
Oscar Lao,1,2 Kate van Duijn,1,2 Paula Kersbergen,2,3 Peter de Knijff,3 and Manfred Kayser1

1Department of Forensic Molecular Biology, Erasmus University Medical Centre Rotterdam, Rotterdam; 2Department of Biology, Netherlands Forensic Institute, The Hague; and 3Forensic Laboratory for DNA Research, Department of Human and Clinical Genetics, Leiden University Medical Centre, Leiden, The Netherlands

The identification of geographic population structure and genetic ancestry on the basis of a minimal set of genetic markers is desirable for a wide range of applications in medical and forensic sciences. However, the absence of sharp discontinuities in the neutral genetic diversity among human populations implies that, in practice, a large number of neutral markers will be required to identify the genetic ancestry of one individual. We showed that it is possible to reduce the amount of markers required for detecting continental population structure to only 10 single-nucleotide polymorphisms (SNPs), by applying a newly developed ascertainment algorithm to Affymetrix GeneChip Mapping 10K SNP array data that we obtained from samples of globally dispersed human individuals (the Y Chromosome Consortium panel). Furthermore, this set of SNPs was able to recover the genetic ancestry of individuals from all four continents represented in the original data set when applied to an independent, much larger, worldwide population data set (Centre d’Etude du Polymorphisme Humain–Human Genome Diversity Project Cell Line Panel). Finally, we provide evidence that the unusual patterns of genetic variation we observed at the respective genomic regions surrounding the five most informative SNPs is in agreement with local positive selection being the explanation for the striking SNP allele-frequency differences we found between continental groups of human populations.

Received November 7, 2005; accepted for publication January 12, 2006; electronically published February 14, 2006.

Address for correspondence and reprints: Dr. Manfred Kayser, Department of Forensic Molecular Biology, Erasmus University Medical Centre Rotterdam, Medical-Genecluster, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: m.kayser@erasmusmc.nl

Am. J. Hum. Genet. 2006;78:680–690. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2006/7804-0014$15.00
admixture throughout their history (Carvalho-Silva et al. 2001). Thus, it seems logical to use autosomal genetic markers in addition to sex-specific markers to correctly identify geographic population structure and genetic ancestry. Elsewhere, it has been shown that individuals from different geographic origins can be classified according to their continental region of sampling by use of the genetic information of several hundred autosomal microsatellites (Rosenberg et al. 2002, 2003) as well as autosomal Alu-insertion polymorphisms and microsatellites (Bamshad et al. 2003). Although the number of microsatellites can be reduced to ~40 when the statistical parameter $I_n$ (“informativeness of assignment” index) is applied to marker ascertainment (Rosenberg et al. 2003), their relatively high mutation rates (Kayser et al. 2000; Holtkemper et al. 2001; Xu and Fu 2004) keeps the number of markers relatively high. In principle, the number of genetic markers could be reduced by using SNPs that mutate ~100,000 times more slowly than do microsatellites (Thomson et al. 2000). Recent studies suggest that there is a considerably large number of autosomal SNPs showing a geographically restricted allele-frequency distribution (Hinds et al. 2005). However, only a small number of populations from a small number of geographic regions have been analyzed so far (HapMap 2003; Hinds et al. 2005).

In this study, we used global whole-genome SNP variation and a newly developed ascertainment algorithm to identify a minimal set of markers with maximal ability to detect geographic population structure and genetic ancestry. In principle, genetic markers with the largest genetic distances between populations—determined, for example, by applying the genetic distance $F_{ST}$ (Weir et al. 2005)—are the best candidates for population differentiation (Shriver et al. 2004). However, the redundancy of ancestry information between markers needs to be considered when aiming to minimize the number of genetic markers. Therefore, we developed a new method that is based on the informativeness of assignment index $I_n$ (Rosenberg et al. 2003) to find a set of markers that tends to maximize the genetic differentiation between populations while minimizing the number of markers needed. This statistic computes the amount of (nonredundant) assignment information that a particular locus or set of loci contains, to differentiate a particular set of groups defined a priori. Since $I_n$ computes the nonredundant amount of ancestry information, we thereby avoid the usually observed ascertainment bias toward markers that only differentiate between African and non-African groups, caused by the fact that genetic differences are usually largest between African and non-African populations (Hinds et al. 2005). This index ranges from 0 (when the frequency of all alleles of one locus are equally distributed between populations) to the natural logarithm of the number of considered populations (when the different alleles are able to unequivocally differentiate the populations). The informativeness of assignment index under the assumption of a population model without admixture ($I_n$) was computed because it has been shown that the $I_n$ statistic produces similar estimates and has similar properties as the informativeness of assignment index under the assumption of a population model with admixture ($I_n$) (Rosenberg et al. 2003). The $I_n$ statistic is preferred over the $I_n$ statistic for defining informative markers, because $I_n$ can produce denominators of 0 when two or more populations have the same allele frequencies, whereas $I_n$ cannot (Rosenberg et al. 2003). We have overcome the problem of extremely large computational efforts needed to consider all possible allele combinations for a large number of loci by applying a genetic algorithm (Haupt and Haupt 2004).

We analyzed >11,500 SNPs, using the Affymetrix 10K Array Xba 131, in 76 human individuals from 21 sampling localities representing six worldwide geographic areas: Africa, South Africa, America, Asia, North Asia, and Europe (Y Chromosome Consortium [YCC] panel). In short, 250 ng of DNA from each individual was digested, ligated, and amplified. PCR products were fragmented and biotin-labeled after pooling and purification. The biotin-labeled DNA fragments were hybridized to the probes on the Affymetrix GeneChip Mapping 10K array. Finally, the arrays were washed, stained, scanned, and analyzed. All procedures were done in accordance with the recommendations of Affymetrix (Sellick et al. 2003; Shriver et al. 2005). SNPs typed in <90% of the individuals or located on the X chromosome were re-

![Figure 1](image-url) Percentage of information explained when the number of markers that are ascertained from 8,491 SNPs by use of the genetic algorithm based on the informativeness of assignment index ($I_n$) is increased from 1 to 10, given four continental groups and the YCC panel (see main text for details). The 95% CI of each SNP combination was computed by resampling the same number of chromosomes from the populations and computing $I_n$ 1,000 times.
moved from the final data set, which resulted in usable genotypes for 8,491 SNPs per individual.

One might expect that the relatively small number of individuals we used per locality (on average, 3.5 individuals) would tend to decrease the observed genetic variance within populations and spuriously increase the amount of genetic variance explained between populations and continents. To test whether the sample size used here could spuriously increase the genetic differences between continents, we applied an analysis-of-molecular-variance approach (Excoffier et al. 1992), using the Arlequin 2.000 software (Schneider et al. 2000) and the 8,491 SNPs and grouping the populations into five continental regions: Europe, Africa, America, Asia, and Oceania. The amount of genetic variance explained within populations was 85.5% ($P < .0005$), which is within the range of the usually observed values of genetic variance within populations when neutral markers are used (Romualdi et al. 2002). This result seems to contradict expectations, but it can be explained by the fact that all SNPs used on the Affymetrix arrays were originally selected from The SNP Consortium repository (Matsuzaki et al. 2004), showing a similar degree of genetic variation based on a small set of population samples from different continents (Hao et al. 2004). In fact, this array has been successfully used for linkage mapping in different human populations (Kelsell et al. 2005). Thus, these SNPs do not represent the true underlying genetic variation of human populations, and it can be expected that the ascertainment bias would tend to increase the genetic variance within populations, compensating for the expected reduction of the genetic variance within populations when small sample sizes are used.

Each individual of the YCC panel could be genetically assigned—by use of all 8,491 SNPs and the STRUCTURE analysis—to one of the four groups considered. These four genetic groups correlate with four geographical regions: western Eurasia, East Asia, Africa, and America. Each group was then considered artificially as a population, and the $I_n$ statistic was computed per marker. Loci with $I_n < 10\%$ of the maximum value (ln 4) were excluded. We applied the genetic algorithm to the remaining set of 977 SNPs. Different runs were performed with an increasing number of SNPs, to quantify how the total amount of ancestry information changes with the number of markers included. The confidence intervals of $I_n$ were determined by resampling the same number of chromosomes in each population for each locus and computing $I_n$. As seen in figure 1, the extra informativeness generated by the addition of new SNPs to the set of achieved markers increased rapidly, with the first eight SNPs already reaching 90% of the maximum informativity value. With only 10 SNPs, 94.6% of the maximum informativity value was obtained (table 1 and fig. 1). Further increasing the number of loci hardly increased the amount of extra information contributed by the additional SNPs. This indicates that almost all the information of genetic ancestry an additional SNP can contain in this data set was already described using the 10 markers considered previously. When the STRUCTURE algorithm (Pritchard et al. 2000) is applied to the YCC genotypes of these 10 SNPs, all individuals become clearly assigned to the correct continental region of sampling (fig. 2).

We then wished to know whether the high ability of the ascertained set of 10 SNPs to identify continental population structure and genetic ancestry persists when they are applied to an independent set of population samples. Therefore, we genotyped these 10 SNPs in the CEPH–Human Genome Diversity Project Cell Line Panel (CEPH-HGDP) samples, using TaqMan technology (for details, see appendix A [online only]). The CEPH-HGDP comprises 1,064 samples from 51 human populations of global distribution, including all continental regions: America, Central and East Asia, Europe, the Middle East, North Africa, sub-Saharan Africa, and

### Table 1

<table>
<thead>
<tr>
<th>Marker</th>
<th>Chromosome</th>
<th>Gene Name</th>
<th>$I_n$ for Four Groups and YCC Panel (%)</th>
<th>$I_n$ for Seven Groups and CEPH-HGDP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs722869</td>
<td>14</td>
<td>VRK1</td>
<td>29.066</td>
<td>7.960</td>
</tr>
<tr>
<td>rs1858465</td>
<td>17</td>
<td>...</td>
<td>25.637</td>
<td>9.228</td>
</tr>
<tr>
<td>rs1876482</td>
<td>2</td>
<td>LOC442008</td>
<td>24.589</td>
<td>10.290</td>
</tr>
<tr>
<td>rs1344870</td>
<td>3</td>
<td>...</td>
<td>22.810</td>
<td>11.074</td>
</tr>
<tr>
<td>rs1363448</td>
<td>5</td>
<td>PCDHGB1</td>
<td>19.418</td>
<td>4.352</td>
</tr>
<tr>
<td>rs952718</td>
<td>2</td>
<td>ABCA12</td>
<td>18.739</td>
<td>9.472</td>
</tr>
<tr>
<td>rs2352476</td>
<td>7</td>
<td>...</td>
<td>18.317</td>
<td>5.603</td>
</tr>
<tr>
<td>rs714857</td>
<td>11</td>
<td>...</td>
<td>18.083</td>
<td>6.157</td>
</tr>
<tr>
<td>rs1823718</td>
<td>15</td>
<td>...</td>
<td>17.845</td>
<td>5.451</td>
</tr>
<tr>
<td>rs735612</td>
<td>15</td>
<td>RYR3</td>
<td>14.315</td>
<td>5.530</td>
</tr>
</tbody>
</table>
The American Journal of Human Genetics

Figure 2  STRUCTURE analysis of the YCC samples, with K = 2, 3, or 4 groups, performed using genotypes of the 10 most informative SNPs ascertained using the genetic algorithm with the total YCC data. STRUCTURE analyses were computed using a model without admixture (A) and a model with admixture (B). Each analysis was repeated five times, after a Markov chain–Monte Carlo (MCMC) burning period of 50,000 and considering the next 200,000 MCMC iterations. In all five runs, good mixing was observed, and similar results were found in accordance with the model used. The natural logarithm of the estimated probability of the data (ln p) is as follows. In panel A, for K = 2, ln p = −762.2; for K = 3, ln p = −629.2; and, for K = 4, ln p = −557.4. In panel B, for K = 2, ln p = −764.9; for K = 3, ln p = −631.2; and, for K = 4, ln p = −559.5.

Figure 3  MDS plot based on the I_n matrix computed between pairs of populations by use of the genotypes of the 10 most informative SNPs in the 51 population samples from CEPH-HGDP. Four clusters of population can be identified: (i) sub-Saharan African populations, (ii) American populations, (iii) Eastern Asian and Oceanian populations, and (iv) European, Middle Eastern, North African, and Central/South Asian populations.
Figure 4  STRUCTURE analysis of the CEPH-HGDP samples, with $K = 2$, 3, 4, or 5 groups, performed using genotypes of the 10 most informative SNPs ascertained using the genetic algorithm with the total YCC data. Two different STRUCTURE analyses were computed: a population model without admixture (A) and a population model with admixture (B). Each analysis was repeated five times after an MCMC burning period of 100,000 and considering the next 10,000 MCMC iterations. In all five runs, good mixing was observed, and similar results were found in accordance with the model used. The $\ln p$, assuming $K$ groups, is as follows. In panel A, for $K = 2$, $\ln p = -11,801.2$; for $K = 3$, $\ln p = -10,977.3$; for $K = 4$, $\ln p = -10,279.2$; and, for $K = 5$, $\ln p = -10,324.9$. In panel B, for $K = 2$, $\ln p = -11,886.2$; for $K = 3$, $\ln p = -11,070.6$; for $K = 4$, $\ln p = -10,345.5$; and, for $K = 5$, $\ln p = -10,436.9$. Cen. Af. Rep. = Central African Republic; S. Afr. = South Africa.
Clustering results using STRUCTURE software separately for each of the four groups detected by previous STRUCTURE analysis of the worldwide HGDP-CEPH samples by use of the 10 most ancestry-informative SNPs. The legend is available in its entirety in the online edition of The American Journal of Human Genetics.

Given that different clustering algorithms can produce different results (Corander et al. 2004), we applied three different ways of detecting clusters of populations in the CEPH panel. In the first approach, the $I_n$ statistic was computed for each pair of populations on the basis of the genotypes of the 10 ascertained SNPs, and the matrix was plotted by means of multidimensional scaling (MDS) (Kruskal and Wish 1978) with the STATISTICA 6.0 software (StatSoft 2001). Although the $I_n$ statistic is an index of the informativeness of markers for ancestral inference, it correlates with classical measures of genetic distances (such as $F_{ST}$) when computed between pairs of populations; thus, it can be considered a measure of genetic distance in this special case (Rosenberg et al. 2003). We were able to detect four differentiated clusters of populations in the graphical representation of the MDS (fig. 3): (i) all sub-Saharan African populations, (ii) all American populations, (iii) all East Asian and Oceanian populations and two Central/South Asian populations (the Uygur from China and the Hazara from Pakistan), and (iv) all European and Middle Eastern populations and all other Central/South Asian populations not in cluster iii, with the North African population somewhat separated but in close proximity. A permutation test was performed to assess the statistical significance of the clustering suggested by the MDS analysis. First, each population was assigned at random to one of the four clusters; then the $I_n$ value, given this new clustering, was computed; and the process was repeated 1,000 times. The resulting $P$ value was highly statistically significant ($P < .0005$), thus supporting the observed clustering. We repeated the permutation test for clusters iii and iv, which appear somewhat close to each other in the MDS plot. The $P$ value was also highly statistically significant ($P < .0005$), suggesting a clear differentiation between both clusters of populations, despite the presence of the two somewhat-intermediate populations of Uygur and Hazara.

The second approach was performed using the individual-based STRUCTURE algorithm, with an increasing number of groups, from $K = 2$ to $K = 5$. STRUCTURE analyses were repeated twice, once using a population model without admixture and once using a population model with admixture; results for each configuration can be seen in figure 4. When the number of groups was two, the algorithm clustered the African, North African, European, Middle Eastern, and Central/South Asian individuals separately from the East Asian, Oceanian, and American individuals, under both models (fig. 4). When three groups ($K = 3$) were specified, the STRUCTURE algorithm yielded the following clusters: (i) African and American individuals; (ii) Oceanian and East Asian individuals, together with the Hazara from Pakistan and the Uygur from China (although the latter two show somewhat more features of cluster iii, yet still belong to cluster ii); and (iii) European, Middle Eastern, and Central/South Asian individuals and North African individuals from Algeria (although the latter show some features of cluster i). When the number of selected groups was four, STRUCTURE clustered the individuals as follows: (i) sub-Saharan African individuals; (ii) American individuals; (iii) Oceanian and East Asian individuals, together with the Hazara and the Uygur (although the latter two show somewhat more features of cluster iv, yet still belong to cluster iii); and (iv) European, Middle Eastern, Central/South Asian, and North African individuals (although the latter show some features of cluster i). Increasing the number of structure groups beyond $K = 4$ did not increase the number of population groups identified. To test for putative presence of population substructure in each of the four genetically defined groups, STRUCTURE analyses were repeated, with one group being considered at a time. No population substructure was detected in the European, Middle Eastern, and Central/South Asian group, the Oceanian and East Asian group, or the African group, when two or more clusters were applied (fig. 5). Only
should be noted that, because the marker selection is
Oceania is detected using the BAPS 3.2 approach). It
in the original YCC data set, such as Oceania (although
ulations that were underrepresented or not considered
 genetic ancestry of HGDP samples from those pop-
 (Turakulov and Easteal 2003; Yang 
used in previous studies to identify a similar geographic
ments by 377 microsatellites (Rosenberg et al. 2002) or with
erved by others previously in HGDP samples with the
all the geographic population structure that was ob-
 were considered (see fig. 6), and the re-
sults for four groups were consistent with MDS and
Almost identical results were observed for this method
for STRUCTURE when two, three, and four groups
Overall, we obtained identical and statistically signifi-
cant clustering results for the population samples when
which suggests consistency in the ability of the 10 as-
certained SNPs to identify population (sub)structure and
genetic ancestry in accordance with four continental
groups.
All these results suggest that it is possible to substan-
tially reduce the amount of markers needed to recover
a particular population structure when carefully ascer-
tained SNPs are used. Applying to the HGDP samples
the 10 most informative SNPs that we ascertained by
means of the genetic algorithm from a set of ∼8,500
SNPs in the YCC panel, we were able to recover almost
all the geographic population structure that was ob-
served by others previously in HGDP samples with the
use of 377 microsatellites (Rosenberg et al. 2002) or with
a subset of the 40 most informative microsatellites (Ro-
osenberg et al. 2003). Thus, the same results were ob-
tained using four times fewer SNPs than microsatellites.
Furthermore, our results demonstrate a reduction in the
number of markers by >10 times the amount of SNPs
used in previous studies to identify a similar geographic
population structure (Turakulov and Eastal 2003; Yang
et al. 2005). Our set of 10 SNPs infers incorrectly only
the genetic ancestry of HGDP samples from those pop-
ulations that were underrepresented or not considered
in the original YCC data set, such as Oceania (although
Oceania is detected using the BAPS 3.2 approach). It
should be noted that, because the marker selection is
based on maximizing the genetic differences between the
populations, the capacity to correctly infer the ancestry
of an individual from a new, previously unanalyzed pop-
ulation, will depend on the degree of genetic variation
shared with the original set of populations. In the case
of human populations, this is of particular relevance
when geographical groups are being considered, because
isolation by distance has played a major role in shaping
human genetic diversity (Harpending and Rogers 2000;
Barbujani and Goldstein 2004).
Since many of the environmental factors tend to be
geographically restricted, genetic markers associated
with a local positively selected genomic region are ex-
pected to show large differences between populations
from different geographic regions because of different
genetic adaptation processes in response to different
vironmental factors (Bamshad and Wooding 2003; Kay-
er et al. 2003; Bamshad et al. 2004). Consequently,
these markers will be informative not only to understand
human genetic adaptation toward environmental factors
but also to infer the genetic ancestry of an individual
and detect geographic population structure. We wished
to know whether the strong continental differences in
allele frequencies we observed at the ascertained SNPs
can be explained by local positive selection in the
different geographic areas or whether these patterns need
to be explained by stochastic processes such as genetic
drift. Therefore, we tested for signatures of positive
selection the genomic regions of the five most ances-
try-informative SNPs from our analysis: rs1344870,
rs1876482, rs952718, rs1858465, and rs722869. Three
of the SNPs (rs1876482, rs952718, and rs722869) fall
within known or predicted genes (table 1). The presence
of footprints of positive selection was tested by analyzing
the genetic diversity of the surrounding region and by
studying how the homozygosity of the haplotypes in
such regions decays when the distance to the putatively
selected region is increased (Sabeti et al. 2002). The SNP
frequencies of the surrounding regions were taken from
the Perlegen database (Hinds et al. 2005), which includes
a large number of SNP genotypes distributed throughout
the whole genome in populations from three of the four
continents considered in our original YCC panel data
set: Africa, Europe, and Asia (although, for the Perlegen
data set, Africans, Europeans, and Asians who reside in

The figure is available in its entirety in the online
North America were used). We computed $I_n$ for each SNP of the surrounding region, considering these three continental groups, and performed a sliding-window approach, considering a window size of 2 kb and computing the mean $I_n$ value of each window. To see whether the mean of the window was significantly different from other regions of the genome, we compared the obtained values with the mean values from windows of the same size and with the same number of markers in 10,145 genes from the entire genome in the Perlegen database (genes on the X chromosome were excluded because of the different effective population size for this chromosome (Schaffner 2004). The haplotype phases were inferred using the PHASE program (Stephens and Donnelly 2003). The extended haplotype analysis was performed using the SWEEP program (kindly provided by P. Sabeti), and the markers defining the core haplotype were ascertained from the window with the largest $I_n$ mean.

In four of the five genomic regions analyzed, we detected unusual patterns of genetic variation that are compatible with the hypothesis of local positive selection—namely, (1) large regions of closely located markers showing high informativeness-of-ancestry scores, statistically significantly higher than the expected scores from other genes from the genome, and (2) large, extended haplotypes in at least one of the three populations analyzed in the haplotype bifurcation analysis as well as in the extended haplotype homozygosity (EHH) analysis (see figs. 7–10). Only in the case of rs1344870 have we not found patterns that are suggestive of local positive selection on the basis of the three populations included in the Perlegen data set. For that region, only a few sliding windows are statistically significant ($P < .05$), there are similar frequencies in the core haplotypes between the three populations, and there is similar decay in the haplotype homozygosity of the different haplotypes in the three populations (see fig. 10). However, the hypothesis of positive selection cannot be rejected, because more continental regions were considered in the YCC samples used for SNP ascertainment than in the Perlegen samples used for analyzing the surrounding regions (i.e., Native Americans were not considered in the Perlegen sample set but were considered in YCC sample set). One of the regions (surrounding marker rs1876482) for which we found unusual patterns of genetic variation in agreement with evidence of local positive selection includes the predicted gene LOC442008 (NCBI GeneID 442008), which shows a highly frequent (frequency 73%) and long extended core haplotype that is practically absent outside the Asian population (fig. 11). Thus, although the SNPs from the whole-genome analyses used to identify ancestry-informative markers were noncoding, our data indicate that the significant population differences of the markers with maximum informativeness of ancestry seem to be shaped by positive selection rather than by genetic drift.

In summary, we have shown that it is possible to substantially reduce the number of markers needed to identify geographic population structure on a continental level by applying carefully ascertained and validated SNPs. With 10 SNPs, we obtained a level of geographic population structure similar to that previously identified using 377 or 40 microsatellites in the same set of samples (Rosenberg et al. 2003), which suggests that carefully ascertained SNPs are more suitable than microsatellites for detecting geographic population structure and identifying genetic ancestry. Furthermore, we can show here that the frequency distributions of SNPs with maximal ability to detect continental population structure and genetic ancestry are most likely shaped by local positive selection rather than by genetic drift. However, our results also show that there is a considerable lack of power when applying ancestry-informative markers ascertained from the original data set to another set of population samples, and the portability of ancestry-informative SNPs depends on the relationship between the populations used. Further studies using more SNPs and/or more population samples from many geographic regions and localities in the world are needed to test whether an ultimate set of SNP markers can be found to identify geographic population structure and genetic ancestry on a more detailed level than the continental level that we achieved here.

Finally, we emphasize that the methodology we have developed here for minimizing the number of genetic markers that are necessary for maximizing the genetic differences between clusters of populations can be applied to any kind of population grouping. In this study, we were interested in finding markers that differentiate...
Figure 11  Sliding-window and haplotype analyses performed on the genomic region that includes SNP rs1876482 (1 of the 10 most informative SNPs identified), which is located in the LOC442008 gene, by use of Perlegen data. A, Sliding-window plot of the mean value observed for each window (the gene is represented by a black bar). B, Associated P value for comparison with an empirical distribution based on >10,000 genes (see main text). The P = .05 cutoff is represented by a black line. C, Bifurcation plots of the main core haplotypes in the three populations considered. D, Extended homozygosity versus genomic distance to the core haplotype. The region of the core haplotype was selected on the basis of the largest region that was statistically significant in the sliding-window analysis (from rs12619554 to rs4832712; see main text for details). Note the high frequency of the third haplotype in the case of Asian populations and the slow decay of the EHH of that haplotype compared with the other haplotypes both within and between populations.

d geographic groups of populations, and we succeeded on the continental level. Clearly, some events in human population history and some forces of local positive selection are expected to enhance the success of finding suitable markers for population differentiation on the continental level. In fact, we have shown here that the surrounding regions of the most ancestry-informative SNP mark-
associated markers that are able to differentiate respective (noncontinental) populations can be found when appropriate samples are used for marker ascertainment. In addition, when neutral markers are applied, the presence of statistically significant genetic differences between populations (Romualdi et al. 2002) allows other sets of informative markers that specify other population clusters to be found, especially if the number of markers is large enough, no matter what the biological sense of the clusters is. Therefore, making inferences from data sets of genetic markers obtained by the procedures used here should be done with care and needs to be conditioned to the biological meaning of the clustering obtained. However, the fact that we were able to identify four continental groups in the CEPH-HGDP samples by using three different clustering algorithms and applying SNP markers that were ascertained in a different set of global populations (YCC panel), for which the SNPs identified the same continental groups, in addition to the obtained evidence of local (continental) positive selection at the respective genomic regions, clearly emphasizes the value of the identified markers in recognizing continental population (sub)structure and continental genetic ancestry.

Acknowledgments

We thank Pardis Sabeti for kindly providing the SWEEP program prior to publication. The Netherlands Forensic Institute and the Erasmus University Medical Center Rotterdam are acknowledged for financial support.

Web Resource

The URL for data presented herein is as follows:


References


Schneider S, Roessli D, Excoffier L (2000) Arlequin, version 2.000: a software for population genetics data analysis. Genetics and Biometry Laboratory, Switzerland
StarSoft (2001) STATISTICA (data analysis software system), release 6.0, Tulsa