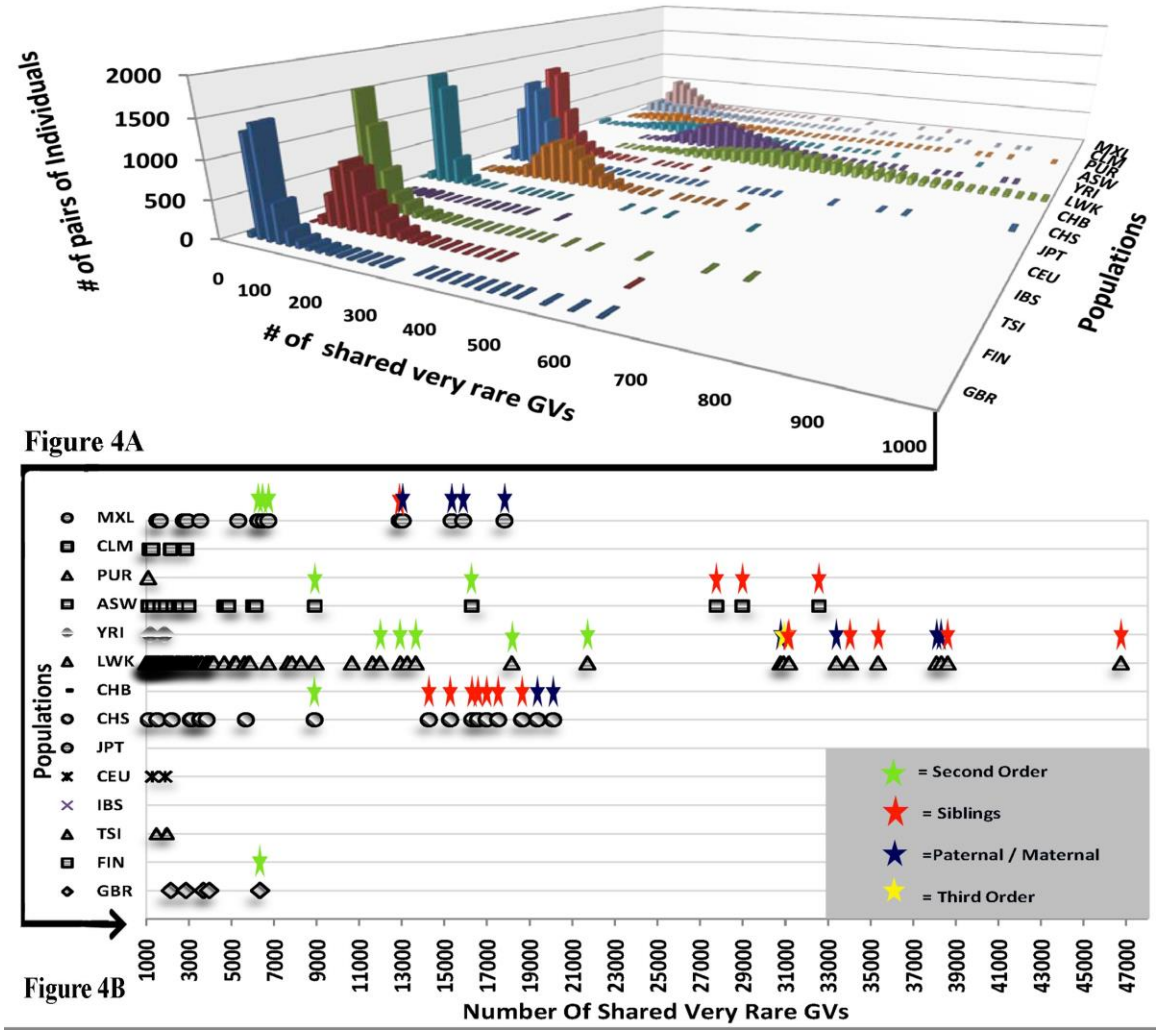

FIGURE 4. Distribution of number of shared very rare genetic variants (vrGVs) between all possible pairs of individuals from the same population. Three populations from Africa (ASW, LWK, YRI), three populations from America (CLM, MXL, PUR), three from Asia (CHB, CHS, JPT), and five from Europe (CEU, FIN, GBR, IBS, TSI) have been examined.

A – Three-dimensional view of the part of the distributions where the majority of pairs are located.

B – Two dimensional view of the tails of the distributions, where pairs are presented by circles, triangles, rectangles, and crosses specific for each population. All pairs with declared genetic relations are marked by stars. The color of a star reflects a specific genetic relationship: red stands for siblings, blue – parent/child pair, green – second order relations, and yellow – third order. Scale for the number of shared vrGVs in the graph 4A is expanded 50 fold compared to 4A.

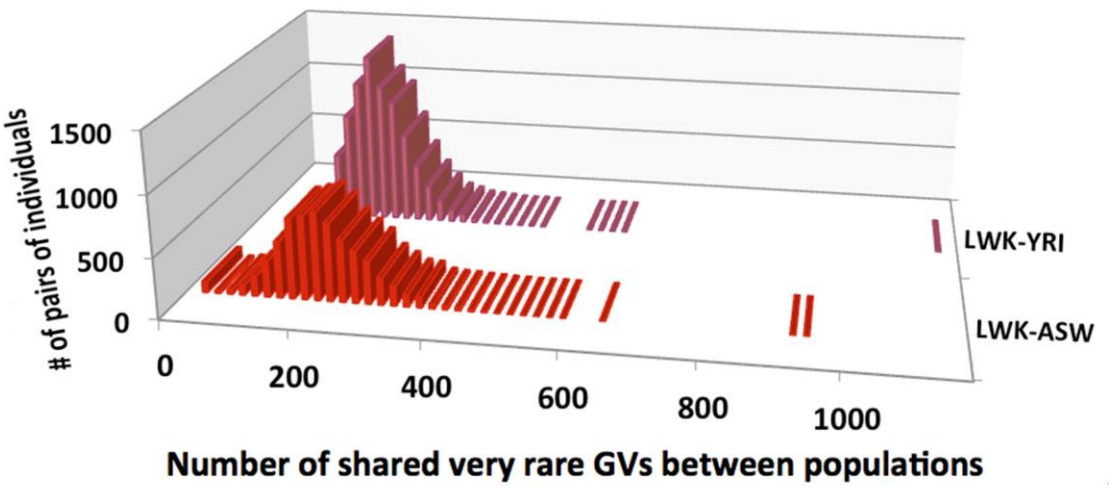

FIGURE 5. Distribution of number of shared vrGVs between pairs of individuals from different African populations. First distribution (shown in red) represents pairs in which one individual belongs to LWK population while another person to ASW. Second distribution (blue) represents pairs in which one individual is from LWK while the other is from YRI.

Literature Cited

Abecasis, G. R., D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin *et al.*, 2010 A map of human genome variation from population-scale sequencing. Nature **467:** 1061-1073.

Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. Nature **491:** 56-65.

Arnheim, N., P. Calabrese and M. Nordborg, 2003 Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. American journal of human genetics **73:** 5-16.

Barbujani, G., A. Magagni, E. Minch and L. L. Cavalli-Sforza, 1997 An apportionment of human DNA diversity. Proceedings of the National Academy of Sciences of the United States of America **94:** 4516-4519.

Boehnke, M., and N. J. Cox, 1997 Accurate inference of relationships in sib-pair linkage studies. American journal of human genetics **61:** 423-429.

Browning, B. L., and S. R. Browning, 2013 Detecting identity by descent and estimating genotype error rates in sequence data. American journal of human genetics **93:** 840-851.

Browning, S. R., and B. L. Browning, 2010 High-resolution detection of identity by descent in unrelated individuals. American journal of human genetics **86:** 526-539.

Conrad, D. F., J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang *et al.*, 2011 Variation in genome-wide mutation rates within and between human families. Nature genetics **43:** 712-714.

Consortium, I. H., 2003 The International HapMap Project. Nature **426:** 789-796.

Durand, E. Y., N. Eriksson and C. Y. McLean, 2014 Reducing Pervasive False-Positive Identical-by-Descent Segments Detected by Large-Scale Pedigree Analysis. Mol Biol Evol **31:** 2212-2222.

Fagny, M., E. Patin, D. Enard, L. B. Barreiro, L. Quintana-Murci *et al.*, 2014 Exploring the occurrence of classic selective sweeps in humans using whole-genome sequencing data sets. Mol Biol Evol **31:** 1850-1868.

Gravel, S., F. Zakharia, A. Moreno-Estrada, J. K. Byrnes, M. Muzzio *et al.*, 2013 Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS genetics **9:** e1004023.

Harris, K., and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS genetics **9:** e1003521.

Hartl, D. L., Clark, A.G., 2007 *Principles of Population Genetics*. Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, USA.

Hochreiter, S., 2013 HapFABIA: identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic Acids Res **41:** e202.

Huff, C. D., D. J. Witherspoon, T. S. Simonson, J. Xing, W. S. Watkins *et al.*, 2011 Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome research **21:** 768-774.

JOBLING, M. A., and P. GILL, 2004 Encoded evidence: DNA in forensic analysis. Nature reviews. Genetics **5:** 739-751.

KEINAN, A., and A. G. CLARK, 2012 Recent explosive human population growth has resulted in an excess of rare genetic variants. Science **336:** 740-743.

KONDRASHOV, A. S., and S. A. SHABALINA, 2002 Classification of common conserved sequences in mammalian intergenic regions. Human molecular genetics **11:** 669-674.

KRUSKAL, W. H. W., W.A., 1952 Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association **47:** 583-621.

LI, H., and R. DURBIN, 2011 Inference of human population history from individual whole-genome sequences. Nature **475:** 493-496.

LI, H., G. GLUSMAN, H. HU, SHANKARACHARYA, J. CABALLERO *et al.*, 2014 Relationship estimation from whole-genome sequence data. PLoS genetics **10:** e1004144.

MOORE, C. B., J. R. WALLACE, D. J. WOLFE, A. T. FRASE, S. A. PENDERGRASS *et al.*, 2013 Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS genetics **9:** e1003959.

ODEGARD, J., and T. H. MEUWISSEN, 2012 Estimation of heritability from limited family data using genome-wide identity-by-descent sharing. Genet Sel Evol **44:** 16.

PARSON, W., and H. J. BANDELT, 2007 Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet **1:** 13-19.

QIU, S., A. MCSWEENY, S. CHOULET, A. SAHA-MANDAL, L. FEDOROVA *et al.*, 2014 Genome evolution by matrix algorithms: cellular automata approach to population genetics. Genome Biol Evol **6:** 988-999.

THOMPSON, E. A., 1975 The estimation of pairwise relationships. Annals of human genetics **39:** 173-188.

WEIR, B. S., A. D. ANDERSON and A. B. HEPLER, 2006 Genetic relatedness analysis: 1

WILLUWEIT, S., A. CALIEBE, M. M. ANDERSEN and L. ROEWER, 2011 Y-STR Frequency Surveying Method: A critical reappraisal. Forensic Sci Int Genet **5:** 84-90.

WRIGHT, S., 1922 Coefficients of inbreeding and relationship. American Naturalist **56:** 330-338.