ARTICLE

Natural positive selection and north–south genetic diversity in East Asia

Chen Suo^{1,12}, Haiyan Xu^{1,12}, Chiea-Chuen Khor², Rick TH Ong^{1,2}, Xueling Sim¹, Jieming Chen², Wan-Ting Tay³, Kar-Seng Sim², Yi-Xin Zeng^{4,5}, Xuejun Zhang^{6,7}, Jianjun Liu², E-Shyong Tai^{8,9}, Tien-Yin Wong^{3,9,10}, Kee-Seng Chia^{1,8} and Yik-Ying Teo^{*,2,8,11}

Recent reports have identified a north–south cline in genetic variation in East and South-East Asia, but these studies have not formally explored the basis of these clinical differences. Understanding the origins of these variations may provide valuable insights in tracking down the functional variants in genomic regions identified by genetic association studies. Here we investigate the genetic basis of these differences with genome-wide data from the HapMap, the Human Genome Diversity Project and the Singapore Genome Variation Project. We implemented four bioinformatic measures to discover genomic regions that are considerably differentiated either between two Han Chinese populations in the north and south of China, or across 22 populations in East and South-East Asia. These measures prioritized genomic stretches with: (i) regional differences in the allelic spectrum for SNPs common to the two Han Chinese populations; (ii) differential evidence of positive selection between the two populations as quantified by integrated haplotype score (iHS) and cross-population extended haplotype homozygosity (XP-EHH); (iii) significant correlation between allele frequencies and geographical latitudes of the 22 populations. We also explored the extent of linkage disequilibrium variations in these regions, which is important in combining genetic association studies from North and South Chinese. Two of the regions that emerged are found in HLA class I and II, suggesting that the HLA imputation panel from the HapMap may not be directly applicable to every Chinese sample. This has important implications to autoimmune studies that plan to impute the classical HLA alleles to fine map the SNP association signals. *European Journal of Human Genetics* (2012) 20, 102–110; doi:10.1038/ejhg.2011.139; published online 27 July 2011

Keywords: positive selection; population genetics; clinal variation; linkage disequilibrium variation

INTRODUCTION

Several recent studies into the population genetics of Han Chinese have unveiled genetic evidence of population structure between northern and southern parts of China,¹ as well as identifying latitudinal clines in genetic variation across China.^{2,3} This is perhaps unsurprising, as numerous European and global studies^{4,5} have previously observed similar correlations between geographical latitudes and variations in the frequencies of alleles that are linked to several human phenotypes, including skin pigmentation^{6–8} salt sensitivity,^{9,10} lactose metabolism^{11,12} and even morphology.^{13–15} A recent bioinformatics investigation into the association between signatures of evolutionary adaptation and candidate genes for common metabolic syndromes also yielded strong evidence of spatially varying patterns of positive natural selection in several metabolic genes, as well as in several SNPs that were previously implicated with the ability to tolerate cold climates.^{16,17}

One striking observation made from the Singapore Genome Variation Project (SGVP), when integrated with genome-wide data from

East Asian populations in the Human Genome Variation Project (HGDP)^{18,19} and in phase 2 of the International HapMap Project (HapMap),²⁰ was that genomic variation in East and South-East Asia appears to follow a strong latitudinal cline (see Figure 1). The HGDP sampled from East and South-East Asian countries which included Cambodia, Japan and the Yakut tribe in East Siberia, as well as 15 distinct ethnic or population groups in China (see Figure 1a for the geographical distribution of the samples). Together with the South-East Asian Malay samples from SGVP (abbreviated MAS), Singapore Chinese with South China ancestries (CHS), Han Chinese from Beijing (CHB) and the Japanese from Tokyo (JPT), the latitudes of these 22 populations span between 3° and 63° north of the equator (Figure 1b). In a principal component analysis (PCA) of the genome-wide genotype data for these populations, the elements of the first axis of variation were found to reflect the latitude the samples originated from (Figure 1c). Although recent literature investigating the use of PCA in population genetics has highlighted the potential that clinical patterns may emerge in the absence of migration-linked gene flow and

¹Centre for Molecular Epidemiology, National University of Singapore, Singapore, Singapore; ²Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore; ³Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore; ⁴State Key Laboratory of Oncology in Southern China, Guangzhou, China; ⁵Department of Experimental Research, Sun Yat-Sen University Cancer Center, Guangzhou, China; ⁶Institute of Dermatology and Department of Dermatology at No. 1 Hospital, Anhui Province, China; ⁷The Key Laboratory of Gene Resource Utilization for Severe Disease, Ministry of Education and Anhui Province, Anhui Medical University, Anhui, China; ⁸Department of Epidemiology and Public Health, National University of Singapore, Singapore, Singapore; ⁹Department of Medicine, National University of Singapore, Singapore, Singapore; ¹⁰Centre for Eye Research Australia, University of Melbourne, Melbourne, Victoria, Australia; ¹¹Department of Statistics and Applied Probability, Faculty of Science, National University of Singapore, Singapore, Singapore

*Correspondence: Professor Y-Y Teo, Department of Statistics and Applied Probability, Faculty of Science, National University of Singapore, Block S16, Level 7, 6 Science Drive 2, Singapore 117546, Singapore. Tel: +65 6516 2760; Fax: +65 6872 3919; E-mail: statyy@nus.edu.sg

¹²These authors contributed equally to this work.

Received 20 January 2011; revised 25 May 2011; accepted 28 June 2011; published online 27 July 2011

Figure 1 Population structure in East and South-East Asian populations. **(a)** Geographical distribution of the 22 East and South-East Asian populations from the International HapMap Project, the Human Genome Diversity Project and the Singapore Genome Variation Project. The colors of the circles have been assigned according to the latitudes of the populations, following the blue–red spectrum with increasing latitude. **(b)** Names of the 22 population groups and their geographical coordinates, where the populations have been ranked according to their latitudes with the corresponding color codes that have been assigned. **(c)** Plot of the first two axes of variations from a principal components analysis of the genetic data from the 22 populations, the first axis of variation has been deliberately set as the vertical axis to reflect the correspondence between the scores of the first axis with latitude. Each circle represents an individual from one of the 22 populations, and the color of the circle defines the population membership according to the color scheme described in **a** and **b**.

is instead a consequence of isolation-by-distance^{21,22} (where gene flow happens between neighboring subgroups), this clinical pattern of genetic variation concurs with an independent finding from a recent pan-Asia study into the migration history across Asia, which revealed evidence of gene flow along a northern migratory route from South-East Asia into East Asia.²³

As a country that spans a considerable latitudinal range, China is one of the few countries that provide a useful model for studying the impact of latitude or geography on genetic variation because of the relative similarity in genetic and cultural histories across the different ethnic and population groups in the country. This is particularly true if the focus is on the Han Chinese ethnic group, which forms the largest population group in China and is the dominant ethnic group in southern provinces, such as Guangdong and Fujian, where the Chinese population in Singapore mainly originated from; in north-eastern provinces, such as Shandong and Jiangsu, where the trade and commerce center Shanghai is located in; and in northern provinces, such as Jilin, Liaoning and Hebei, where the capital, Beijing, is located in. Although genetic drift is likely to explain most of the subtle genetic variations in these populations, some of the larger differences between

North and South Chinese may be the result of evolutionary adaptations as a consequence of environmental influences, including the effects of seasonality and climate, agricultural distribution across the country, or varying prevalence of infectious diseases.

The advent of inexpensive large-scale genotyping across the human genome offers unprecedented opportunities to survey interpopulation genetic variation, particularly when integrated with the suite of statistical and bioinformatics tools that are available for assessing population differences. At the SNP level, the Wright's²⁴ F_{ST} offers a single metric for quantifying the variation in allele frequencies, whereas sophisticated methodologies, such as the iHS ²⁵ and $XP-EHH$ ²⁶ statistics, for identifying the putative genomic signatures of positive natural selection allow interpopulation comparisons to be made at the haplotypic level. Here we leverage on these bioinformatic approaches to discover genomic regions that are most differentiated (i) between North and South Chinese; or (ii) across 22 populations in East and South-East Asia, subject to the condition that these regions exhibit consistent evidence across several bioinformatic metrics. In addition, we also investigate the extent of linkage disequilibrium (LD) variations in these regions, which have downstream implications on

integrating data from genetic association studies from North and South Chinese.

MATERIALS AND METHODS

Datasets

Our analyses relied on genome-wide genotype data from three primary sources: (i) the East Asian panel of phase 2 of the International HapMap Project (abbreviated subsequently as HapMap);²⁰ (ii) the HGDP;^{18,19} (iii) the SGVP.¹ The data from the HapMap consists of 3 821 888 autosomal SNPs that have been genotyped in 45 unrelated Han Chinese individuals from Beijing located in North-East China (abbreviated CHB) and 45 unrelated Japanese individuals from Tokyo (abbreviated JPT). Of the 1074 samples in the HGDP that are assayed on the Illumina HumanMap 650K BeadChip (Illumina, San Diego, CA, USA), we only considered the 228 unrelated samples from 18 population groups in East and South-East Asia. The SGVP database consists of 268 unrelated individuals from three population groups in Singapore that have been assayed on both the Affymetrix SNP6.0 (Affymetrix, Santa Clara, CA, USA) and Illumina 1M arrays. Our current analyses only consider the 96 Han Chinese individuals with ancestries originating from southern China (abbreviated CHS), and the 89 Malay individuals with ancestries from Peninsula Malaysia and Indonesia (abbreviated MAS, see reference 1 for a detailed description of the CHS and MAS samples), where 1 584 040 and 1 580 905 autosomal SNPs remained after quality checks, respectively. To validate the findings on the correlation between allele frequencies and latitudes, the genotype data of Chinese control samples from four independent genome-wide association studies conducted in Singapore (2434 Chinese population controls from the Singapore Prospective Study Program^{27,28} and 2542 Malay population controls from the Singapore Malay Eye Study),^{29,30} Guangzhou (980 control samples)² and Shandong province (181 control samples)² were used.

Analysis with 22 East and South-East Asian populations

Correlation between allele frequencies and latitude. To identify clinical variations in allele frequencies, we calculated the Pearson correlation coefficient R between the allele frequencies of each SNP and the geographical latitudes of the 22 populations at the 610 437 autosomal SNPs that are common across the HGDP, HapMap and SGVP databases. These populations consist of the 18 groups in East and South-East Asia from HGDP, the two East Asian populations from HapMap (CHB, JPT), and the Chinese (CHS) and Malay (MAS) samples from SGVP. The geographical locations (latitudes and longitudes) for the samples from HGDP are available online (<http://www.cephb.fr/en/hgdp/table.php>), whereas for the HapMap populations, we used the latitudes corresponding to Beijing and Tokyo. As the Chinese samples in Singapore are descended mainly from migrants originating from the Fujian and Guangdong provinces in China, we took the average of the latitudes for these provinces. The latitude for the Malay samples was obtained as the average latitude between Malaysia and Singapore. The P -value for the Pearson correlation coefficient R between the allele frequencies and latitudes for the 22 populations is calculated with the test statistic

$$\frac{R}{\sqrt{(1-R^2)/20}}$$

which follows an approximate Student's t -distribution with 20° of freedom.

Population structure analysis with PCA. For the 22 populations (18 from HGDP, 2 from HapMap and 2 from SGVP), we selected a thinned set of 101 704 SNPs out of the 610 437 common autosomal SNPs by choosing every sixth SNP in order to minimize the use of correlated SNPs. We performed an eigenanalysis on this set of thinned SNPs with the *pca* option that is distributed as part of the *eigenstrat* software.³¹ To calculate the contribution of each SNP to the resultant principal components from the eigenanalysis, suppose the genotype of individual j at SNP i is defined as $g_{ij} \in \{0, 1, 2, \text{NULL}\}$. Let g_{ij}' denote the normalized genotype, calculated as $(g_{ij} - \bar{g}_i) / \sqrt{p_i(1-p_i)}$, where $\bar{g}_i$ denotes the average of g_{ij} across the individuals with non-NULL genotypes and p_i denotes the allele frequency for SNP i . The loadings for SNP i for the k th principal component, γ_i^k , is subsequently calculated as $\gamma_i^k = \sum_j a_{ij}^k g_{ij}'$, where a_{ij}^k is the corresponding element for individual j for the k th principal component.

We do not use the SNP loadings for discovering regions of interest, but only as an additional source of evidence to corroborate the findings at interesting regions identified by the other metrics. We cross-reference every region that has been identified by the four approaches by checking whether there is at least one SNP in the region that lies in the top 0.1 or 0.5% of the distribution of the SNP loadings across the genome.

Comparisons between two populations in North and South China

Quantifying north-south population variation in China with F_{ST} . To assess whether there are considerable differences in the allelic architecture between populations with ancestries that are predominantly found in North China (CHB) and South China (CHS), we quantified the extent of the disparity in the allele frequencies at each SNP with the F_{ST} statistic.²⁴ There are a total of 1 248 469 autosomal SNPs that are common between CHB and CHS, and the SNP level F_{ST} is calculated as

$$F_{ST} = \frac{(p_1 - p_2)^2}{(p_1 + p_2)(2 - p_1 - p_2)}$$

following Rosenberg *et al*³² for two populations, where p_1 and p_2 denote the allele frequencies of a chosen allele at a particular SNP in CHB and CHS, respectively.

North-south variation in signatures of positive natural selection. We used the iHS statistic²⁵ and the XP-EHH metric²⁶ to identify genomic signatures of positive natural selection in the CHB and CHS samples. The software used in the iHS and XP-EHH calculations are downloaded from <http://hgdp.uchicago.edu/Software/>.³³

The iHS calculations are performed independently in each of the two populations, except that the iHS analysis of CHB is performed on a similar set of SNPs that the CHS database contains, to avoid differential signals that are attributed entirely to different SNP densities from the HapMap and SGVP databases. We used the recombination rates that are averaged across all the four HapMap phase 2 populations, and we normalized the raw iHS statistics in 20 derived allele frequency bins, each spanning 5%. The iHS signals are used to discover regions of interest if the iHS score in either one population is found in the top 0.1% but not in the top 1% of the other population.

The XP-EHH analysis was performed on the set of 1 102 122 SNPs common to CHB and CHS, and the resultant XP-EHH statistics were subsequently normalized to have a zero mean and unit variance. A clustering of SNPs displaying large positive values of the normalized XP-EHH statistic suggests that a selection event is likely to have occurred in the first population (CHB) relative to the second population (CHS), whereas a clustering of large negative values suggests a selection event is likely to have occurred in the second population relative to the first population. As such, we used the XP-EHH analysis between CHB and CHS to identify regions of interest, defined as regions with normalized XP-EHH signals in the top 0.01% of either tails of the genome-wide distribution of the XP-EHH scores, and noting the direction of these signals as this indicates whether the candidate selection event occurred in CHB or CHS.

Additional methods on quantifying interpopulation LD differences and further details of quantifying regional evidence of: (i) the correlation between allele frequencies and geographical latitude; and (ii) high F_{ST} can be found in the Supplementary Material.

RESULTS

We used four mechanisms to discover genomic regions experiencing north-south clinical genetic variation in the East Asian populations from HapMap, HGDP and SGVP: (i) stretches of high F_{ST} SNPs between the 1 248 469 SNPs that are common to the HapMap Han Chinese from Beijing (CHB) and the Singapore Chinese samples with genetic ancestries from South China (CHS); (ii) regional evidence of SNPs found in the 22 East and South-East Asian populations where the allele frequencies are significantly correlated with the corresponding latitudes of the populations; (iii) genomic stretches where there are significant evidence of differential positive natural selection signals between CHB and CHS, when assessed using the XP-EHH metric;

Table 1 A description of the bioinformatic metrics used to discover and validate genomic regions that are differentiated along a north–south cline

Criteria	Populations	Discovery criterion	Validation criterion
F_{ST} Overrepresentation of SNPs in a genomic region with high F_{ST} relative to genome-wide distribution of F_{ST} scores	CHB vs CHS	Top 0.1% of genome-wide distribution of regional evidence where: – Region defined by window sizes of 100 and 500 kb – Evidence defined by the P -value of the exact Binomial test for the proportion of SNPs with F_{ST} in the top 1st and 0.1st percentile, respectively, of the genome-wide distribution of F_{ST} scores	Discovered region containing regional evidence found in the top 1% of the genome-wide distribution
Correlation between allele frequency and latitude	In all, 22 East and South-East Asian population groups from HapMap, HGDP and SGVP	Top 0.1 or 0.5% of genome-wide distribution of regional evidence where: – Region defined by window size of 500 kb – Evidence defined by the P -value of the exact Binomial test for the proportion of SNPs with Pearson correlation coefficient P -values $<10^{-4}$	Existence of at least one SNP in discovered region with Bonferroni corrected P -value for the Pearson correlation coefficient test <0.05
XP-EHH	CHB vs CHS	Top 0.01% of the genome-wide distribution of the normalized XP-EHH scores	Existence of at least one SNP in discovered region in the top 0.5% of genome-wide distribution of the normalized XP-EHH scores
Differential signals of iHS for CHB and CHS iHS calculated independently from CHB and CHS genotype data, with SNPs for CHB thinned to similar density as CHS	CHB vs CHS	SNP with iHS score in top 0.1% of the normalized genome-wide distribution in first population, but absent in the top 1% of normalized genome-wide distribution in second population	Discovered region containing at least one SNP with iHS score in top 1% of normalized genome-wide distribution, but absent in the top 1% of normalized genome-wide distribution in second population
PCA SNP loadings for first axis of variation from Figure 1	In all, 22 East and South-East Asian population groups from HapMap, HGDP and SGVP	No discovery mechanism from this	Existence of at least one SNP in discovered region with PCA SNP loadings at least in the top 0.5% of the genome-wide distribution

Abbreviations: CHB, Han Chinese from Beijing; CHS, Singapore Chinese with South China ancestries; HGDP, Human Genome Variation Project; iHS, integrated haplotype score; PCA, principal component analysis; SGVP, Singapore Genome Variation Project; XP-EHH, cross-population extended haplotype homozygosity. The populations that each metric is applied on are also stated.

(iv) genomic regions where there are conflicting evidence of positive natural selection when assessed using the iHS metric in CHS and CHB. To avoid spurious findings from the use of a single discovery metric, we require each identified region to be supported by evidence from at least one of the other metrics, or to contain SNPs that are found to contribute significantly to the north–south cline as evident in the first axis of the principal component analysis in Figure 1 (see Table 1 for a summary of discovery and validation metrics, and Materials and Methods for the details of these metrics).

Clinical variation in allele frequencies with latitude

In the discovery phase, we identified five regions with an overrepresentation of SNPs exhibiting evidence of correlation (defined as a Pearson test of correlation P -value $<10^{-4}$) between allele frequencies and the latitudes of 22 populations (see Table 2, Figure 2 and Supplementary Figures S1–S5). Each of these five regions displayed concordant evidence of population differentiation between northern and southern Chinese populations in at least one other validation metric, which perhaps unsurprisingly, almost always included SNPs with high loadings for the first axis of variation in the PCA from Figure 1 (Table 2).

One of the two regions in the top 0.1% of the genome-wide distribution spans a series of *HLA* genes between 32.61 and 33.11 Mb in class II of the major histocompatibility complex (MHC) region on chromosome 6, including *-DRB1*, *-DQA1*, *-DQA2*, *-DOB*, *-DMB*, *-DMA* and *-DOA*. Our analysis of this region

reveals strong evidence of positive natural selection in both Han Chinese populations from Beijing (CHB) and Singapore (CHS), with iHS metrics in the top 0.01% of the genome-wide distributions for each of these two populations (Supplementary Figure S1), as well as concordant evidence from both XP-EHH and F_{ST} . The other region identified in the top 0.1% spans the *NRG1* gene, and exhibited evidence of positive natural selection in both northern and southern Chinese with both iHS and XP-EHH (Supplementary Figure S2). The emergence of this region is perhaps unsurprising, as a detailed survey of the genetic variation at this gene in 39 populations has previously revealed significant differences in the frequency spectrum of alleles and haplotypes in intronic SNPs, which correlated with the geographical locations of the 39 populations.³⁴ This region similarly emerged as one of the top regions in the human genome exhibiting evidence of regional variation in patterns of LD when assessed across all the HapMap phase 2 populations.³⁵

One of the three regions found in the top 0.5% encompasses a cluster of genes between 39.04 and 39.54 Mb on chromosome 3 (Supplementary Figure S3) with associations to phenotypes and functions such as tumor suppression (*TTC21A*, *AXUD1* and *LAMR1*), HIV progression with immunological tolerance and inflammation roles (*CX3CR1*), pyridoxine-refractory sideroblastic anemia in humans, while functionally responsible for anemic phenotype in an animal model with zebrafish embryos (*SLC25A38*), and a hereditary cardiomyopathy (arrhythmogenic right ventricular dysplasia) that

Table 2 Regions identified across the genome which contains an overrepresentation of SNPs that exhibit strong correlations between allele frequencies and latitude in 22 East and South-East Asian populations in the HapMap, HGDP and SGVP

Chr	Start (Mb)	End (Mb)	MAF latitude correlation P ^a (rsID)	F _{ST} (CHB vs CHS)	XP-EHH ^b (direction)	iHS (CHB)	iHS (CHS)	SNP loadings (rsID)	Genes
<i>Top 0.1%</i>									
6	32.610	33.110	2.1×10 ⁻⁵ (rs6901084)	Top 0.5%	Top 0.5% (positive)	Top 0.01%	Top 0.01%	Top 0.1% (rs9268832)	<i>HLA-DRB1, HLA-DQA1-2, HLA-DOB, PSMB9, BRD2, TAP2, PSMB8, TAP1, HLA-DMB, HLA-DMA, HLA-DOA</i>
8	32.155	32.655	2.0×10 ⁻⁴ (rs4489283)	No evidence	Top 0.5% (positive)	Top 0.5%	Top 0.5%	Top 0.1% (rs4489283)	<i>NRG1</i>
<i>Top 0.5%</i>									
3	39.038	39.538	6.6×10 ⁻⁵ (rs2370969)	No evidence	Top 0.1% (negative)	Top 0.5%	Top 0.1%	Top 0.1% (rs1464047)	<i>WDR48, GORASP1, TTC21A, AXUD1, CMYA1, CX3CR1, CCR8, SLC25A38, LAMR1, MOBP</i>
3	136.038	136.538	9.3×10 ⁻⁴ (rs6762261)	No evidence	No evidence	Top 0.1%	Top 0.5%	Top 0.5% (rs6788931)	<i>EPHB1</i>
6	18.610	19.110	9.5×10 ⁻⁴ (rs986148)	No evidence	Top 0.1% (positive)	Top 0.1%	No evidence	No evidence	NA

Abbreviations: CHB, Han Chinese from Beijing; CHS, Singapore Chinese with South China ancestries; HGDP, Human Genome Variation Project; iHS, integrated haplotype score; SGVP, Singapore Genome Variation Project; XP-EHH, cross-population extended haplotype homozygosity.

^aBonferroni corrected *P*-value for the test of correlation between allele frequencies and latitude of the 22 East and South-East Asian populations from HapMap, HGDP and SGVP. The Bonferroni correction is performed by multiplying the empirical *P*-value by the number of SNPs found in each region.

^bXP-EHH between CHB and CHS, with positive indicating evidence of positive selection in CHB, whereas negative indicating evidence of positive selection in CHS.

The table highlights the genomic stretches found in the top 0.1 and 0.5% of the genome-wide distribution for regional evidence of clinical variation in allele frequencies that are supported by concordant information from the SNP loadings of the first axis of variation in a principal component analysis of the 22 populations and from other bioinformatic evidences from the comparisons between CHB and CHS (F_{ST}, XP-EHH and differential signals of iHS). For each region, the SNP with the strongest evidence of MAF latitude correlation is reported.

Figure 2 Genomic regions identified with evidence of clinical genetic variation. Five regions emerged with regional evidence of significant correlations between the allele frequencies of SNPs and the geographical latitudes of 22 East and South-East Asian populations, according to the order as described in Table 2: (a) across the *HLA* gene cluster in class II of the MHC on chromosome 6; (b) the region on chromosome 8 encompassing the *NRG1* gene; (c) between 39.04 and 39.54 Mb on chromosome 3 encompassing a cluster of genes; (d) the region on chromosome 3 encompassing the *EPHB1* gene; (e) a gene desert between 18.61 and 19.11 Mb on chromosome 6. SNPs with correlation *P*-values less significant than 10⁻⁴ are represented by blue circles, while yellow diamonds represent SNPs with 10⁻⁵ ≤ *P*-values < 10⁻⁴; orange diamonds represent SNPs with 10⁻⁶ ≤ *P*-values < 10⁻⁵; red diamonds represent SNPs with *P*-values ≤ 10⁻⁶. The SNPs exhibiting the strongest evidence of clinical variation in allele frequencies and SNP loadings of the first axis of variation in the PCA are also shown. Green bars at the top of each plot indicate the locations of genes in the region, and horizontal dotted lines linking to each bar indicate that the gene spans beyond the region shown in the figure.

Table 3 Regions identified across the genome by different discovery mechanisms using the three bioinformatic metrics calculated from the CHB and CHS genome-wide data from HapMap and SGVP

Discovery mechanism	Chr	Start (Mb)	End (Mb)	F_{ST}^a (window size)	XP-EHH ^b (direction)	iHS (CHB)	iHS (CHS)	MAF latitude correlation P^c (rsID)	SNP loadings (rsID)	Genes
iHS	3	189.512	190.012	No evidence	Top 0.5% (positive)	Top 0.01%	No evidence	2.5×10^{-3} (rs16863396)	Top 0.5% (rs3817462)	<i>LPP</i>
F_{ST} , iHS	4	100.552	101.052	Top 0.1% (100 kb, 500 kb)	Top 0.5% (positive)	Top 0.1%	No evidence	7.7×10^{-3} (rs13150247)	Top 0.1% (rs13150247)	<i>ADH</i> gene cluster, <i>RG9MTD2</i> , <i>MTTP</i> , <i>DAPP1</i> , <i>MAP2K1IP1</i> , <i>DNAJB14</i>
iHS	6	18.610	19.110	No evidence	Top 0.1% (positive)	Top 0.1%	No evidence	9.5×10^{-4} (rs986148)	No evidence	NA
F_{ST} , XP-EHH	6	29.795	29.895	Top 0.01% (100 kb, 500 kb)	Top 0.01% (negative)	No evidence	No evidence	1.3×10^{-2} (rs1633021)	Top 0.1% (rs3131020)	<i>HLA-F</i> , <i>HLA-G</i>
F_{ST}	11	61.189	61.689	Top 0.1% (100 kb, 500 kb)	Top 0.1% (positive)	Top 0.5%	No evidence	No evidence	Top 0.5% (rs1495941)	<i>FEN1</i> , <i>FADS1-3</i> , <i>RAB3IL1</i> , <i>BEST1</i> , <i>FTH1</i> , <i>INCENP</i>
XP-EHH	12	71.358	73.069	Top 0.01% (100 kb)	Top 0.01% (negative)	No evidence	Top 0.5%	8.6×10^{-3} (rs2102755)	Top 0.5% (rs10879537)	NA
F_{ST} , XP-EHH	13	97.957	98.957	Top 0.1% (100 kb, 500 kb)	Top 0.01% (negative)	No evidence	Top 0.5%	9.1×10^{-2} (rs11069349)	No evidence	<i>STK24</i> , <i>SLC15A1</i> , <i>DOCK9</i> , <i>PHGDHL1</i> , <i>GPR18</i> , <i>EBI2</i>

Abbreviations: CHB, Han Chinese from Beijing; CHS, Singapore Chinese with South China ancestries; HGDP, Human Genome Variation Project; iHS, integrated haplotype score; SGVP, Singapore Genome Variation Project; XP-EHH, cross-population extended haplotype homozygosity.

^aRegional evidence from the F_{ST} metric, where the size of the region containing evidence is defined in the parentheses.

^bXP-EHH between CHB and CHS, with positive indicating evidence of positive selection in CHB, whereas negative indicating evidence of positive selection in CHS.

^cBonferroni corrected P -value for the test of correlation between allele frequencies and latitude of the 22 East and South-East Asian populations from HapMap, HGDP and SGVP. The Bonferroni correction is performed by multiplying the empirical P -value by the number of SNPs found in each region.

These metrics utilizing the discovery populations of CHB and CHS are described in Table 1.

causes sudden death in the young.³⁶ Another region on chromosome 3 (136.04–136.54 Mb, see Supplementary Figure S4) encompasses the ephrin receptor *EPHB1* where a strong correlation was established between EphB expression and degree of malignancy in colorectal cancer progression.³⁷ The region identified on chromosome 6 was particularly intriguing given the absence of any genes in the vicinity (Supplementary Figure S5), as there were consistent evidence of positive selection occurring in North Chinese compared with South Chinese represented by a positive XP-EHH signal in the top 0.1% and an iHS signal in the top 0.1% in CHB, but absent even in the top 1% of the CHS signals.

Population differentiation between CHB and CHS

The availability of larger sample sizes from the Chinese populations in HapMap (45 CHB samples) and SGVP (96 CHS samples) allows the use of population genetics metrics to quantify the differences in the allelic spectrum and genomic signatures of positive natural selection between the two populations. By prioritizing genomic regions that emerged with consistent evidence of extreme differentiation between the two populations, we identified seven regions, of which the region on chromosome 6 between 18.61 and 19.11 Mb was previously seen with strong evidence of a latitudinal cline in allele frequency variation (see Table 3, Figure 3, and Supplementary Figures S7–S11).

Of the six additional regions, the region on chromosome 3 between 189.51 and 190.01 Mb encompassed the lipoma-preferred partner (*LPP*) gene that was recently implicated with celiac disease in numerous studies^{38–40} and was previously reported to have an important role in tumor metastasis,^{41–43} including in acute myeloid leukemia.^{44,45} This region displayed consistent evidence of differential signals of positive natural selection that was only present in CHB and not in CHS (Figure 3a), an observation that was corroborated by the XP-EHH signals in the East Asian population groups from the HGDP Selection Browser (<http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/>),³³ which displayed stronger evidence of positive selection in

the populations from the north (Supplementary Figure S6). The discovered region also contained several SNPs, including rs16863396 (Figure 3b), that displayed significant evidence of a latitudinal cline in allele frequency variation (for rs16863396: empirical P -value = 1.6×10^{-5} , Bonferroni corrected P -value = 2.5×10^{-3}). The latter observation of the latitudinal cline in allele frequency variations was supported even after the inclusion of four additional populations with considerably larger sample sizes that are located at latitudes of between 3° north (Peninsula Malaysia) and 37° north (Shandong province; empirical P -value = 9.2×10^{-6} ; Figure 3b).

Another region that emerged with strong evidence from two discovery mechanisms (F_{ST} , iHS), demonstrating signs of positive selection in CHB in the top 0.1% of the iHS signals across the genome but not even in the top 1% in CHS, encompassed the cluster of genes responsible for alcohol metabolism (alcohol dehydrogenase *ADH* gene cluster) on chromosome 4 (100.55–101.05 Mb). Strong corroborating evidence was observed from all other metrics (Table 3, Supplementary Figure S7), with the same SNP (rs13150247) observed to contribute significantly to the SNP loadings of the first PC in Figure 1 and also to display consistent evidence of a latitudinal cline in allele frequencies (empirical P -value = 7.3×10^{-5} , Bonferroni corrected P -value = 7.7×10^{-3} , Supplementary Figure S7).

The *HLA-F* and *HLA-G* region in class I of the MHC on chromosome 6 also emerged as a region with numerous high F_{ST} SNPs and with XP-EHH signals in the top 0.01% of the genome (Table 3, Supplementary Figure S8). Two other intronic regions on chromosomes 11 and 13 were similarly identified with consistent evidence of population differentiation between CHB and CHS by F_{ST} and XP-EHH (Supplementary Figures S9, S11). The former region is putatively selected in CHB and encompasses genes implicated in cancer pathogenesis (*FEN1*)^{46,47} and iron metabolism (*FTH1*);^{48,49} the latter region appears to be selected in CHS and contains the genes involved in pancreatic cancer inhibition (*SLC15A1*)⁵⁰ and bipolar disorder (*DOCK9*).⁵¹

Figure 3 Evidence of genetic differentiation between CHB and CHS around the *LPP* gene on chromosome 3. (a) Evidence of population differentiation between CHB and CHS from three discovery mechanisms looking at differential evidence of positive natural selection from iHS (top panel); regional clustering of SNPs with considerably different allelic spectrum between CHB and CHS (as quantified by the F_{ST} metric) relative to the genome, where the top 0.5% of the F_{ST} distribution corresponds to an empirical F_{ST} score of 2.7, top 0.1% corresponds to an empirical F_{ST} of 3.8% and the top 0.01% corresponds to an empirical F_{ST} of 17.0% (middle panel); XP-EHH signals comparing CHB and CHS that are found in either tails of the genome-wide distribution (bottom panel), with the diamonds representing signals in the top 0.5% (yellow), top 0.1% (orange) and top 0.01% (red) of the distribution. (b) Scatter plot of the frequencies of allele A for rs16863396, located at 189715374 bp on chromosome 3, across 22 populations in East and South-East Asia. The size of each circle represents the sample size of the population, and the color follows the assignment in Figure 1. The Pearson correlation and the corresponding P -value are calculated from the 22 populations. Four additional independent populations are shown in circles with decreasing shades of gray (with increasing latitude) for validating the clinical relationship between allele frequency and latitude.

DISCUSSION

The availability of at least 1.25 million SNPs, that is common to CHB and CHS, offered unprecedented opportunities to survey the genetic landscape between two Han Chinese population groups with genetic ancestries from North and South China. By including the 18 East Asian populations from HGDP, the HapMap Japanese samples and the South-East Asian Malays, we have a unique opportunity to survey the genetic variability in East and South-East Asia that is directly correlated to geography, an observation that has been reported in several similar studies performed in Europe,^{52–54} the Pacific islands,⁵⁵ East Asia,^{1,2,56} South Asia⁵⁷ and Africa.⁵⁸ Regions that emerged in our survey include the alcohol dehydrogenase (*ADH*) gene cluster, the *HLA* regions in the MHC, and the regions on chromosomes 3 and 8 that encompass the genes *LPP* and *NRG1*, respectively (see Supplementary Material for additional discussion on these regions).

The observation of a north–south cline in genetic variation in China by us¹ and others^{2,3,23} was made with the use of autosomal SNPs. This appears to be discordant with earlier findings from the use

of mitochondrial DNA (mtDNA) and chromosome Y (chrY), which established a more complex migration pattern across China,^{59,60} including a west–north passage,⁶¹ a east–west passage⁶² and a postglacial migration into East Asia from the north.⁶³ The inference on migration and population demography with mtDNA and chrY is expected to be superior to the use of autosomal SNPs, as the lack of recombination allows the genealogy of individuals from different populations to be estimated more accurately. However, although there have been numerous reports on the complexity of the probable migration patterns, we noticed that even the literature from mtDNA and chrY is consistent in reporting the genetic diversity along a south–north migration cline.^{60,62,64–67} In this article, we specifically focus on identifying the genomic regions that exhibited the strongest evidence of north–south diversity rather than to infer any migration and demographic patterns.

The analyses with the five bioinformatic metrics discovered 11 regions that were substantially differentiated between North and South Chinese populations. A natural extension is to evaluate the

implications of these differences in medical genetics. We observed that all 11 regions displayed evidence of LD variation between CHB and CHS in the extreme 5% of the genome-wide distribution of LD differences, as quantified by the varLD statistic (see Supplementary Material and Table S1). The current strategy in genome-wide association studies aims to replicate the lead SNP exhibiting the strongest signal from each region in other populations. Regions containing strong evidence of LD variation between two populations have previously been found to exhibit larger differences in the statistical evidence at the index SNPs,³⁵ which can confound meta-analyses of association studies from North and South Chinese populations. Conversely, fine mapping the unknown functional polymorphisms in these regions are likely to be more successful, as the different LD patterns are likely to imply the presence of different core haplotypes that are carrying the functional allele.⁶⁸ Leveraging on these diverse haplotype patterns is expected to be an important feature when attempting to localize the possible candidates for the causal variants, as long-range LD that has benefited the discovery phase of GWAS is likely to confound the fine mapping phase by producing numerous perfect surrogates that are statistically indistinguishable from the true causal variant.

We have used three different bioinformatic metrics that are commonly used in population genetics to quantify population differences and identify signatures of positive natural selection. Two additional metrics looked at clinical patterns of genetic differentiation across 22 populations, as assessed by the correlation between allele frequencies and geographical latitudes, and by identifying SNPs that possess higher loadings in a PCA of genetic variation across these populations (Supplementary Table S2). Although the sample sizes in HGDP are particularly small for certain population groups for accurate inference of the allele frequencies, we have used four independent cohorts from large-scale genetic studies to validate the findings of geographical clines in the allele frequencies of the discovered SNPs.

One caveat with the use of these mechanisms for discovery and validation is that these metrics essentially prioritized regions in the tail of the genome-wide distributions, and the regions that emerged may not necessarily be functionally important or relevant. However, given that there is clear evidence of genetic variation between these populations from previous studies, we have sought to discover the genomic regions that may explain these interpopulation differences. In searching for regional evidence of population differences, we have searched for an overrepresentation of SNPs within each genomic region that either displayed high F_{ST} values or exhibited strong correlations between the allele frequencies and the latitudes. Although this avoids the problem of false positives introduced from isolated SNPs displaying strong evidence of population differentiation, the approach to search for a clustering of SNPs with strong evidence may inevitably be confounded by the presence of LD. However, as we require concordant evidence from multiple metrics, including iHS and XP-EHH, which use genetic distances for calculating the test statistics, and are thus more robust to effects of LD, we do not expect the regions that have emerged to be artifacts due to LD. A recent article describing a composite metric for identifying regions undergoing positive selection also showed that correlations between F_{ST} , iHS and XP-EHH are generally weak even in selected regions, particularly with increasing distance from the causal polymorphism.⁶⁹ This further suggests it is unlikely our findings are due to chance occurrences of the same regions appearing in the tail of the distributions. However, it is important to recognize that these bioinformatic measures only provide an approach to prioritize genomic regions for downstream investigations, and our approach is

not meant to provide conclusive evidence on the biological relevance and consequences.

This study has extended previous observations of geography-linked genetic variation to East and South-East Asia, and through a systematic survey of population genetics data from two Han Chinese populations, identified genomic regions that contribute to explain the observed north–south cline in genetic differences in China. Although most of the findings are association driven, this study highlights the potential of integrating genomic evidence at the level of population and evolutionary genetics for the science of anthropology, and in mapping the geographical variations in the incidences of diseases and complex human traits.⁷⁰ With considerable variance in the incidences of major diseases across the different geographical regions,⁷¹ China presents a unique opportunity for exploring the effects of geography and climate on human genetics. The increasing availability of genome-wide data for multiple populations worldwide, including China, may finally herald the progression from anecdotal and observational evidence of population differences toward a more precise quantification of the genetic basis behind interpopulation variations.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for their constructive comments, which have greatly improved the article. This project acknowledges the support of the Yong Loo Lin School of Medicine from the National University of Singapore, National Medical Research Council, 0796/2003, Singapore and the Biomedical Research Council, 09/1/35/19/616, Singapore. The study used data generated by the International HapMap Consortium, the Singapore Genome Variation Project and the Human Genome Diversity Project. YYT acknowledges support from the National Research Foundation, NRF-RF-2010-05, Singapore.

AUTHOR CONTRIBUTIONS

YYT and KSC jointly conceived, designed and directed the experiment; YYT, CS and HX wrote the paper; YYT, CS, HX, XS, JC, RTHO and KSS analyzed the data; YXX, XZ, JL, EST and TYW contributed samples.

- 1 Teo YY, Sim X, Ong RT *et al*: Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res* 2009; **19**: 2154–2162.
- 2 Chen J, Zheng H, Bei JX *et al*: Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am J Hum Genet* 2009; **85**: 775–785.
- 3 Xu S, Yin X, Li S *et al*: Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am J Hum Genet* 2009; **85**: 762–774.
- 4 Beckman G, Birgander R, Sjalander A *et al*: Is p53 polymorphism maintained by natural selection? *Hum Hered* 1994; **44**: 266–270.
- 5 Cavalli-Sforza LL, Menozzi P, Piazza A: *History and Geography of Human Genes*. Princeton University Press: Princeton, New Jersey, 1994.
- 6 Jablonski NG, Chaplin G: The evolution of human skin coloration. *J Hum Evol* 2000; **39**: 57–106.
- 7 Lamason RL, Mohideen MA, Mest JR *et al*: SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 2005; **310**: 1782–1786.
- 8 Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M: Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Ann Hum Genet* 2007; **71**: 354–369.
- 9 Thompson EE, Kuttub-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A: CYP3A variation and the evolution of salt-sensitivity variants. *Am J Hum Genet* 2004; **75**: 1059–1069.
- 10 Young JH, Chang YP, Kim JD *et al*: Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet* 2005; **1**: e82.
- 11 Bersaglieri T, Sabeti PC, Patterson N *et al*: Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004; **74**: 1111–1120.
- 12 Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG: The origins of lactase persistence in Europe. *PLoS Comput Biol* 2009; **5**: e1000491.
- 13 Allen JA: The influence physical conditions in the genesis of species. *Radical Rev* 1877; **1**: 108–140.

- 14 Katzmarzyk PT, Leonard WR: Climatic influences on human body size and proportions: ecological adaptations and secular trends. *Am J Phys Anthropol* 1998; **106**: 483–503.
- 15 Roberts DF: Body weight, race and climate. *Am J Phys Anthropol* 1953; **11**: 533–558.
- 16 Hancock AM, Witonsky DB, Gordon AS *et al*: Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet* 2008; **4**: e32.
- 17 Novembre J, Di Rienzo A: Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet* 2009; **10**: 745–755.
- 18 Li JZ, Absher DM, Tang H *et al*: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; **319**: 1100–1104.
- 19 Rosenberg NA, Pritchard JK, Weber JL *et al*: Genetic structure of human populations. *Science* 2002; **298**: 2381–2385.
- 20 Frazer KA, Ballinger DG, Cox DR *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
- 21 Novembre J, Stephens M: Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 2008; **40**: 646–649.
- 22 Reich D, Price AL, Patterson N: Principal component analysis of genetic data. *Nat Genet* 2008; **40**: 491–492.
- 23 Abdulla MA, Ahmed I, Assawamakin A *et al*: Mapping human genetic diversity in Asia. *Science* 2009; **326**: 1541–1545.
- 24 Wright S: Genetical structure of populations. *Nature* 1950; **166**: 247–249.
- 25 Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. *PLoS Biol* 2006; **4**: e72.
- 26 Sabeti PC, Varilly P, Fry B *et al*: Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007; **449**: 913–918.
- 27 Nang EE, Khoo CM, Tai ES *et al*: Is there a clear threshold for fasting plasma glucose that differentiates between those with and without neuropathy and chronic kidney disease?: the Singapore Prospective Study Program. *Am J Epidemiol* 2009; **169**: 1454–1462.
- 28 Tan JT, Ng DP, Nurbaya S *et al*: Polymorphisms identified through genome-wide association studies and their associations with type 2 diabetes in Chinese, Malays, and Asian-Indians in Singapore. *J Clin Endocrinol Metab* 2010; **95**: 390–397.
- 29 Foong AW, Saw SM, Loo JL *et al*: Rationale and methodology for a population-based study of eye diseases in Malay people: the Singapore Malay eye study (SiMES). *Ophthalmic Epidemiol* 2007; **14**: 25–35.
- 30 Wong TY, Chong EW, Wong WL *et al*: Prevalence and causes of visual impairment and blindness in an urban Malay population: the Singapore Malay Eye Study. *Arch Ophthalmol* 2008; **126**: 1091–1099.
- 31 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.
- 32 Rosenberg NA, Li LM, Ward R, Pritchard JK: Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 2003; **73**: 1402–1422.
- 33 Pickrell JK, Coop G, Novembre J *et al*: Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 2009; **19**: 826–837.
- 34 Gardner M, Gonzalez-Neira A, Lao O, Calafell F, Bertranpetit J, Comas D: Extreme population differences across Neuregulin 1 gene, with implications for association studies. *Mol Psychiatry* 2006; **11**: 66–75.
- 35 Teo YY, Fry AE, Bhattacharya K, Small KS, Kwiatkowski DP, Clark TG: Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res* 2009; **19**: 1849–1860.
- 36 Asano Y, Takashima S, Asakura M *et al*: Lamr1 functional retroposon causes right ventricular dysplasia in mice. *Nat Genet* 2004; **36**: 123–130.
- 37 Batlle E, Bacani J, Begthel H *et al*: EphB receptor activity suppresses colorectal cancer progression. *Nature* 2005; **435**: 1126–1130.
- 38 Amundsen SS, Rundberg J, Adamovic S *et al*: Four novel coeliac disease regions replicated in an association study of a Swedish-Norwegian family cohort. *Genes Immun* 2010; **11**: 79–86.
- 39 Dubois PC, Trynka G, Franke L *et al*: Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* 2010; **42**: 295–302.
- 40 Hunt KA, Zernakova A, Turner G *et al*: Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* 2008; **40**: 395–402.
- 41 Dahlen A, Mertens F, Rydholm A *et al*: Fusion, disruption, and expression of HMG2 in bone and soft tissue chondromas. *Mod Pathol* 2003; **16**: 1132–1140.
- 42 Grunewald TG, Pasedag SM, Butt E: Cell adhesion and transcriptional activity - defining the role of the novel protooncogene LPP. *Transl Oncol* 2009; **2**: 107–116.
- 43 Rogalla P, Lemke I, Kazmierczak B, Bullerdiek J: An identical HMGIC-LPP fusion transcript is consistently expressed in pulmonary chondroid hamartomas with t(3;12)(q27-28;q14-15). *Genes Chromosomes Cancer* 2000; **29**: 363–366.
- 44 Daheron L, Veinstein A, Brizard F *et al*: Human LPP gene is fused to MLL in a secondary acute leukemia with a t(3;11)(q28;q23). *Genes Chromosomes Cancer* 2001; **31**: 382–389.
- 45 Sweetser DA, Chen CS, Blomberg AA *et al*: Loss of heterozygosity in childhood de novo acute myelogenous leukemia. *Blood* 2001; **98**: 1188–1194.
- 46 Kucherlapati M, Yang K, Kuraguchi M *et al*: Haploinsufficiency of Flap endonuclease (Fen1) leads to rapid tumor progression. *Proc Natl Acad Sci USA* 2002; **99**: 9924–9929.
- 47 Zheng L, Dai H, Zhou M *et al*: Fen1 mutations result in autoimmunity, chronic inflammation and cancers. *Nat Med* 2007; **13**: 812–819.
- 48 Pham CG, Bubici C, Zazzeroni F *et al*: Ferritin heavy chain upregulation by NF-kappaB inhibits TNFalpha-induced apoptosis by suppressing reactive oxygen species. *Cell* 2004; **119**: 529–542.
- 49 Shi H, Bencze KZ, Stemmler TL, Philpott CC: A cytosolic iron chaperone that delivers iron to ferritin. *Science* 2008; **320**: 1207–1210.
- 50 Mitsuoka K, Kato Y, Miyoshi S *et al*: Inhibition of oligopeptide transporter suppress growth of human pancreatic cancer cells. *Eur J Pharm Sci* 2010; **40**: 202–208.
- 51 Detera-Wadleigh SD, Liu CY, Maheshwari M *et al*: Sequence variation in DOCK9 and heterogeneity in bipolar disorder. *Psychiatr Genet* 2007; **17**: 274–286.
- 52 Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K: An Icelandic example of the impact of population structure on association studies. *Nat Genet* 2005; **37**: 90–95.
- 53 Novembre J, Johnson T, Bryc K *et al*: Genes mirror geography within Europe. *Nature* 2008; **456**: 98–101.
- 54 Pappu BP, Borodovsky A, Zheng TS *et al*: TL1A-DR3 interaction regulates Th17 cell function and Th17-mediated autoimmune disease. *J Exp Med* 2008; **205**: 1049–1062.
- 55 Friedlaender JS, Friedlaender FR, Reed FA *et al*: The genetic structure of Pacific Islanders. *PLoS Genet* 2008; **4**: e19.
- 56 Yamaguchi-Kabata Y, Nakazono K, Takahashi A *et al*: Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet* 2008; **83**: 445–456.
- 57 Reich D, Thangaraj K, Patterson N, Price AL, Singh L: Reconstructing Indian population history. *Nature* 2009; **461**: 489–494.
- 58 Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL: Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* 2005; **102**: 15942–15947.
- 59 Karafet T, Xu L, Du R *et al*: Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet* 2001; **69**: 615–628.
- 60 Kong QP, Sun C, Wang HW *et al*: Large-scale mtDNA screening reveals a surprising matrilineal complexity in east Asia and its implications to the peopling of the region. *Mol Biol Evol* 2011; **28**: 513–522.
- 61 Deng W, Shi B, He X *et al*: Evolution and migration history of the Chinese population inferred from Chinese Y-chromosome evidence. *J Hum Genet* 2004; **49**: 339–348.
- 62 Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP: Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet* 2002; **70**: 635–651.
- 63 Zhong H, Shi H, Qi XB *et al*: Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. *Mol Biol Evol* 2011; **28**: 717–727.
- 64 Kivisild T, Tolk HV, Parik J *et al*: The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol* 2002; **19**: 1737–1751.
- 65 Wen B, Li H, Gao S *et al*: Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol Biol Evol* 2005; **22**: 725–734.
- 66 Xue Y, Zerjal T, Bao W *et al*: Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics* 2006; **172**: 2431–2439.
- 67 Zhang F, Su B, Zhang YP, Jin L: Genetic studies of human diversity in East Asia. *Philos Trans R Soc Lond B Biol Sci* 2007; **362**: 987–995.
- 68 Teo YY, Ong RT, Sim X, Tai ES, Chia KS: Identifying candidate causal variants via trans-population fine-mapping. *Genet Epidemiol* 2010; **34**: 653–664.
- 69 Grossman SR, Shylakhter I, Karlsson EK *et al*: A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 2010; **327**: 883–886.
- 70 Conrad DF, Jakobsson M, Coop G *et al*: A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 2006; **38**: 1251–1260.
- 71 He J, Gu D, Wu X *et al*: Major causes of death among men and women in China. *N Engl J Med* 2005; **353**: 1124–1134.

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