HUMAN MUTATION 29(10), 1209–1216, 2008

RESEARCH ARTICLE

An Unusual Haplotype Structure on Human Chromosome 8p23 Derived From the Inversion Polymorphism

Libin Deng,[1,2] Yuezheng Zhang,[1,2] Jian Kang,[3] Tao Liu,[2,4] Hongbin Zhao,[1,2] Yang Gao,[1,2] Chaohua Li,[1] Hao Pan,[1] Xiaoli Tang,[1] Dunmei Wang,[1] Tianhua Niu,[5] Huanming Yang,[1] and Changqing Zeng[1]*

[1]Beijing Institute of Genomics, Chinese Academy of Sciences, P.R. China; [2]Graduate School of the Chinese Academy of Sciences, Beijing, P.R. China; [3]Department of Mathematical Sciences, Tsinghua University, Beijing, P.R. China; [4]Institute of Biophysics, Chinese Academy of Sciences, Beijing, P.R. China; [5]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts

Communicated by David N. Cooper

Chromosomal inversion is an important type of genomic variations involved in both evolution and disease pathogenesis. Here, we describe the refined genetic structure of a 3.8-Mb inversion polymorphism at chromosome 8p23. Using HapMap data of 1,073 SNPs generated from 209 unrelated samples from CEPH—Utah residents with ancestry from northern and western Europe (CEU); Yoruba in Ibadan, Nigeria (YRI); and Asian (ASN) samples, which were comprised of Han Chinese from Beijing, China (CHB) and Japanese from Tokyo, Japan (JPT)—we successfully deduced the inversion orientations of all their 418 haplotypes. In particular, distinct haplotype subgroups were identified based on principal component analysis (PCA). Such genetic substructures were consistent with clustering patterns based on neighbor-joining tree reconstruction, which revealed a total of four haplotype clades across all samples. Metaphase fluorescence in situ hybridization (FISH) in a subset of 10 HapMap samples verified their inversion orientations predicted by PCA or phylogenetic tree reconstruction. Positioning of the outgroup haplotype within one of YRI clades suggested that Human NCBI Build 36-inverted order is most likely the ancestral orientation. Furthermore, the population differentiation test and the relative extended haplotype homozygosity (REHH) analysis in this region discovered multiple selection signals, also in a population-specific manner. A positive selection signal was detected at *XKR6* in the ASN population. These results revealed the correlation of inversion polymorphisms to population-specific genetic structures, and various selection patterns as possible mechanisms for the maintenance of a large chromosomal rearrangement at 8p23 region during evolution. In addition, our study also showed that haplotype-based clustering methods, such as PCA, can be applied in scanning for cryptic inversion polymorphisms at a genome-wide scale. Hum Mutat 29(10), 1209–1216, 2008.

KEY WORDS: chromosome 8p23; inversion polymorphism; haplotype; natural selection; SNPs

INTRODUCTION

Chromosome 8p23 in the human genome is an intriguing region because of its unusual genomic structure. Giglio et al. [2001] first reported an inversion segment flanked by two low-copy repeats (LCRs) at this region. The length of the inversion segment without LCRs was estimated to be 3.8 Mb, which made 8p23 one of the largest inversions identified in autosomes [Sugawara et al., 2003]. Although this structural variation has been detected across various ethnic groups [Giglio et al., 2001; Sugawara et al., 2003], its frequency distributions among populations and the evolutionary history of this region remain enigmatic.

The completion of the International HapMap Project has spurred tremendous achievements in our understanding of structural variations, including deletion, insertion, copy number variation, and inversion [International HapMap Consortium, 2005; Feuk et al., 2006]. For instance, a scanning of large inversion polymorphisms was recently reported based on unusual linkage disequilibrium (LD) patterns [Bansal et al., 2007]. Despite its failure to identify 8p23 inversion and its limitation of false-positive signals due to the heterogeneity of recombination rate, this first genome-scale survey provided candidates for inversion polymorphism in the human genome.

The Supplementary Material referred to in this article can be accessed at http://www.interscience.wiley.com/jpages/1059-7794/suppmat.

Received 10 October 2007; accepted revised manuscript 6 February 2008.

*Correspondence to: Changqing Zeng, Beijing Institute of Genomics, Airport Industrial Zone, B-6, Beijing, 101300, China. E-mail: czeng@genomics.org.cn

Grant sponsor: Chinese Academy of Sciences (Century Program); Grant sponsor: National Natural Science Foundation of China; Grant number: 30225017; Grant sponsor: Ministry of Science and Technology; Grant number: 2002BA711A09.

DOI 10.1002/humu.20775

Published online 12 May 2008 in Wiley InterScience (www.interscience.wiley.com).

Genetic studies in *Drosophila* have demonstrated great potentials in using haplotype configurations to reveal the genetic features in segments with chromosomal rearrangements. Given the considerable suppression of recombination in heterozygous status at inversion regions, mutations arising independently would contribute to the formation of haplotype subgroups (i.e., "clades") [Pritchard and Przeworski, 2001]. Surveys of the nucleotide variations within such regions have uncovered significant genetic differentiations among haplotypes formed by gene arrangements in fruit flies [Munte et al., 2005; Rozas et al., 2001]. Similarly, Stefansson et al. [2005] recently estimated the frequency of an inversion at human chromosome 17q21.31 in Europeans based on regional haplotype groups. However, the relationship between SNPs and inversion polymorphisms has seldom been addressed.

In this study, we unveiled notable haplotype subgroups corresponding to different alleles of the 8p23 inversion polymorphism in HapMap populations, and experimentally verified such a correlation by metaphase fluorescence in situ hybridization (FISH). Further phylogenetic analysis with an outgroup haplotype revealed the ancestral orientation of this structural rearrangement. Our results of haplotype analysis also suggested that various selection patterns might provide possible explanations for the maintenance of this inversion polymorphism in a population-specific manner.

MATERIALS AND METHODS

Genotype Data of Different Populations

Genotype data of 209 unrelated HapMap individuals, including 60 parents of the 30 trio samples from CEPH in Utah residents with ancestry from northern and western Europe (CEU), 60 parents of the 30 trio samples from Yoruba in Ibadan, Nigeria (YRI), 45 unrelated Han Chinese from Beijing, China (CHB), and 44 Japanese from Tokyo, Japanese (JPT) individuals, were used in all analysis. Due to the significant genetic similarity between Chinese and Japanese groups, we pooled CHB and JPT data and denoted these individuals as Asian (ASN) as seen in other studies [Voight et al., 2006].

HapMap phase I data were downloaded from the website (HapMap Phase I/rel♯16c data files; www.hapmap.org). Among over one million SNPs passed HapMap QC criteria, and 629,958 autosomal markers that were segregating in all three HapMap samples were applied for analysis. Among these, 48,863 are in the entire chromosome 8 and 1,073 SNPs are located at the 8p23 region. SNP data of chromosome 8p were generated by BeadStation 500 (Illumina, San Diego, CA) at Beijing Institute of Genomics for the International HapMap Project.

Population Genetics Analysis

Heterozygosity (*HET*) analysis was applied to measure the genetic diversity at each SNP site. Genetic differentiation between populations was evaluated by pairwise F_{ST} on each SNP. *HET* and F_{ST} values were calculated using the Arlequin package (version 2.0; http://lgb.unige.ch/arlequin) [Schneider et al., 2000]. The regional genetic patterns were represented by averaged *HET* and F_{ST} values.

To identify high-F_{ST} outliers on chromosome 8, we set F_{ST} thresholds of 0.61, 0.41, and 0.56 to compare pairs of ASN and YRI (ASN_YRI), ASN and CEU (ASN_CEU), and CEU and YRI (CEU_YRI), respectively. Based on the F_{ST} distribution along the entire chromosome, these thresholds corresponded to an empirical type I error (α) = 0.025 in each comparison.

Haplotype Analysis

Haplotype phase estimation was performed by using PHASE v2.1.1 (www.stats.ox.ac.uk/mathgen/software) [Stephens and Donnelly, 2003]. In each region analyzed, 418 distinctive long-range haplotypes (LRHs): 178 from ASN and 120 each from CEU and YRI, were obtained.

To construct an outgroup haplotype of 8p23 in the phylogenetic tree, FASTA sequences of SNPs (dbSNP, NCBI) were aligned with the draft build of chimpanzee genome using BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat) [Kent, 2002]. For each SNP, we selected the best alignments and to consider the allele of chimp as the ancestral state as described [Voight et al., 2006]. Then we built the outgroup haplotype based on the ancestral status of these SNPs within the inversion region.

For haplotype clustering, principal component analysis (PCA) was carried out and the neighbor-joining (NJ) tree algorithm was used. PCA based on haplotype data was performed using Matlab 7.0 (The Mathworks, Natick, MA). Haplotype genetic distances were estimated using the allele sharing distance (ASD) method, and the haplotype tree was then constructed by means of the NJ algorithm implemented in MEGA version 3.0 [Kumar et al., 2001].

Identification of Candidate Genes Subjected to Selection

All SNPs of chromosome 8 were mapped to genic regions by searching the NCBI Entrez Gene database (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = gene) for population differentiation. SNPs were further analyzed if they were located in a gene region, including the 5′ upstream, 5′ untranslated region (UTR), exonic, intronic, 3′ UTR, or 3′ downstream regions. If a gene contained at least one high-F_{ST} outlier in any population pair under comparison, it was considered as a gene under selection [Akey et al., 2002].

The LRH test was applied to scan the "core haplotype" [Sabeti et al., 2002]. The scanned "core" was determined as the haplotype block defined by Gabriel et al. [2002]. For all the possible "core haplotypes" (minor haplotype frequency >0.05), the value of the relative extended haplotype homozygosity (REHH) was calculated at about 0.25 centiMorgans (cM) away from the core on both sides. All REHH values were then placed into 20 bins based on the haplotype frequency (ranges of 0–5%, 5–10%, …, 95–100%). The P value for each REHH was estimated by comparing its value with all REHH in that bin. A "core haplotype" was considered as a candidate if its REHH value exceeded the 99.9th percentile (corresponding to an empirical $\alpha = 0.001$). If a gene contained at least one candidate "core haplotype" in any population, it was considered as a gene under selection.

Fluorescence In Situ Hybridization (FISH)

HapMap cell lines were obtained from Coriell Repositories (Camdem, NJ). Human BAC clones (RP11-399J23 and RP11-589N15) were obtained from the BACPAC Resource Center at the Children's Hospital Oakland Research Institute (Oakland, CA). Probe and slide preparation, DNA hybridization, and image analysis were performed as described [Sugawara et al., 2003].

RESULTS

Population-Specific Haplotype Substructures at 8p23

To disclose the haplotype substructures of the ASN (i.e., CHB+JPT), CEU, and YRI populations, we carried out haplotype clustering by considering this 3.8-Mb inversion segment as a specific polymorphism frame. In total, 1,073 SNPs in the 8p23

region were utilized and 418 haplotypes were phased for all the 209 unrelated HapMap individuals. The haplotype substructure was then examined by PCA. In contrast to ASN haplotypes, which were clustered as one group with the exception for only two outliers, two distinct subgroups were revealed in the CEU population (Fig. 1A, red). The 51 CEU haplotypes in one subgroup were actually mingled with ASN ones (Fig. 1A, green). Two discernible haplotype subgroups were also found in the YRI population (Fig. 1A, blue), although the internal genetic differentiation appeared not as extensive as seen in the CEU population.

When focusing on the score of the first principal component (PC), which explained 21.1% of the total variation, a well-separated bimodal distribution was revealed in the European population (Fig. 1B). The standard deviation (SD) of this first PC score in the CEU haplotypes was 3.32, nearly three times larger than the corresponding values of ASN (0.952) and YRI (1.26). Interestingly, despite the dramatic diversity within the CEU samples, haplotype diversities between populations were not that significant. Two CEU haplotype subgroups overlapped with the ASN and YRI groups, respectively. Mann-Whitney's *U* (MWU) test on the first PC suggested that the average score of CEU haplotypes was not significantly different from that of YRI ($P = 0.284$).

To compare the haplotype configuration at 8p23 with that in the rest part of the human genome, we carried out bootstrap tests based on 1,000 3.8-Mb regions randomly selected across all autosomes. Each region contained more than 400 SNPs in the simulation. Then we performed PCA to infer haplotype substructures in these regions. In most bootstrap samples, the YRI and non-YRI haplotypes were well separated by the first PC score. At the $\alpha = 0.01$ level in the MWU test, a significant difference between scores of CEU and YRI was shown in 95.7% bootstrap samples. Only 2% bootstrap samples were found with P values larger than that of the MWU test between CEU and YRI haplotypes as seen in 8p23 (0.284; Supplementary Table S1, available online at http://www.interscience.wiley.com/jpages/1059-7794/suppmat), suggesting an uncommon colocalization of CEU and YRI haplotypes at 8p23 in the human genome. Moreover, results of the bootstrap test also confirmed a large SD in the first PC score of the CEU haplotypes. To preclude any bias caused by heterogeneity across different genomic regions, the ratios of SD values of CEU haplotypes to that of other populations were used as the metrics. For both comparisons, the bootstrap resulted in $P = 0.001$ (Supplementary Table S1). Taken together, in comparison with other genomic regions, a distinctive evolution history of 8p23 is suggested.

Haplotype Groups Correlating to the Chromosomal Rearrangement

To further categorize the haplotypes of each population, we set up a demarcation line at the first PC score of 1.4 (Fig. 1B). We then designated five haplotype subgroups as CEU_H, CEU_L, YRI_H, YRI_L, and ASN_L according to their scores of high (H, first PC score ≥ 1.4) or low (L, first PC score < 1.4). Based on this categorization, an individual's diplotype at the 8p23 region could be either H/H, L/L, or H/L. For instance, the European population shows the largest heterozygosity, with 17 homozygotes of CEU_H and eight homozygotes of CEU_L haplotypes, and 35 heterozygotes. The YRI population contains 34 homozygotes of YRI_H, three homozygotes of YRI_L haplotypes, and 23 heterozygotes. Interestingly, ASN individuals are almost all ASN_L (87 homozygotes vs. two heterozygotes). Such a distinctive grouping of haplotypes in this region indicates that the population-specific genetic substructures were most likely caused by the inversion polymorphism.

To experimentally determine the correlation between haplotype structure and chromosomal inversion status, and to particularly verify our haplotype grouping results from PCA, we then carried out FISH to uncover the orientation status of the 8p23 region in 10 HapMap cell lines. BAC clones RP11-399J23 and RP11-589N15, localized at each end of the inversion segment in ~8 Mb and ~11 Mb of chromosome 8 [Sugawara et al., 2003], were chosen to make FISH probes. Based on the reference order of Human NCBI Build 36, we defined the telomere-to-centromere orientation of RP11-399J23 (Green) and RP11-589N15 (Red) as "Build 36-noninverted" and the reverse orientation as "Build 36-inverted" (Fig. 2). As shown in Table 1, FISH results indicated that each of the haplotype groups (H and L) corresponded to the experimentally determined chromosomal orientations perfectly. Three homozygotes of CEU_H (GM06985, GM06994, and GM07055), with their FISH patterns shown in Fig. 2A, were in the "Build 36-inverted" orientation, whereas two homozygotes of CEU_L, GM12815 and GM12249, were in the "Build 36-noninverted" orientation (Fig. 2B). Indeed, two heterozygotes of GM12156 and GM06993 were found with both the "Build 36-inverted" and the "Build 36-noninverted" orientations (Fig. 2C). Similarly, for a homozygous individual of YRI_H (GM18502), a heterozygous individual of YRI_H and YRI_L (GM18517), and a heterozygous ASN individual (GM18620), the predicted orientations by haplotype classification perfectly matched with their FISH results (Table 1).

Taken together, we found that the haplotypes with high and low first PC scores corresponded exactly to the inversion status in the 8p23 region. Moreover, assuming that the inversion orientations can be predictive by haplotype clustering analysis, we are able to estimate the frequencies of Build 36-inverted 8p23 as ~60% (69/120) in CEU and ~76% (91/120) in YRI populations. In the same region, almost all ASN chromosomes are in the Build 36-noninverted orientation.

Evolution of the Structural Variation at 8p23

The correlations of haplotype subgroups with the inversion orientations allowed us to uncover the evolution history of this structural polymorphism by "haplotype tree". We clustered all haplotypes with an outgroup haplotype reconstructed from the corresponding DNA sequence of the chimpanzee (*Pan troglodytes*) Genome Project [Chimpanzee Sequencing and Analysis Consortium, 2005]. In an attempt to obtain a complete sequence comparison, we utilized a subset of 1,041 genotypes out of the 1,073 HapMap SNPs at 8p23. In total, four haplotype clades were resolved in the ASD tree (Fig. 3), demonstrating a nearly perfect match with the results of PCA. The only bias is that an ASN haplotype, which was grouped as one of two high-score outliers in PCA (Fig. 1A), was localized between the ASN clade and a subgroup of YRI haplotypes in the ASD tree (Fig. 3).

The outgroup haplotype was clustered together with one clade that encompassed all YRI_H haplotypes exclusively, indicating that this YRI subgroup ("Build 36-inverted" status) most likely represents the ancient type (i.e., "founder") (Fig. 3). Thus, we denoted this group as Clade_1. Evidently, the clade containing 69 CEU_H haplotypes and one ASN outlier had the shortest genetic distances and the same orientation with the founder Clade_1. We therefore named it as Clade_2, and the remaining two clades,

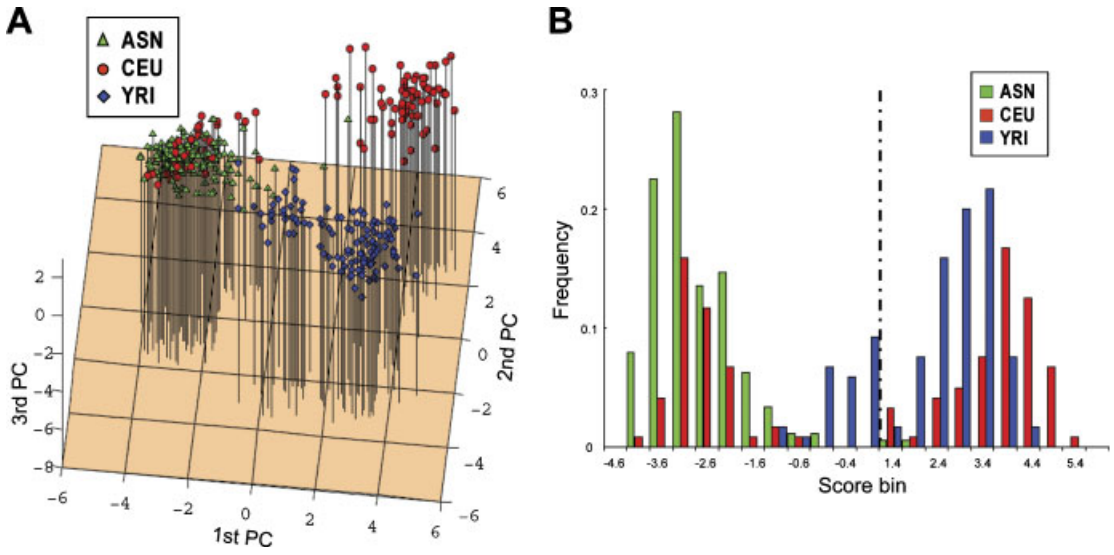

FIGURE 1. **Haplotype clustering in the 8p23 region by PCA. A:** Plot of the first three PCs of 178 ASN (green), 120 CEU (red), and 120 YRI (blue) haplotypes generated from 209 HapMap samples. Each point represents a haplotype. The line below each sample symbol shows the measurement of each haplotype on the third PC. **B:** The distribution of the first PC scores of three geographic populations. Dashed line indicates the demarcation point (first PC score = 1.4) that classifies all haplotypes into high- and low-scoring groups.

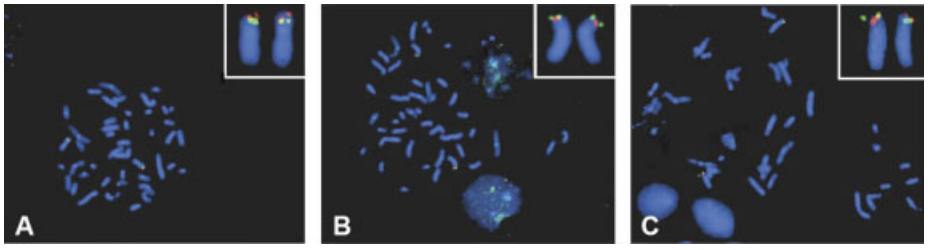

FIGURE 2. **Metaphase FISH resulting from three HapMap individuals with each of the three possible genotypes inferred by PCA. DNA probes were made from BAC clones RP11-399J23 (Green) and RP11-589N15 (Red). A:** A CEU_H homozygote (GM07055) in an NCBI Build 36-inverted orientation. **B:** A CEU_L homozygote (GM12249) with the opposite orientation of (A). **C:** A heterozygous individual (GM06993).

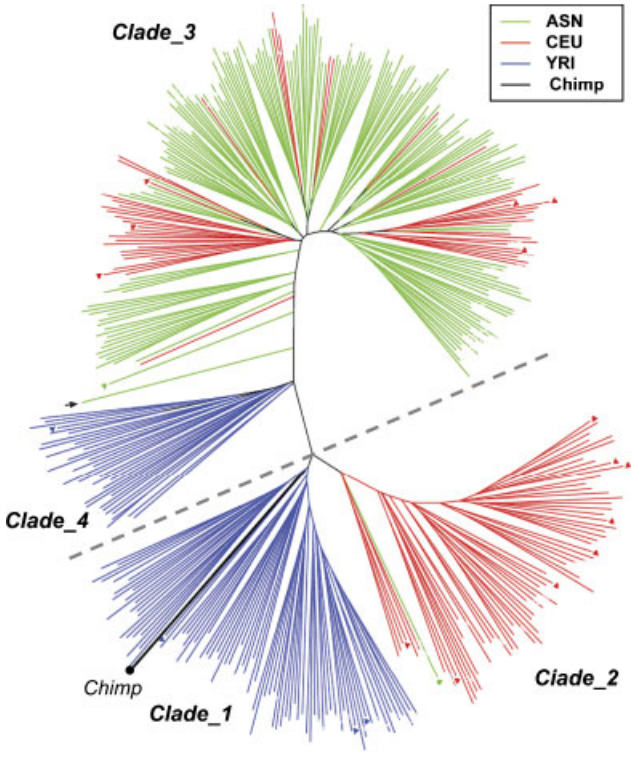

TABLE 1. **Chromosomal Arrangement Status at the 8p23 Region Shown by FISH**

Population	GM_ID[a]	FISH[b]	Haplotype groups based on first PC score	
CEU	GM12815	N/N	CEU_L	CEU_L
	GM12249	N/N	CEU_L	CEU_L
	GM06993	N/I	CEU_H	CEU_L
	GM12156	N/I	CEU_H	CEU_L
	GM07055	I/I	CEU_H	CEU_H
	GM06994	I/I	CEU_H	CEU_H
	GM06985	I/I	CEU_H	CEU_H
YRI	GM18517	N/I	YRI_H	YRI_L
	GM18502	I/I	YRI_H	YRI_H
ASN	GM18620	N/I	ASN_L	ASN_H

[a]Cell line ID in Coriell Repositories.
[b]N, build 36-noninverted order; I, build 36-inverted order.

FIGURE 3. **NJ tree of LRHs from HapMap samples and an outgroup haplotype. With each branch representing a haplotype, clusters were constructed from 418 chromosomes at 8p23 and the outgroup haplotype was assembled from the chimpanzee DNA sequence in the corresponding region. The haplotypes tested by FISH (Table 1) are labeled as triangles at the branch ends. A total of four distinct clades, Clade_1, Clade_2, Clade_3, and Clade_4, were detected. The chimpanzee haplotype resides in Clade_1, together with YRI_H, exclusively. The dashed line denotes the boundary for inversion statuses. The lower part consists of the haplotypes with high 1st PC scores. The ASN haplotype at the bottom of Clade_3 is localized at the center of three clusters by PCA (Fig. 1), which was difficult to be classify.**

TABLE 2. *HET* and Pairwise F_{ST} at the Inversion Region and the Rest of Chromosome 8

SNP number	Rest of chr. 8 (n = 47,790)	8p23 (n = 1,073)	P value*
HET			
ASN	0.347 ± 0.155	0.279 ± 0.176	$< 10^{-35a}$
CEU	0.362 ± 0.140	0.374 ± 0.140	1.16×10^{-3}
YRI	0.343 ± 0.148	0.345 ± 0.141	9.46×10^{-1}
F_{ST}			
ASN_CEU	0.091 ± 0.115	0.172 ± 0.172	$< 10^{-35a}$
ASN_YRI	0.156 ± 0.172	0.248 ± 0.254	$< 10^{-35a}$
CEU_YRI	0.143 ± 0.162	0.116 ± 0.141	1.65×10^{-5}

*MWU test.
[a]Statistical significance was confirmed by the bootstrap test ($P < 0.05$).

which were in opposite chromosomal orientations, were named Clade_3 and Clade_4, respectively.

Distinctive Regional Genetic Patterns of Inversion Polymorphism

To unveil potentially unusual genetic patterns caused by the inversion polymorphism, we compared the population genetic measures *HET* and F_{ST} between 8p23 and the rest of chromosome 8. As shown in Table 2, a considerable deficit in genetic diversity at 8p23 was seen in the ASN samples. The average *HET* of the ASN population in this region was 0.279, dramatically lower than the value in the rest of the chromosome (0.347, $P < 10^{-35}$). The bootstrap test also detected significantly reduced levels of genetic variation in the ASN population ($P = 0.028$; Supplementary Table S1). In multiple-population analysis, the respective average F_{ST} values for ASN_CEU and ASN_YRI were significantly larger than their corresponding values in the rest of chromosome 8 ($P < 0.001$). These significances were also verified by the bootstrap test with P values less than 0.05 (Supplementary Table S1). Supporting evidence also came from the distribution of high-F_{ST} outliers. As shown in Table 3, by comparing ASN_CEU and ASN_YRI population pairs, over five times of more outlier SNPs by proportion was seen in the 8p23 region in comparison with the rest of chromosome 8 ($P < 10^{-13}$).

Next, we performed window sliding in *HET* and F_{ST} analysis. For each SNP, a 500-kb window including all SNPs in 250-kb flanking regions was set up. Then we plotted the averages of *HET* and F_{ST} values of each window against the corresponding chromosomal positions. As shown in Fig. 4, consistent with the results of the averages in single-locus analysis, a lower population diversity was seen within the ASN samples. Also, a higher genetic differentiation between ASN and YRI samples was revealed at the region of inversion polymorphism.

Genes Under Natural Selection at 8p23

The above regional analysis suggested a positive selection in the ASN population at 8p23. Especially, the highest F_{ST} peak and the lowest *HET* value at 11.06 Mb implied that genes nearby were the targets of positive selection (Fig. 4). To identify candidate genes under selection, we performed analyses using F_{ST} and REHH in gene regions of the whole chromosome 8. For population differentiation analysis, we scanned a total of 633 genes and identified 202 candidates that contained at least one high-F_{ST} outlier in any population pair under comparison. As shown in Table 4, in the 8p23 region, 10 candidates (*THEX1*, *PPP1R3B*, *MSRA*, *XKR6*, *C8orf15*, *C8orf16*, *MTMR9*, *BLK*, *GATA*, and *FDFT1*) were found out of 20 scanned genes. Most of these genes (8/10) showed positive signatures by comparing pairs of ASN with other populations.

TABLE 3. Distribution of High F_{ST} Outliers in Different Regions

Region	SNP number	Outlier proportions		
		ASN_CEU	ASN_YRI	CEU_YRI
Rest Chr. 8	47,790	0.022	0.021	0.024
8p23	1,073	0.122	0.121	0.017
P value*		2.46×10^{-14a}	8.80×10^{-14a}	1.19×10^{-1}

*χ^2 test, df = 1.
[a]Statistical significance was confirmed by the bootstrap test ($P < 0.05$).

Considering the dramatic contrasts in LD patterns between ASN and CEU populations (Supplementary Fig. S1), the impact on recombination caused by the inversion polymorphism appears very low in the ASN population because almost all individuals are in the same inverted status. Thus, we calculated the REHH to scan for recent positive selection signatures in ASN samples, assuming only little influence of inversion polymorphism in this LRH test. A total of 2,026 "cores" across the entire chromosome 8 were scanned. By using a 99.9th percentile threshold of the REHH value, 86 candidate "core haplotypes" were found (Supplementary Fig. S2). In 8p23, we mapped four candidate "core haplotypes" to the *PPP1R3B*, *TNKS*, *MSRA*, and *XKR6* genes (Fig. 4; Table 4).

We then combined the results of the F_{ST} and REHH test above, as well as the integrated Haplotype Score (*iHS*) values in the same region reported previously [Voight et al., 2006]. Among all SNPs in candidate cores of the REHH test, only three SNPs, all localized in *XKR6*, showed not only large F_{ST} between ASN and other samples, but also remarkably high *iHS* values in the ASN population (Supplementary Table S2). These results suggested that the *XKR6* gene had been subjected to positive selection specifically in Asians.

DISCUSSION

Distribution of 8p23 Inversion in Human Populations

By means of both haplotype-based computational analyses and FISH experiments, we inferred and verified the orientation statuses of alleles in the 8p23 region in HapMap populations. The highest frequency of the "Build 36-inverted" allele (76%) was observed in the YRI population. Moreover, in light of the ASD tree, which was constructed by haplotypes of the HapMap and the corresponding outgroup, we deduced that the "Build 36-inverted" allele represented the ancestral state (i.e., "founder"). However, it should be pointed out that a potential bias on this conclusion may come from the ascertainment of SNPs. First, our estimation of genetic distance among haplotypes was based on HapMap SNPs, but not on all genetic variations in that region. Second, our knowledge of the chimpanzee sequence is still limited to infer the ancestral orientation, since at this moment we can not rule out the possibility that the corresponding genomic region might also be polymorphic in chimpanzees. Nevertheless, these findings suggested that this inversion polymorphism has been maintained for a long time in humans, and that the chromosomal inversion event in 8p23 region occurred before the migration out of Africa.

Population-Specific Positive Selection in the 8p23 Region

As shown in previous studies, the frequencies of chromosome rearrangements were often subjected to strong selective forces [Hamblin and Veuille, 1999; Hoffmann et al., 2004]. On the other hand, a large proportion of non-African allele could also be interpreted by an explicit spatial model of human demography [Currat et al., 2006]. However, based on the inversion frequencies

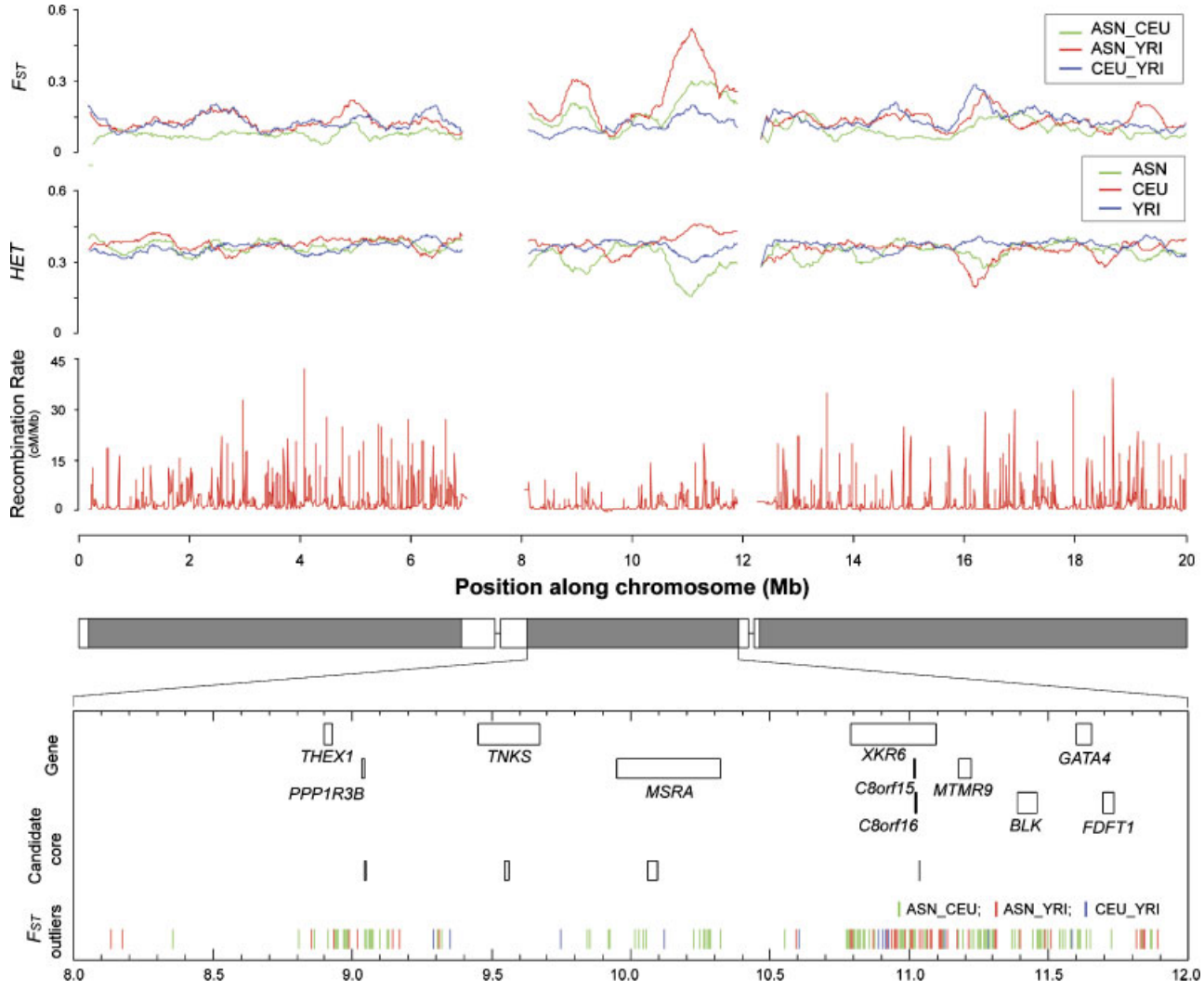

FIGURE 4. **The population genetic patterns of 8p23 with its flanking sequences, and the distribution of positive selection signals across the inversion region. From top to bottom: The averages of F_{ST}, *HET* with 500-kb sliding windows; the recombination rates (www.hapmap.org); the schematic diagram illustrating the regions able to be HapMapped (gray); and the segmental duplications (white) flanking two unfilled sequencing gaps (line), the distribution of genes, haplotype blocks, and F_{ST} outliers subjected to positive selection across 8p23. Note a significantly lower intensity (MWU test $P < 10^{-35}$) and a smaller number of recombination hotspots (χ^2 test, df = 1, $P < 0.05$) in this region compared with those of the rest.**

TABLE 4. **Candidate Genes Under Selection in the 8p23 Region***

				Outlier F_{ST} SNPs				Candidate cores			
RefSeq ID	Genes Symbol	Tran_start	Tran_end	SNP	A_C	A_Y	C_Y	Start	End	SNP	REHH
NM_153332.2	*THEX1*	8897859	8925898	9	1	1	0	–	–	–	–
NM_024607.2	*PPP1R3B*	9032915	9045616	7	1	1	0	9044683	9050324	3	7.62
NM_003747.2	*TNKS*	9450854	9677265	53	0	0	0	9549064	9562959	3	19.71
NM_012331.2	*MSRA*	9949235	10323803	105	16	2	1	10062158	10082523	7	32.81
NM_173683.3	*XKR6*	10791890	11096285	91	25	50	7	11036040	11038711	3	3.99
NM_001033662.2	*C8orf15*	11018300	11020984	1	1	1	0	–	–	–	–
NM_001014439.1	*C8orf16*	11021389	11025155	3	1	1	0	–	–	–	–
NM_015458.3	*MTMR9*	11179409	11223062	19	2	2	0	–	–	–	–
NM_001715.2	*BLK*	11388929	11459516	17	3	1	0	–	–	–	–
NM_002052.2	*GATA4*	11599161	11654918	17	5	0	0	–	–	–	–
NM_004462.3	*FDFT1*	11697598	11734226	15	1	0	0	–	–	–	–

*A_C, A_Y, and C_Y represent the comparisons of ASN_CEU, ASN_YRI, and CEU_YRI population pairs, respectively.

and the genetic structure of this region, we concluded that in the ASN population, the role of positive selection appears to be the most possible mechanism resulting in an increased frequency of the "derived" allele (i.e., ASN_L haplotypes). In addition to our data of low genetic diversity in ASN samples and high population differentiation (Table 2), several previous genome-wide studies also detected clustering signals of positive selection at the 8p23 region, as summarized in Supplementary Table S3 [Tang et al.,

2007; Voight et al., 2006; Wang et al., 2006]. Furthermore, we also identified several candidate genes subjected to positive selection. Among these candidates, *XKR6* showed the strongest signal in the ASN population by all tests, including F_{ST}, REHH, and *iHS* analyses. *XKR6* (XK, Kell blood group complex subunit-related family, member 6), a member of XK-related superfamily genes (XRG), has been detected to transcribe in various human tissues [Appel et al., 2002]. Multiple messenger RNAs have been identified in the XKR6 region and the functions of each transcript and their products await further elucidation.

Haplotype Analysis to Discover Chromosome Status in Population Genetic Research

In our study, regional haplotype analyses provided an effective strategy to unveil the inversion status of each individual in the 8p23 region. As experimentally verified by FISH, distinct haplotype subgroups revealed by PCA and ASD tree analyses in silico corresponded to different alleles of the inversion polymorphism. PCA with bootstrap tests further provided the quantification of genetic background for each haplotype. Hence, strategies used in this regional haplotype analysis could be readily applied to systematically scan for possible inversion polymorphisms on a genome-wide scale.

The cryptic haplotype substructures coming from inversion polymorphisms might pose false results in association studies. Based on haplotype analysis, one could develop strategies to search the potential relationship between the inversion status and the disease candidates, or to make proper adjustments for the association statistics of each marker in consideration of structural variations. It is notable that a set of proxy SNPs can be selected to capture the information for the inversion status of each haplotype effectively. For instance, using a greedy-discard method based on PCA [Lin and Altman, 2004], we assembled a set of 13 proxy SNPs, out of 1,073 variations in the 8p23 region, that could exactly indicate the inversion status for each chromosome in the CEU population (data not shown), showing that PCA could be an efficient tool for large-scale studies of inversion polymorphism.

In conclusion, using both computational and experimental approaches, we consistently detected distinct haplotype subgroups in HapMap populations that corresponded to specific alleles of inversion polymorphism at 8p23. By treating this structural variation as a polymorphism frame, our approach successfully revealed a complex evolution history in this genomic region, as well as signals of specific positive selection in the ASN population. In addition, this study demonstrated that haplotype-based clustering methods (in particular, PCA) could be utilized to search for inversion polymorphisms on a genome-wide scale.

ACKNOWLEDGMENTS

We are thankful for the significant efforts of the HapMap group and the Genotyping platform at the Beijing Genomics Institute for SNP data production of chromosome 8p. We are grateful to Dr. Jurg Ott for helpful comments, and to Drs. Qingjie Liu and Dacheng He for providing experimental facilities. Our thanks also extend to Dake Zhang, Xiaowen Hao, Gwen Zahner, and Jessica Pang for discussion and proofreading. This study was supported by grants from the National Natural Science Foundation of China (30225017), the Ministry of Science and Technology (2002BA711A09), and the Chinese Academy of Sciences (Century Program) (all to C.Z.).

REFERENCES

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814.

Appel S, Filter M, Reis A, Hennies HC, Bergheim A, Ogilvie E, Arndt S, Simmons A, Lovett M, Hide W, Ramsay M, Reichwald K, Zimmermann W, Rosenthal A. 2002. Physical and transcriptional map of the critical region for keratolytic winter erythema (KWE) on chromosome 8p22-p23 between D8S550 and D8S1759. Eur J Hum Genet 10: 17–25.

Bansal V, Bashir A, Bafna V. 2007. Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Res 17: 219–230.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Currat M, Excoffier L, Maddison W, Otto SP, Ray N, Whitlock MC, Yeaman S. 2006. Comment on "Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in *Homo sapiens*" and "Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans". Science 313:172a.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. Nat Rev Genet 7:85–97.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Giglio S, Broman KW, Matsumoto N, Calvari V, Gimelli G, Neumann T, Ohashi H, Voullaire L, Larizza D, Giorda R, Weber JL, Ledbetter DH, Zuffardi O. 2001. Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. Am J Hum Genet 68:874–883.

Hamblin MT, Veuille M. 1999. Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics 153:305–317.

Hoffmann AA, Sgro CM, Weeks AR. 2004. Chromosomal inversion polymorphisms and adaptation. Trends Ecol Evol 19:482–488.

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res 12:656–664.

Kumar S, Tamura K, Jakobsen IB, Nei M. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17: 1244–1245.

Lin Z, Altman RB. 2004. Finding haplotype tagging SNPs by use of principal components analysis. Am J Hum Genet 75:850–861.

Munte A, Rozas J, Aguade M, Segarra C. 2005. Chromosomal inversion polymorphism leads to extensive genetic structure: a multilocus survey in *Drosophila subobscura*. Genetics 169:1573–1581.

Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14.

Rozas J, Gullaud M, Blandin G, Aguade M. 2001. DNA variation at the rp49 gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics 158:1147–1155.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, AcKerman HC, Campbell SJ, Altschuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Schneider S, Roessli D, Excoffier L. 2000. Arlequin: a software for population genetics data analysis. Version 2.000. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Geneva, Switzerland.

Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N,

Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, Stefansson K. 2005. A common inversion under selection in Europeans. Nat Genet 37: 129–137.

Stephens M, Donnelly P. 2003. A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169.

Sugawara H, Harada N, Ida T, Ishida T, Ledbetter DH, Yoshiura K, Ohta T, Kishino T, Niikawa N, Matsumoto N. 2003. Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. Genomics 82:238–244.

Tang K, Thornton KR, Stoneking M. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol 5:e171.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. PLoS Biol 4:e72.

Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. Proc Natl Acad Sci USA 103:135–140.