Geographical Affinities of the HapMap Samples

Miao He, Jane Gitschier, Tatiana Zerjal, Peter de Knijff, Chris Tyler-Smith, Yali Xue

OPEN ACCESS Freely available online

PLoS one

Geographical Affinities of the HapMap Samples

Miao He[1], Jane Gitschier[1,2], Tatiana Zerjal[1,3], Peter de Knijff[4], Chris Tyler-Smith[1], Yali Xue[1]*

1 The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom, **2** Department of Medicine and Pediatrics, University of California San Francisco, San Francisco, California, United States of America, **3** Station de Génétique Végétale, Ferme du Moulon, Gif-sur-Yvette, France, **4** Department of Human Genetics, Center for Human and Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands

Abstract

Background: The HapMap samples were collected for medical-genetic studies, but are also widely used in population-genetic and evolutionary investigations. Yet the ascertainment of the samples differs from most population-genetic studies which collect individuals who live in the same local region as their ancestors. What effects could this non-standard ascertainment have on the interpretation of HapMap results?

Methodology/Principal Findings: We compared the HapMap samples with more conventionally-ascertained samples used in population- and forensic-genetic studies, including the HGDP-CEPH panel, making use of published genome-wide autosomal SNP data and Y-STR haplotypes, as well as producing new Y-STR data. We found that the HapMap samples were representative of their broad geographical regions of ancestry according to all tests applied. The YRI and JPT were indistinguishable from independent samples of Yoruba and Japanese in all ways investigated. However, both the CHB and the CEU were distinguishable from all other HGDP-CEPH populations with autosomal markers, and both showed Y-STR similarities to unusually large numbers of populations, perhaps reflecting their admixed origins.

Conclusions/Significance: The CHB and JPT are readily distinguished from one another with both autosomal and Y-chromosomal markers, and results obtained after combining them into a single sample should be interpreted with caution. The CEU are better described as being of Western European ancestry than of Northern European ancestry as often reported. Both the CHB and CEU show subtle but detectable signs of admixture. Thus the YRI and JPT samples are well-suited to standard population-genetic studies, but the CHB and CEU less so.

Citation: He M, Gitschier J, Zerjal T, de Knijff P, Tyler-Smith C, et al. (2009) Geographical Affinities of the HapMap Samples. PLoS ONE 4(3): e4684. doi:10.1371/journal.pone.0004684

Editor: Philip Awadalla, University of Montreal, Canada

Received October 31, 2008; **Accepted** January 17, 2009; **Published** March 4, 2009

Funding: This work was supported by The Wellcome Trust (MH, CT-S, YX), the Howard Hughes Medical Institute (JG) and a grant from the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (PdK). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ylx@sanger.ac.uk

Introduction

The International HapMap Project was established in 2002 with the primary aim of determining the common patterns of DNA sequence variation in the human genome in order to facilitate the discovery of sequence variants that affect common diseases [1]. It was based on 270 individuals from four sources: YRI (Yoruba in Ibadan, Nigeria), CHB (Han Chinese in Beijing, China), JPT (Japanese in Tokyo, Japan) and CEU (CEPH Utah residents with ancestry from northern and western Europe). Over 3.1 million SNPs were genotyped in these samples and the patterns of linkage disequilibrium (LD) defined [2,3]; these patterns, and the SNPs necessary to tag them have been shown to be similar in a broader set of populations, e.g. [4]. As a result, our understanding of the genetic factors influencing common diseases has accelerated considerably [5]. In addition, the availability of cell lines from these samples has allowed many additional studies to be performed, including analyses of copy number variation [6,7] and gene expression [8,9], while whole-genome resequencing is now under way (http://www.1000genomes.org/page.php). Moreover, the HapMap samples have been extensively used in studies searching for signals of population differentiation and natural selection, e.g. [10–12]. It is therefore no exaggeration to consider the HapMap samples the most intensively studied genetic samples ever.

Yet these samples, and the way in which they were collected, differ significantly from the samples used more commonly by population and evolutionary geneticists. Geneticists interested in the events that have shaped human populations over the last 50,000 years or so have usually preferred to sample individuals living in the same location as their ancestors (indigenous people), often excluding individuals whose grandparents do not all come from the same local area, or whose ancestors are known to have migrated during historical times [13]. By these criteria, the CHB and CEU samples would have been excluded. Geneticists have also generally analysed samples from different locations independently, but the CHB and JPT are often combined into a single Asian sample sometimes abbreviated 'ASN', e.g. [14]. What effect would the different sampling and grouping criteria introduce?

We set out to compare the HapMap samples with those more commonly used by population, evolutionary and forensic geneticists [e.g. 15,16,17]. We performed genomewide analyses based on published autosomal SNP genotypes [3,18] to obtain an overall view, and supplemented these with Y-chromosomal analyses

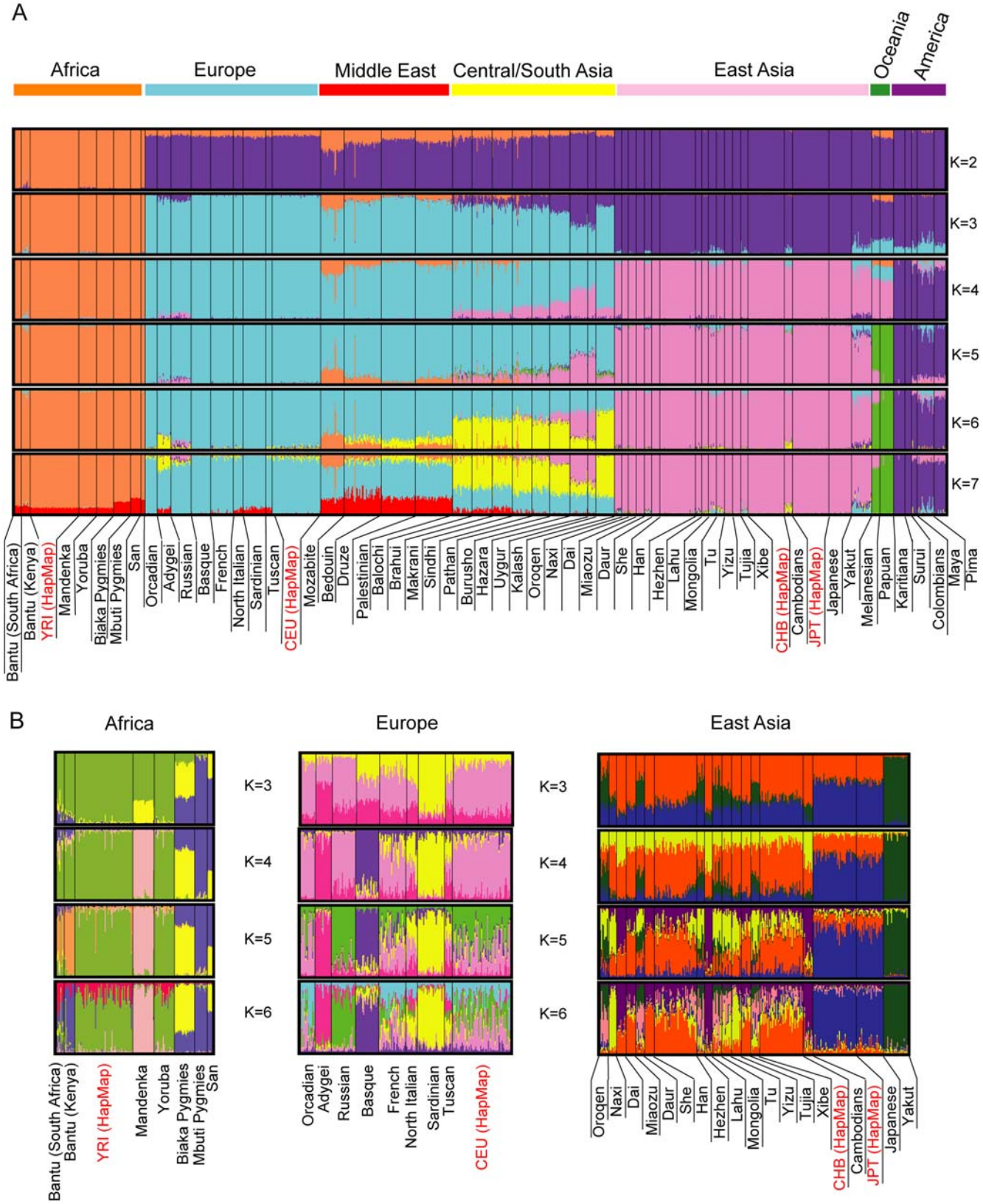

Figure 1. STRUCTURE analysis of the HapMap and HGDP-CEPH panels using 5,254 unlinked SNPs. A. Full dataset. B. Subsets of the panels from restricted geographical regions.
doi:10.1371/journal.pone.0004684.g001

Table 1. Comparison of frequencies of genetic clusters identified by STRUCTURE (K = 6) in HapMap samples and the most similar HGDP-CEPH sample.

Comparison	K	Cluster	p-value
YRI-Yoruba	6	1	0.108
		2	0.697
		3	0.891
		4	0.360
		5	0.235
		6	0.686
JPT-Japanese	6	1	0.460
		2	0.067
		3	0.139
		4	0.686
		5	0.367
		6	0.335
CHB-Han	6	1	0.030
		2	0.435
		3	0.086
		4	0.075
		5	0.140
		6	<0.001*
CEU-French	6	1	0.012
		2	0.005*
		3	0.045
		4	<0.001*
		5	0.021
		6	0.011

*Significant difference after Bonferroni correction for six tests.

doi:10.1371/journal.pone.0004684.t001

because of the uniquely powerful geographical information carried by this locus [19]. We show that, while all the HapMap samples do indeed show the general affinities expected from their ancestral origins, the paternal geographical ancestry of the CEU is slightly different from the 'northern and western Europe' suggested by the HapMap, and both the CHB and CEU differ in subtle ways from samples collected using more standard criteria.

Results

The program STRUCTURE allows individuals to be clustered on the basis of their genetic information [20]. It has previously been applied to genome-wide STR and SNP datasets from the HGDP-CEPH panel of 52 worldwide populations and identified clusters of individuals corresponding to specific geographical regions which appear to be robust and largely independent of the set of markers used [18,21]. We performed STRUCTURE analysis on a set of genome-wide SNP genotypes from the combined HGDP-CEPH and HapMap panels using 5,254 SNPs [18] that were located $\geq$0.5 Mb apart and thus expected to show little LD. The STRUCTURE program requires that a number of clusters, K is specified in advance, but allows K to be varied between runs. As K was increased from 2 to 7 in different runs, clusters corresponding to finer geographical subdivisions of the world were identified, as seen when the HGDP-CEPH panel was used alone [18]. At this worldwide level of resolution, the HapMap samples always lay in the cluster expected from their ancestry (Figure 1A). We then refined the analysis by examining sub-Saharan Africa, East Asia and Europe individually (Figure 1B). In these more detailed comparisons, the YRI were still indistinguishable from the HGDP-CEPH Yoruba, and the JPT from the HGDP-CEPH Japanese (Figure 1B, Table 1). In contrast, both the CHB and CEU were distinguishable from all the HGDP-CEPH samples at higher values of K (Figure 1B). The CHB appeared most similar to the HGPD-CEPH Han or Tujia, and the CEU to the HGDP-CEPH French, but still showed visible differences in the frequency of one or more clusters (Figure 1B), and these were confirmed as statistically significant by a Mann-Whitney test after Bonferroni correction (Table 1). However, because of the limited population representation in the HGDP collection, it is possible that these samples would be more similar to other populations that had not been sampled.

In order to investigate their genetic relationships further, we turned to the locus that provides the highest geographical resolution, and for which large geographically-structured datasets are available: the Y chromosome. We typed the DNAs with a widely-used set of Y-STRs (Table S1), calculated population pairwise genetic distances, and compared the HapMap to the HGDP-CEPH set to provide a worldwide perspective. A Multidimensional Scaling (MDS) plot of these distances showed considerable geographical structure (Figure 2A), although not complete separation of continental regions. Nevertheless, the YRI lay closest to the HGDP-CEPH Yoruba in a cluster of African populations. The CHB and JPT lay close together near the centre of the East Asian cluster, near the Han, Yizu, Dai, Tujia and HGDP-CEPH Japanese. The CEU were located outside the main cluster of European populations, but between this cluster and the Basques who are often observed as an outlier in population-genetic studies [22]. Thus this analysis also revealed overall similarities between the HapMap samples and traditionally-ascertained samples with ancestry from the same regions.

It was possible to investigate these relationships further for East Asian and European samples due to the availability of additional published Y-chromosomal datasets for populations from these regions. We therefore compared the CHB and JPT to a set of 27 populations from East Asia, largely independent of the HGDP-CEPH collection [17]. The JPT again lay closest to the Japanese sample (Figure 2B), and the genetic distance between them was not significantly greater than zero, although the distance between each of the Japanese samples and all the other samples was significant (Table S2). The conclusions about the CHB were somewhat different. They lay well within the East Asian cluster. However, based on their origin in Beijing in Northern China, they would be expected to lie within the Northern cluster of East Asian populations (blue in Figure 2B). Instead, they lie at the border between the Northern and Southern clusters. Examination of the genetic distances between the CHB and the other populations revealed that they were not significantly different from 11 of the others, an unusually large number since the mean value was 3.7, SD = 3.8. The geographical distribution of these 'similar' samples is broad (Figure 3A), and while the Xibe and Han (Xinjiang) populations in the West are known to result from migration within the last few centuries [23], the similar populations include both Northern and Southern populations that cannot all be explained by recent migration.

The CEU were compared with a set of 81 European populations [24]. In the MDS plot they lie at the edge of the Western European cluster (Figure 2C). Interestingly, they shared with the CHB the feature of showing an unusually large number of populations with genetic distances that were not significantly

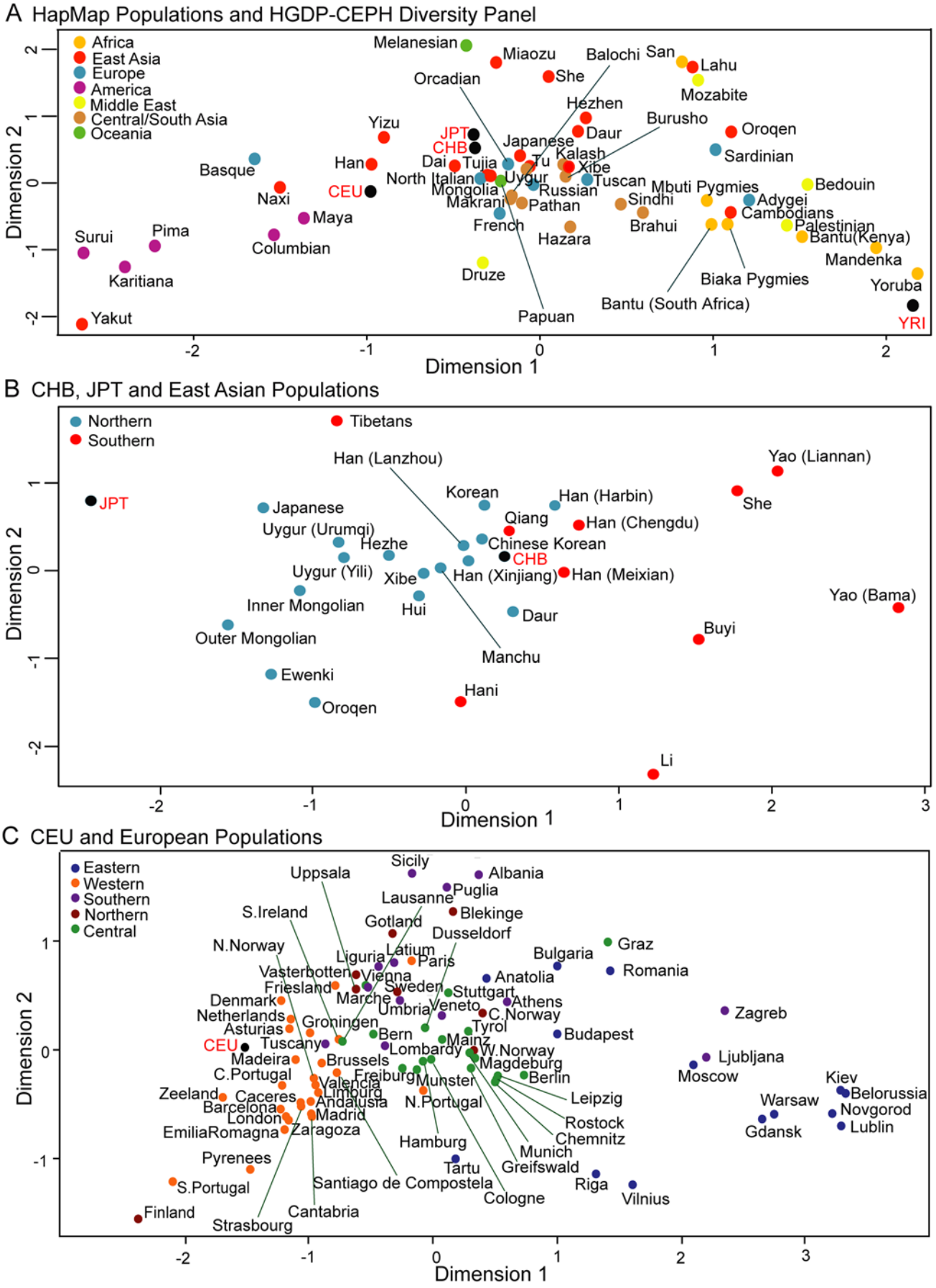
A HapMap Populations and HGDP-CEPH Diversity Panel
Africa
East Asia
Europe
America
Middle East
Central/South Asia
Oceania
Melanesian
Miaozu
Balochi
San
Lahu
Mozabite
Orcadian
She
Hezhen
Burusho
Oroqen
Yizu
JPT
CHB
Japanese
Daur
Sardinian
Basque
Han
Dai
Tujia
Tu
Kalash
Xibe
North Italian
Uygur
Tuscan
Mbuti Pygmies
Bedouin
Naxi
CEU
Mongolia
Russian
Sindhi
Adygei
Makrani
Pathan
Cambodians
Maya
French
Brahui
Palestinian
Pima
Hazara
Bantu(Kenya)
Surui
Columbian
Mandenka
Druze
Biaka Pygmies
Karitiana
Yoruba
Bantu (South Africa)
Yakut
Papuan
YRI
Dimension 1
Dimension 2
B CHB, JPT and East Asian Populations
Northern
Southern
Tibetans
Han (Lanzhou)
Yao (Liannan)
JPT
Korean
Han (Harbin)
She
Japanese
Qiang
Han (Chengdu)
Uygur (Urumqi)
Hezhe
Chinese Korean
CHB
Uygur (Yili)
Xibe
Han (Xinjiang)
Han (Meixian)
Yao (Bama)
Inner Mongolian
Hui
Daur
Buyi
Outer Mongolian
Manchu
Ewenki
Hani
Oroqen
Li
Dimension 1
Dimension 2
C CEU and European Populations
Eastern
Western
Southern
Northern
Central
Sicily
Uppsala
Albania
Puglia
Lausanne
S.Ireland
Blekinge
Gotland
Dusseldorf
Graz
N.Norway
Latium
Bulgaria
Liguria
Paris
Romania
Vasterbotten
Vienna
Anatolia
Friesland
Sweden
Stuttgart
Denmark
Marche
Athens
Umbria
Veneto
C.Norway
Zagreb
Netherlands
Groningen
Tyrol
Asturias
Bern
Budapest
CEU
Tuscany
Mainz
Lombardy
W.Norway
Ljubljana
Madeira
Brussels
Magdeburg
Moscow
C.Portugal
Freiburg
Berlin
Kiev
Valencia
Munster
Zeeland
Caceres
Limburg
Leipzig
Belorussia
Barcelona
Andalusia
N.Portugal
Warsaw
London
Madrid
Rostock
Novgorod
EmiliaRomagna
Zaragoza
Chemnitz
Gdansk
Lublin
Hamburg
Pyrenees
Munich
Tartu
Greifswald
S.Portugal
Santiago de Compostela
Riga
Vilnius
Cologne
Finland
Cantabria
Strasbourg
Dimension 1
Dimension 2

Figure 2. Genetic distances between populations based on Y-STR haplotypes. A. Complete HapMap and HGDP panels using 17 loci (*DYS19*, *DYS189I*, *DYS389II*, *DYS390*, *DYS391*, *DYS392*, *DYS393*, *DYS438*, *DYS439*, *DYS437*, *DYS448*, *DYS456*, *DYS458*, *DYS635*, *Y GATA H4*). B. CHB, JPT and East Asian populations using 10 loci (*DYS19*, *DYS389I*, *DYS389b*, *DYS390*, *DYS391*, *DYS392*, *DYS393*, *DYS437*, *DYS438*, *DYS439*). C. CEU and European populations using seven loci (*DYS19*, *DYS389I*, *DYS389b*, *DYS390*, *DYS391*, *DYS392*, *DYS393*).
doi:10.1371/journal.pone.0004684.g002

greater than zero: in this case 33, compared with a mean of 15.9 and SD of 10.5. As expected from the MDS plot, the geographical distribution of these similar populations was mostly from Western Europe, with only three from Northern Europe (Figure 3B).

Discussion

In this study we compared the HapMap samples with population samples ascertained according to more standard sampling protocols, using both autosomal and Y-chromosomal datasets. We found that they do broadly resemble other samples from the same geographical region (YRI, CHB, JPT) or with similar ancestry (CEU, Europeans). In particular, the YRI and JPT were indistinguishable from independent Yoruba and Japanese samples, respectively, by all the criteria used, but were distinct from other available samples from their regions. A detailed study of over 7,000 samples from the Japanese archipelago using >140,000 SNPs found limited substructure within this region, and also confirmed that the HapMap JPT fell into the major 'Hondo' cluster [25]. The CHB and CEU did not resemble in detail any of the HGDP populations when analysed with autosomal markers (Figure 1B, Table 1), but showed similarities to unusually large numbers of neighbouring populations with Y-chromosomal markers. We now consider CHB and CEU findings in more detail, and a number of implications for the use of the HapMap samples.

The lack of detailed similarity between the genome-wide autosomal genotypes of the CHB and CEU samples and the HGDP-CEPH panel could reflect the combination of high discriminatory power from such a large number of SNPs and the small number of comparison populations. In a more detailed comparison of the CEU with 2,457 individuals from 23 European populations, individual's SNP genotypes were clustered using principal component analysis [26]. Individuals from each European population generally clustered together and although the populations formed overlapping clusters, the broad North, South, East and West geographical areas of Europe were readily separated. In this analysis, the CEU were most similar to samples from the Netherlands and the UK, in agreement with the Y-chromosomal data, but in contrast were quite distinct from Spanish and Portuguese samples, which were not significantly different at the Y-chromosomal level (c.f. Figure 3B). We compared the number of samples that showed different or not different Y-chromosomal distances from the CEU in Central, Northern, Southern, Eastern and Western Europe with, in each case, the rest of Europe, using a Fisher exact test and found a striking enrichment of similar samples in Western Europe ($p<0.000001$) but in no other region. Some differences between a single locus and the combination of a large number of loci is unsurprising, but may also reflect the limited number of Y-STRs available for the detailed European comparison and the similarities in Y-chromosomal haplotypes throughout much of Western Europe, where haplogroup R1b predominates, being common in both Britain and Iberia [27,28], for example. Together, these results show that the CEU, in contrast to the HapMap recommended descriptor 'Utah residents with ancestry from northern and western Europe' (http://www.hapmap.org/citinghapmap.html) are not appropriately described as having Northern European ancestry; Western or North-western Europe ancestry would be more accurate. A similarly detailed comparison of the CHB with additional East Asian samples would be of interest, but would require additional data, which are not yet available.

The Y-chromosomal genetic similarity of both the CHB and CEU to an unusually large number of other populations is likely to reflect their mixed origins. The CHB samples were collected from volunteers at Beijing Normal University [1],which hosts 16,000 students originating from many parts of China and including 2,000 from overseas (http://www.bnu.edu.cn/eng/about_bnu/facts_of_bnu.htm). The CEU were recruited in Utah, USA, and are descendants of Europeans whose ancestry is not well documented, but could well include more than one European country.

Finally, we emphasise one obvious point: the CHB and JPT are readily distinguished from one another with both autosomal and Y-chromosomal markers, and conclusions derived from a combined 'ASN' population should be interpreted with caution. For example, when we constructed an artificial mixture of equal numbers of CHB and JPT Y chromosomes, the mixture showed different characteristics from both HapMap samples and resembled five populations, including Koreans and Chinese Koreans (results not shown). While Korea is geographically intermediate, it would clearly be inappropriate to regard a HapMap sample as Korean. The HapMap study is currently being extended to additional more diverse populations in a Phase 3 (http://www.hapmap.org/index.html.en), and several of these samples also differ from conventional samples in having recently admixed and/or migrant origins, so the interpretation of the results from this phase of the project would be enhanced by including studies of the kind performed here.

Materials and Methods

Datasets

The genome-wide SNP genotypes of the 270 individuals in the International HapMap Project were downloaded from www.hapmap.org (Schema: rel22_NCBI_Build36), and after removing the children in the YRI and CEU samples all analyses were performed on 210 samples. Genotypes of 940 individuals from 52 populations in the HGDP-CEPH Diversity Panel (Stanford University HGDP-CEPH SNP Genotyping Data [18]) were downloaded from http://www.cephb.fr/hgdp-cephdb/. These were based on the commonly-used H952 subset [29], omitting individuals with insufficient data. Autosomal loci in common between the two datasets were then identified using a pair of Perl scripts (Script S1 and Script S2), and 5,254 loci separated by ≥0.5 Mb (and thus probably unlinked) were chosen from this list.

Y-STR data for 17 markers were generated from the HapMap and HGDP-CEPH males, again excluding the YRI and CEU sons, using the AmpFℓSTR® Yfiler® PCR amplification kit (Applied Biosystems) (*DYS19*, *DYS189I*, *DYS389II*, *DYS390*, *DYS391*, *DYS392*, *DYS393*, *DYS385I/II*, *DYS438*, *DYS439*, *DYS437*, *DYS448*, *DYS456*, *DYS458*, *DYS635 and Y GATA H4*) [30]. Additional Y-STR data were obtained from public sources: 16 markers from 980 individuals belonging to 27 East Asian populations [17], or seven markers from over 12,700 samples from 91 locations in Europe which were downloaded from the Y-STR Haplotype Reference Database (YHRD, http://www.yhrd.org).

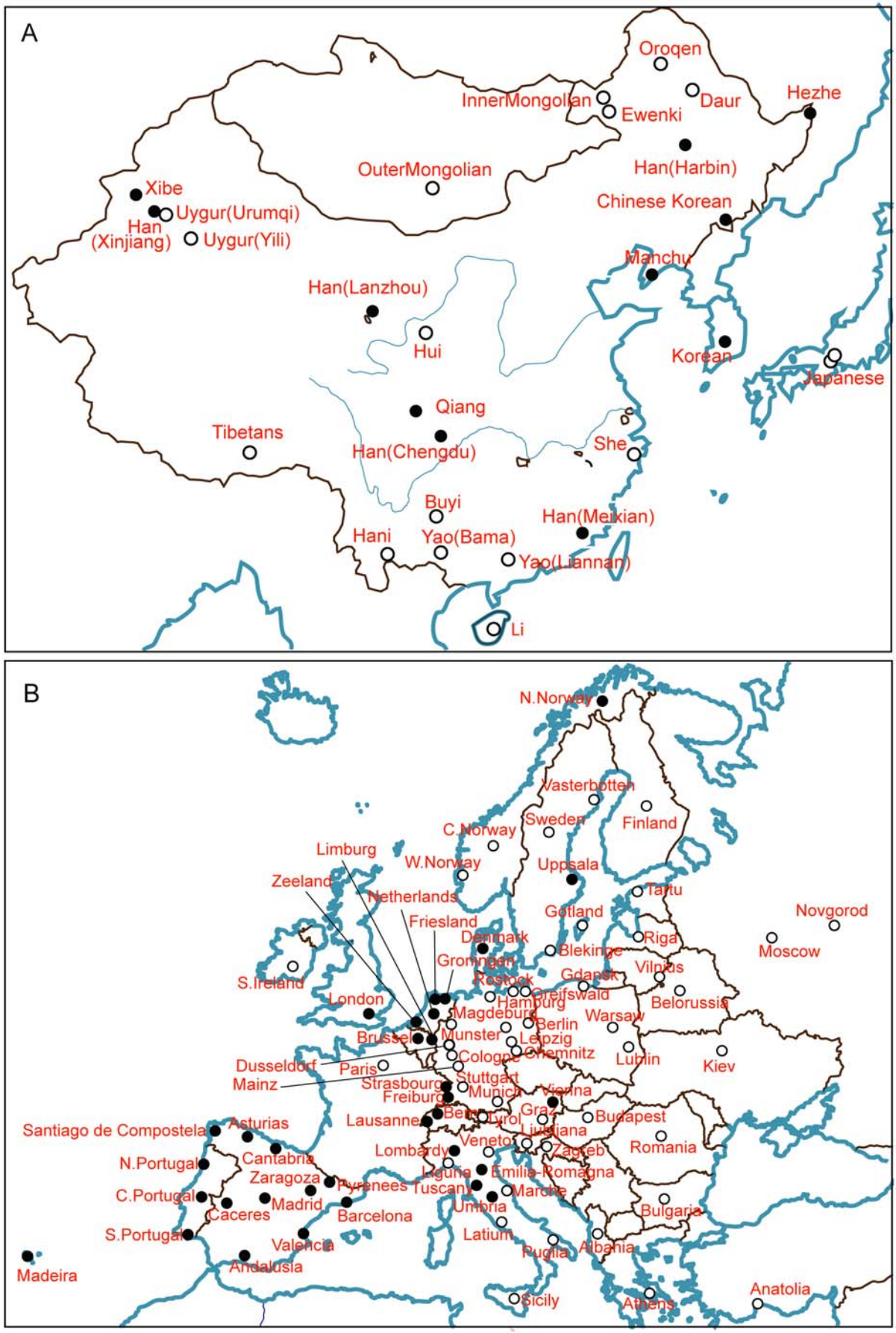
A
Oroqen
Daur
Hezhe
InnerMongolian
Ewenki
Han(Harbin)
OuterMongolian
Xibe
Uygur(Urumqi)
Han
(Xinjiang)
Uygur(Yili)
Chinese Korean
Manchu
Han(Lanzhou)
Hui
Korean
Japanese
Qiang
Tibetans
Han(Chengdu)
She
Buyi
Han(Meixian)
Hani
Yao(Bama)
Yao(Liannan)
Li
B
N.Norway
Vasterbotten
Sweden
Finland
C.Norway
Limburg
W.Norway
Uppsala
Zeeland
Netherlands
Tartu
Friesland
Gotland
Novgorod
Denmark
Riga
Moscow
Blekinge
Groningen
Vilnius
S.Ireland
Rostock
Gdansk
Greifswald
London
Hamburg
Belorussia
Magdeburg
Warsaw
Berlin
Brussels
Munster
Leipzig
Chemnitz
Lublin
Dusseldorf
Paris
Cologne
Kiev
Mainz
Strasbourg
Stuttgart
Freiburg
Munich
Vienna
Graz
Budapest
Santiago de Compostela
Asturias
Lausanne
Bern
Tyrol
Ljubljana
Veneto
Romania
Cantabria
Lombardy
Zagreb
N.Portugal
Zaragoza
Liguria
Emilia-Romagna
Pyrenees
Tuscany
Marche
C.Portugal
Madrid
Umbria
Caceres
Barcelona
Bulgaria
S.Portugal
Latium
Valencia
Albania
Puglia
Andalusia
Madeira
Sicily
Anatolia
Athens

Figure 3. Geographical distributions of regional populations analysed with Y-STRs. A. East Asia; filled circles represent populations that are not significantly different from the CHB. B. Europe; filled circles represent populations that are not significantly different from the CEU.
doi:10.1371/journal.pone.0004684.g003

Analyses of the different datasets used all Y-STRs except *DYS385* for the HGDP-CEPH dataset, or the subsets in common, consisting of 10 Y-STRs for the East Asian (*DYS19*, *DYS389I*, *DYS389b*, *DYS390*, *DYS391*, *DYS392*, *DYS393*, *DYS437*, *DYS438*, *DYS439*) and seven Y-STRs for the European YHRD dataset (*DYS19*, *DYS389I*, *DYS389b*, *DYS390*, *DYS391*, *DYS392*, *DYS393*).

Statistical analyses

Population structure was investigated using the program STRUCTURE version 2.1 [20] with an admixture model. For each run, the number of clusters, K, needs to be specified in advance and values in the range 2–7 was used. Numbers of iterations in the burn-in period and MCMC replication were 4,000 and 6,000, respectively, for the runs of world-wide populations, and both 10,000 for runs of sub-regions. STRUCTURE output was processed with CLUMPP [31] and distruct (http://rosenberglab.bioinformatics.med.umich.edu/distruct.html). Cluster frequencies were compared between pairs of populations using a Mann-Whitney U test implemented in SPSS 16.0. Population pairwise genetic distances (Φ_{ST} values) were calculated from Y-STR haplotypes using the Arlequin package (http://lgb.unige.ch/arlequin/) and their significance was assessed from 1,000 bootstrap simulations, except for the European dataset where that these calculations did not reach completion and R_{ST} values were used. MDS analysis of population pairwise distances was carried out using SPSS 16.0. RSQ and stress values were: HGDP, 0.81 and 0.23; East Asia, 0.89 and 0.17; Europe, 0.95 and 0.13.

Electronic database information

1000 Genomes: http://www.1000genomes.org/page.php
Arlequin: http://lgb.unige.ch/arlequin/
Distruct: http://rosenberglab.bioinformatics.med.umich.edu/distruct.html
HapMap data: www.hapmap.org
HGDP-CEPH data: http://www.cephb.fr/hgdp-cephdb/
STRUCTURE: http://pritch.bsd.uchicago.edu/structure.html
YHRD: http://www.yhrd.org

Supporting Information

Table S1 HapMap YSTR haplotypes
Found at: doi:10.1371/journal.pone.0004684.s001 (0.04 MB XLS)

Table S2 Population pairwise comparisons
Found at: doi:10.1371/journal.pone.0004684.s002 (0.18 MB XLS)

Script S1 Perl script 1
Found at: doi:10.1371/journal.pone.0004684.s003 (0.00 MB TXT)

Script S2 Perl script 2
Found at: doi:10.1371/journal.pone.0004684.s004 (0.00 MB TXT)

Acknowledgments

We thank all the individuals who made this study possible by donating blood samples, Howard Cann for HGDP-CEPH DNAs, and Ni Huang for advice on Perl scripts.

Author Contributions

Conceived and designed the experiments: JG TZ PdK CTS YX. Performed the experiments: JG PdK. Analyzed the data: MH JG YX. Contributed reagents/materials/analysis tools: PdK. Wrote the paper: CTS YX.

References

1. The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796.
2. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
3. The International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
4. Mueller JC, Lohmussaar E, Magi R, Remm M, Bettecken T, et al. (2005) Linkage disequilibrium patterns and tagSNP transferability among European populations. Am J Hum Genet 76: 387–398.
5. McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: past, present and future. Hum Mol Genet 17: R100–101.
6. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–454.
7. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38: 75–81.
8. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853.
9. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4: e1000214.
10. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.
11. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
12. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340–345.
13. Crawford MH, ed (2007) Anthropological Genetics: Theory, Methods and Applications. Cambridge, UK: Cambridge University Press.
14. Deng L, Zhang Y, Kang J, Liu T, Zhao H, et al. (2008) An unusual haplotype structure on human chromosome 8p23 derived from the inversion polymorphism. Hum Mutat 29: 1209–1216.
15. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–262.
16. Willuweit S, Roewer L (2007) Y chromosome haplotype reference database (YHRD): update. FSI Genet 1: 83–87.
17. Xue Y, Zerjal T, Bao W, Zhu S, Shu Q, et al. (2006) Male demography in East Asia: a north-south contrast in human population expansion times. Genetics 172: 2431–2439.
18. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.
19. Jobling MA, Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet 4: 598–612.
20. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
21. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.
22. Jobling MA, Hurles ME, Tyler-Smith C (2004) Human Evolutionary Genetics. New York and Abingdon: Garland Science.
23. Powell GT, Yang H, Tyler-Smith C, Xue Y (2007) The population history of the Xibe in northern China: a comparison of autosomal, mtDNA and Y-chromosomal analyses of migration and gene flow. FSI Genetics 1: 115–119.
24. Roewer L, Croucher PJ, Willuweit S, Lu TT, Kayser M, et al. (2005) Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum Genet 116: 279–291.
25. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, et al. (2008) Japanese population structure, based on SNP genotypes from 7003

individuals compared to other ethnic groups: effects on population-based association studies. Am J Hum Genet 83: 445–456.
26. Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, et al. (2008) Correlation between genetic and geographic structure in Europe. Curr Biol 18: 1241–1248.
27. Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, et al. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 67: 1526–1543.
28. Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, et al. (2000) The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. Science 290: 1155–1159.
29. Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70: 841–847.
30. Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, et al. (2006) Development and validation of the AmpFℓSTR® Yfiler™ PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. J Forensic Sci 51: 64–75.
31. Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23: 1801.