Annals of human genetics

doi: 10.1111/j.1469-1809.2012.00730.x

Exploring Genomic Structure Differences and Similarities between the Greek and European HapMap Populations: Implications for Association Studies

Vasileios Stathias[1], Georgios R. Sotiris[1], Iordanis Karagiannidis[1], Georgios Bourikas[2], Georgios Martinis[2], Dimitrios Papazoglou[3], Anna Tavridou[4], Nikolaos Papanas[3], Efstratios Maltezos[3], Marios Theodoridis[5], Vassilios Vargemezis[5], Vangelis G. Manolopoulos[4], William C. Speed[6], Judith R. Kidd[6], Kenneth K. Kidd[6], Petros Drineas[7] and Peristera Paschou[1*]

[1]*Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli, Greece*
[2]*Department of Hematology, University Hospital of Alexandroupoli, Faculty of Medicine, Democritus University of Thrace, Alexandroupoli, Greece*
[3]*Second Department of Internal Medicine, University Hospital of Alexandroupoli, Faculty of Medicine, Democritus University of Thrace, Alexandroupoli, Greece*
[4]*Laboratory of Pharmacology, Faculty of Medicine, Democritus University of Thrace, Alexandroupoli, Greece*
[5]*Department of Nephrology, University Hospital of Alexandroupoli, Faculty of Medicine, Democritus University of Thrace, Alexandroupoli, Greece*
[6]*Department of Genetics, School of Medicine, Yale University, New Haven CT, USA*
[7]*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA.*

Summary

Studies of the genomic structure of the Greek population and Southeastern Europe are limited, despite the central position of the area as a gateway for human migrations into Europe. HapMap has provided a unique tool for the analysis of human genetic variation. Europe is represented by the CEU (Northwestern Europe) and the TSI populations (Tuscan Italians from Southern Europe), which serve as reference for the design of genetic association studies. Furthermore, genetic association findings are often transferred to unstudied populations. Although initial studies support the fact that the CEU can, in general, be used as reference for the selection of tagging SNPs in European populations, this has not been extensively studied across Europe. We set out to explore the genomic structure of the Greek population (56 individuals) and compare it to the HapMap TSI and CEU populations. We studied 1112 SNPs (27 regions, 13 chromosomes). Although the HapMap European populations are, in general, a good reference for the Greek population, regions of population differentiation do exist and results should not be light-heartedly generalized. We conclude that, perhaps due to the individual evolutionary history of each genomic region, geographic proximity is not always a perfect guide for selecting a reference population for an unstudied population.

Keywords: Population structure, Southern Europe, Greek population, PCA

Introduction

The genomic structure of the Greek population has not been studied to date, despite its central position in influencing genomic variation throughout Europe. Greece has served as a gateway to migrations from Anatolia, as well as a key refuge of Northern European populations retreating to the South during the last glacial maximum (Semino et al., 2000; Di Giacomo et al., 2003; Semino et al., 2004; Novelletto, 2007). It is in Greece that the earliest signs of Neolithic farmers in Europe are found, about 7000 cal BC, with the founding of a fully fledged farming community at Knossos on the Island of Crete, followed slightly later in the Northwestern

*Corresponding author: Dr. Peristera Paschou, Department of Molecular Biology and Genetics, Democritus University of Thrace, Panepistimioupoli, Dragana, Alexandroupoli 68100, Greece. Tel: +30 25510 30658; Fax: +30 25510 30613; E-mail: ppaschou@mbg.duth.gr

Peloponnese of mainland Greece (Efstratiou, 2005; Perlès, 2005). These populations were undoubtedly crucial to expanding farming to the rest of Europe (Di Giacomo et al., 2004; King et al., 2008). Furthermore, during the period of the Magna Graecia, the sea served rather as a bridge among populations than as a barrier, with Greek traders forming settlements throughout the coasts of Italy, France, and Spain (King et al., 2011).

The patterns of genetic variation across different populations, shaped by history, environment, and stochastic processes, have long been studied in order to infer population relationships and uncover the origins of the human species (Cavalli-Sforza et al., 1994; Tishkoff & Kidd, 2004). Early in the 21st century, the premise of genome-wide association studies (GWAS) was built upon the notion that common variation in the human genome could be tagged and interrogated by a small number of carefully selected single nucleotide polymorphisms (SNPs; the so-called tagging SNPs or tSNPs for short) (Daly et al., 2001; Johnson et al., 2001). This same notion motivated the HapMap project, aiming to characterize the linkage disequilibrium (LD) structure of the human genome (International HapMap Consortium, 2003, 2005; 2007). The HapMap project has become an incredibly valuable resource for investigators around the world, guiding their studies of the genetic background of human disease.

At the same time, the study of genetic structure within Europe has proven to be a lot more complex than until recently appreciated, and this could also be reflected upon the use of HapMap reference samples for the study of European populations. Two major axes of variation are observed within Europe, namely from North to South and from East to West (Lao et al., 2008; Novembre et al., 2008; Paschou et al., 2008; Drineas et al., 2010). The HapMap phase 1 project only included one European population, the CEPH Europeans (actually collected in Utah, USA) and shown to have northwestern European ancestry (International HapMap Consortium, 2003). Since the North-to-South cline of variation was discovered, a second population (Italians from the region of Tuscany) was selected, presumably as representatives of Southern European descent. It is worth mentioning that the Tuscan population, particularly during the Bronze Age and the Apennine Culture, had extensive trading relationships with the Minoan and Myceanean civilizations of Greece (Barker & Rasmussen, 2000).

A topic of considerable debate since the start of the HapMap project has been the degree to which the HapMap populations can actually be considered representative of unstudied populations. A large number of studies have already investigated this issue, although a large portion of the world still remains unstudied (Conrad et al., 2006; González-Neira et al., 2006; De Bakker et al., 2006; Gu et al., 2007, 2008; Paschou et al., 2007a; Hu et al., 2008; Javed et al., 2011). Most studies support the finding that the HapMap populations can indeed serve as reference for unstudied populations, assuming that a closely related or geographically neighboring population is used. Within Europe, most studies comparing the genetic structure of HapMap European populations have been performed for Northern European populations (Mueller et al., 2005; Montpetit et al., 2006; Willer et al., 2006; Lundmark et al., 2008; Pardo et al., 2009). Only a handful of studies focused on Southern Europe [namely Spain (Laayouni et al., 2010; Rodríguez-Ezpeleta et al., 2010)], a population isolate from Croatia (Navarro et al., 2010), and the Italian population (Mueller et al., 2005). In general, it has been shown that the HapMap CEU still capture the most significant portion of variation of other studied populations. However, a couple of studies (Mueller et al., 2005; Pardo et al., 2009) point out the fact that such results are highly dependent on the studied region and are not uniform throughout the genome.

Here, we present for the first time an extensive study of the genetic structure of the Greek population, in comparison to the HapMap reference European populations of Northern Europe (CEPH Europeans– CEU) and Italy (Tuscan Italians–TSI). We study a total of 1112 SNPs spread across 27 regions of the genome. We show that the HapMap reference populations should be expected to serve as a good reference of genetic structure in the Greek population if detailed analysis per region is not required. Regions of population differentiation do exist and results cannot be easily generalized for the entire genome. Furthermore, our results indicate that, perhaps due to complex population relationships and environmental pressures, geographic proximity is not always a perfect guide for selecting a reference population for an unstudied population.

Methods

Samples and Genotypes

We studied samples from 56 unrelated Greeks collected in Alexandroupoli, a city that lies in the northeastern corner of Greece. Participating volunteers were students of the Democritus University of Thrace (originating from many different regions of Greece), or healthy blood donors from the local University Hospital. Informed consent was taken from every participating individual. Self-reported ancestry was considered Greek if the individual reported all four grandparents to be of Greek ancestry and to have been born in Greece. DNA was extracted from whole blood using the Qiagen Puregene kit (Qiagen, Valencia, CA, USA).

Genotyping for 1813 SNPs across 27 different chromosomal regions was performed using an Illumina genotyping custom chip (Illumina, San Diego, CA, USA). The 27 regions across 13 chromosomes represent genomic regions

Table 1 A detailed description of the 27 chromosomal regions (a list of all SNPs is available in the online supplement). The last column is an indication of the "normality" of the region in terms of outlier SNPs, that is, SNPs with abnormally high or abnormally low PCA scores (see Methods section for details). Note that the vast majority of the regions fall below 40 and 60%, a strong indication that the selected regions have normal behavior.

Region	Chr	Start pos.	End pos.	No. of SNPs	Avg. intermarker distance	Avg. PCA score
1	1	75,816,780	76,166,738	26	13,459	31.28
2	2	136,199,825	136,453,930	33	7700	36.67
3	3	46,283,476	46,601,593	26	12,235	38.17
4	3	115,120,114	115,426,997	38	8075	49.77
5	4	100,195,669	100,638,912	31	14,298	41.52
6	5	9,615,669	9,816,915	28	7187	50.86
7	7	27,035,016	27,320,429	31	9206	57.58
8	7	122,306,920	122,549,775	20	12,142	60.99
9	7	141,211,249	141,404,927	25	7747	48.49
10	10	42,894,942	43,205,226	15	20,685	47.86
11	10	106,615,962	107,003,242	34	11,390	44.65
12	11	4,962,261	5,265,654	33	9193	47.32
13	11	5,580,324	5,688,798	19	5709	54.49
14	11	112,565,679	112,840,927	25	11,009	44.98
15	12	6,742,152	6,946,143	18	11,332	58.75
16	12	101,677,417	101,840,870	14	11,675	42.31
17	12	110,643,349	110,824,226	16	11,304	67.84
18	16	88,403,211	88,677,423	35	7834	46.22
19	17	35,175,268	36,858,740	91	18,499	48.45
20	17	37,725,506	38,449,934	13	55,725	59.39
21	17	38,780,091	40,817,264	197	10,340	44.07
22	17	40,905,309	41,615,467	30	23,671	12.88
23	17	41,699,679	46,291,238	320	14,348	52.17
24	17	47,324,368	47,593,466	18	14,949	38.23
25	17	73,570,277	73,876,088	30	10,193	47.06
26	19	50,833,893	51,050,518	20	10,831	47.61
27	22	18,240,875	18,409,878	31	5451	48.11

that have been studied extensively at the laboratory of Drs. Kenneth and Judith Kidd at Yale University over the past few decades, in studies of human population structure around the world (Table 1, Supporting Information Table S1, and online supplement at http://www.cs.rpi.edu/~drinep/GREEKS/). We should note that these regions were *a priori* selected to include genes (Table S1), so we expect our results to be most informative of genic regions throughout the genome. The SNPs were selected to be informative (i.e. nonmonomorphic) and selection was based on distance (i.e. an attempt was made to create maps of markers at equal intermarker distance in each region). We should also note that the genotyped SNPs were explicitly chosen to be polymorphic, so our study is focused on common SNPs. For instance, only 9 and 15.6% of the studied SNPs in the CEU have a rare allele frequency below 10 and 15%, respectively (see online supplement for details on all studied SNPs in all three populations). Among the studied regions, there was one exceptionally large region on Chromosome 17 (about 4.6 Mb and 320 genotyped SNPs at an average intermarker distance of 14.3 kb), whereas for the rest of the regions, the average size was about 420 kb, covered on average by 35 SNPs at an average intermarker distance of 13.1 kb (Table 1).

Of the 1813 SNPs that were genotyped for our Greek population, we were able to find 1112 SNPs genotyped in both the HapMap CEU (112 individuals) and TSI (88 individuals). Only unrelated HapMap individuals were retained. We extracted the relevant data from the HapMap phase 3 database and produced a joint data set of 1112 SNPs and 256 individuals from three European populations that became the focus of our analysis.

Analysis of LD and Population Structure

We visualized the data via pricipal components analysis (PCA), a well-known dimensionality reduction technique.

In prior work (Paschou et al. 2007b; Paschou et al. 2008), we have extensively described how to encode and mean-center genotypic data in order to apply PCA. In our setting, PCA represents all samples with respect to the top two principal components (eigenSNPs). Our choice of two eigenSNPs stems from extensive prior work on the analysis of European genotypic data (Paschou et al., 2008; Drineas et al., 2010).

In order to characterize the profile of the studied regions in terms of population differentiation, we used data from the POPRES (population reference) sample (Nelson et al., 2008). The subset of the POPRES data set that we analyzed comprises 1200 individuals from 11 European populations and has been described in detail previously (Novembre et al., 2008). For each SNP, we computed its correlation with the top two principal components of the data set (PCA-scores), which have been shown to capture the most significant axes of genetic variation within Europe (Lao et al., 2008; Novembre et al., 2008). PCA-scores were computed as we have previously described (Paschou et al., 2007b; Paschou et al., 2008), and they were compared to the distribution of PCA-scores for all available SNPs in the POPRES data set (447,212 SNPs).

Pairwise linkage disequilibrium tests, as well as tagging SNP (tSNP) selection and haplotype block definition, were performed using the algorithms implemented in Haploview (Barrett et al., 2005). For tSNP selection, the Tagger algorithm (as integrated in Haploview) was used without the multimarker testing option. The Gabriel et al. (2002) definition was applied here in order to define "haplotype blocks" across the studied regions. Haplot was used in order to visualize block boundaries across the three studied populations (Gu et al., 2005). We also measured the percentage of "block overlap" (see Supplementary Methods for details) between pairs of populations, in order to evaluate the similarity of haplotypic blocks between the Greek samples and the TSI samples (and vice versa), as well as the Greek samples and the CEU samples (and vice versa).

Results

Allele Frequencies and Population Differentiation over Studied Regions

A total of 1112 SNPs from 27 chromosomal regions were included in our analysis. The PCA of analyzed genotypes for the three populations is shown in Figure S1. In order to characterize the profile of allele frequencies over the studied regions in comparison to the entire genome, we used the POPRES genome-wide data set of Europeans and compared the PCA-scores of SNPs across the studied regions to that of the remaining genome (Table 1). In order to compute the PCA-scores, we effectively calculated the correlation of each SNP with the top two principal components of the data set; these components have been previously shown to correlate with ancestry across Europe (Lao et al., 2008; Novembre et al., 2008; Drineas et al., 2010). For each region we studied here, we estimated the average PCA-score of all available SNPs within the region. Table 1 shows the percentage of SNPs in the genome with a higher PCA-score than the average PCA-score in a particular region. As we have analyzed in detail in earlier work, a high PCA-score is expected for SNPs that show high association with population ancestry.

Our results show that most regions that are included in our analysis are "average" in terms of allele frequencies and correlation to ancestry. Thus, our findings here can be considered, at least to some extent, representative for most regions of the genome as well, when it comes to the level of population differentiation. However, we would like to briefly comment on the top five population-differentiating regions included in our analysis (less than 40% of genome-wide SNPs show a higher PCA-score than the average PCA-score of the respective region). The first is the region of 17q21 in Chromosome 17, encompassing the *MAPT* gene and the recently identified inversion haplotype H2 (Stefansson et al., 2005). As we have recently shown, this 17q21 inversion, often thought to be found at levels of ~20% throughout Europe, actually shows a great range of frequencies within Europe (ranging from 5% up to 37.5%) (Donnelly et al., 2010). The inverted H2 haplotype is actually most frequent around the Mediterranean and decreases outward in all directions. The second region spans the *SLC44A5* and *ACADM* genes and has also been previously shown to account for high population differentiation as a possible candidate region for recent positive selection (Voight et al., 2006; Zhong et al., 2010). The third region spans the *LCT* gene, well known for its involvement in population differentiation across Europe (Bersaglieri et al., 2004; Campbell et al., 2005). The fourth region encompasses the *CCR5* and neighboring genes. CCR5 is the coreceptor that HIV most commonly uses to enter target cells, and, in fact, specific variants of this gene have been associated with protection from HIV infection (de Silva & Stumpf, 2004). This region is well known to show population differentiation and has also been implicated as a candidate locus for natural selection (Novembre et al., 2005; Sabeti et al., 2005; Edo-Matas et al., 2011). Finally, the fifth region with average PCA-scores above the 40th percentile includes the *CA10* gene. Variation across this particular gene has not been previously suggested as population differentiating, even though it lies approximately 6 Mb away from the aforementioned *MAPT* region.

Allele frequencies across studied SNPs for the three European populations are highly correlated, as shown in the scatter plots of Figures 1(A) and (B). As expected, higher correlation is observed between the Greek and the Tuscan population

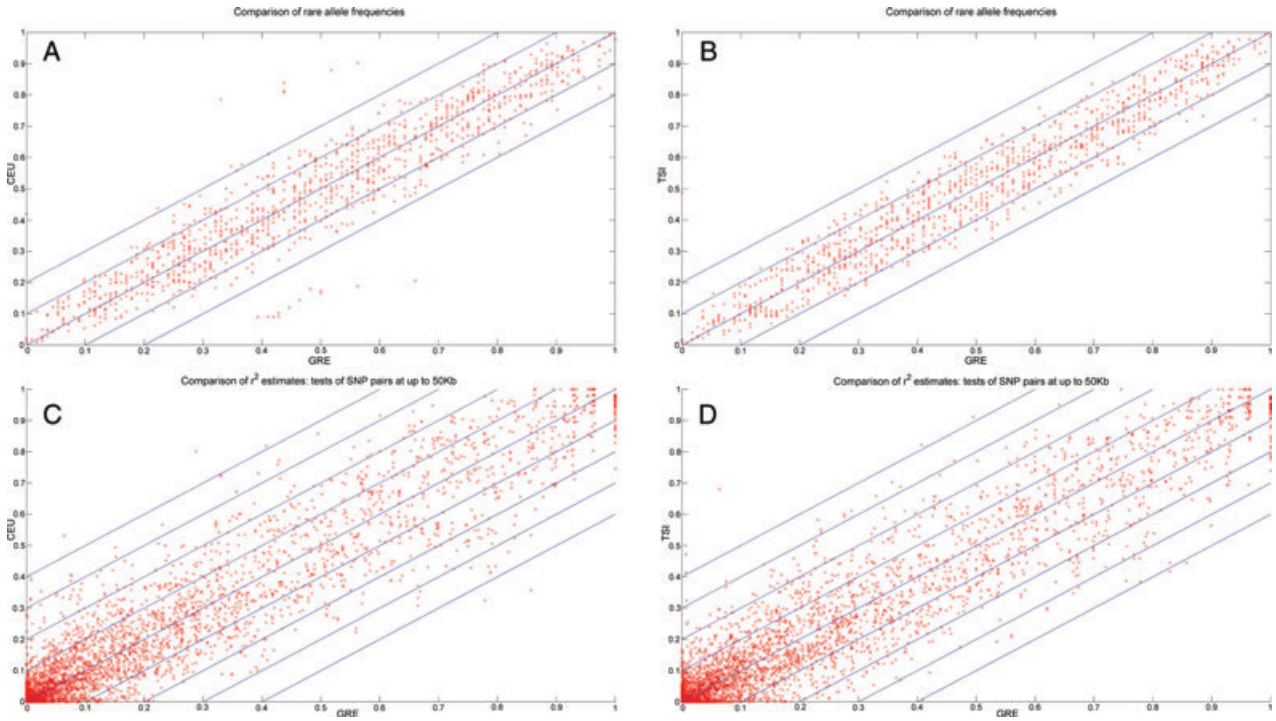

Figure 1 (A) Scatter plot of rare allele frequencies for all available SNPs between the GRE and the CEU populations. Forty-six outlier SNPs are observed (see also Table S2A). (B) Scatter plot of rare allele frequencies for all available SNPs between the GRE and the TSI populations. Seven outlier SNPs are observed (see also Table S2C). (C) Scatter plot of r^2 values between all pairs of SNPs (within 50 kb windows) in the GRE and the CEU populations. Clearly, SNP pairs that are in high LD in the CEU are also in high LD in the GRE, with 104 outlier pairs, as discussed in the Results section (a list of all 104 outlier pairs of SNPs is available in the online supplement). (D) Scatter plot of r^2 values between all pairs of SNPs (within 50 kb windows) in the GRE and the TSI populations. Clearly, SNP pairs that are in high LD in the TSI are also in high LD in the GRE, with 83 outlier pairs, as discussed in the Results section (a list of all 83 outlier pairs of SNPs is available in the online supplement).

with a Pearson correlation coefficient $r = 0.9643$. The respective value for the correlation between the Greek and the CEU population is $r = 0.9387$. In order to identify the most population differentiating SNPs in our data set, we calculated the Informativeness (I_n) of each studied SNP, as defined by Rosenberg et al. (2003) (Fig. S2). Outliers were defined as SNPs whose I_n value exceeds the mean plus three standard deviations. Two SNPs were thus identified as outliers between the Greeks and the Tuscans (residing in the *CD4* and *ADH4* regions, respectively) and 23 SNPs were identified as outliers between the Greeks and the CEPH Europeans. Of the latter 23 SNPs, the top one resides in the *CD4* region and the remaining 22 are found across the *LCT* region.

LD Structure in Greeks Compared to the HapMap 3 European Populations

We measured the extent of LD between all SNP pairs at a distance of 50 kb maximum in the Greek population and compared it to the respective estimates in the CEU and TSI populations. As shown in the scatter plots of Figures 1(C) and (D), most SNP pairs show a high degree of correlation between Greeks and HapMap European populations, with the average correlation coefficient being 0.9622 for comparisons between Greeks and TSI and 0.9614 for comparisons between Greeks and CEU. However, even though the average is high, isolated regions of lower correlation should not be overlooked. An SNP pair was defined to be an outlier if its residual fit to the diagonal (perfect correlation) exceeded the average residual plus three standard deviations. A total of 83 such discordant pairs were found in the Greeks to TSI comparison, whereas 104 such pairs were found in the Greeks to CEU comparison (see Supporting Information and online supplement at http://www.cs.rpi.edu/~drinep/GREEKS/ for considerations on the characteristics of the outlier pairs as well as a list of those pairs).

In order to further study the LD structure of the regions in the three European populations, we defined haplotype blocks in each of the studied regions using the criteria proposed by

Gabriel et al. (2002) as implemented in Haploview. Results of this analysis are shown in Figure S3. A total of 143 blocks over the studied regions were found in the Greek population, whereas 169 and 176 blocks were found in the TSI and CEU populations, respectively. The average block size was 34.2 kb in Greeks, 31 kb in the TSI, and 34.2 kb in the CEU.

In an effort to quantify the degree of similarity between the haplotype blocks and LD structure in the Greek population compared to the HapMap European populations, we estimated the overlap of blocks defined in the Greek population and those defined in each of the TSI and CEU populations. The (average) block overlap values from the GRE samples to the CEU samples (and vice versa) and the (average) block overlap values from the GRE samples to the TSI samples (and vice versa) are shown in Table 2 for each of the 27 studied regions. The overall average overlap values are very similar: block overlap from GRE to CEU is (on average) 83%; block overlap from CEU to GRE is (on average) 65%; block overlap from GRE to TSI is (on average) 76%; and, block overlap from TSI to GRE is (on average) 73%. The average $F1$ statistics for the two pairs of populations are essentially the same: 71% for the GRE and CEU pair, and 72% for the GRE and TSI pair.

Selecting tSNPs in the Greek versus the HapMap European Populations

Next, we investigated the degree to which the HapMap European populations could serve as good reference samples for the selection of tSNPs in the Greek population. In order to do so, we selected tSNPs in each of the HapMap European reference samples (CEU and TSI) and proceeded to test the coverage and efficiency achieved by the selected tSNPs in the Greek population in all studied regions. Results were compared to coverage and efficiency of tSNPs selected from the Greek sample and are shown in Figures 2 and 3. Coverage is defined as the percentage of "untyped" SNPs in the studied region with $r^2 > 0.8$ with a tSNP in the Greek population.

Overall, both the TSI and CEU, when used as reference for tSNP selection, achieve very good coverage of variation in the Greek population (Fig. 3). The TSI tSNPs capture, on average, 94.7% of "untyped" SNPs in Greeks, whereas the CEU capture a somewhat smaller percentage at 92.4%. For about half (13) of the studied regions, both TSI and CEU achieve exactly the same coverage, with exactly the same efficiency (number of selected tSNPs) in nine of these 13 regions. Interestingly, for eight of the regions we studied (regions 1, 2, 4, 6, 8, 9, 18, and 22) both the TSI and CEU achieve perfect (100%) coverage of variation in the Greek population. For three of these regions (2, 4, and 22), the CEU are actually more efficient as reference for Greeks, with fewer tSNPs needed; for the remaining regions, the same number of tSNPs is selected in both populations.

The TSI outperform the CEU as reference for the Greek population in 10 regions. However, there are four regions where the CEU are actually better reference samples than the TSI, contrary to what one might expect based on geographic proximity of the populations. Among them, the most notable are a region of chromosome 7 (100% coverage using the CEU as reference vs. 91.7% using the TSI as reference) and the chromosomal region around *COMT* (100% coverage using the CEU as reference vs. 93% coverage using the TSI as reference). The Chromosome 7 region spans the *TAS2R38* gene (responsible for the PTC taster/nontaster phenotype), as well as the *CLEC5A* gene. The latter gene has been found to have a role in immune response and interact with dengue virus. Finally, the *SLC44A5* regions (discussed in previous sections as one of the most population-differentiating regions in our study) were also captured more accurately in Greeks when the CEU were used as the reference population as opposed to the TSI.

Of the 27 studied regions, five regions resulted in coverage less than 90% when the TSI were used as reference. Of these five regions, one is captured at a percentage of less than 80% (79.3% coverage). This region is the *LCT* region, which is well known to be correlated with population differentiation between Southern and Northern European populations. It is not surprising that this same region results in the lower coverage when the CEU population is used as reference for the Greek population (a coverage of only 65% is achieved). Using the CEU population as reference for the Greek population results in two more regions with coverage below 80%; interestingly, these two regions were the most population-differentiating regions in our sample, according to the PCA-scores-based analysis that we described earlier. More specifically, these regions are: (i) a Chromosome 12 region spanning a large number of genes including *CD4* (66.6% coverage), and (ii) the Chromosome 17 *CA10* region (73.3% coverage). A total of seven of the 27 studied regions cannot be covered at a percentage higher than 90% when the CEU population is used as the reference population (Fig. 3).

Overall, for regions that show less than 95% coverage in Greeks with TSI or CEU tSNPs, LD patterns are typically more complicated in the Greek population. Indeed, more tSNPs would have been selected if the Greek population was used as reference for itself (64.7% of the available SNPs were selected, on average, as tSNPs in such regions in the TSI, whereas 71.4% of the available SNPs were selected, on average, as tSNPs in the Greeks for the same regions; the respective numbers for such regions in the CEU and Greeks comparison were 64.4 and 69.4%).

Finally, we also performed the opposite experiment. We selected tSNPs in Greeks and attempted to see if they could

Table 2 Average block overlap per region for both pairs of populations (GRE and CEU and vice versa, as well as GRE and TSI and vice versa). The *F*1 statistic (see Methods section for details) summarizes the two measurements for each pair of populations. The correlation coefficient between the GRE-to-CEU block overlaps and their reciprocals is equal to 0.73; the correlation coefficient between the GRE-to-TSI block overlaps and their reciprocals is equal to 0.57.

Region	GRE to CEU	CEU to GRE	F1 (CEU, GRE)	GRE to TSI	TSI to GRE	F1 (TSI, GRE)
1	69.75	45.78	0.55	31.88	22.64	0.26
2	98.23	96.94	0.98	92.55	96.75	0.95
3	78.57	71.1	0.75	37.83	38.89	0.38
4	84.61	87.24	0.86	46.93	80	0.59
5	100	100	1	91.58	100	0.96
6	100	91.57	0.96	100	84.3	0.91
7	100	60.48	0.75	95.47	74.01	0.83
8	87.13	84.81	0.86	52	100	0.68
9	67.9	65.71	0.67	67.9	68.22	0.68
10	0	0	*NA*	0	0	*NA*
11	98.52	76.03	0.86	84.78	80.23	0.82
12	100	85.56	0.92	77.01	75.59	0.76
13	100	57.85	0.73	55.58	43.27	0.49
14	100	75.06	0.86	100	100	1
15	77.93	53.16	0.63	100	87.79	0.94
16	38.92	22.28	0.28	38.92	100	0.56
17	100	100	1	100	100	1
18	100	19.54	0.33	100	25.69	0.41
19	94.96	61.3	0.75	78.94	63.2	0.7
20	100	100	1	100	100	1
21	78.79	76.07	0.77	88.9	84.91	0.87
22	74.95	28.93	0.42	74.95	62.04	0.68
23	87.63	74.85	0.81	84.95	81.39	0.83
24	91.36	60.65	0.73	100	62.37	0.77
25	39.6	29.15	0.34	75.89	80.3	0.78
26	100	82.6	0.9	96.25	100	0.98
27	72.6	55.08	0.63	82.98	60.52	0.7
Average	83.01	65.25	0.71	76.12	73.04	0.72

serve as good reference for the TSI and CEU (Fig. 3B). Our premise was that if the Greek population has a more complex structure than the TSI and the CEU, it could serve as a better reference population for European genetic variation. Indeed, the Greek population appears to be a better reference for the CEU population than the CEU is for the Greek population, at least for the regions studied here. On average the Greek tSNPs covered 96.2% of the CEU variation across the studied regions versus the 92.4% coverage achieved by the reverse comparison. On the other hand, the comparison between the Greek and TSI population yields equivocal results (94.6% coverage of TSI variation with Greek tSNPs vs. 94.7% for Greek variation coverage with TSI tSNPs). Nevertheless, upon closer examination of results, we observe that more regions in the TSI are captured with greater than 95% coverage with the Greek tSNPs than the other way around. With the Greek tSNPs 13 regions in the TSI data are covered at 100%, whereas 18 regions are covered at above 95%. For the reverse experiment, (TSI tSNPs applied to the Greeks) the numbers were 9 and 16 regions, respectively.

Discussion

It has long been understood that allele and haplotype frequencies, as well as LD patterns and haplotype block structure, differ across worldwide populations (Sawyer et al., 2005). Within Europe, genetic variation has been shown to be distributed across two major axes (Novembre et al., 2008). However, we are only now beginning to appreciate the extent to which genomic structure differs among European populations and how this structure could possibly affect the design and interpretation of GWAS.

The HapMap project has created an extremely valuable resource for the worldwide scientific community, providing the most comprehensive catalog of genetic variation to date across multiple populations (International HapMap Consortium,

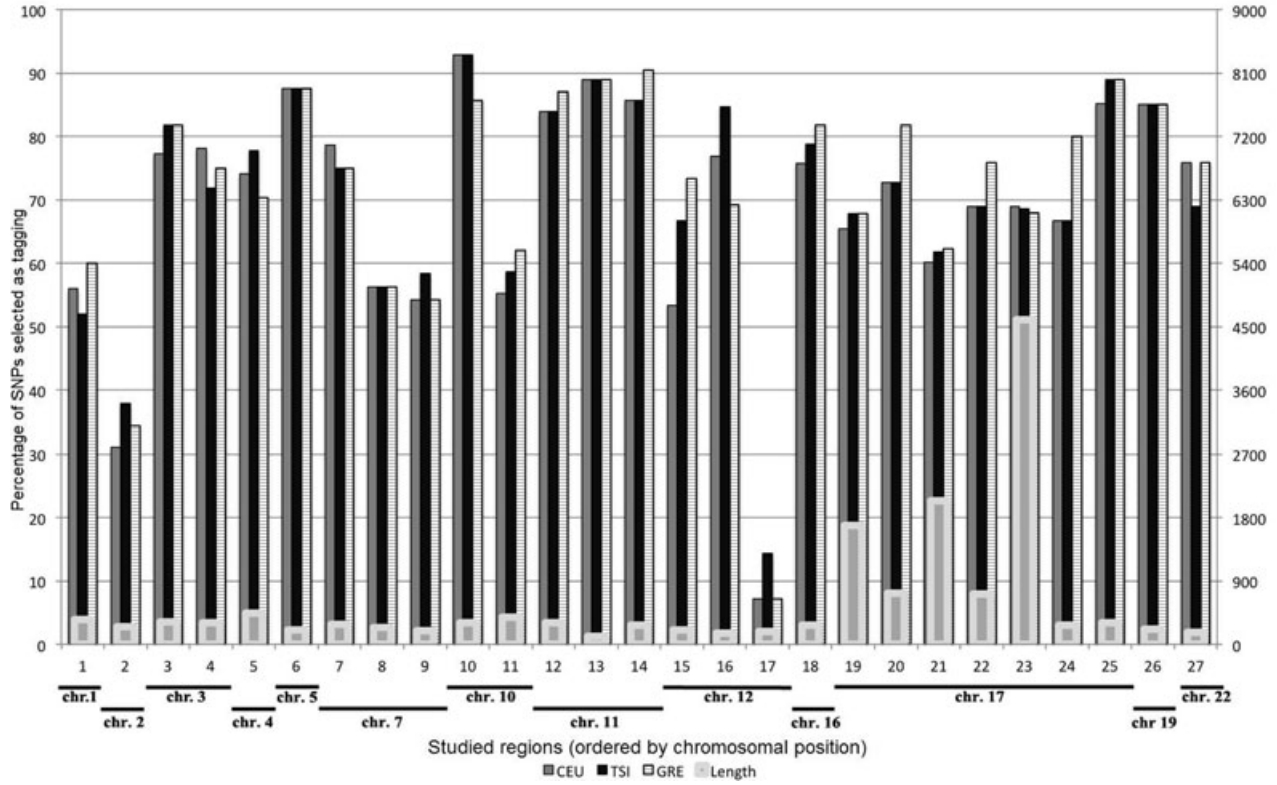

Figure 2 Percentage of SNPs selected as tagging SNPs using Tagger in each region, for each of the three European populations (GRE, CEU, and TSI; see the ruler at the left-hand side of the figure for the percentages). The bold white bars indicate the length of each region (in thousands of bps; see the ruler at the right-hand side of the figure for region lengths).

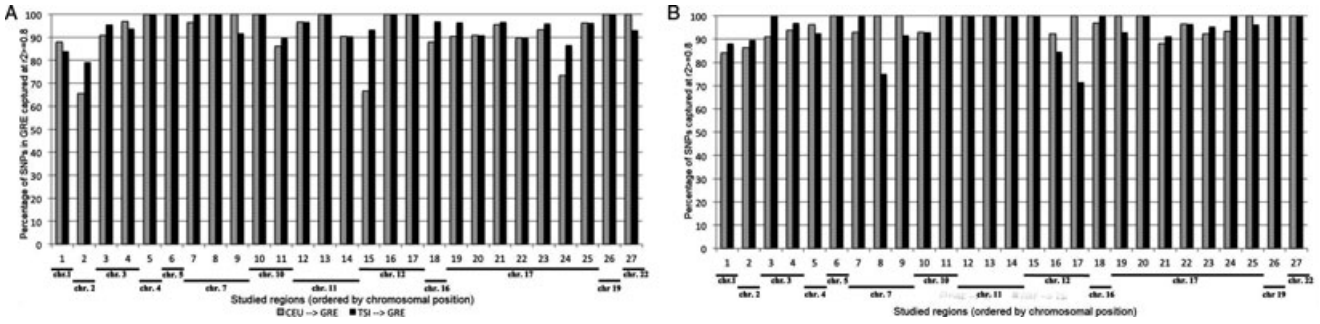

Figure 3 (A) Tagging SNP transferability from the CEU (TSI) population to the GRE population. The CEU (TSI) → GRE bar indicates the percentage of SNPs in the GRE population that were in high LD (r^2 exceeding 0.8) with the tSNPs selected in the CEU (TSI) population.(B) Tagging SNP transferability from the GRE population to the CEU (TSI) population. The GRE → CEU (TSI) bar indicates the percentage of SNPs in the CEU (TSI) population that were in high LD (r^2 exceeding 0.8) with the tSNPs selected in the GRE population.

2007). It is generally accepted that the HapMap data set is hampered by a selection bias in the studied populations and by the inclusion of common SNPs, overlooking rare population variants. However, a number of studies have supported the intuitive fact that, when designing an association study in an unstudied population (i.e. not included in the HapMap project), and if interested in common variation, the optimal solution is to select the geographically closer HapMap population as reference, suggesting this would yield satisfactory coverage of the unknown population (Gu et al., 2007, 2008). At the same time, a couple of studies have attempted to alert researchers to the fact that genomic evolution is not homogeneous:

different stochastic and biological factors may influence regions of the genome in a different way, thus producing unexpected patterns (Mueller et al., 2005; Pardo et al., 2009).

Here, we are presenting for the first time a large-scale study of genomic structure of the Greek population, in comparison to the HapMap reference populations of Northern (CEU) and Southern Europe (TSI). We studied 27 regions across the genome and 1112 SNPs, attempting to get a glimpse of the genomic structure of the Greek population and determine which of the HapMap reference populations could best represent variation in Greeks. So far, studies of genetic structure of the populations of Southern and Eastern Europe have been scarce and largely limited to Y chromosome and mitochondrial variation (Semino et al., 2000; Parreira et al., 2002; Richards et al., 2002; Di Giacomo et al., 2003; Malyarchuk et al., 2003; Robino et al., 2004). Little is known about the genomic architecture of these populations, despite their value in understanding the genomic structure of the European gene pool. The Balkan peninsula and Greece have been the gateway of human migrations to Europe during the Paleolithic and Neolithic ages, whereas the Bronze and Iron Ages were marked by the influence of the Greek culture and trading (Bosch et al., 2006). It is therefore clear that understanding the genetic structure of these populations could provide important insights into the genetic structure of the whole of Europe.

In general, after our detailed comparison of the Greek population with the HapMap Europeans, the patterns of our results are not as straightforward as one might have predicted based on intuition alone. It is clear that when thinking in terms of "averages," both the TSI and the CEU can be considered good reference populations for the Greek population. However, we should emphasize that there exist genomic regions that will be captured poorly by either the TSI or the CEU. Furthermore, again based on "averages" and geographic proximity, one might be quick to conclude that the TSI are a better reference for the genomic variation of the Greek population compared to the CEU. However, our results reveal a rather surprising finding: despite an overall greater degree of similarity among Southern European populations, there exist regions of the genome where the CEU, a population of Northern European descent, is actually more representative of the Greek population than the Southern European Italian Tuscan population (TSI). Our results underline the fact that the evolution of the human genome is extremely complex and generalizations should be avoided.

Great care should be taken when interpreting GWAS results from a population that was not among the HapMap populations and, even more, when attempting to transfer GWAS findings from one population to another. Our findings indicate that the existence of local and population-specific LD variation in the human genome could significantly impair the design of association studies in the Greek population, if, for example, the variant in question lies in a region that is poorly covered by the HapMap reference samples. This could also be the case for other Balkan populations, but also for any population that is not included in the HapMap project. As the quest for the missing heritability component of common disorders becomes more pressing, it becomes apparent that such differentiating regions could have a negative impact in the accuracy of existing association studies. Moreover, their presence suggests the hypothesis that, by using the HapMap populations as reference, researchers may fail to appreciate population-specific variation, a deficit which can only be addressed by large-scale studies of a large number of carefully defined populations.

Acknowledgements

This work was supported, in part, by two Tourette Syndrome Association (TSA) Research Grant Awards to PP; a National Science Foundation (NSF) CAREER award to PD; and a European Molecular Biology Organization Short-Term Fellowship to PD. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

References

Barker, G. & Rasmussen, T. (2000) *The Etruscans*. Malden, MA: Blackwell Publishers.

Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. (2005) Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E. & Hirschhorn, J. N. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**, 1111–1120.

Bosch, E., Calafell, F., González-Neira, A., Flaiz, C., Mateu, E., Scheil, H. G., Huckenbeck, W., Efremovska, L., Mikerezi, I., Xirotiris, N., Grasa, C., Schmidt, H. & Comas, D. (2006) Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Ann Hum Genet* **70**(Pt 4), 459–487.

Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G. & Hirschhorn, J. N. (2005) Demonstrating stratification in a European American population. *Nat Genet* **37**, 868–872.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994) *The history and geography of human genes*. Princeton, NJ: Princeton University Press.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A. & Pritchard, J. K. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**, 1251–1260.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. (2001) High-resolution haplotype structure in the human genome. *Nat Genet* **29**, 229–232.

De Bakker, P. I., Graham, R. R., Altshuler, D., Henderson, B. E., & Haiman, C. A. (2006) Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. *Pac Symp Biocomput*, 478–486.

Di Giacomo, F., Luca, F., Anagnou, N., Ciavarella, G., Corbo, R. M., Cresta, M., Cucci, F., Di Stasi, L., Agostiano, V., Giparaki, M., Loutradis, A., Mammi', C., Michalodimitrakis, E. N., Papola, F., Pedicini, G., Plata, E., Terrenato, L., Tofanelli, S., Malaspina, P. & Novelletto, A. (2003) Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. *Mol Phylogenet Evol* **28**, 387–395.

Di Giacomo, F., Luca, F., Popa, L. O., Akar, N., Anagnou, N., Banyko, J., Brdicka, R., Barbujani, G., Papola, F., Ciavarella, G., Cucci, F., Di Stasi, L., Gavrila, L., Kerimova, M. G., Kovatchev, D., Kozlov, A. I., Loutradis, A., Mandarino, V., Mammi', C., Michalodimitrakis, E. N., Paoli, G., Pappa, K. I., Pedicini, G., Terrenato, L., Tofanelli, S., Malaspina, P. & Novelletto, A. (2004) Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Hum Genet* **115**, 357–371.

Donnelly, M. P., Paschou, P., Grigorenko, E., Gurwitz, D., Mehdi, S. Q., Kajuna, S. L., Barta, C., Kungulilo, S., Karoma, N. J., Lu, R. B., Zhukova, O. V., Kim, J. J., Comas, D., Siniscalco, M., New, M., Li, P., Li, H., Manolopoulos, V. G., Speed, W. C., Rajeevan, H., Pakstis, A. J., Kidd, J. R. & Kidd, K. K. (2010) The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am J Hum Genet* **86**, 161–171.

Drineas, P., Lewis, J. & Paschou, P. (2010) Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers. *PLoS ONE* **5**, e11892.

Edo-Matas, D., Lemey, P., Tom, J. A., Serna-Bolea, C., van den Blink, A. E., van 't Wout, A. B., Schuitemaker, H. & Suchard, M. A. (2011) Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intrahost evolutionary processes: Efficient hypothesis testing through hierarchical phylogenetic models. *Mol Biol Evol* **28**, 1605–1616.

Efstratiou, N. (2005) Tracing the story of the first farmers in Greece–A long and winding road. In: *How did farming reach Europe?* (ed. C. Lichter), pp. 143–153. Istanbul: Deutsches Archäologisches Institut, BYZAS BAND 2.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. & Altshuler, D. (2002) The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

González-Neira, A., Ke, X., Lao, O., Calafell, F., Navarro, A., Comas, D., Cann, H., Bumpstead, S., Ghori, J., Hunt, S., Deloukas, P., Dunham, I., Cardon, L. R., & Bertranpetit, J. (2006) The portability of tagSNPs across populations: A worldwide survey. *Genome Res* **16**, 323–330.

Gu, C. C., Yu, K., Ketkar, S., Templeton, A. R. & Rao, D. C. (2008) On transferability of genome-wide tagSNPs. *Genet Epidemiol* **32**, 89–97.

Gu, S., Pakstis, A. J. & Kidd, K. K. (2005) HAPLOT: A graphical comparison of haplotype blocks, tagSNP sets and SNP variation for multiple populations. *Bioinformatics* **21**, 3938–3939.

Gu, S., Pakstis, A. J., Li, H., Speed, W. C., Kidd, J. R. & Kidd, K. K. (2007) Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations. *Eur J Hum Genet* **15**, 302–312.

Hu, C., Jia, W., Zhang, W., Wang, C., Zhang, R., Wang, J., Ma, X. & Xiang, K. (2008) An evaluation of the performance of HapMap SNP data in a Shanghai Chinese population: Analyses of allele frequency, linkage disequilibrium pattern and tagging SNPs transferability on chromosome 1q21-q25. *BMC Genet* **9**, 19.

International HapMap Consortium. (2003) The International HapMap Project. *Nature* **426**, 789–796.

International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437**, 1299–1320.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.

Javed, A., Drineas, P., Mahoney, M. W. & Paschou, P. (2011) Efficient genomewide selection of PCA-correlated tSNPs for genotype imputation. *Ann Hum Genet* **75**, 707–722.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G. & Todd, J. A. (2001) Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233–237.

King, R. J., DiCristofaro, J., Kouvatsi, A., Triantaphyllidis, C., Scheidel, W., Myres, N. M., Lin, A. A., Eissautier, A., Mitchell, M., Binder, D., Semino, O., Novelletto, A., Underhill, P.A. & Chiaroni, J. (2011) The coming of the Greeks to Provence and Corsica: Y-chromosome models of archaic Greek colonization of the western Mediterranean. *BMC Evol Biol* **11**, 69.

King, R. J., Ozcan, S. S., Carter, T., KalfoÄŸlu, E., Atasoy, S., Triantaphyllidis, C., Kouvatsi, A., Lin, A. A., Chow, C. E., Zhivotovsky, L. A., Michalodimitrakis, M. & Underhill, P. A. (2008) Differential Y-chromosome Anatolian influences on the Greek and Cretan Neolithic. *Ann Hum Genet* **72**(Pt 2), 205–214.

Laayouni, H., Calafell, F. & Bertranpetit, J. (2010) A genome-wide survey does not show the genetic distinctiveness of Basques. *Hum Genet* **127**, 455–458.

Lao, O., Lu, T. T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L. A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A. G., Gieger, C., Wichmann, H. E., Rüther, A., Schreiber, S., Becker, C., Nürnberg, P., Nelson, M. R., Krawczak, M. & Kayser, M. (2008) Correlation between genetic and geographic structure in Europe. *Curr Biol* **18**, 1241–1248.

Lundmark, P. E., Liljedahl, U., Boomsma, D. I., Mannila, H., Martin, N. G., Palotie, A., Peltonen, L., Perola, M., Spector, T. D. & Syvänen, A. C. (2008) Evaluation of HapMap data in six populations of European descent. *Eur J Hum Genet* **16**, 1142–1150.

Malyarchuk, B. A., Grzybowski, T., Derenko, M. V., Czarny, J., Drobnic, K., & Miścicka-Śliwka, D. (2003) Mitochondrial DNA variability in Bosnians and Slovenians. *Ann Hum Genet* **67**(Pt 5), 412–425.

Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Remm, M., Cardon, L., Hudson, T.J. & Metspalu, A. (2006) An

evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet* **2**, e27.

Mueller, J. C., Lõhmussaar, E., Mägi, R., Remm, M., Bettecken, T., Lichtner, P., Biskup, S., Illig, T., Pfeufer, A., Luedemann, J., Schreiber, S., Pramstaller, P., Pichler, I., Romeo, G., Gaddi, A., Testa, A., Wichmann, H. E., Metspalu, A. & Meitinger, T. (2005) Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* **76**, 387–398.

Navarro, P., Vitart, V., Hayward, C., Tenesa, A., Zgaga, L., Juricic, D., Polasek, O., Hastie, N. D., Rudan, I., Campbell, H., Wright, A. F., Haley, C. S. & Knott, S. A. (2010) Genetic comparison of a Croatian isolate and CEPH European founders. *Genet Epidemiol* **34**, 140–145.

Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G., Vollenweider, P., Oksenberg, J. R., Hauser, S. L., Stirnadel, H. A., Kooner, J. S., Chambers, J. C., Jones, B., Mooser, V., Bustamante, C. D., Roses, A. D., Burns, D. K., Ehm, M. G. & Lai, E. H. (2008) The Population Reference Sample, POPRES: A resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* **83**, 347–358.

Novelletto, A. (2007) Y chromosome variation in Europe: Continental and local processes in the formation of the extant gene pool. *Ann Hum Biol* **34**, 139–172.

Novembre, J., Galvani, A. P. & Slatkin, M. (2005) The geographic spread of the CCR5 Delta32 HIV-resistance allele. *PLoS Biol* **3**, e339.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M. & Bustamante, C. D. (2008) Genes mirror geography within Europe. *Nature* **456**, 98–101.

Pardo, L., Bochdanovits, Z., de Geus, E., Hottenga, J. J., Sullivan, P., Posthuma, D., Penninx, B. W., Boomsma, D. & Heutink, P. (2009) Global similarity with local differences in linkage disequilibrium between the Dutch and HapMap-CEU populations. *Eur J Hum Genet* **17**, 802–810.

Parreira, K. S., Lareu, M. V., Sánchez-Diz, P., Skitsa, I. & Carracedo, A. (2002) DNA typing of short tandem repeat loci on Y-chromosome of Greek population. *Forensic Sci Int* **126**, 261–264.

Paschou, P., Drineas, P., Lewis, J., Nievergelt, C. M., Nickerson, D. A., Smith, J. D., Ridker, P. M., Chasman, D. I., Krauss, R. M. & Ziv, E. (2008) Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet* **4**, e1000114.

Paschou, P., Mahoney, M. W., Javed, A., Kidd, J. R., Pakstis, A. J., Gu, S., Kidd, K. K. & Drineas, P. (2007a). Intra- and interpopulation genotype reconstruction from tagging SNPs. *Genome Res* **17**, 96–107.

Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W. & Drineas, P. (2007b) PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet* **3**, 1672–1686.

Perlès, C. (2005) From the near East to Greece: Let's reverse the focus. Cultural elements that didn't transfer. In: *How did farming reach Europe?* (ed. C. Lichter), pp. 275–290. Istanbul: Deutsches Archäologisches Institut, BYZAS BAND 2.

Richards, M., Macaulay, V., Torroni, A. & Bandelt, H. J. (2002) In search of geographical patterns in European mitochondrial DNA. *Am J Hum Genet* **71**, 1168–1174.

Robino, C., Varacalli, S., Gino, S., Chatzikyriakidou, A., Kouvatsi, A., Triantaphyllidis, C., Di Gaetano, C., Crobu, F., Matullo, G., Piazza, A. & Torre, C. (2004) Y-chromosomal STR haplotypes in a population sample from continental Greece, and the islands of Crete and Chios. *Forensic Sci Int* **145**, 61–64.

Rodríguez-Ezpeleta, N., Alvarez-Busto, J., Imaz, L., Regueiro, M., Azcárate, M. N., Bilbao, R., Iriondo, M., Gil, A., Estonba, A. & Aransay, A. M. (2010) High-density SNP genotyping detects homogeneity of Spanish and French Basques, and confirms their genomic distinctiveness from other European populations. *Hum Genet* **128**, 113–117.

Rosenberg, N., Li, L., Ward, R. & Pritchard, J. (2003) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73**, 1402–1422.

Sabeti, P. C., Walsh, E., Schaffner, S. F., Varilly, P., Fry, B., Hutcheson, H. B., Cullen, M., Mikkelsen, T. S., Roy, J., Patterson, N., Cooper, R., Reich, D., Altshuler, D., O'Brien, S. & Lander, E. S. (2005) The case for selection at CCR5-Delta32. *PLoS Biol* **3**, e378.

Sawyer, S. L., Mukherjee, N., Pakstis, A. J., Feuk, L., Kidd, J. R., Brookes, A. J. & Kidd, K. K. (2005) Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* **13**, 677–686.

Semino, O., Magri, C., Benuzzi, G., Lin, A. A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P., Oefner, P. J., Zhivotovsky, L. A., King, R., Torroni, A., Cavalli-Sforza, L. L., Underhill, P. A. & Santachiara-Benerecetti, A. S. (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: Inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* **74**, 1023–1034.

Semino, O., Passarino, G., Oefner, P. J., Lin, A. A., Arbuzova, S., Beckman, L. E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., Marcikiae, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A. S., Cavalli-Sforza, L.L. & Underhill, P. A. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: A Y chromosome perspective. *Science* **290**, 1155–1159.

de Silva, E. & Stumpf, M. P. (2004) HIV and the CCR5-Delta32 resistance allele. *FEMS Microbiol Lett* **241**, 1–12.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., Desnica, N., Hicks, A., Gylfason, A., Gudbjartsson, D. F., Jonsdottir, G. M., Sainz, J., Agnarsson, K., Birgisdottir, B., Ghosh, S., Olafsdottir, A., Cazier, J. B., Kristjansson, K., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., Kong, A. & Stefansson, K. (2005) A common inversion under selection in Europeans. *Nat Genet* **37**, 129–137.

Tishkoff, S. A. & Kidd, K. K. (2004) Implications of biogeography of human populations for 'race' and medicine. *Nat Genet* **36**(Suppl 11), S21–S27.

Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. (2006) A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72.

Willer, C. J., Scott, L. J., Bonnycastle, L. L., Jackson, A. U., Chines, P., Pruim, R., Bark, C. W., Tsai, Y. Y., Pugh, E. W., Doheny, K. F., Kinnunen, L., Mohlke, K. L., Valle, T. T., Bergman, R. N., Tuomilehto, J., Collins, F. S. & Boehnke, M. (2006) Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet Epidemiol* **30**, 180–190.

Zhong, M., Lange, K., Papp, J. C. & Fan, R. (2010) A powerful score test to detect positive selection in genome-wide scans. *Eur J Hum Genet* **18**, 1148–1159.

Supporting Information

Additional supporting information may be found in the online version of this article:

Table S1 List of the genes that appear in each of the 27 studied chromosomal regions.

Table S2 Lists of outlier SNPs in population comparisons.

Figure S1 PCA plot of the three populations.

Figure S2 Informativeness (I_n) scores for all SNPs for the CEU-GRE and the CEU-TSI comparisons.

Figure S3 Haplotype blocks in each of the 27 studied chromosomal regions.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganised for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received: 12 March 2012
Accepted: 13 July 2012