Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes

Larisa Fedorova[1,*], Shuhao Qiu[2,3], Rajib Dutta[4] and Alexei Fedorov[2,3,*]

[1]GEMA-Biomics, Ottawa Hills, OH,

[2]Program in Bioinformatics and Proteomics/Genomics, University of Toledo

[3]Department of Medicine, University of Toledo

[4]Program in Biomedical Sciences, University of Toledo

*Corresponding author: E-mail: lvfedorova3@gmail.com; Alexei.fedorov@utoledo.edu.

Accepted: February 13, 2016

Abstract

A novel computational method for detecting identical-by-descent (IBD) chromosomal segments between sequenced genomes is presented. It utilizes the distribution patterns of very rare genetic variants (vrGVs), which have minor allele frequencies <0.2%. Contrary to the existing probabilistic approaches our method is rather deterministic, because it considers a group of very rare events which cannot happen together only by chance. This method has been applied for exhaustive computational search of shared IBD segments among 1,092 sequenced individuals from 14 populations. It demonstrated that clusters of vrGVs are unique and powerful markers of genetic relatedness, that uncover IBD chromosomal segments between and within populations, irrespective of whether divergence was recent or occurred hundreds-to-thousands of years ago. We found that several IBD segments are shared by practically any possible pair of individuals belonging to the same population. Moreover, shared short IBD segments (median size 183 kb) were found in 10% of inter-continental human pairs, each comprising of a person from sub-Saharan Africa and a person from Southern Europe. The shortest shared IBD segments (median size 54 kb) were found in 0.42% of inter-continental pairs composed of individuals from Chinese/Japanese populations and Africans from Kenya and Nigeria. Knowledge of inheritance of IBD segments is important in clinical case–control and cohort studies, since unknown distant familial relationships could compromise interpretation of collected data. Clusters of vrGVs should be useful markers for familial relationship and common multifactorial disorders.

Key words: DNA, bioinformatics, evolution, genealogy, inheritance, biomarker.

Introduction

Studies of genetic relatedness rely on the fundamental concept of identical-by-descent (IBD) for inheritance of genetic material (Powell et al. 2010; Browning and Browning 2012; Carmi et al. 2013, 2014; Thompson 2013). Genome of every individual is a mosaic of IBD segments inherited from previous generations. Real human populations have limited sizes and have frequently experienced admixtures. Thus, even genealogically unrelated individuals from the same geographical region frequently share one or several IBD genomic segments transmitted from their common distant ancestors. Investigation of peculiarities in IBD segment inheritance is critical for understanding fundamental questions regarding human evolution and demographic history, as well as for practical purposes including individualized medicine and clinical association studies. However, precise detection of IBD segments, even when shared by not-very-distant genetic relatives, has several problems. Whole-genome SNP analysis on gene arrays frequently produces erroneous results (Browning and Browning 2011; Huff et al. 2011; Durand et al. 2014; Li et al. 2014). The widely used shotgun next-generation sequencing does not confidently distinguish maternal and paternal genomic portions (the so-called "phasing" of sequenced DNA). Numerous phasing errors in distinguishing between parental chromosomes have led to frequent incorrect IBD segments detection (Kong et al. 2008). Characterization of IBD segments by current methods depends on complex statistical algorithms, multiple assumptions, and probabilistic approaches (Su et al. 2012; Browning and Browning 2013; Durand, et al. 2014). Hence, false positive and false negative

predictions often take place in establishing distant genetic relatedness.

Our group recently presented a novel and simple computational method for detecting shared IBD segments (Al-Khudhair et al. 2015). This method utilizes the distribution patterns of very rare genetic variants (vrGVs), which have minor allele frequencies <0.2%, and does not require phasing of genomic sequences. Since all living species experience an intense influx of mutations in their genomes, vrGVs are very abundant. Any given human being has 50–100 *de novo* DNA changes, on average (Conrad et al. 2011; Li and Durbin 2011; Kondrashov and Shabalina 2002). Due to this intense mutagenesis, vrGVs occur by the tens of thousands in every individual and their patterns along chromosomes are exceptional clues and signs of their most recent evolution. Usefulness of rare SNPs has been acknowledged in several publications (Hochreiter 2013; Moore et al. 2013). We showed that shared vrGVs between two individuals are clustered in a single or a few genomic loci. This article introduces and defines clusters of vrGVs and presents a new approach to distinguish between identical-by-state (IBS) versus IBD chromosomal segments. When two people share five adjacent vrGVs located in the same region, the probability of this event occurring by random coincidence (the so-called IBS event) is equal to 0.002^5, which is less than one in 10^{13}. Therefore, these clusters of shared vrGVs are credible markers of IBD genomic segments. Five or more adjacent shared vrGVs are called as rare variant clusters (RVCs). Characterization of shared RVCs gives a remarkable reliability for IBD segment identification and, at the same time, precise localization of IBD segments on the chromosome. In this article, we characterized the entire set of shared RVCs for every possible pair of 1,092 sequenced individuals (1,191,372 pairs) and demonstrated that distribution of shared RVCs perfectly matches human history and migration routes during the last 9,500 years.

Materials and Methods

Datasets

We used data from the 1000 Genomes Project, phase 1, that are available through public ftp site ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/, last accessed March 1, 2016 (Abecasis et al. 2012). Specifically, Variant Call Format (VCF) files version 4.1 that contain 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.

We have defined vrGVs as polymorphisms in which minor allele counts have <0.2% frequencies in the 1000 Genomes data (Al-Khudhair et al. 2015). By processing VCF files of 1,092 individuals we identified 16,326,219 vrGVs in total for mutant alleles, which represent minor allele counts, and only 17,611 vrGVs for reference alleles. This asymmetry exists due to the fact that the reference human genome has been created based on the pooled information on six sequenced individuals (Lander et al. 2001). Since the vrGVs corresponding to reference alleles represent only 0.1% of vrGVs corresponding to mutant alleles, we omitted the former and processed only the vrGVs from mutant alleles.

Algorithms, programs, parameters, and calculated probabilities for false-positive detection of RVCs are presented in the supplementary file "M&M", Supplementary Material online.

Availability

All our programs, their instruction manuals and notes, supporting files, and results files are available in supplementary file "Data", Supplementary Material online, and also from our web site (http://bpg.utoledo.edu/~afedorov/lab/atlas_vrGV.html, last accessed March 1, 2016). Supplementary file "Data", Supplementary Material online, is an archive file in the ".tar.gz" format, which contains five folders: ProgramsInstructions, InputData, OutputDataWindow20, OutputDataWindow9, and Modeling. The execution of the entire pipeline of Perl programs for processing 1,092 genomes takes <3 h on a modern desktop Linux workstation using a single CPU (no parallelization required). The execution of our modeling program *IBDsimulator.pl* takes about 1–2 min per one primogenitor genome on a modern desktop computer. To repeat this program 500,000 times we ran it in parallel on 64 cores simultaneously for several weeks.

Statistics

Statistical analyses of the distribution of RVC lengths between two pairs of populations were performed using R package (R Development Core Team 2010).

Results

Creation of vrGV Databases

Complete sets of vrGVs for each of 1,092 individuals have been created and available as supplementary file "Data", Supplementary Material online (folder InputData). An example of such individual-specific vrGV database is shown in table 1, which represents an arbitrarily chosen individual (HG01365) from Colombian population (CLM). Each minor allele of vrGV in this individual-specific database is present in the genome of the person HG01365 and, often, in one to three other genomes of the 1,092 sequenced individuals. When a vrGV from an individual-specific vrGV database is shared by two, three, or four people, the identifiers of all individuals who have this minor allele are also present in the database (columns 6–9, table 1). The number of vrGVs of a particular person depends on the population that person belongs to (Al-Khudhair et al. 2015). The highest number of vrGVs is seen in the African populations (average number is 67,000 vrGVs per person; SD = 7,500), followed by American (average 24,600 vrGVs;

Table 1

An Exemplified Segment of an Individual-Specific Database for the Individual HG01363 from CLM Population

chr	Position	Identifier	ref	mut	Person-1	Person-2	Person-3	Person-4
CHR1	3438563	rs185156707	G	A	CLM_HG01365			
CHR1	3503010	rs141463795	C	T	CLM_HG01365	TSI_NA20813		
CHR1	3567024	rs184518958	A	G	CLM_HG01365			
CHR1	3669552	rs186811888	C	T	FIN_HG00355	CLM_HG01365		
CHR1	4022297	rs185199014	C	T	CLM_HG01365			
CHR1	4393739	rs116739584	T	G	FIN_HG00190	CLM_HG01365	CEU_NA11829	MXL_NA19762
CHR1	4530544	rs187766509	G	A	CLM_HG01365	CEU_NA11829	CHB_NA18749	JPT_NA18987
CHR1	4722235	rs148197646	T	G	CLM_HG01365	CEU_NA11829		
CHR1	4937903	rs185870613	G	C	CLM_HG01365	ASW_NA19922		
CHR1	4978507	rs183610263	A	T	CLM_HG01365	ASW_NA19922		
CHR1	5219590	rs191615351	C	T	GBR_HG00096	GBR_HG00106	GBR_HG00120	CLM_HG01365
CHR1	***5343566***	***rs146986028***	***C***	***T***	***GBR_HG00106***	***CLM_HG01365***		
CHR1	***5481720***	***rs190394368***	***G***	***A***	***CLM_HG01365***	***CEU_NA11931***		
CHR1	***5551303***	***rs186874087***	***G***	***A***	***CLM_HG01365***	***IBS_HG01625***	***TSI_NA20797***	
CHR1	***5553504***	***rs180676356***	***G***	***A***	***CLM_HG01365***	***IBS_HG01625***	***TSI_NA20797***	
CHR1	***5559272***	***rs192278468***	***G***	***C***	***CLM_HG01365***	***IBS_HG01625***	***TSI_NA20797***	
CHR1	***5560643***	***rs146515020***	***G***	***A***	***CLM_HG01365***	***IBS_HG01625***	***TSI_NA20529***	***TSI_NA20797***
CHR1	***5561084***	***rs187249140***	***G***	***A***	***CLM_HG01365***	***IBS_HG01625***	***TSI_NA20797***	
CHR1	***5576119***	***rs142071781***	***T***	***C***	***CLM_HG01365***	***IBS_HG01625***	***TSI_NA20795***	***TSI_NA20813***
CHR1	***5710524***	***rs189964921***	***G***	***A***	***CLM_HG01365***			
CHR1	5713296	rs189528396	C	A	CLM_HG01365			

In a window of nine consecutive rows (italics entries) the person HG01363 shares five vrGVs with the individual TSI_NA20797 and also six vrGVs with the individual IBS_HG01625. We have named such chromosomal regions with five or more shared neighboring vrGVs inside a scanning window as RVCs. The default size parameter for a scanning window is 20 consecutive rows.

SD = 4,500), Asian (24,100 ± 4,100), and European (16,200 ± 2,700) populations. The number of shared vrGVs for a particular pair of individuals also depends on the populations these two persons belong to (tables 2 and 3). When a pair of individuals shares several vrGVs, these shared vrGVs are usually grouped in one locus or a few loci (table 1 italicized entries). Five or more shared adjacent vrGVs are called RVCs. In order to characterize shared RVCs we created a Perl program *RVC.pl*, which scans an individual-specific vrGV database and identifies all shared RVCs inside it. For all 1,092 processed genomes, an RVC contains 12.6 vrGVs per cluster on average. The distribution of clusters along chromosomes is seemingly random, so no obvious patterns in their genomic locations have been observed. A segment of the output file from the *RVC.pl*, representing a complete list of RVCs the individual under analysis shares with the other 1,092 sequenced individuals, is demonstrated in table 4. Such output files were obtained for each of the 1,092 individuals and they provide the information on the number of RVCs an individual shared with other people and also the length of sharing clusters. These datasets are available in the supplementary file "Data", Supplementary Material online (folder OutputDataWindow20). By computational processing of these datasets, we created the complete table of shared RVCs for each of the possible 1,092 × 1,092 pairs (tables S1 and S2 in supplementary file "Data", Supplementary Material online, folder OutputDataWindow20). The heat-map representation of it is shown in figure 1.

Table 2. Intra-Population Sharing of RVCs

Population	Number of shared RVC/pair	Length RVC/pair (Mb)
CEU	1.56	1.73
FIN	6.67	2.77
GBR	2.21	2.74
IBS	5.04	1.69
TSI	2.64	2.23
CHB	2.41	1.26
CHS	3.75	2.14
JPT	10.9	1.53
ASW	11.7	0.88
LWK	25.3	0.89
YRI	19.1	0.74
CLM	6.45	3.97
MXL	2.85	3.87
PUR	8.17	4.44

Analysis of genetic relations based on the number and length of shared RVCs

Sharing of RVC Within the Same Population

The highest number of shared RVCs between two individuals is observed, unsurprisingly, when the two persons belong to the same population (fig. 1). The average numbers and lengths of shared RVC within 14 studied populations are

Table 3. Inter-Population Sharing of RVC for 1,092 Individuals from 14 Populations

Pop 1	Pop 2	Number of VRC/pair	Median Length	Average Length	Pop 1	Pop 2	Number of VRC/pair	Median Length	Average Length
CEU	ASW	0.293	686.5	1161	MXL	CEU	0.446	556.5	971
CHB	ASW	0.016	241.5	610	MXL	CHB	0.026	287	673
CHB	CEU	0.015	405.5	760	MXL	CHS	0.024	307	527
CHS	ASW	0.021	308	630	MXL	CLM	0.927	696	1050
CHS	CEU	0.005	244	458	MXL	FIN	0.210	528.5	897
CHS	CHB	2.440	830	1164	MXL	GBR	0.440	616	1000
CLM	ASW	1.224	313	485	MXL	IBS	1.224	841	1257
CLM	CEU	0.550	577.5	993	MXL	JPT	0.017	301	501
CLM	CHB	0.016	397	715	MXL	LWK	0.544	212	439
CLM	CHS	0.009	323	524	PUR	ASW	1.676	324	520
FIN	ASW	0.140	647	1038	PUR	CEU	0.605	578	985
FIN	CEU	0.770	760	1306	PUR	CHB	0.014	265	528
FIN	CHB	0.033	365	676	PUR	CHS	0.010	151	469
FIN	CHS	0.014	317	796	PUR	CLM	1.163	669	1087
FIN	CLM	0.277	569.5	949	PUR	FIN	0.294	573	956
GBR	ASW	0.293	717.5	1210	PUR	GBR	0.634	617.5	1035
GBR	CEU	1.404	897	1435	PUR	IBS	1.421	810	1237
GBR	CHB	0.013	427	764	PUR	JPT	0.006	147.5	256
GBR	CHS	0.004	322	793	PUR	LWK	1.340	258.5	464
GBR	CLM	0.576	581	937	PUR	MXL	0.858	650	1086
GBR	FIN	0.730	725	1249	TSI	ASW	0.200	416.5	792
IBS	ASW	0.305	416.5	699	TSI	CEU	0.846	604	1059
IBS	CEU	0.858	647	1041	TSI	CHB	0.020	352.5	659
IBS	CHB	0.005	207	306	TSI	CHS	0.010	184.5	554
IBS	CHS	0.006	116	202	TSI	CLM	0.654	497.5	891
IBS	CLM	1.731	1018	1416	TSI	FIN	0.410	554	1029
IBS	FIN	0.454	573.5	955	TSI	GBR	0.816	593	1072
IBS	GBR	1.031	671	1098	TSI	IBS	0.847	551.5	947
JPT	ASW	0.008	190	465	TSI	JPT	0.005	122	200
JPT	CEU	0.005	232	402	TSI	LWK	0.132	218	392
JPT	CHB	1.355	641	1042	TSI	MXL	0.521	467	857
JPT	CHS	1.088	542	911	TSI	PUR	0.758	513	943
JPT	CLM	0.004	223	505	YRI	ASW	12.71	498	556
JPT	FIN	0.021	350.5	714	YRI	CEU	0.026	158	354
JPT	GBR	0.002	150	440	YRI	CHB	0.002	82	114
JPT	IBS	0.005	117	218	YRI	CHS	0.003	42	82
LWK	ASW	8.750	360	421	YRI	CLM	1.150	285	466
LWK	CEU	0.041	216	389	YRI	FIN	0.003	194.5	395
LWK	CHB	0.006	93	230	YRI	GBR	0.016	144	258
LWK	CHS	0.008	42.5	66	YRI	IBS	0.115	153	273
LWK	CLM	0.995	253	432	YRI	JPT	0.001	30	44
LWK	FIN	0.009	167.5	294	YRI	LWK	8.182	364	442
LWK	GBR	0.031	189	366	YRI	MXL	0.597	236.5	463
LWK	IBS	0.171	190	343	YRI	PUR	1.576	282	470
LWK	JPT	0.005	33	62	YRI	TSI	0.049	170.5	348
MXL	ASW	0.712	283	517					

Median and Average sizes of IBD segments are shown in kb.

shown in table 2. African and African-American individuals have the highest number of shared RVCs per pair within their populations followed by the Japanese and the Puerto-Ricans. In European groups, the highest cluster sharing is observed among the Finns (on average, 6.7 shared RVCs per pair) while the lowest—1.6 RVCs—is found in the Utah white population (CEU). Among Asian people, the average number of shared RVCs also broadly varies from 10.9 for Japanese (JPT) to 2.4 for Chinese (CHB) (table 2). These results are congruent to Frazer et al. (2007) that an average pair of individuals from the same population shares ~0.5% of their genomes through recent IBD.

Table 4. Example of a segment of the output file CEU_NA12763_dat4

Individual host	Individuals with shared RVCs	Number of shared vrGV clusters	Total length of clusters (bp)	Total number of vrGVs
CEU_NA12763	CEU_NA12286	4	3,585,741	39
CEU_NA12763	MXL_NA19779	1	304,700	5
CEU_NA12763	ASW_NA20317	1	374,390	11
CEU_NA12763	FIN_HG00382	1	497,473	10
CEU_NA12763	GBR_HG00141	2	1,242,733	15
CEU_NA12763	FIN_HG00280	4	5,300,628	28
CEU_NA12763	CEU_NA12827	3	1,612,501	16
CEU_NA12763	FIN_HG00361	1	744,208	6
CEU_NA12763	CLM_HG01271	1	511,489	9
CEU_NA12763	FIN_HG00173	1	1,376,231	8
CEU_NA12763	GBR_HG00106	1	1,216,725	5
CEU_NA12763	CLM_HG01134	3	991,604	25
CEU_NA12763	FIN_HG00266	1	771,106	9
CEU_NA12763	FIN_HG00344	1	676,386	8

Sharing of RVC by Individuals from Neighboring Populations

Continental inter-population RVC sharing is correlated well with the geographic distances between the populations (table 3). In Europe, the lowest number of shared RVC is observed between Finnish people (FIN) and South European populations TSI (Toscani in Italia) and IBS (Iberian Population in Spain), which are geographically distant and historically have not intensively intersected with each other (on average, TSI–FIN and IBS–FIN pairs have ~0.4 shared RVCs per pair). RVC sharing between all other groups with European origin is higher. Particularly, the number of shared RVCs per pair of individuals belonging to any two European (non-FIN) populations ranges from 0.82 to 1.03. The highest number of RVC sharing (1.03) is observed between pairs of people from Utah (CEU) and Britain (GBR) inhabiting different geographic regions but originated from the same founder populations.

In Asia, Chinese Han people from South and Beijing (CHS and CHB) share on average 2.4 RVCs per pair between themselves, and approximately twice less with geographically remote Japanese people (1.1–1.4 RVCs). African YRI (Yoruba in Ibadan, Nigeria) and LWK (Luhya in Webuye, Kenya) groups share on average 7.8 RVCs per pair. However, such high numbers of shared RVCs between African populations may be partially due to the fact that they have ~3 and 4 times more vrGVs than Asian and European populations, respectively.

Sharing of RVC by Individuals from Different Continents

The lowest numbers of shared RVCs are detected for pairs of individuals inhabiting distant parts of the Old World. A majority of these inter-continental pairs of individuals do not share RVCs at all. Thus, these areas in figure 1 are predominantly white (see table 3 for details). The lowest RVC sharing is found in Asia–Africa pairs (0.0042 shared RVC per pair) and specifically for YRI–JPT populations (where only nine pairs among all 7,832 possible pairs have shared RVCs with median RVC length of merely 44 kb). The second lowest inter-continent RVC sharing is observed between people from Asia and Europe. Only 0.2–3.3% of these inter-continental pairs have shared RVCs. Interestingly, among these Asian–European pairs, the highest admixture is observed between both Chinese groups and two Europeans—FIN and TSI (0.01–0.03 shared RVC per pair). Japanese people share <0.005 RVC per pair with all Europeans except with Finns (0.02). Such enrichment of Asian RVC among the Finns presumably is an effect of belonging of the Finns, unlike other Western Europeans, to the Finno–Ugric population of the Uralic family of the Northern Eurasians (Lahermo et al. 1996; Lappalainen et al. 2006; Rootsi et al. 2007).

The highest Old World intercontinental admixture of RVC is depicted between Africa and Europe. Our data are in accordance with previously reported increasing gradient of admixture of African genes from Northern Europe to Southern (Adams et al. 2008; Moorjani et al. 2011; Cerezo et al. 2012; Botigue et al. 2013). We also found that all European groups share more RVCs with LWK than with YRI (see table 3 for details), thus supporting the hypothesis of gene exchange between Europe and Africa through Near Eastern migration routes rather than Trans-Saharan (Cavalli-Sforza et al. 1994; Richards et al. 2000; Currat and Excoffier 2005). The Italian (TSI) population has 2.7 times more shared RVCs with Kenyan (LWK) than with Nigerian (YRI) populations (0.132 vs. 0.049 RVCs per pair, respectively). Northern Europeans (CEU, FIN, and GBR) also share more RVCs with LWK than with YRI (table 3). This prevalence of LWK over YRI in shared IBD chromosomal segments in Southern European populations is statistically significant according to the Chi-squared test with P-value $<10^{-15}$.

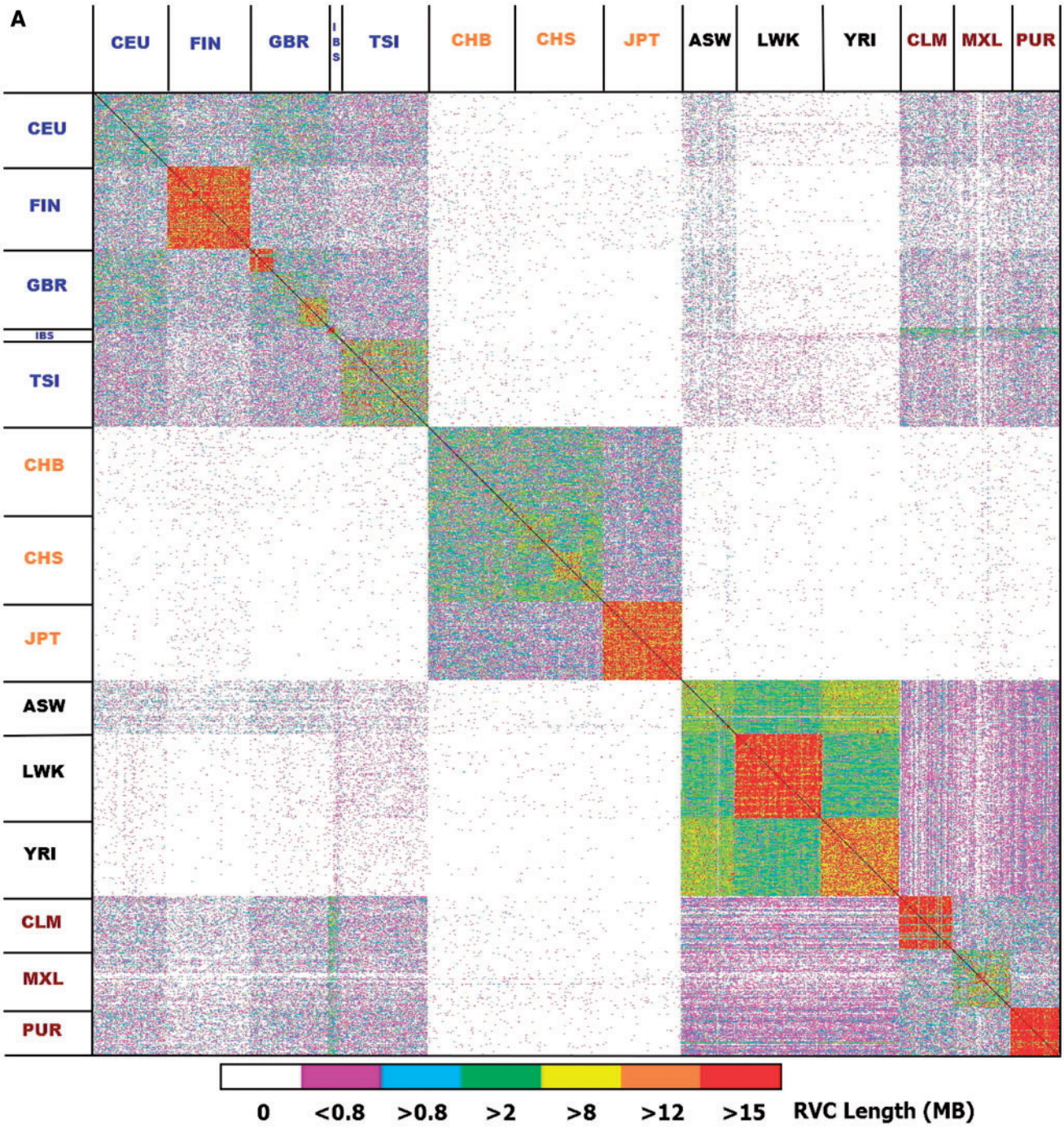

Fig. 1.—Heat-map table (1,092 × 1,092) presenting the total length (A) or number (B) of shared vrGV clusters for every possible pair of 1,092 individuals. Populations are grouped by the continent they originated from and labeled by different colors according to the Olympic scheme. Five populations with European origin are labeled in blue (CEU, FIN, GBR, IBS, and TSI); three Asian populations—in yellow (CHB, CHS, and JPT); three African populations—in black (ASW, LWK, and YRI); and three American populations—in red (CLM, MXL, and PUR). If a pair of individuals does not share an IBD segment, the corresponding square is present in white. The squares corresponding to pairs that share IBD segment(s) are colored according to a rainbow scheme. The smallest segments, for which total length shared by a pair does not exceed 900 kb are shown in violet, while the largest segments with total length per pair exceeding 10 Mb are shown in red.

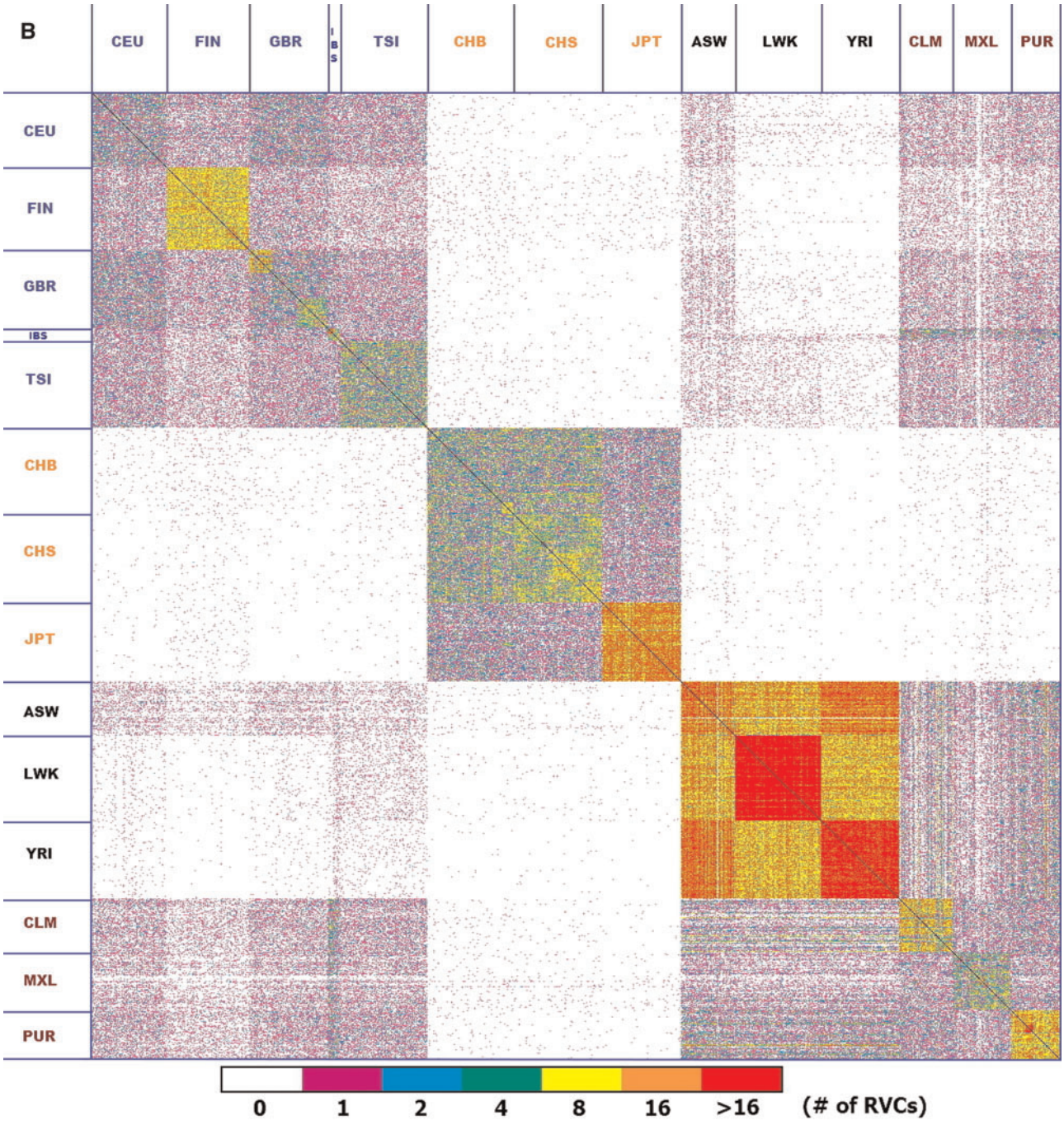

Fig. 1.—Continued.

RVC sharing among people inhabiting New World perfectly reflects recent global human demographic events and migration routes. All three Caribbean populations (CLM, MXL, and PUR—Colombians, Mexican from Los Angeles, and Puerto Ricans, respectively) share considerable amount of RVCs with African and European populations (on average ~0.9 and 0.5 shared RVCs per pair and 254 and 620 kb of average median RVC size, respectively). However, these numbers considerably vary from population-to-population (e.g., compare MXL–GBR vs. PUR–GBR, or PUR–YRI vs. MXL–YRI in table 3). American South-West Black population (ASW) represents another good example of recent admixture. Figure 1 demonstrates extensive presence of RVCs from European and American populations in ASW genomes. However, the admixture of African and American populations is still nonhomogeneous and there are multiple strips in the corresponding

segments of the heat-map in figure 1. Interestingly, one person from ASW (NA20314) shares 10 times less RVCs with LWK and YRI than any other ASW representative (for details compare ASW_NA20314_dat3 and ASW_NA20314_dat4 files with other "dat3" and "dat4" files from the ASW population available from the supplementary file "Data", Supplementary Material online (folder OutputDataWindow20). This person also shares the minimal number of RVCs among all possible pairs within ASW population. NA20314 is presented by a tiny white line across African populations in figure 1. Possibly, some errors might have occurred in population identification of this individual.

RVC sharing between Caribbean and European populations revealed at least twice fewer admixtures of the Caribbean with the British, Italian, and Finns than with the Spanish thus reflecting a rich Spanish colonial exploration of the region. Due to well-known social restrictions, the African–American genomes share only 0.3 RVC per pair on average with the Spanish, the British, and Utah whites. Genomes of all Caribbean groups are enriched with African RVC (1.0–1.5 RVC per pair in CLM and PUR and 0.5 in MXL).

All three Caribbean populations share considerable amount of RVCs with Africans and the impact of YRI and LWK is even. These data are consistent with the database of the slave-trading voyages (www.slavevoyages.org) and also the Atlas of the Transatlantic Slave Trade (Eltis and Richardson 2010). According to this book, three million people were taken from the Bight of Benin—a native land of YRI population. In addition, about five million people were taken from West Central Africa, which people belong to Bantu linguistic/ethic group, the same as LWK population (Gomez et al. 2014). However, a considerable portion of slaves from West Central Africa were brought to Brazil. Finally, African–Americans from the South–West (ASW) share more RVCs with Western African (YRI) than with Eastern African (LWK). Impact of gene flow from Asia to the American continent is the least profound (0.04–0.26 shared RVCs per Asian–American pair).

Modeling the number and size of IBD segment inheritance in generations

Computer simulations in population genetics have several advantages over mathematical modeling (Qiu et al. 2014). We created a program *IBDsimulator.pl* that models an inheritance of IBD autosomal segments from an initial person (primogenitor) along a chain of his/her descendants in multiple successive generations. The program uses real distribution of meiotic recombination sites along the human genome from HapMap table of genetic versus physical distances in human chromosomes (Frazer et al. 2007). In order to obtain reliable statistics, this program has been repeated independently 500,000 times. The distribution of average numbers and sizes of inherited computer-simulated IBD segments from a primogenitor in successive generations are shown in figure 2*A* and *B*, respectively, while the data from the program are available in the supplementary file "Data", supplementary material online, folder Modeling.

An offspring (generation G_1) of a primogenitor inherits 22 IBD segments (22 autosomes) from the parent (fig. 2*A*). In the next generation (G_2), the average number of IBD segments inherited from this primogenitor reaches its maximum value (28.5 IBD segments per grand-child). In the following generations, the average number of IBD chromosomal segments inherited from the primogenitor drops monotonously. The tenth generation (the G_{10} progeny of the primogenitor) retains, most often, one or zero IBD segment (on average, 0.37 IBD segments per G_{10}-descendant). At the 20th generation only 7 out of 10,000 G_{20}-descendants inherit an IBD segment from their particular primogenitor, while the rest 99.93% of G_{20}-descendants do not possess any genetic material from this particular G_{20}-primogenitor. The length of IBD segments shortens dramatically during the first few generations. Then, the diminution of the IBD segments length starts slowing down (fig. 2*B*). However, the distribution of the sizes of IBD segments in these generations is very wide (fig. 3). This phenomenon is due to the very uneven distribution of meiotic recombination rates along human chromosomes (Arnheim et al. 2003). In many occasions, the length of an IBD segment in the G_{20}-descendant might be longer than an IBD segment

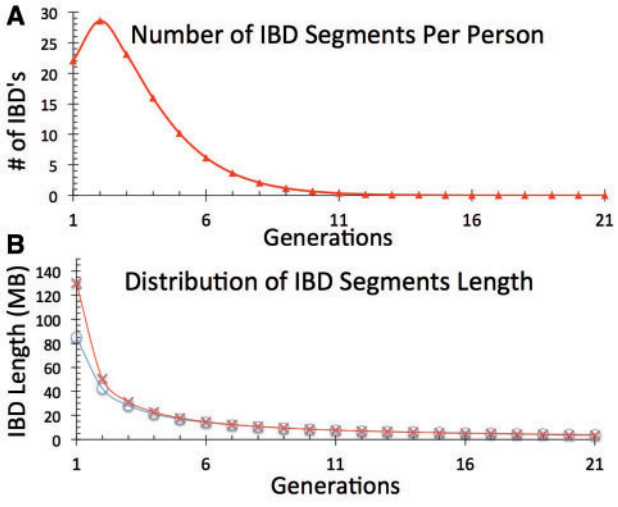

Fig. 2.—Distribution of length and number of IBD autosomal segments inherited from a model primogenitor in consecutive generations calculated by a computer simulation program *IBDsimulator.pl*. (*A*) Average number of primogenitor's IBD segments per descendant. First generation contains one copy of 22 primogenitor's autosomes (22 IBD segments). (*B*) Average size of primogenitor's IBD segments per descendant obtained by *IBDsimulator.pl* (red curve). Blue curve (open circles) shows the average size of primogenitor's IBD segments calculated from equation (1) with the following parameters: recombination rate value is $r = 0.0118\,\mathrm{Mb}^{-1}$.

in the G_5-descendant (fig. 3). Therefore, a particular length of an IBD segment does not allow accurately determining the generation of the inherited person. For this reason, many papers use genetic distances (measured in centimorgans, cM) rather than physical IBD length in nucleotides (e.g., Browning and Browning 2013). In this article, we use physical distances of IBD segments because measurement of genetic distances of human chromosomes is still not very accurate and based on the HapMap tables (Frazer et al. 2007), which are not continues and have many gaps.

The diminution of average IBD segment size in generations is described by the formula (Browning and Browning 2012):

$$L = 1/(r \times g), \quad (1)$$

where g is the consecutive generation number (or equivalently number of meioses) and r is the recombination rate

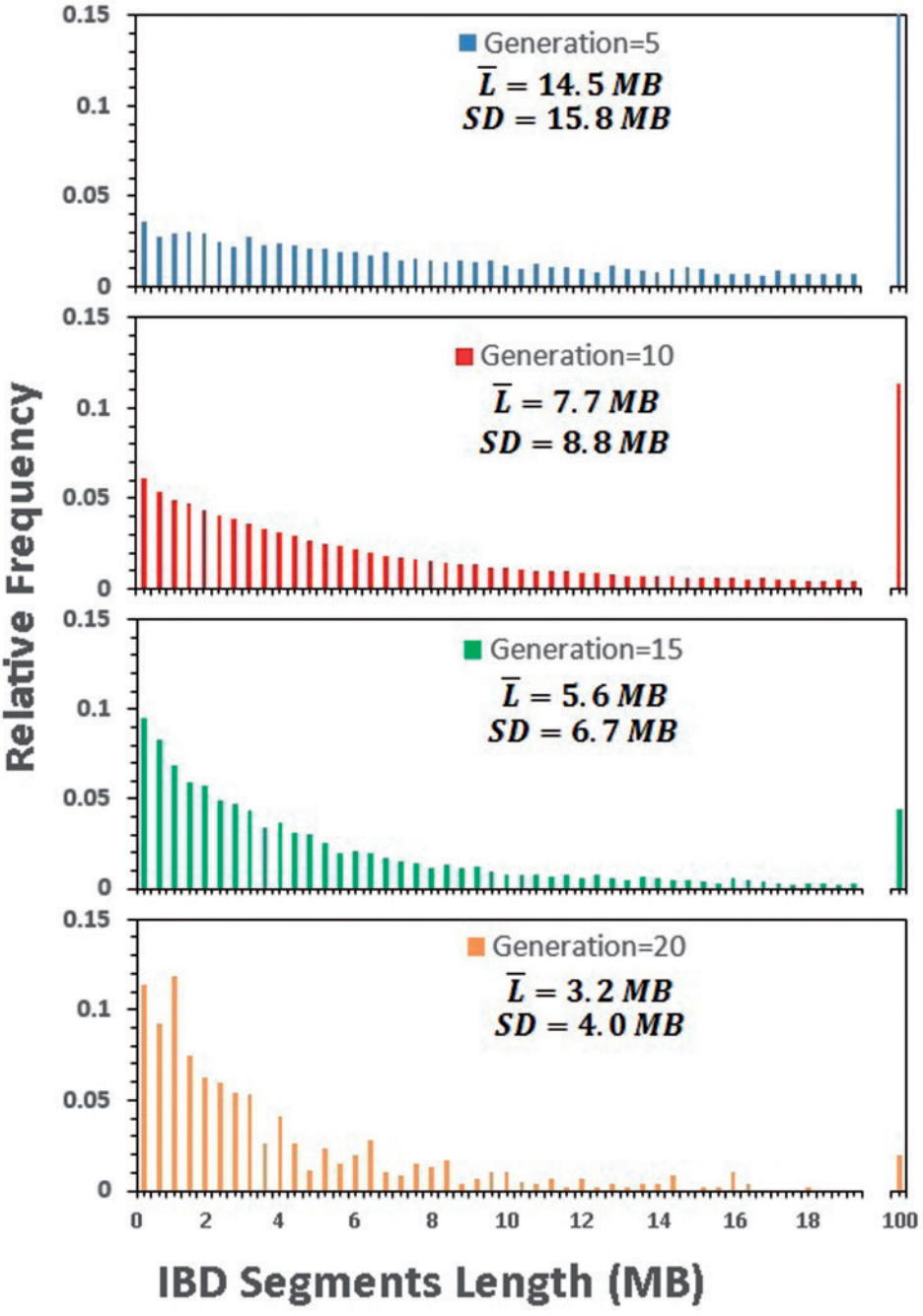

Fig. 3.—Distribution of model primogenitor's IBD segments by their lengths at 5th, 10th, 15th, and 20th generations. Number of IBD segments within particular ranges of lengths was calculated for 0.4 Mb bins. The last bin represents the number of segments with length > 10 Mb.

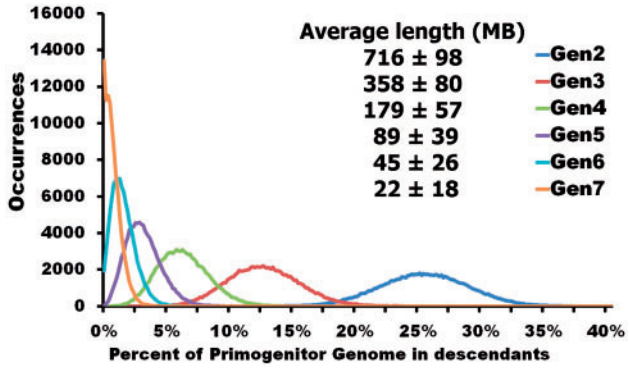

Fig. 4.—Distribution of proportion of model Primogenitor's genetic materials inherited by descendants in six successive generations. The figure presents 100,000 computational simulation experiments performed with the *IBDsimulator.pl* program. The *x*-axis presents the percentage of the total autosomal length of Primogenitor in the descendants. The *y*-axis shows the number of occurrences of different proportions of Primogenitors' genetic material in different generations out of 100,000 experiments.

($r = 0.0118\,\mathrm{Mb}^{-1}$, one event per 85 Mb). In order to get the most accurate estimation about the time of last common ancestor (in generations) based on the length of shared IBD segment, one should use the local recombination rate (r) inside equation (1).

The computer-simulated curve for IBD segments length on figure 2*B* (red line) is almost the same as the theoretical one based on equation (1) (blue line). Equation (1) allows us to estimate the time of population admixture/separation below.

Along a genealogical lineage, first degree relatives (e.g., parent-offspring) share on average 50% of genetic material, second degree relatives (grandparent–grandchild) 25%, third degree relatives (great grandparent–great grandchild) 12.5%, and so on according to the formula $100\% \times 2^{-n}$, where n is a degree of relationship in generations. However, due to a limited number of meiotic recombination events per gamete (on average, 36), which are distributed very unevenly along the genome, the inheritance of the primogenitor's chromosomal material in generations may be very uneven (Consortium 2003). Our computation modeling with real distributions of human meiotic recombination sites based on HapMap dataset (Frazer et al. 2007) generated statistics of such unevenness of autosomal material inheritance (fig. 4). For example, a grandchild does not always get exactly 25% of genetic material from a grandparent. This amount frequently varies between 20% and 33% interval. An explanation to figure 4 is provided in the "Discussion" section and figure 5.

Discussion

Genealogical and Genetic Relatedness

From a genealogical viewpoint, every human being has two biological parents, four grandparents, and so on in

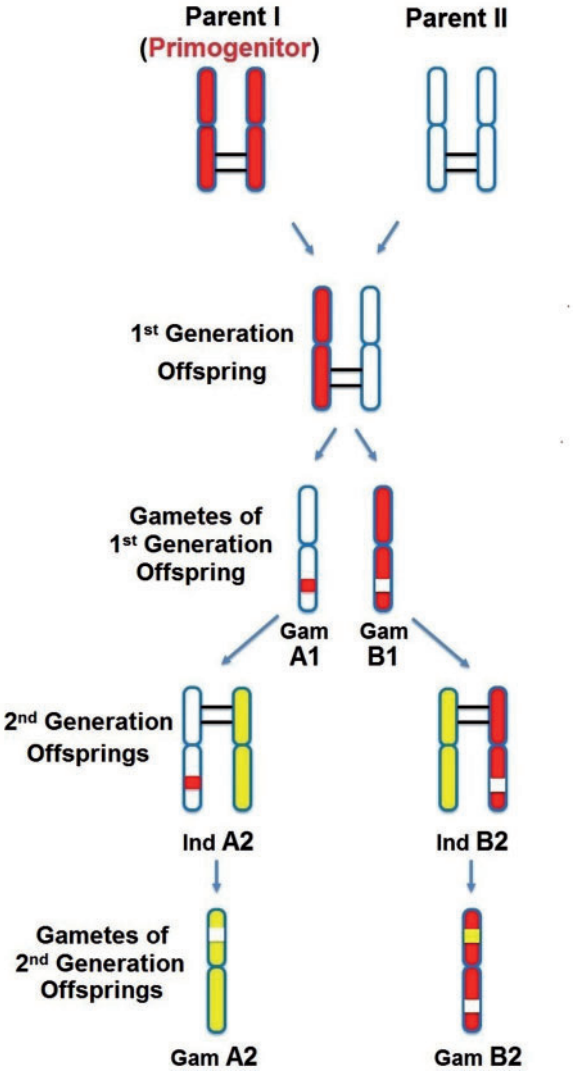

Fig. 5.—Randomness and unevenness of the inheritance pattern of the primogenitor's chromosomal material in generations. The chromosomal material of primogenitor (parent I) is shown in red. The chromosomal material of the mating partner of primogenitor (parent II) is shown in white. First generation offspring inherits one copy of primogenitor's chromosome. Since meiotic recombination events (black horizontal lines) are few and random, different gametes from the progeny could obtain different amount of primogenitor's IBD segments. Gamete "gam A1", which creates an second generation offspring "Ind A2" has only ~10% of primogenitor's (red) chromosome and 90% of the second parent's (white) chromosome, while another gamete B1 transfers ~90% of primogenitor's (red) chromosome in three IBD segments and only 10% of the other parent's (white) chromosome to the other second generation offspring "Ind B2". The yellow chromosome in both individuals A2 and B2 is contributed by their second parent (i.e., the mating partner of the first generation offspring). Even the third generation could easily lose all primogenitor's chromosomal material (red) via gamete A2, or inherit a majority of primogenitor's chromosomal material (red) in several IBD segments via gamete B2.

geometrical progression. At the *n*th generation a person has 2^n direct genealogical ancestors. When $n = 20$ the number of ancestors becomes 1,048,576 while when $n = 40$, it becomes 1,099,511,627,776. Therefore, at generation ~20–30 a majority of people from the same geographical region are distant genealogical relatives to each other. Rohde et al. (2004) examined human genealogical relations and estimated that the last common genealogical ancestor for all modern humans (presumably from the same continent) lived ~76 generations ago (~2,000 years ago). How genetic material is transmitted through generations along a genealogical tree determines the genetic relatedness. The transmission occurs via gametes, which are created, from pieces of maternal and paternal chromosomes via meiotic recombination. On average, 22 human autosomes have 34.5 recombination events per gamete and these recombination sites are distributed very unevenly along chromosomes. The inheritance of genetic material is random and may be uneven (fig. 5). Due to immense variations in recombination rates along the genome, the spread of IBD segment sizes is very wide (fig. 3). During transmission of IBD segments from generation-to-generation they become smaller and smaller (fig. 2*B*). After the tenth generation, a majority of direct genealogical descendants have lost all genetic material from their particular G_{10}-primogenitor. However, since human populations have limited sizes, individuals often share multiple short IBD segments from their common distant ancestors. The patterns (numbers and lengths) of shared IBD segments across human populations significantly vary from population-to-population depending on their size, mating traditions, migration, admixture, and other parameters. Knowledge of inheritance of IBD genomic segments is important for medicine, specifically in case–control association and cohort studies, since unknown distant familial relationships could potentially compromise interpretation of collected data.

IBD Segments Identification with Modern Approaches

In Al-Khudhair et al. (2015), our team has discussed various methods used to detect IBD familial relationships with up to tenth degree of relatedness. In a nutshell, even for close relatives, modern algorithms have very high level of errors in IBD segments identification (Huff et al. 2011; Durand et al. 2014; Li et al. 2014). Recent papers extrapolated statistical analyses of SNP distributions to much older events for intercontinental population admixture, and even for the relationship between modern humans and other, now extinct, archaic hominid groups (Reich et al. 2010; Meyer et al. 2012; Castellano et al. 2014; Lazaridis et al. 2014). These sophisticated statistical methods have been recently reviewed by Racimo et al. (2015). They include Patterson's D statistic (Green et al. 2010; Durand et al. 2011; Patterson et al. 2012); analysis of incomplete lineage sorting from introgressed haplotypes seen by increased long-range linkage disequilibrium (LD) using *S** statistic (Wall et al. 2009; Vernot and Akey 2014); probabilistic hidden Markov model (Prufer et al. 2014; Seguin-Orlando et al. 2014) and conditional random field model (Sankararaman et al. 2014). Yet, the accuracy and the reliability of these methods cannot be directly verified. Importantly, anthropologists have presented reasonable doubts of modern statistical methods for evaluating population admixtures and evolution, showing that statistical conclusions "go so much against the well-known evolutionary realities. . ." (Weiss and Lambert 2014). A major problem in statistical approaches for revealing genetic relatedness exists in pipeline of approximations that may amplify errors in a vicious cycle. For example, to calculate LD, the phasing of genomic sequences is required, which is prone to errors. Calculated LD values, with already-embedded errors, are frequently used for nucleotide imputations of *de novo* sequenced genomes, as well as for their phasing. This cycling may result in progressive multiplication of the initial errors. In addition, nucleotide sequence imputations often do not consider many important biological processes (e.g., biased gene conversion) that are often involved in haplotype changes and alteration of LD values.

A direct comparison between our approach and the computer predictions of shared IBD segments by two popular algorithms (GERMLINE and PLINK) is possible using table 5 from Gusev et al. (2009). This table 5 of Gusev et al. presents the IBD data for pairs from 45 unrelated individuals from Japan (JPT) and also Chinese people from Beijing (CHB). On the other hand, our table 2 contains the distribution of IBD segments predicted by RVCs within 89 Japanese (JPT) and 97 Han people from Beijing (CHB). For Japanese population, the detected mean IBD segment length is 1.53 Mb for our RVC algorithm, is 1.8 Mb for GERMLINE, and 4.8 for PLINK. For CHB population the mean IBD segment length is 1.26 Mb (RVC), 2.1 Mb (GERMLINE), and 4.8 Mb (PLINK). Thus, we detected on average shorter IBD segments. Table 2 provides the mean number of IBD segments per pair (2.41 for CHB and 10.9 for JPT) and the mean length of shared IBD segment (1.26 Mb for CHB and 1.53 Mb for JPT). According to these records, the expected total number of shared IBD segments for a group of 45 people would be 4,772 (CHB) and 21,582 (JPT), while the total length of all shared IBD segments for 45 people would be 6,012 Mb (CHB) and 33,020 Mb (JPT). These numbers are many times higher than the corresponding numbers for JPT and CHB populations in table 5 of Gusev et al. (2009). Hence, our RVC approach allows to detect several times more shared IBD segments than GERMLINE and PLINK. In addition, GERMLINE and PLINK are used for predicting relatively long shared IBD segments (>1 Mb) that originated over the last 10 generations (e.g., on the order of second to ninth cousins) (Gusev et al. 2009; Henn et al. 2012; Zhuang et al. 2012). Hence, the advantage of our RVC approach is in the ability to detect short IBD segments (down to 30 Kb) that share common ancestors down to 378 generations ago (or

~9,500 years) (see the section "Estimations of Time for Common Ancestors from Shared RVC" below). In the supplementary file "Data", Supplementary Material online (folder OutputDataWindow20) we provide exhaustive details about the distribution of RVCs for every possible pair of 1,092 sequenced genomes, so our results can be compared with any competitive programs.

There is a clear difference between STRUCTURE software (Falush et al. 2003) and our RVC algorithm. The main use of STRUCTURE is for assigning individuals to populations, inferring the presence of distinct populations, identifying migrants and admixed individuals. For these purposes, STRUCTURE is heavily based on much more frequent SNPs or other genetic markers. In contrast, our approach is aimed at revealing most distant cryptic genetic relatedness among pairs of individuals.

Population Analysis Using RVCs

Contrary to the probabilistic approaches, our method is rather deterministic because we consider a group of very rare events which, practically speaking, cannot happen together only by chance. Indeed, our threshold probability for sharing of clusters of vrGVs between individuals is 0.5×10^{-9} for the default search parameters (five shared vrGVs in a consecutive window of 20 individual-specific vrGVs, see supplementary file "M&M", Supplementary Material online). Our approach allowed detection of genetic relatedness among people from remote geographic regions. It is in good agreement with the known human population history. Moreover, it allowed clarifying some debated issues. For example, our data (fig. 1) clearly demonstrate that the Finns, which migrated from Northern Eurasia several thousand years ago, deeply admixtured with the European populations and now share the majority of their RVCs with the Europeans. According to the analysis of the mtDNA haplogroups and several autosomal markers, the Finns are undistinguishable from other Europeans (Lahermo et al. 1996). On the contrary, the Y-chromosome investigations show high prevalence (>50%) of North Eurasian-specific N3 haplogroups among the Finns, which also present in China and Japan (Lappalainen et al. 2006; Rootsi et al. 2007). Thus, elevated number of shared short sized RVCs between Finns and both the Chinese and Japanese, compared with other Europeans, supports the Y chromosome data of ancient origin of Finns from Asia. Another example is the distribution of the African RVCs among Europeans. Higher levels of African admixture in Southern (especially South-Western) European compare with Northern have been identified by analysis of Y-chromosome and mtDNA haplogroups as well as by autosomal SNP distribution and IBD sharing (Adams et al. 2008;Moorjani et al. 2011; Cerezo et al. 2012; Botigue et al. 2013). African (sub-Saharan) ancestry was estimated to be around slightly <3% in Iberia and ~1% in Northern Italy (Moorjani et al. 2011) or <1% for Iberia and TSI (Botigue, et al. 2013). However the authors did not find the difference in IBD segments sharing between YRI and LWK and European populations. Hence, the source and the routes of the delivery of African genomes to the Europeans have been debated. Our data demonstrate significantly higher number of the Kenyan (LWK) RVCs than the Western African (YRI) in all European populations, thus, supporting the hypothesis of the Near Eastern rather than the trans-Saharan route of gene exchange between the Africans and the Europeans.

Estimations of Time for Common Ancestors from Shared RVC

There are two obstacles for the estimation of time for the last common ancestors for people sharing RVCs. First, we detect only the intersections of IBD segments between pairs of individuals. Since the intersections occur randomly, the whole IBD segments may be considerably larger than their intersections. According to figure 3, there is a great variation in the IBD segment sizes that may vary dozens of times for the same generation. The size of intersection of a large and a short fragment never exceeds the shortest one. Second, RVC approach characterizes not the whole IBD segments but only the inner part of them bordered by the two extreme rightmost and leftmost vrGV positions. Even though a human being on average bears ~30,000 vrGVs, still there are some areas with no rare variant differences between individuals. Therefore, it is necessary to make an adjustment (calibration) for the estimation of time for the last common ancestors for the pairs with shared RVCs. It can be achieved using well-known admixture of Spanish–Americans populations starting in 1492 (about 21 generations ago), for which median size of shared intersected RVCs detected by our approach is 890 kb (table 3). However, according to figure 2*B*, the median size of entire IBD segments after 21st generation should be 3.55 Mb, 4 times larger than the detected value. This shows that although the expected IBD segments after 21st generation is around 3.55 Mb in two individuals, the average length (L) of the RVCs intersection is only 890 kb. Based on this Spanish–American data, we made calibration of RVC length $L_{cal} = (3.55/0.89) \times L$, which can be placed in equation (1) to calculate the time of common ancestors between populations as the following: $g = 1/(L_{cal} * r)$.

Thus, we can calculate the time when the common ancestors to the European–African pairs (median size of shared RVCs is 180 kb) lived using the following:$g_{\text{Af–Eu}} = 1/(L_{cal} \times r)$, where $L_{cal} = 4 \times 180$ kb and $r = 0.0118\ \text{Mb}^{-1}$. It gives us $g_{\text{Af–Eu}} = 118$ generations ago or ~2,950 years ago. This valuation is congruent to previous estimations (Moorjani et al. 2011). Using the same approach, we estimated that the last common ancestors for the shortest shared RVCs that are observed for Asian–African pairs (median 54 kb) probably lived ~378 generations ago or ~9,500 years ago.

Future Directions

For a broad public usage, a precise definition and cataloging of vrGVs are required. Creation of a public database of human vrGVs is in our nearest plans. With a massive flood of genome sequencing in the next few years, hundreds of millions of novel vrGVs will be available. Hence, the size of the vrGV database should be enormous. (Theoretically, seven billion people on the planet may have up to ten billion SNPs.) Therefore, it would be sensible to generate a database of frequent genetic variants, which are NOT-vrGVs. Any genetic variant that is absent in the NOT-vrGVs database should be considered as a very rare one. According to our preliminary data, the size of the NOT-vrGV is only 22 million genetic variants based on the phase 1 dataset of "1000 Genomes Project". This number should not increase much with further sequenced genomes because adding new people will not generate novel frequent genetic variants. While considering vrGVs across multiple populations, a variant may have a total frequency of $<0.2\%$, yet local frequency of the same variant in a particular population might be considerably high (e.g., 5%). We would rather exclude counting such variants as vrGVs if their frequency in a particular population is above a certain threshold (e.g., 1%). Human populations vary significantly in the number of vrGVs per person. However, these variations should not noticeably influence the detection of cryptic relatedness since rare variants are spread over a vast genomic regions and the probability of sharing of five or more vrGVs within a particular locus depends only on their frequencies and the window size for registration of RVCs according to the equation (2) from the "Materials and Methods" section (which is merely 0.5×10^{-9} for our default parameters).

Due to the simplicity and computational speed, our method may be used for large cohort and GWAS studies where thousands of sequenced genomes will be available. Proper identification of genetic relationships is essential for forensic identification, in criminal investigations, inheritance claims, and in other areas of human life.

Conclusion

Inheritance of genetic materials creates an intricate fractal mosaic of IBD chromosomal segments in the genome. Close familial relationships are presented by shared long IBD segments that in turn are mosaics of shorter IBD segments from previous generations. Further, each IBD segment is built from smaller pieces inherited from distant ancestors. Identification of shared vrGV clusters presents a powerful tool for characterization of long and short IBD segments and for evaluation of population stratification. Proper recognition of genetic relationships is essential for individualized medicine, forensic identification, criminal investigations, inheritance claims, and in other areas of human life.

Acknowledgments

We are grateful to Dr. Robert Blumenthal, University of Toledo Health Science Campus, for his insightful discussion of the project. We also appreciate the financial support from the Department of Medicine to conduct our research.

Disclosure

All our programs and datasets are available to public without any restrictions.

Supplementary Material

Supplementary_File_DATA and Supplementary_File_M&M are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Literature Cited

Abecasis GR, et al. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.
Adams SM, et al. 2008. The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am J Hum Genet. 83:725–736.
Al-Khudhair A, et al. 2015. Inference of distant genetic relations in humans using "1000 genomes". Genome Biol Evol. 7:481–492.
Arnheim N, Calabrese P, Nordborg M. 2003. Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. Am J Hum Genet. 73:5–16.
Botigue LR, et al. 2013. Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc Natl Acad Sci U S A. 110:11791–11796.
Browning BL, Browning SR. 2013. Detecting identity by descent and estimating genotype error rates in sequence data. Am J Hum Genet. 93:840–851.
Browning SR, Browning BL. 2011. Population structure can inflate SNP-based heritability estimates. Am J Hum Genet. 89:191–193. author reply 193–195.
Browning SR, Browning BL. 2012. Identity by descent between distant relatives: detection and applications. Annu Rev Genet. 46:617–633.
Carmi S, et al. 2013. The variance of identity-by-descent sharing in the Wright-Fisher model. Genetics 193:911–928.
Carmi S, Wilton PR, Wakeley J, Pe'er I. 2014. A renewal theory approach to IBD sharing. Theor Popul Biol. 97:35–48.
Castellano S, et al. 2014. Patterns of coding variation in the complete exomes of three Neandertals. Proc Natl Acad Sci U S A. 111:6666–6671.
Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of humangenes. Princeton, NY: Princeton University Press.
Cerezo M, et al. 2012. Reconstructing ancient mitochondrial DNA links between Africa and Europe. Genome Res. 22:821–826.
Conrad DF, et al. 2011. Variation in genome-wide mutation rates within and between human families. Nat Genet. 43:712–714.
Consortium IH 2003. The International HapMap Project. Nature 789:96.
Currat M, Excoffier L. 2005. The effect of the Neolithic expansion on European molecular diversity. Proc Biol Sci R Soc. 272:679–688.
Durand EY, Eriksson N, McLean CY. 2014. Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis. Mol Biol Evol. 31:2212–2222.
Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28:2239–2252.

Eltis D, Richardson D. 2010. Atlas of the Transatlantic Slave Trade: Yale University Press.
Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587.
Frazer KA, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.
Gomez F, Hirbo J, Tishkoff SA. 2014. Genetic variation and adaptation in Africa: implications for human evolution and disease. Cold Spring Harb Perspect Biol. 6:a008524.
Green RE, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.
Gusev A, et al. 2009. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19:318–326.
Henn BM, et al. 2012. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One 7:e34267.
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 CD, et al. 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res. 21:768–774.
Kondrashov AS, Shabalina SA. 2002. Classification of common conserved sequences in mammalian intergenic regions. Hum Mol Genet. 11:669–674.
Kong A, et al. 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet. 40:1068–1075.
Lahermo P, et al. 1996. The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. Am J Hum Gen. 58:1309–1322.
Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.
Lappalainen T, et al. 2006. Regional differences among the Finns: a Y-chromosomal perspective. Gene 376:207–215.
Lazaridis I, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409–413.
Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493–496.
Li H, et al. 2014. Relationship estimation from whole-genome sequence data. PLoS Genet. 10:e1004144.
Meyer M, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222–226.
Moore CB, et al. 2013. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS Genet. 9:e1003959.
Moorjani P, et al. 2011. The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS Genet. 7:e1001373.
Patterson N, et al. 2012. Ancient admixture in human history. Genetics 192:1065–1093.
Powell JE, Visscher PM, Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet. 11:800–805.
Prufer K, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49.
Qiu S, et al. 2014. Genome evolution by matrix algorithms: cellular automata approach to population genetics. Genome Biol Evol. 6:988–999.
R Development Core Team. 2010. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E. 2015. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. 16:359–371.
Reich D, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.
Richards M, et al. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet. 67:1251–1276.
Rohde DL, Olson S, Chang JT. 2004. Modelling the recent common ancestry of all living humans. Nature 431:562–566.
Rootsi S, et al. 2007. A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. Eur J Hum Genet. EJHG 15:204–211.
Sankararaman S, et al. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507:354–357.
Seguin-Orlando A, et al. 2014. Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. Science 346:1113–1118.
Su SY, et al. 2012. Detection of identity by descent using next-generation whole genome sequencing data. BMC Bioinformatics 13:121.
Thompson EA. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194:301–326.
Vernot B, Akey JM. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. Science 343:1017–1021.
Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. Mol Biol Evol. 26:1823–1827.
Weiss KM, Lambert BW. 2014. What type of person are you? Old-fashioned thinking even in modern science. Cold Spring Harb Perspect Biol. 6:021238.
Zhuang Z, Gusev A, Cho J, Pe'er I. 2012. Detecting identity by descent and homozygosity mapping in whole-exome sequencing data. PLoS One 7:e47618.

Associate editor: Partha Majumder