

ORIGINAL ARTICLE

Nucleotide polymorphism of the *TNF* gene cluster in six Chinese populations

Yongbiao Zhang^{1,2}, Feng Zhang², Hongbin Lin², Lei Shi¹, Panpan Wang², Li Shi¹, Qiang Gong², Xin Li², Mei Wang², Songnian Hu², Jiayou Chu¹ and Duen-Mei Wang²

DNA variants in a 31-kb region of the human major histocompatibility complex, encompassing the tumor necrosis factor (*TNF*) gene cluster, were surveyed by direct sequencing of 283 unrelated individuals from six Chinese populations. A total of 273 polymorphic sites were identified, with nearly half of them novel. We observed an excess of rare variants and negative values of selection tests of the region, implying either that these populations experienced a historical expansion or that the surveyed region was subjected to natural selection. Different characteristics of the sequence variation in the six populations outline the genetic differentiation between Northern and Southern Chinese populations. The distributions of recombination rates are similar among all the populations, with variation in the magnitude and/or in the fine location of hot spots. Tag single-nucleotide polymorphisms (SNPs) selected from HapMap (Phase II) CHB data accounted for an average of 64% of common SNPs from the six Chinese populations. We also observed a limited transferability of tag SNPs between Chinese populations on the 31-kb region with an excess of untaggable SNPs and ragged linkage disequilibrium blocks. It suggested that the design and interpretation of future association studies should be more cautious, and that a resequencing approach may refine tag SNP selection on Chinese-specific disease mapping.

Journal of Human Genetics (2010) 55, 350–357; doi:10.1038/jhg.2010.33; published online 27 May 2010

Keywords: Chinese population; DNA polymorphism; selection; tag SNP; *TNF* gene cluster

INTRODUCTION

Human genetic studies have become much easier since the availability of annotated DNA sequences from the Human Genome Project and cataloged single-nucleotide polymorphisms (SNPs) of major ethnic groups from the HapMap project.^{1,2} High-throughput of genotyping and sequencing technologies has also facilitated the related research. To investigate genetic variation, association studies are frequently used to identify DNA variants conferring phenotypes such as physical appearance, disease susceptibility and drug metabolism.^{3–5} However, obstacles resulting from population substructure,⁶ variation in linkage disequilibrium (LD) patterns^{7,8} and local natural selection⁹ have affected the progression of the study. Poor replication of genotype–phenotype associations furthers the problem while evaluating the true effect of a genetic variant.^{10,11} All these can be improved with information on the underlying population differentiation by deep sequencing with more individuals and more populations.

The International HapMap Project released a collection of over 3.1 million SNPs taken from individuals on three continents (Africa,

Europe and Asia).¹² However, the polymorphisms of populations other than the HapMap populations are inadequately represented.^{13,14} Two concerns arise from population-related studies on the Chinese population with the use of HapMap data. First, the SNP discovery preceding the HapMap genotyping project was mostly conducted in a small sample panel, predominantly of European descent.^{15,16} The derived genetic information from the HapMap database is often applied to a larger sample set. As a result, ascertainment bias may be introduced while conducting association studies with Asian populations. Second, the genetic complexity of Chinese populations cannot be solely represented by the 45 HapMap Hans. China has a significant proportion of the world population with over 1 billion people, and is composed of 56 official nationalities.¹⁷ Population genetic studies revealed that Chinese minorities are distinct and have multiple origins: Southeast Asia, Northeast Asia (Altaic) and Mid-Asia.^{18–21} Furthermore, Han, the major ethnic group of the Chinese, can be divided into northern and southern groups.^{18,19,22} Questions regarding the underlying population substructure of the Chinese are still awaiting investigation.

¹Institute of Medical Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Kunming, PR China and ²CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, PR China

Correspondence: Professor J Chu, Institute of Medical Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, 379#, Jiaoling Road, Kunming, Yunnan 650118, PR China.

E-mail: chujy@imbcams.com.cn or Professor D-M Wang, CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, 7#, Beitucheng West Road, Beijing 100029, PR China.

E-mail: wangdm@big.ac.cn

Received 28 November 2009; revised 23 February 2010; accepted 18 March 2010; published online 27 May 2010

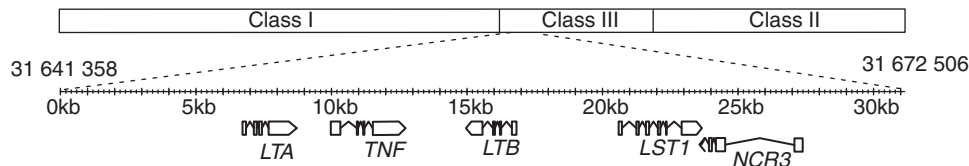


Figure 1 Map of the 31-kb region comprising the *TNF* gene cluster on human major histocompatibility complex.

Despite the differences in sampled populations, genotyped markers and examined genomic regions, previous studies found that at least four major Chinese groups exist.^{18,22,23} To be properly representative of Chinese populations, six populations with 283 individuals were selected in this study for their geographic location and association in major linguistic subfamilies. Each of the aforementioned four Chinese groups was exemplified by at least one of our study populations. The six populations are Han Shandong (from Northern China), Han Guangdong (from Southern China), Li, Yi, Tibetan and Mongolian. We sequenced a 31-kb region, comprising the tumor necrosis factor (*TNF*) gene cluster, in the class III region of the human histocompatibility complex on chromosome 6p21.3 (Figure 1). Five immune-related genes (*LTA*, *TNF*, *LTB*, *LST1* and *NCR3*) are located in the region with known connections to various human diseases (<http://www.ncbi.nlm.nih.gov/omim>). In particular, *TNF*, a proinflammatory cytokine, has an important role in innate immunity as the first line of host defense to infection and is implicated in infectious disease, metabolic disorder and cancer.

By resequencing this gene-rich and immune-related region, we attempt to uncover genetic differentiation at the DNA sequence level for Chinese populations. The objective is to gain knowledge for future association studies of disease mapping with Chinese populations.

MATERIALS AND METHODS

Sampled populations

Six Chinese populations (283 unrelated individuals in total) including Han Shandong, Han Guangdong, Li, Yi, Tibetan and Mongolian were used in this study with sample sizes of 45, 47, 51, 46, 46 and 47, respectively. Mongolian and Han Shandong populations reside in North China; Han Guangdong and Li populations reside in South China; Yi and Tibetan populations reside in Southwest China. All DNA samples were extracted from cultured cells using a phenol–chloroform method, measured by NanoDrop ND-1000 (Thermo-Fisher Scientific, Waltham, MA, USA) and adjusted to a concentration of 30 ng μl^{-1} . The cell lines were deposited at the Immortalize Cell Bank of Chinese Nationalities, which is supported by the Chinese Human Genome Diversity Project. The donor signed the written informed consent for cell line establishment and subsequent studies. Our project was reviewed and approved by the Ethics Committee at the Chinese Academy of Medical Sciences and Peking Union Medical College.

PCR amplification, DNA sequencing and identification of SNPs

Primer sets for PCR and sequencing were designed using Primer-Premier 5.0 (PREMIER Biosoft International, Palo Alto, CA, USA). All PCR reactions were performed with a touchdown method and PCR products were purified using an AcroPrep 384 Multi-well Filter 30k plate (Pall, Port Washington, NY, USA). Sequencing reactions were conducted using Applied Biosystems Big Dye Terminator chemistry, and the products were resolved on ABI Prism 3730XL DNA Analyzers (Applied Biosystems, Carlsbad, CA, USA). Sequence trace files were analyzed using Phred/Phrap/Polyphred/Consed (University of Washington, Seattle, WA, USA) software. The base-quality value threshold was set to 20 in Phred (that is, a 99% probability that the base is accurate), and all polymorphic sites were manually inspected by at least two individuals. To

validate every polymorphic site, the 31-kb region was sequenced on both strands at least once. Singletons and doubletons were further sequenced with additional primer sets for verification. To maintain data integrity, a polymorphic site was retained for those cases for which the confirmed genotype rate at the site was over 95%.

Population genetic analysis

To ensure that only unrelated individuals were included in each population, the Hardy–Weinberg equilibrium test was used on all identified polymorphic sites. Only 8 out of 1638 tests showed $P < 0.05$ (χ^2 -test) in a sporadic manner, which may be caused by the high number of repeated tests or by a small sample size for each population. The SNPs corresponding to the aforementioned eight tests were not removed from subsequent analysis because of their negligible impact on results. Population pairwise G_{ST} (Supplementary Table 1), rather than F_{ST} , was calculated with DnaSP v5,²⁴ because Nei's G_{ST} provided more reliable estimates (based on our simulation results) in this study. To obtain an empirical P -value of the metric, a permutation-based procedure was used. For each paired population, identities of individuals were swapped at random and G_{ST} was calculated for each new configuration. One thousand permutations were performed and the percentage of G_{ST} greater than the observed value was then taken as the P -value. An unrooted neighbor-joining tree was constructed using observed G_{ST} as the genetic distance in MEGA 4.²⁵ θ_W and π , two parameters that describe nucleotide diversity, were estimated with ARLEQUIN 3.1.1.²⁶ Selection tests, including Tajima's test and Fu and Li's test, were performed using DnaSP v5.²⁴ The empirical distribution of these test statistics was generated by a coalescent simulation method, and the P -value corresponded to the percentiles of the observed statistic against an empirical distribution. A sliding-window analysis of Tajima's test was also performed to inspect its variation along the sequence. Fu's F_s statistic, which is sensitive to population growth,²⁷ was also calculated.

For haplotype analysis, 26 polymorphic sites with a minor allele frequency (MAF) greater than 0.2 in the combined population (283 individuals) were used (Supplementary Table 2). Population-specific haplotypes were constructed using the Bayesian statistical method implemented in PHASE 2.1.²⁸ The best-fitting haplotype set of each population was obtained with five different seeds of inferences. Finally, the Unweighted Pair Group Method with the Arithmetic Mean (UPGMA) clustering method was used to evaluate the relationships of all haplotypes. The sex- and time-averaged population recombination rate was estimated with the subprogram Interval of LDhat 2.1,⁷ assuming an effective population size of 10 000 for the two Han populations and 5700 for the other four populations.²⁹ Haplotype blocks and tag SNPs were obtained using Haploview 4.0.³⁰ Tag SNPs were selected using the pairwise mode with r^2 thresholds of 0.8. Only biallelic SNPs were used for the above analyses. Insertion/deletion (Indel) and short tandem repeat were not considered. The homologous sequences of one chimpanzee and one rhesus macaque were downloaded from the NCBI database to serve as outgroups.

RESULTS

Sequence variation in the 31-kb region

With sequences of 283 unrelated individuals from the six Chinese populations, a total of 273 polymorphic sites were identified in the 31-kb region (8.5 sites per kb). SNPs represented the majority of the polymorphisms (89%), whereas Indel and short tandem repeat made up the rest (10 and 1%, respectively). A summary of sequence variants of the region in the six Chinese populations is shown in Table 1. The

Mongolian population had the highest number of variants (177) among the six populations; Li and Yi populations were less polymorphic with low numbers of variants (121 and 122, respectively). Percentages of singletons and doubletons differed significantly among populations. In particular, singletons varied from 7% in Li to 29% in Mongolian populations. These differences in variation distribution among the six Chinese populations further substantiate the known genetic differentiation among Chinese populations.

A total of 20 SNPs were identified in the coding region with six synonymous SNPs and 14 nonsynonymous SNPs. Furthermore, two novel SNPs found in the genic region only appeared in the Li population: one (MAF=0.13) was located in the 3'UTR region of *TNF* and the other (MAF=0.06) was in the exon 2 region of *NCR3* (as a nonsynonymous SNP). Among all identified SNPs, eight SNPs are associated with diseases according to OMIM (Table 2). We observed a one- to five-fold variation in allele frequencies of the eight SNPs in the six populations. Such frequency variation of disease-associated SNPs may reflect differential disease susceptibilities among populations.

Population subdivision

To measure genetic differentiation of the six populations, we used Nei's pairwise G_{ST} along with an empirical P -value to substantiate the measure. Most of the pairwise G_{ST} values between populations were statistically significant ($P < 0.05$), except the value between Han Shandong and Mongolian populations ($G_{ST} = 0.00108$; $P = 0.176$). An

neighbor-joining tree constructed with G_{ST} shows that the grouping of these six populations agrees fairly well with their geographic distributions (Figure 2). We noticed that the geographic proximity of the Han Guangdong and Li populations, as well as that of the Yi and Tibetan populations, reflects their genetic distances in terms of G_{ST} fairly well. Furthermore, Han Shandong and Han Guangdong populations are clearly genetically differentiated. Our previous study, including the same six populations with 10 short tandem repeat markers from chromosome 3, exhibited a similar pattern of genetic relationships (H Lin *et al.*, unpublished data). It is reasonable to believe that geographic isolation and gene flow between adjacent populations drove the genetic characteristic of Chinese populations within the same region to become more similar.

Selection tests

To reveal possible selection processes in the six populations, allele frequency-based selection tests were performed with DnaSP. As shown in Table 3, the Mongolian population is the most polymorphic, with the highest measures of nucleotide diversity: π and θ_{W} . Li and Yi populations have the lowest θ_{W} s owing to their low numbers of polymorphic sites. Both populations exhibit similar values of π and θ_{W} s whereas the other populations have distinct values. Significant negative values of selection tests observed in Han Shandong (Fu and Li's $D/F/F^*$ tests, $P = 0.031$, 0.042 and 0.036 , respectively) and Mongolian (Tajima's D and Fu and Li's F^* tests, $P = 0.046$ and 0.045 , respectively) populations indicate the deviation from neutrality and the possibility of selection. Nonetheless, population demographic history, such as population expansion, could confound this deviation,

Table 1 Summary of sequence variants across the 31-kb region in six Chinese populations

Population	Number of polymorphic sites				Total
	Singleton	Doubleton	MAF < 0.05	MAF \geq 0.05	
Han Shandong	44	23	89	61	150
Han Guangdong	31	26	77	71	148
Li	8	4	53	68	121
Yi	24	22	51	71	122
Tibetan	36	25	83	71	154
Mongolian	50	23	96	81	177
Total	39	16	198	72	270 ^a

Abbreviations: MAF, minor allele frequency; STR, short tandem repeat.
^aSTR not included.



Figure 2 Unrooted phylogenetic neighbor-joining tree of the 31-kb region in six Chinese populations. The tree was constructed with Nei's G_{ST} as genetic distance. The scale of the G_{ST} value is 1/1000.

Table 2 Disease-associated SNPs in six Chinese populations

rs#	Mutation	Frequency of risk alleles						Associated diseases ^a
		Han Shandong	Han Guangdong	Li	Yi	Tibetan	Mongolian	
rs2239704	<i>LTA</i> , +80C>A	0.46	0.33	0.39	0.48	0.34	0.43	Leprosy
rs909253	<i>LTA</i> , +252A>G	0.40	0.46	0.43	0.23	0.34	0.31	Psoriatic arthritis, myocardial infarction
rs1041981	<i>LTA</i> , +804C>A	0.40	0.45	0.40	0.23	0.34	0.31	Myocardial infarction, coronary artery disease
rs1800630	<i>TNF</i> , -863C>A	0.12	0.18	0.14	0.27	0.23	0.14	Alzheimer disease
rs1799724	<i>TNF</i> , -850C>T	0.15	0.05	0.19	0.17	0.11	0.13	Vascular dementia, Alzheimer disease
rs1800750	<i>TNF</i> , -376G>A	0	0	0	0	0	0.01	Cerebral malaria
rs1800629	<i>TNF</i> , -308G>A	0.10	0.03	0.14	0.04	0.04	0.07	Immunodeficiency virus dementia, asthma, migraine without aura, psoriatic arthritis, systemic lupus erythematosus, malaria
rs361525	<i>TNF</i> , -238G>A	0.02	0.03	0.04	0	0.06	0.10	Cerebral malaria

Abbreviations: SNP, single-nucleotide polymorphism; TNF, tumor necrosis factor.
^aData from OMIM (<http://www.ncbi.nlm.nih.gov/omim>).

Table 3 Statistics of sequence diversity and natural selection tests

Population	Han Shandong	Han Guangdong	Li	Yi	Tibetan	Mongolian
$\theta_W (\times 10^{-4})$	8.4 ± 0.7	8.4 ± 0.7	6.7 ± 0.7	6.9 ± 0.7	8.7 ± 0.8	10 ± 0.8
$\pi (\times 10^{-4})$	6.3 ± 0.3	6.7 ± 0.3	6.6 ± 0.3	6.9 ± 0.2	7.3 ± 0.3	7.4 ± 0.3
Tajima's <i>D</i>	-0.84	-0.67	-0.05	0.03	-0.54	-0.87*
Fu and Li's <i>D</i> *	-1.35	-0.22	1.49**	-0.11	-0.58	-1.01
Fu and Li's <i>F</i> *	-1.36*	-0.49	1.00	-0.06	-0.68	-1.14*
Fu and Li's <i>D</i> ^a	-1.47*	0.05	1.66**	0.12	-0.45	-0.69
Fu and Li's <i>F</i> ^a	-1.45*	-0.35	1.06	0.09	-0.59	-0.94
Fu's <i>F</i> _s	-8.25	-5.82	2.95*	-4.43	-4.15	-5.19

The coalescent simulation method was used to obtain an empirical *P*-value. The number of replicates was 10000.

**P* < 0.05.

***P* < 0.01.

^aFu and Li's test with a chimpanzee as the outgroup.

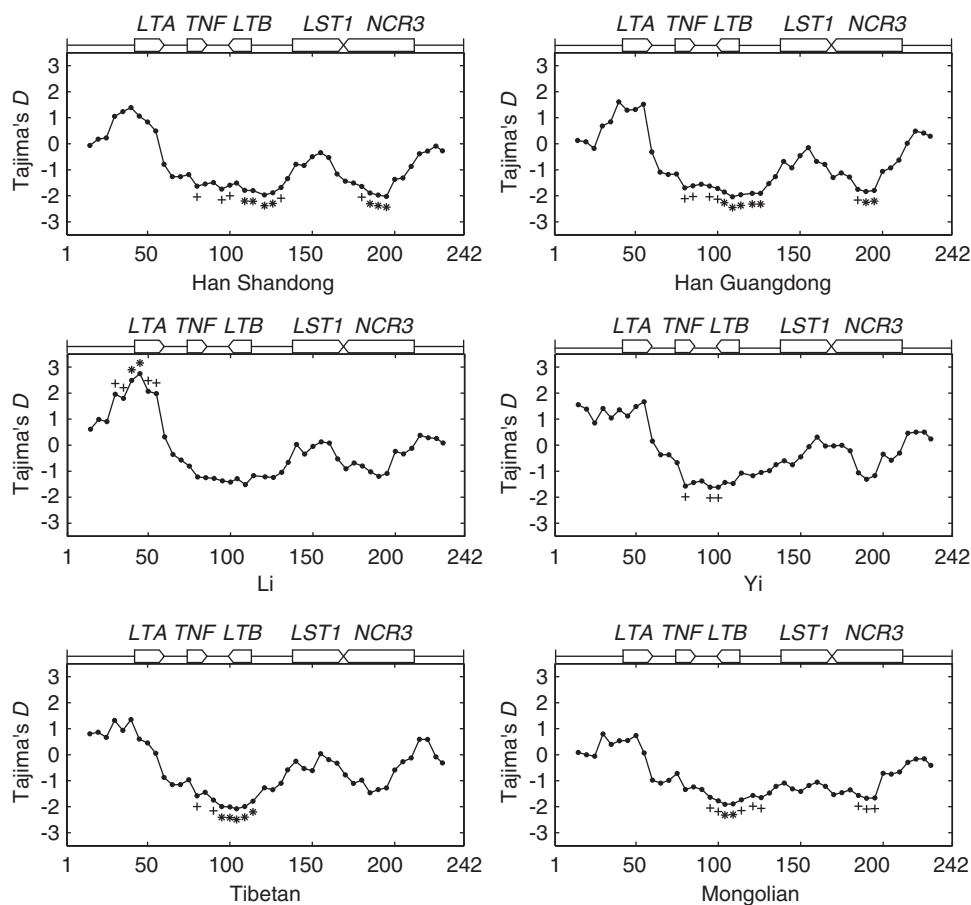


Figure 3 Sliding-window analysis with Tajima's *D* test in the 31-kb region. Tajima's *D* statistics are calculated in 30-SNP sliding windows with 5-SNP steps. The *x* axis represents the ordered 242 SNPs along the 31-kb region. The respected position of each dot on the *x* axis marks the middle of a sliding window. The *y* axis represents the value of Tajima's *D*. A display of corresponding genes is on the top of each panel. The symbol '+' represents $0.05 < P < 0.1$, and '*' represents $P < 0.05$.

for negative values of *F*_s were observed in these two populations. In contrast, the significant positive values of Fu and Li's *D*/*D** test (both $P=0.004$) and Fu's *F*_s test ($P=0.015$) in the Li population indicate that the Li population may have either experienced a balancing selection or had a constant size of population for a long period.

Instead of generalizing the selection test results with pooled genetic data of the 31-kb region as above, Tajima's *D* test across the region was

also conducted with a sliding-window method (Figure 3). The six populations have similar patterns of Tajima's *D*, except for differences in *D* values and their associated significance levels. At the *LTA* gene, all populations displayed positive *D* values, but were only significant in the Li population. As for the *TNF* and *LTB* genes, Han Shandong, Han Guangdong, Tibetan and Mongolian populations had significant negative *D* statistics. The *NCR3* gene was also influenced by selection,

for the two Han populations have significantly negative values. Although demographic events such as population expansion can confound the results of selection tests, they always function on all loci of the human genome, whereas natural selection functions on specific regions, such as the genic region.³¹ Therefore, observed negative *D* values can mostly be ascribed to selection at the genic regions of the 31-kb sequence.

Haplotype reconstruction and recombination rate estimation

To simplify the haplotype analysis, only SNPs with an MAF > 0.2 in the combined population were used. As shown in Figure 4, 36 haplotypes were derived from the 26 analyzed SNPs, and they clustered into two major clades. The ancient clade (40% of total chromosomes) was defined by cogrouping with the haplotypes from chimpanzee and rhesus macaque. Except for the haplotype distribution of the Li population slightly departing from those of the other populations, there were no significant differences in haplotype distributions among the six populations or between the three closely related population pairs (Han Shandong and Mongolian, Han Guangdong and Li, and Yi and Tibetan) defined by *G*_{ST}.

Recombination is the driving event to shape an LD block. To investigate the variation in the LD patterns of these populations, we used all SNPs obtained in the 31-kb region to estimate local recombination rates and to map recombination hot spots. In total, eight haplotypes were found in all six populations. These eight haplotypes exceed the average number of four common haplotypes per LD block with HapMap CHB+JPT data.³² Frequent recombination events in the region may contribute to this excessive number of haplotypes. The overall population recombination rate ranged widely from 3.1 in the Li population to 18.5 in the Han Shandong population. However, the patterns of recombination rate were similar among the six populations (Figure 5), marked with three recombination hot spots at around the 10-, 22- and 30-kb sites. Subtle differences in recombination patterns, such as the position and magnitude of hot spots, were observed among populations. For example, hot spot shifting was noticed for the Li population from the 10-kb site to a 5-kb site. In the case of the Tibetan population, the 10-kb hot spot was replaced with a wide-spread hot spot at the 20-kb site. Because of active recombination, only a small 7-kb block was observed at the beginning of the 31-kb region for Han Shandong and Mongolian

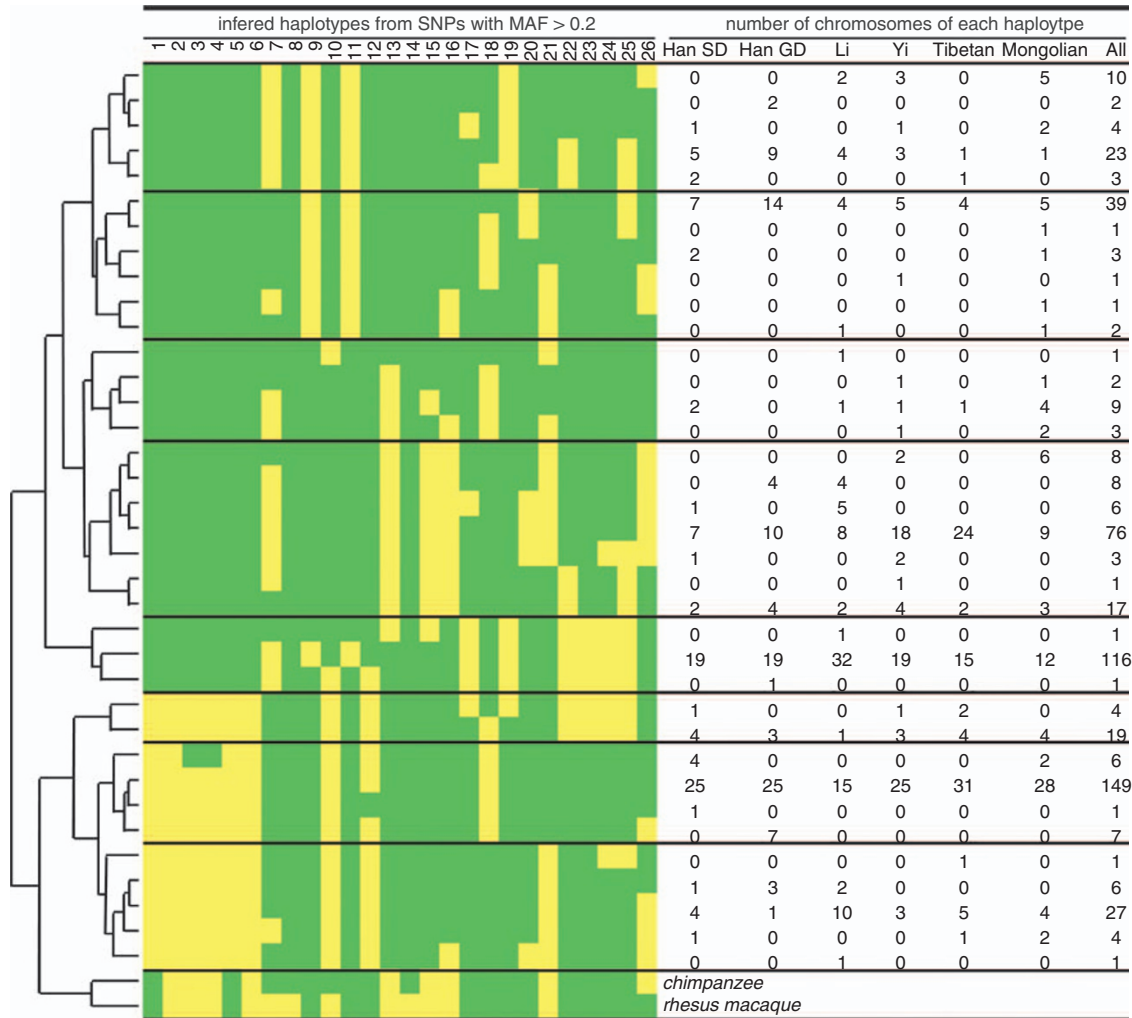


Figure 4 Haplotype clades and distributions of the six Chinese populations. The blue square represents the major allele of an SNP. The yellow square represents its minor allele. A total of 26 SNPs (MAF > 0.2) were used to infer haplotypes by PHASE. The derived haplotypes were clustered with the UPGMA method, giving rise to eight clades separated by darker lines. The number of chromosomes in each haplotype from the six populations is shown on the right. Han SD represents Han Shandong; and Han GD represents Han Guangdong. Chimpanzee and rhesus macaque sequences were used as the outgroup.

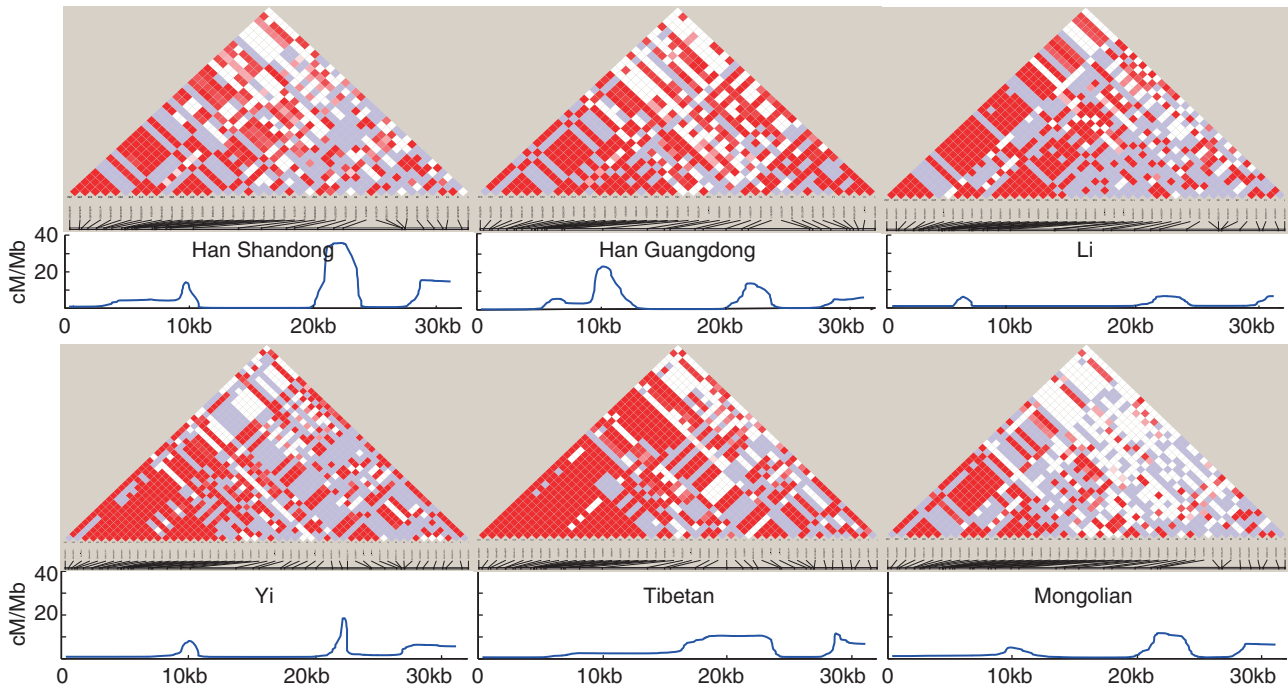


Figure 5 LD blocks and recombination patterns of the 31-kb region among the six Chinese populations. The LD block was constructed from SNPs with $MAF \geq 0.1$ by Haploview. The colors of the D' plots indicate the following—red: $D'=1$ and $LOD \geq 2$; pink: $D' < 1$ and $LOD \geq 2$; blue: $D'=1$ and $LOD < 2$; and white: $D' < 1$ and $LOD < 2$. The SNPs used to construct LD blocks are displayed under each D' plot. The panel beneath each D' plot shows the estimated population recombination rates (in cM/Mb, by LDhat) versus the physical position (in base pair) of all 242 SNPs along the *TNF* gene cluster region.

Table 4 Coverage of common SNPs by tag SNPs from the HapMap CHB and six Chinese populations

Tag SNPs-derived population	No. of common SNPs		No. of SNPs can be tagged in each population						Average percentage
	SNPs	No. of tags	Han Shandong	Han Guangdong	Li	Yi	Tibetan	Mongolian	
Han Shandong	50	24		51 (86%) ^a	48 (85%)	53 (89%)	46 (76%)	41 (59%)	79%
Han Guangdong	59	23	36 (72%)		49 (87%)	52 (88%)	45 (75%)	39 (57%)	76%
Li	56	21	34 (68%)	44 (74%)		42 (72%)	44 (73%)	41 (59%)	69%
Yi	59	19	35 (70%)	50 (84%)	42 (75%)		39 (65%)	35 (50%)	69%
Tibetan	60	24	40 (80%)	45 (77%)	46 (82%)	50 (84%)		47 (68%)	78%
Mongolian	69	29	42 (84%)	49 (83%)	50 (89%)	50 (84%)	56 (93%)		87%
HapMap CHB	27	15	29(57%)	39(66%)	31(55%)	50(84%)	44(73%)	32(46%)	64%

Abbreviation: SNP, single-nucleotide polymorphism.
^aThe proportion of coverage is shown in parentheses.

populations. In a disease association study, this type of LD blocks, flanked by active recombination hot spots, will have coverage problems with tag SNPs from populations other than the same ethnicity.

Tag SNP transferability

To evaluate tag SNP transferability of the 31-kb region in Chinese populations, we compared the efficiency of a population-specific tag SNP set in capturing genotypic information of other populations with the use of our resequencing data and Phase II HapMap data. The HapMap CHB-derived tag SNPs displayed an averaged coverage rate of 64% for common SNPs from the six populations (Table 4). The coverage rate was not improved by using HapMap CHB+JPT-derived tag SNPs (54% on average). Although it is known that the HapMap CHB sample is a Han collection obtained from Northern China,² in this study, the coverage ability of the CHB-derived tag SNPs

performed better in the Han Guangdong than in the Han Shandong (66 vs 57%; $P=0.003$ obtained by a re-sampling method) population. This unexpected result may be caused by limited sample size.¹³ The low coverage of CHB-derived tags for Li (55%) and Mongolian (46%) populations may be ascribed to the high number of population-specific SNPs in the Li population and the high nucleotide diversity of the Mongolian population, respectively. Tag SNPs selected from the Mongolian population could account for 87% of the common SNPs of the other five populations, whereas tag SNP sets of both Li and Yi populations exhibited low transferability with a coverage of 69%. This finding suggests that the Mongolian population is a good proxy for diverse Chinese populations and a preferred population for tag SNP selection with a Chinese-based association study. Moreover, when considering only Han populations, Northern Han is the preferred population for tag SNP selection, for we observed that tag SNP

transferability of the Han Shandong population is higher than that of the Han Guangdong population (Table 4). Limited tag SNP transferability in the 31-kb region among these populations suggests that an association study on Chinese populations should be designed cautiously, especially when selecting tag SNPs from HapMap CHB.

DISCUSSION

Previous genetic profiles of the Chinese population revealed the following: (1) existence of genetic differences between the Northern and Southern populations, even to the extent that Han shared this geography-related distinction;^{17,18,23,33} (2) formation of a divided continuum of the Chinese population through migration, geographic isolation and indigenous convention.^{19,20,22,34} However, the detailed structure of Chinese populations remains elusive.^{17,20,22} The information available to date was collected mostly by genotyping markers on mtDNA, Y chromosome and autosomal chromosomes. Our approach was the use of autosomal DNA with the resequencing method, which was seldom used in past studies with Chinese samples. From the sequence polymorphism of the 31-kb region, we observed similar findings as in previous reports, namely, that our six Chinese populations are genetically different from each other, including the two Han populations, in terms of heterozygosity and genetic distance. Even in this small number of populations, three major groups can be defined by a G_{ST} -based phylogenetic tree. In Figure 2, the geographically adjacent populations, such as the Han Shandong and Mongolian populations, and the Yi and Tibetan populations, are grouped together. It supports the second notion above that populations residing in proximate areas are generally closer in genetic relationship.^{19,22} Therefore, our results demonstrated that Chinese populations not only have genetic distinctions between the north and south but can also be further divided into more demographic groups, and local gene flow may contribute to the genetic similarity that we observed.

TNF gene cluster is conserved during evolution,³⁵ and encodes inflammatory cytokines crucial to host defense against infection.³⁶ Their roles in the immune system make the genes susceptible to balancing or negative selection.³⁷ Dependent on the geographic environment, traditional customs and infectious agents, populations may encounter different selection pressures, leading to changes in relevant genes. The six populations used here are representative of the diversity of Chinese populations and reflect sequence changes in the *TNF* gene cluster, in part, to account for the distinct environments that they live in.

Our sliding-window analysis of Tajima's D test revealed that most values were negative for all the six populations, with statistical significance on the genic regions of *TNF*, *LTB* and *NCR3*. The only region that had positive values for all populations was *LTA*, but it was only statistically significant in the Li population. In contrast to the other five populations, the Li population stood out from the populations in our study by being the least polymorphic and by virtue of its positive F_s statistics. It is known that the Li population migrated to Hainan Island during the Neolithic period³⁸ and was isolated until recent years (in agreement with our finding of a positive F_s statistic). Furthermore, Hainan Island has been one of the major areas of malaria epidemics in China.³⁹ The associations of malaria infection with the human sickle-cell trait and certain genotypes of *G6PD* were demonstrated to be a function of balancing selection.^{40–42} As shown by Zhou's study, the percentage of $\alpha(+)$ thalassemia carriers in the Li population is very high (56.7%) to counteract the malaria epidemic.⁴³ Furthermore, an SNP located in the *TNF* promoter region is associated with cerebral malaria, with high levels of circulating TNF.^{44,45} It is reasonable to speculate that the unusual polymorphic pattern of

LTA in the Li population may result from balancing selection. Because of their proximity to the immediate upstream region of *TNF*, polymorphic sites in *LTA* can affect the level of TNF expression and collaborate with *TNF* to have an important role in resistance to malaria in the Li population. This may explain the unusual pattern of selection on Li.

The analysis of fine-scale recombination of the 31-kb region revealed three recombination hot spots. They mostly locate in intergenic regions and share a similar physical distribution, with variation in the intensity and exact location among the six populations. A recent report from Coop *et al.*⁴⁶ suggested that LD-based hot spots were used for 60% of actual crossovers, and also stressed that variation in hot spot usage is extensive and heritable among individuals. The CCTCCCT-derived recombination motifs were found to mainly scatter in three regions (8–12, 17–24 and 27–30 kb; Figure 5), which coincide with the observed hot spots.⁴⁷ Population-specific polymorphic sites found in certain motifs can reflect the intensity of a hot spot.⁴⁸ This observation supports the fluidity of recombination hot spots. The recombination pattern is clearly different among the populations, but it is also confounded by selection and demographic history.⁷ It is of interest to use sperm typing or pedigrees for direct testing of population-specific hot spot usage, as well as for understanding the evolution of a recombination hot spot.^{46,49}

Tag SNP transferability of the HapMap populations was evaluated for association study.^{13,50–53} Most studies agreed that the population-specific HapMap tag SNPs can be applied to other populations,⁵⁴ but are inferior while being evaluated against resequencing data.^{52,53} Related studies indicated the existence of SNP-deposition bias in the dbSNP and HapMap database, for the majority of SNPs are common in frequency and shared among populations.¹⁶ To avoid this SNP ascertainment bias, resequencing data are more desirable for selection of tag SNPs and uncovering of the population substructure. Tag SNP evaluation with our resequencing data revealed that 16–54% of common SNPs for the six Chinese populations are not accounted for by the HapMap CHB-derived tag SNPs. In this 31-kb region, the genetic information of the Li and Mongolian populations is least covered by HapMap CHB tags, not only for active recombination in the region but also for the unique polymorphism distribution of the Li population and the highly polymorphic nature of the Mongolian population. Therefore, to improve this shortcoming while using Chinese populations without the knowledge of the underlying population substructure, we propose to raise the r^2 threshold for more tag SNP inclusion, to apply multipoint imputation methods for further information extraction,⁵⁵ and/or to use ethnic-mixed data for tag selection.^{13,53} Furthermore, if candidate genes (or regions) of an association study contain known recombination hot spots, direct sequencing of hot spots is a preferred way to ensure a proper SNP coverage of the area.¹²

The 31-kb DNA region in human major histocompatibility complex has the characteristic of being highly polymorphic, gene rich, subjected to selection and ragged in LD block. As it represents only a part of the genome, some of our conclusions, such as tag SNP transferability, may not apply to every genomic region. Our finding would be most useful for an association study with a candidate-gene approach on genomic regions that have shared characteristics with the *TNF* gene cluster. Our results also support the conclusion that there are distinct genetic compositions between Northern and Southern Chinese populations, with further divisions according to their geographic locations. By gaining knowledge of this underlying substructure of Chinese populations, we may improve the efficiency of disease mapping by Chinese-based association studies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation of China Grants 30470946 (Professor Wang) and by China Medical Board, CMB 04-805, CHINESE GENOMIC DIVERSITY (Professor Chu).

- 1 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
- 2 The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
- 3 Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S. *et al.* Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.* **40**, 584–591 (2008).
- 4 Livingston, R. J., von Niederhausern, A., Jegga, A. G., Crawford, D. C., Carlson, C. S., Rieder, M. J. *et al.* Pattern of sequence variation across 213 environmental response genes. *Genome Res.* **14**, 1821–1831 (2004).
- 5 Zhou, S. F., Liu, J. P. & Chowbay, B. Polymorphism of human cytochrome P450 enzymes and its clinical impact. *Drug Metab. Rev.* **41**, 89–295 (2009).
- 6 He, Y., Jiang, R., Fu, W., Bergen, A. W., Swan, G. E. & Jin, L. Correlation of population parameters leading to power differences in association studies with population stratification. *Ann. Hum. Genet.* **72**, 801–811 (2008).
- 7 McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. & Donnelly, P. The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
- 8 Li, J., Zhang, M. Q. & Zhang, X. A new method for detecting human recombination hotspots and its applications to the HapMap ENCODE data. *Am. J. Hum. Genet.* **79**, 628–639 (2006).
- 9 Barreiro, L. B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has driven population differentiation in modern humans. *Nat. Genet.* **40**, 340–345 (2008).
- 10 Zondervan, K. T. & Cardon, L. R. Designing candidate gene and genome-wide case-control association studies. *Nat. Protoc.* **2**, 2492–2501 (2007).
- 11 Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G. *et al.* Replicating genotype-phenotype associations. *Nature* **447**, 655–660 (2007).
- 12 The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
- 13 de Bakker, P. I., Burt, N. P., Graham, R. R., Guiducci, C., Yelensky, R., Drake, J. A. *et al.* Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38**, 1298–1303 (2006).
- 14 de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J. & Altshuler, D. Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223 (2005).
- 15 Teo, Y. Y., Small, K. S. & Kwiatkowski, D. P. Methodological challenges of genome-wide association analysis in Africa. *Nat. Rev.* **11**, 149–160 (2010).
- 16 Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005).
- 17 Cavalli-Sforza, L. L. The Chinese human genome diversity project. *Proc. Natl Acad. Sci. USA* **95**, 11501–11503 (1998).
- 18 Chu, J. Y., Huang, W., Kuang, S. Q., Wang, J. M., Xu, J. J., Chu, Z. T. *et al.* Genetic relationship of populations in China. *Proc. Natl Acad. Sci. USA* **95**, 11763–11768 (1998).
- 19 Deng, W., Shi, B., He, X., Zhang, Z., Xu, J., Li, B. *et al.* Evolution and migration history of the Chinese population inferred from Chinese Y-chromosome evidence. *J. Hum. Genet.* **49**, 339–348 (2004).
- 20 Ding, Y. C., Wooding, S., Harpending, H. C., Chi, H. C., Li, H. P., Fu, Y. X. *et al.* Population structure and history in East Asia. *Proc. Natl Acad. Sci. USA* **97**, 14003–14006 (2000).
- 21 Jin, L. & Su, B. Natives or immigrants: modern human origin in East Asia. *Nat. Rev.* **1**, 126–133 (2000).
- 22 Xue, Y., Zerjal, T., Bao, W., Zhu, S., Shu, Q., Xu, J. *et al.* Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics* **172**, 2431–2439 (2006).
- 23 Xue, F., Wang, Y., Xu, S., Zhang, F., Wen, B., Wu, X. *et al.* A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. *Eur. J. Hum. Genet.* **16**, 705–717 (2008).
- 24 Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics (Oxford, England)* **25**, 1451–1452 (2009).
- 25 Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
- 26 Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol. Bioinform Online* **1**, 47–50 (2005).
- 27 Ramos-Onsins, S. E. & Rozas, J. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* **19**, 2092–2100 (2002).
- 28 Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462 (2005).
- 29 Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
- 30 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England)* **21**, 263–265 (2005).
- 31 Biswas, S. & Akey, J. M. Genomic insights into positive selection. *Trends Genet.* **22**, 437–446 (2006).
- 32 The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
- 33 Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T. & Zhang, Y. P. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am. J. Hum. Genet.* **70**, 635–651 (2002).
- 34 Wen, B., Li, H., Lu, D., Song, X., Zhang, F., He, Y. *et al.* Genetic evidence supports demic diffusion of Han culture. *Nature* **431**, 302–305 (2004).
- 35 Deakin, J. E., Papenfuss, A. T., Below, K., Cross, J. G., Coggill, P., Palmer, S. *et al.* Evolution and comparative analysis of the MHC Class III inflammatory region. *BMC Genomics* **7**, 281 (2006).
- 36 Skoog, T., van't Hooft, F. M., Kallin, B., Jovinge, S., Boquist, S., Nilsson, J. *et al.* A common functional polymorphism (C→A substitution at position –863) in the promoter region of the tumour necrosis factor-alpha (TNF-alpha) gene associated with reduced circulating levels of TNF-alpha. *Hum. Mol. Genet.* **8**, 1443–1449 (1999).
- 37 Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nat. Rev. A*, 99–111 (2003).
- 38 Li, D. N., Ying, D. J. & Ou, C. Y. Establishment and preservation of immortal lymphoblastoid cell lines of the Ha-Li, the Qi-Li and the Bendi-Li ethnic groups in Hainan Island, China. *Hai Nan Yi Xue Yuan Xue Bao* **10**, 6–8 (2004).
- 39 Lin, S. Q. The History of resistance of Malaria in Hainan Island. *Hainan Med. J.* **2**, 70–72 (1986).
- 40 Verrelli, B. C., McDonald, J. H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A. *et al.* Evidence for balancing selection from nucleotide sequence analyses of human G6PD. *Am. J. Hum. Genet.* **71**, 1112–1128 (2002).
- 41 Timmann, C., Evans, J. A., Konig, I. R., Kleensang, A., Ruschendorf, F., Lenzen, J. *et al.* Genome-wide linkage analysis of malaria infection intensity and mild disease. *PLoS Genet.* **3**, e48 (2007).
- 42 Hill, A. V. The immunogenetics of resistance to malaria. *Proc. Assoc. Am. Physicians* **111**, 272–277 (1999).
- 43 Zhou, D. F., Wang, Z. & Cai, L. J. Investigation into the deletion types of alpha-thalassaemia in Han and Li Nationalities in Hainan Province. *CHINA Trop. Med.* **6**, 1549–1551 (2006).
- 44 McGuire, W., Hill, A. V., Allsopp, C. E., Greenwood, B. M. & Kwiatkowski, D. Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria. *Nature* **371**, 508–510 (1994).
- 45 Flori, L., Delahaye, N. F., Iraqi, F. A., Hernandez-Valladares, M., Fumoux, F. & Rihet, P. TNF as a malaria candidate gene: polymorphism-screening and family-based association analysis of mild malaria attack and parasitemia in Burkina Faso. *Genes Immun.* **6**, 472–480 (2005).
- 46 Coop, G., Wen, X., Ober, C., Pritchard, J. K. & Przeworski, M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* **319**, 1395–1398 (2008).
- 47 Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
- 48 Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40**, 1124–1129 (2008).
- 49 Evans, D. M. & Cardon, L. R. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am. J. Hum. Genet.* **76**, 681–687 (2005).
- 50 Gonzalez-Neira, A., Ke, X., Lao, O., Calafell, F., Navarro, A., Comas, D. *et al.* The portability of tagSNPs across populations: a worldwide survey. *Genome Res.* **16**, 323–330 (2006).
- 51 Service, S., Sabatti, C. & Freimer, N. Tag SNPs chosen from HapMap perform well in several population isolates. *Genet. Epidemiol.* **31**, 189–194 (2007).
- 52 Bhargava, T. R., Rieder, M. J. & Nickerson, D. A. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat. Genet.* **40**, 841–843 (2008).
- 53 Xu, Z., Kaplan, N. L. & Taylor, J. A. Tag SNP selection for candidate gene association studies using HapMap and gene resequencing data. *Eur. J. Hum. Genet.* **15**, 1063–1070 (2007).
- 54 Need, A. C. & Goldstein, D. B. Genome-wide tagging for everyone. *Nat. Genet.* **38**, 1227–1228 (2006).
- 55 Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).

Supplementary Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)