European Journal of Human Genetics (2014) 22, 248–253

www.nature.com/ejhg

ARTICLE

A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese

Pengfei Qin[1,2,7], Zhiqiang Li[3,7], Wenfei Jin[1,2], Dongsheng Lu[1,2], Haiyi Lou[1,2], Jiawei Shen[4], Li Jin[*,2,5], Yongyong Shi[*,6] and Shuhua Xu[*,1,2]

Population stratification acts as a confounding factor in genetic association studies and may lead to false-positive or false-negative results. Previous studies have analyzed the genetic substructures in Han Chinese population, the largest ethnic group in the world comprising ~20% of the global human population. In this study, we examined 5540 Han Chinese individuals with about 1 million single-nucleotide polymorphisms (SNPs) and screened a panel of ancestry informative markers (AIMs) to facilitate the discerning and controlling of population structure in future association studies on Han Chinese. Based on genome-wide data, we first confirmed our previous observation of the north–south differentiation in Han Chinese population. Second, we developed a panel of 150 validated SNP AIMs to determine the northern or southern origin of each Han Chinese individual. We further evaluated the performance of our AIMs panel in association studies in simulation analysis. Our results showed that this AIMs panel had sufficient power to discern and control population stratification in Han Chinese, which could significantly reduce false-positive rates in both genome-wide association studies (GWAS) and candidate gene association studies (CGAS). We suggest this AIMs panel be genotyped and used to control and correct population stratification in the study design or data analysis of future association studies, especially in CGAS which is the most popular approach to validate previous reports on genetic associations of diseases in post-GWAS era.

European Journal of Human Genetics (2014) **22,** 248–253; doi:10.1038/ejhg.2013.111; published online 29 May 2013

Keywords: Han chinese; genetic association studies; ancestry informative markers; population structure; false association; correction for population stratification

INTRODUCTION

Population stratification due to genetic ancestry is likely to impact the outcome of genotype–phenotype studies such as genome-wide association studies (GWAS), which are designed to identify the risks of common diseases in human populations in which the presence of uncontrolled population structure may lead to false-positive or false-negative results.[1–4] Especially, as only a small number of SNPs are genotyped in candidate gene association studies (CGAS), which do not provide sufficient ancestry information, an independent set of ancestry informative markers (AIMs) is necessary to detect and control potential population stratification. To discern the ancestry of Europeans or European Americans, multiple sets of AIMs have been established that allow correction for population stratification in association studies using Europeans.[5–7]

Recently, a great number of genetic association studies on various diseases have been conducted on non-European populations. Especially in China, hundreds of human gene-disease association studies have been reported using Han Chinese population, the largest ethnic group in the world comprising about 20% of the global human population. However, population substructures are expected to exist in Han Chinese because of its complex ancestral origin, long history of interaction with many surrounding ethnic groups and recent migrations. Indeed, our previous study[8] and some other recent genome-wide studies[9,10] have revealed the complexity in Han Chinese population structure, particularly the north–south stratification. Therefore, a set of AIMs is required to discern the population stratification of Han Chinese and reduce the spurious associations. In this study, we collected 5540 Han Chinese samples in total, most of which were genotyped using Affymetrix 6.0. Among these, 757 samples were used for structure analysis and screening for AIMs, and 4783 samples were used to validate the performance of our AIMs panel. Population structure analyses showed that the main

[1]Max Planck Independent Research Group on Population Genomics, Chinese Academy of Sciences and Max Planck Society Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Shanghai, China; [2]Chinese Academy of Sciences Key Laboratory of Computational Biology, Chinese Academy of Sciences and Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China; [3]Shanghai Genome Pilot Institutes for Genomics and Human Health, Shanghai, China; [4]Changning Mental Health Center, Shanghai, China; [5]Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai, China; [6]Bio-X Center, MOE Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, Shanghai, China

*Correspondence: Dr S Xu, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology, 320 Yueyang Road, Shanghai 200031, China. Tel: +86 21 5492 0479; Fax: +86 21 5492 0451; E-mail: xushua@picb.ac.cn

or Dr Y Shi, Bio-X Center, MOE Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Shanghai Jiao Tong University, Shanghai 200030, China. Tel: +86 21 6293 3338; Fax: +86 21 6293 2059; E-mail: shiyongyong@gmail.com

or Dr L Jin, Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China. Tel: +86 21 6564 3714; Fax: +86 21 6564 3714; E-mail: lijin.fudan@gmail.com.

[7]These authors contributed equally to this work.

Received 28 February 2013; revised 10 April 2013; accepted 16 April 2013; published online 29 May 2013

substructure in Han Chinese is the differentiation between northern and southern populations, supporting previous results. According to this, we established a panel of 150 validated SNPs that were highly informative in distinguishing northern Han (N-Han) from southern Han (S-Han) Chinese. Our analysis showed that this set of AIMs had sufficient power to correct population stratification, which could be useful especially in CGAS where only a few loci or genes are genotyped or sequenced.

MATERIALS AND METHODS

Population samples

In total, 5540 Han Chinese samples were collected in this study, which included 97 Han Chinese from Beijing (CHB) and 100 Han Chinese from southern China (CHS) from the 1000 Genomes Project[11] (1KG), 90 Han Chinese from metropolitan Denver, CO, USA (CHD) from the International HapMap Project (HapMap) phase III,[12] 470 Han Chinese collected via Fudan University and 4783 Han Chinese collected via Shanghai Jiao Tong University. Taken together, these Han Chinese samples represented majority of the geographical areas where Han Chinese reside (including 27 out of the 34 administrative areas in China). These 27 areas can be classified into northern and southern regions, with the Yangtze River as a geographical boundary. The sampling areas and sample size for each regional population are shown in Supplementary Figure S1. For the following analysis, 757 samples were used for selection of AIMs screening and 4783 samples were used for validation.

Genotyping, data assembly and quality control

SNP data in 1KG and HapMap projects were downloaded from 1KG (http://www.1000genomes.org) and HapMap (http://www.hapmap.org) websites, respectively. In total, about 1 312 343 out of 36 820 992 SNPs were common among the tested samples in 1KG and HapMap. Considering the low coverage of 1KG data that could result in high sequencing errors, we made additional quality control by comparing the data of identical samples between the 1KG and HapMap (Supplementary Text S1). SNPs with discordant strands or genotypes were either corrected or removed. We treated CHB and CHS samples from 1KG as N-Han and S-Han, respectively. 1KG samples whose geographical origins were discordant with PCA clusters were presumed to be outliers and were excluded from the analyses (Supplementary Text S2).

All the other Han Chinese samples were genotyped with Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Inc., Santa Clara, CA, USA) that contains 934 969 SNPs. SNP calling from the raw data of all samples was processed by Affymetrix Power Tools 1.10.2. (Affymetrix, Inc.) Quality control was performed with 'apt-geno-qc' and genotype calling was performed with 'apt-probeset-genotype' in birdseed algorithm.[13] Only samples with call rate >0.86 were included in the downstream analyses.

Some further filterations were made for the combined data. Especially, individuals with $>10\%$ missing genotypes were removed, SNPs with missing samples $>10\%$ or in Hardy–Weinberg disequilibrium ($P<0.001$) were also removed. Finally, we obtained data for 5520 Han Chinese individuals sharing 738 937 autosomal SNPs.

Population structure analysis of Han Chinese

Population structure of Han Chinese was examined primarily by principle component analysis (PCA) and FRAPPE.[14] PCA was performed with EIGENSOFT version 3.0[14,15] using 101 038 SNPs, which were selected from 738 937 autosomal SNPs with inter-marker distance >25 Kb to avoid high linkage disequilibrium. FRAPPE analysis based on a 'frequentist' maximum likelihood for clustering was also performed with the same number of SNPs (101 038) as used in PCA with iterations set at 10 000. To estimate the genetic distance between N-Han and S-Han populations, we calculated unbiased estimates of F_{ST} according to Weir and Cockerham.[16]

Selection of AIMs for distinguishing N-Han from S-Han

AIMs are genetic variants that exhibit substantially different frequencies between populations from different geographical regions. There are at least two essential criteria to categorize a SNP as an AIM. These include (1) SNPs among populations to be highly different and (2) the distance between two contiguous AIMs to be large enough to avoid strong linkage disequilibrium. Various statistics have been proposed to measure ancestry information of genetic markers. A previous study has compared those statistics using both simulations and emperical data,[17] which showed that F_{ST} and I_{n}[18] gave estimation of ancestry information with lower bias and mean square error compared with the other ones. So these statistics are chosen to measure ancestry informativeness of markers in this study. Both F_{ST} and I_{n} utilize information of allele frequency based on genetic polymorphism data. F_{ST} measures population differentiation or relatedness. I_{n} is a mutual information-based statistics. From a likelihood perspective, I_{n} gives the expected logarithm of the likelihood ratio that an allele is assigned to one of the populations compared with a hypothetical 'average' population whose allele frequencies equal the mean allele frequency across sub-populations.

Based on the frequency of 738 937 autosomal SNPs in N-Han and S-Han populations, we first calculated unbiased F_{ST} for each locus. Then we screened each autosome and dropped those SNPs with little difference between the two clusters of Han Chinese (here the set threshold F_{ST} value was <0.01). For each pair of contiguous SNPs with interval distance smaller than 500 Kb, we retained the one with higher F_{ST} value. Finally, we ranked the markers that satisfied these criteria in descending order based on their F_{ST} values, so that markers with high values could be used for downstream analyses according to the desired cutoff. In addition, we calculated I_{n} value for each SNP between N-Han and S-Han for reference. I_{n} was calculated using the following equation:

$$I_{\mathrm{n}}(Q;J)=\sum_{j=1}^{N}\left(-p_j\ log\ p_j+\sum_{i=1}^{K}\frac{p_{ij}}{K}\log p_{ij}\right)$$

where K is the number of populations or groups, p_{ij} is the relative frequency for allele j in population i. Average frequency of allele j is defined as p_j. I_{n} measures the amount of information about ancestry Q contained in genotype J.

Statistical power of AIMs for distinguishing N-Han from S-Han

We designed a stepwise procedure to estimate the performance of the AIMs panel in the classification of N-Han and S-Han. We first ranked all AIMs in descending order by their F_{ST} values. Then, we examined and evaluated performance of different number of AIMs to differentiate N-Han from S-Han. The number of AIMs was increased from 10–500, in increments of 10 and the change of PCA clustering was monitored. Statistical power of these AIMs in sample classification was evaluated using the maximum Matthews correlation coefficient (MCC) based on the formula

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives.

Validations were compiled into data sets of 757 samples and 4783 samples, respectively.

Evaluation of the AIMs panel performance in simulated CGAS

The main application of this AIMs panel is that it can be genotyped in CGAS to discern and control population stratification to reduce false positive. Therefore, we simulated Han Chinese population data with different levels of population stratification for CGAS to evaluate the performance of AIMs in reducing false-positive rates.

We first simulated gene regions with allelic distribution of frequency and divergence between N-Han and S-Han being similar to those from empirical whole-genome data. To simulate a gene with 1000 loci, we sorted empirical SNPs of the whole genome in ascending order based on their F_{ST} values and split them into 1000 bins, and then sampled random loci from these bins with one locus from each bin. For each simulated gene, we assigned 5000 cases and 5000 controls based on the allele frequency in empirical Han samples. We also integrated 150 AIMs to provide ancestral origins and 20 risk alleles with odds ratios at four different levels ranging from 1.2–2.0. Considering the different degrees of population stratification that presumably existed in samples of association studies, controls were sampled only from N-Han while cases were a mix of N-Han and S-Han samples at varying degrees. Overall, we provided 11 different scenarios with proportions of S-Han in cases ranging from 0 to 100% with increments of 10%. One hundred genes were simulated and the mean and SEM values were calculated from these 100 repeats. We used Armitage trend χ^2-statistics[19] to detect associations between loci and phenotypes. *P*-values were calculated with one degree of freedom and were controlled by Bonferroni single-step method.[20]

We implemented two commonly used methods to correct population stratification in genetic association studies using the 150 AIMs that classified Han samples well. One is genomic control[21] and the other is a PCA-based method implemented in EIGENSTRAT.[22] For genomic control, we evaluated χ^2 inflation factor λ in the association study. The value of λ was computed as the median χ^2 statistic divided by 0.456, which is the predicted median χ^2 if there was no inflation. We first estimated the value of λ for each scenario of stratification using only genotypes of those 150 AIMs. Then χ^2 statistics were adjusted by the corresponding λ. For PCA-based method, we first inferred principal components for each individual using only the 150 AIMs, and then calculated χ^2 statistic of the markers excluding AIMs using adjusted genotypes and phenotypes. The quantile–quantile (Q–Q) plots of *P*-values, with and without correction for population stratification, were plotted for comparison. False positives, with and without correction for population stratification, were calculated for all scenarios. In addition, power to detect risk alleles with various odds ratios was also estimated.

RESULTS

Analysis of Han Chinese population structure using genome-wide data

One data set including 757 Han samples was used for population structure analysis. After controlling data quality (see Methods) and removing outliers based on PCA analysis, we obtained 504 Han Chinese samples with 738 937 SNPs in whole genomes. Previous studies have revealed one-dimensional 'north–south' population structure and no discernible east–west pattern were observed.[8–10] The north–south population structure is consistent with the historical migration and expansion pattern of the Han Chinese population.[23] Our Han Chinese samples were widely spread over PC1 (Figure 1a), suggesting a cryptic stratification in Han Chinese population. In addition, the north–south pattern became more pronounced when the CHB and CHS from 1KG, which had passed strict quality control (Supplementary Text S1, Supplementary Table S1) and outliers filtering (Supplementary Text S2, Supplementary Figure S2), were marked (Figure 1a). The north–south pattern of our samples was mainly explained by PC1 while other PCs were much less informative, and no discernible structure in the other combinations of PCs other than the top two PCs (Supplementary Figure S3). Considering the intermarriages of Han Chinese from different parts of China and the fact that parts of our samples were from metropolitan cities, Anhui and Jiangsu which are located in mid-China, it was not easy to distinguish samples among N-Han, S-Han and the highly mixed central Han.[8]

We used several strategies including geographical locations, PCA clustering and ingredient proportion in FRAPPE analysis to distinguish N-Han from S-Han. We first classified all the remaining samples into two groups based on the natural separation of northern and southern Mainland China by the Yangtze River. Then we removed those genetically mixed individuals that were too ambiguous to be clustered into either N-Han or S-Han based on PCA clustering results and ingredient proportion analysis in FRAPPE. At last, we obtained 467 samples including 250 N-Han and 217 S-Han.

Population structure of the 467 samples is described in PCA plot (Figure 1b) and FRAPPE (Figure 1c). The genetic difference between N-Han and S-Han was estimated by Weir and Cockerham's F_{ST}. Average F_{ST} value from the whole genome loci is 0.00126 (SD = 0.0027), which is lower than the European population groups (0.0033).[24] The estimated F_{ST} value from this study is very close to a previously reported result (0.00116).[8] In addition, the average F_{ST} between CHB and CHS from 1KG is 0.00145 (SD = 0.00634), which is similar to our Affy6.0 data set. To identify genomic regions with highly differentiated allele frequencies between N-Han and S-Han, we examined the F_{ST} distributions for all SNPs over the entire genome. Some genes associated with the most different SNPs were labeled on the Manhattan plot (Supplementary Figure S4). The most highly differentiated region is an intron on *FADS2* gene located on

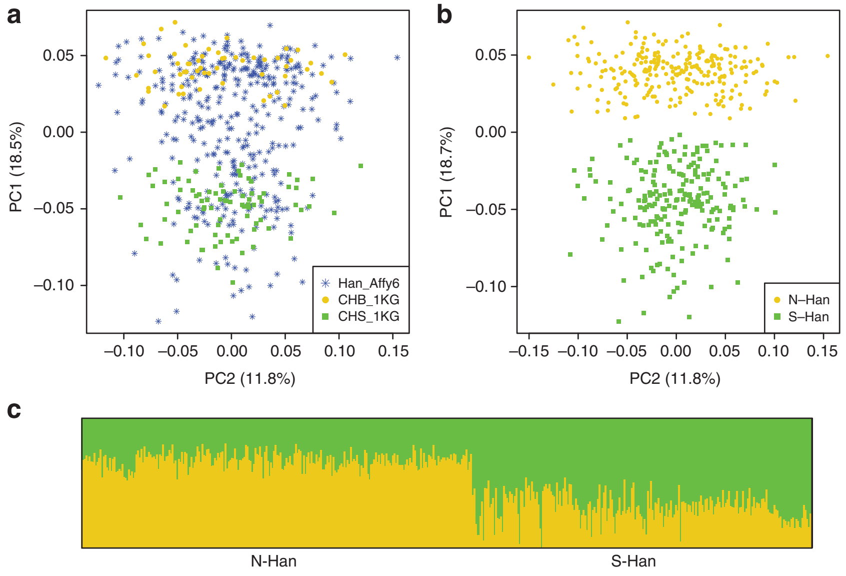

Figure 1 Population structure of Han Chinese. All PCA plots and FRAPPE analyses were based on 101 038 SNPs randomly chosen from genome-wide data. (**a**) PCA plot of all Han Chinese samples (1410 Han samples plot after quality control and outlier filtering). The 56 CHB and 83 CHS from 1000 Genome Project representing N-Han and S-Han, respectively, are highlighted. (**b**) PCA plot of 467 Han samples containing 250 N-Han and 217 S-Han. (**c**) Structure of 467 Han samples analyzed by FRAPPE when $K = 2$.

chromosome 11 (at Chr11:61353788), which is associated with the fatty acid composition in phospholipids and arachidonic acid levels, involved in inflammation and immunity processes and related disease. This region could have been a target of natural selection considering environmental differentiation such as climate and agriculture between northern and southern China, but further study is needed to confirm this result, as well as our hypothetical conclusion.

AIMs selection and validation

F_{ST} and I_n are commonly used approaches to measure the ancestral information of SNPs. We found a high correlation between I_n and F_{ST} ($R^2=0.996$) values (Supplementary Figure S5). To screen for AIMs, we used 467 Han samples including 250 from N-Han and 217 from S-Han with 738 937 autosomal SNPs. We first screened each autosome and removed the SNPs with low F_{ST} values (≤ 0.01). Second, for each pair of contiguous SNPs separated by a distance smaller than 500 Kb, we retained the one with a higher F_{ST} value. Following these criteria, we identified more than 3000 markers in our data sets and then ranked them based on their F_{ST} values in descending order. The top 1000 markers ($F_{ST}> =0.0149$) (Supplementary Table S2) were used for validations.

Validations of our AIMs panel to distinguish N-Han from S-Han were conducted in two data sets. One included 250 N-Han and 217 S-Han, which passed strict quality control and filtering procedure as described in Methods. The other was a much larger data set, including 2779 N-Han and 2004 S-Han, which were filtered based on their geographical origins. We followed a stepwise procedure as described in Methods to select a small panel of markers sufficient to distinguish the two clusters of Han Chinese. Maximum MCC was used to describe clustering performance. We started validation with a minimal of 10 AIMs and added 10 more for each analysis. MCC value increased with the increasing number of AIMs (Figure 2a). Using the top 150 AIMs, we obtained a perfect classification in PCA plots (Figure 2b) with MCC value equal to 1, which suggested that at least 150 AIMs were needed for the classification of N-Han and S-Han. Besides, ingredient proportion analysis of FRAPPE using these 150 AIMs also clearly distinguished N-Han from S-Han with $K=2$ (Supplementary Figure S6). When we repeated the stepwise procedure for the second data set with 4783 samples, a similar pattern in MCC plot was observed (Figure 2c). However, a perfect classification (MCC = 1) was not reached, which was most likely to be due to the high population shift of Chinese and the inclusion of metropolitan samples in this data set. Nevertheless, MCC value was 0.97 with 150 AIMs, which clearly separated the two clusters (Figure 2d).

Correcting population stratification in genetic association studies

In both GWAS and CGAS, population stratification could cause false associations between markers with different frequencies across subpopulations, if there were ancestral differences between cases and controls. Therefore, small panels of AIMs are required to accurately predict the ancestry of individuals especially in CGAS and fine mapping or sequencing studies. Here, we simulated a series of CGAS to determine the occurrence of the number of false-positive associations if the effects of Han Chinese population substructures were not corrected in a case–control association study. In addition, we wanted to determine whether our AIMs panel could efficiently correct the

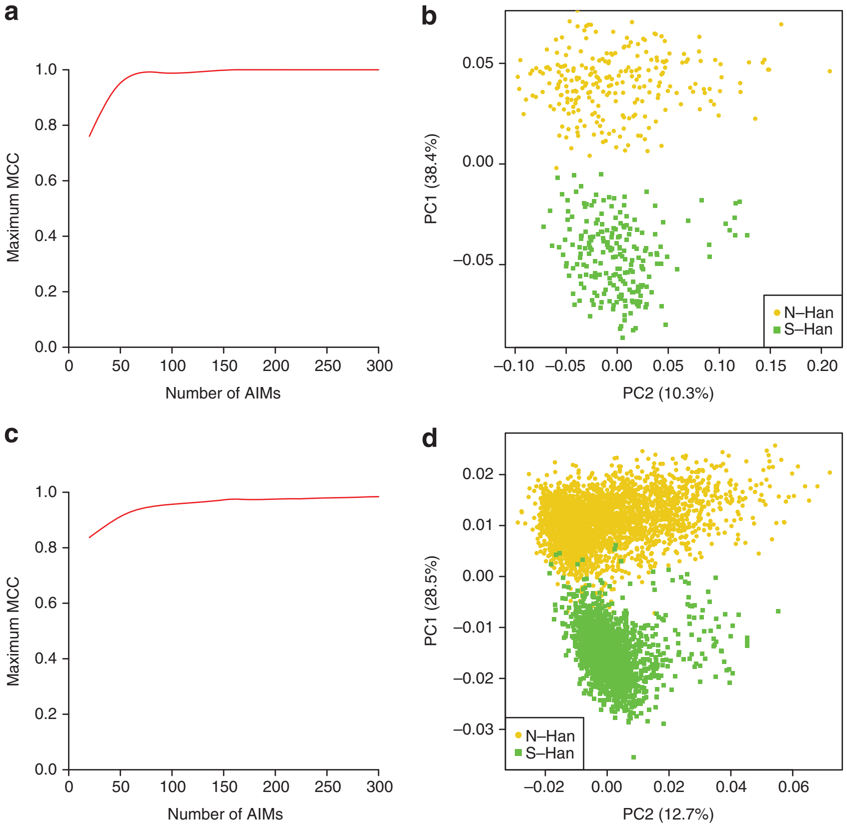

Figure 2 Validation of our AIMs panel in two Han Chinese data sets. (**a**) Classification of 467 Han samples using varying numbers of AIMs in the first data set. (**b**) PCA plot of 467 Han samples in the first data set based on the genotypes of top 150 AIMs. (**c**) Classification of 4783 Han samples using varying numbers of AIMs in the second data set. (**d**) PCA plot of 4783 Han samples in the second data set based on the genotypes of top 150 AIMs.

false-positive associations due to population stratification. We simulated 5000 cases and 5000 controls in each study. To create different degrees of stratification, control samples were randomly selected only from N-Han while case samples were a mix of N-Han and S-Han with proportions of S-Han ranging from 0 to 100%. Loci of each gene region were simulated according to description in Methods.

Results of the simulated CGAS showed that spurious associations are likely to be generated if the impact of population stratification was not corrected especially in studies with strong stratification (Figures 3 and 4a). Values of inflation factor λ increased exponentially with increasing degrees of stratification (Supplementary Figure S7). In addition, false-positive rates would be much higher due to population stratification (Supplementary Table S3). The various odds ratios for risk alleles did not impact the false-positive rates because of the large sample size we used here such as 5000 cases and 5000 controls (Figure 4a). Procedures for correcting the impact of stratification were necessary for CGAS in Han Chinese population.

We conducted two commonly used methods, the PCA-based method EIGENSTRAT and the genomic control method, to correct the impact of population stratification in association studies using the AIMs panel selected in this study. In our simulations, PCA-based method reasonably corrected the *P*-values from χ^2-statistics with population stratification (Figure 3 and Supplementary Figure S8). Our analysis also showed that most of the false-positive rates could be corrected regardless of the odds ratios and stratifications (Figure 4a). However, when stratification degree was extreme with 100% cases from S-Han, this method could neither adjust *P*-values (Supplementary Figure S8) nor correct false positives (Figure 4a). Moreover, we also determined the power for detecting risk alleles in CGAS with stratification in samples. Without applying corrections for stratification, power to detect risk alleles depends on the level of odds ratio that is not influenced by stratification (Figure 4b). Risk alleles with higher odds ratio levels are much easier to detect. The power to detect risk alleles after applying corrections for stratification was satisfactory, especially in studies with high levels of odds ratio (> 1.4) or low degrees of stratification (< 50%), although part of the power was suppressed (Figure 4b).

Genomic control, however, was not powerful and suitable for correcting population stratification in CGAS using AIMs. We first

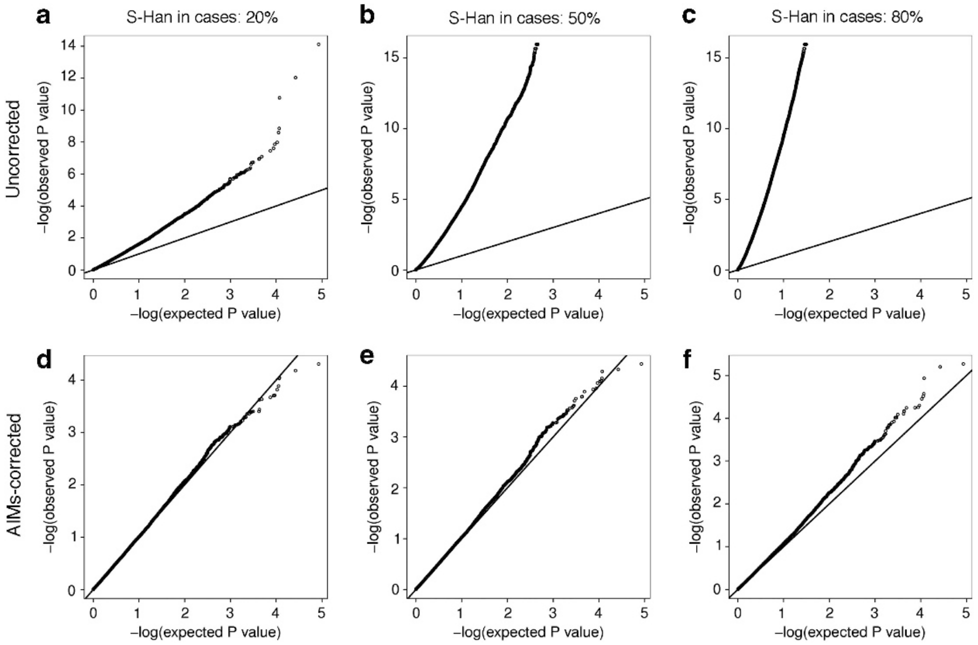

Figure 3 Q-Q plots of the *P*-values from simulated association studies with or without correction for population stratification using top 150 AIMs. Columns correspond to degrees of stratification of 5000 case and 5000 control samples. Rows correspond to plots of uncorrected and AIMs-corrected (PCA-based method) *P*-values.

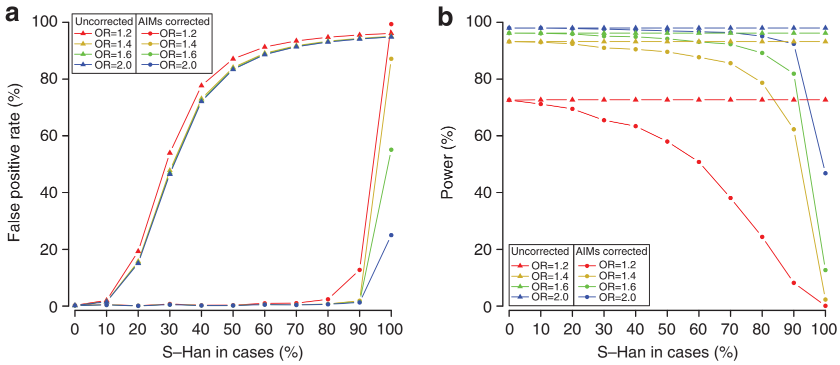

Figure 4 False-positive rates and statistical power to detect risk alleles before and after correction for stratification using top 150 AIMs. Different colors represent different odds ratio levels. Lines with triangles represent values before correction while lines with circles represent values after correction.

estimated the inflation factor λ for each degree of stratification using the 150 AIMs genotypes. The value of λ increased exponentially with increasing degrees of stratification (Supplementary Figure S7). In our simulation, the estimated value of λ without stratification was 1.07 while it increased to 641.7 in cases with samples only from S-Han. Two reasons could have resulted in such a large λ value. One was the large sample size (10 000) used for χ^2 statistics and the other was the inclusion of the top 150 most differentiated markers in N-Han and S-Han for estimation, which could have resulted in an over estimation of χ^2-statistics compared with markers from whole genomes. As a result, *P*-values for both normal and risk alleles in χ^2-statistics could be grossly over-adjusted (Supplementary Figure S8). Power to detect risk alleles could be totally lost due to the over-adjusted *P*-values (Supplementary Table S4). We thus suggest that the inflation factor λ should be estimated based on whole-genome markers rather than just AIMs, and it would be not suitable to correct stratification using AIMs in genomic control procedure.

DISCUSSION

We previously developed an AIMs panel that included 5000 SNPs to discern the N-Han and S-Han Chinese population.[25] However, that panel of AIMs was selected from a data set of small sample size consisting only 162 N-Han and 74 S-Han, which was likely to result in a bias for allele frequency estimation, and the performance of previous AIMs panel was not good enough to cluster N-Han and S-Han samples used in this study (Supplementary Figure S9). In this study, taking advantage of the large sample size (757 for screening AIMs and 4783 for validation) and genome-wide high-density SNP data (931 000 markers), we were able to provide a set of high quality AIMs with improved performance in distinguishing and clustering N-Han and S-Han. We demonstrated that the AIMs panel developed here is currently the best for clustering N-Han and S-Han populations, and stratification adjustment in case–control association studies.

The panel of 150 informative markers to predict Han Chinese ancestry could be used in small-scale studies for genotyping and in addition, permitting correction of population stratification in Han Chinese at a reasonably low cost for a genome-wide scan. We propose two applications[5,26] for this AIMs panel: (1) to evaluate study design before starting GWAS, for example, by genotyping cases and controls on this panel, we can remove unsuitable cases and controls; (2) to be used for genotyping in targeted association studies, such as CGAS or replication studies following GWAS, in which variants are targeted in a large number of samples that have not been densely genotyped. Our AIMs panel can be used to efficiently correct for stratification using methods such as EIGENSTRAT not genomic control procedure, to ensure that the observed associations are not spurious without relying on self-reported ancestry.

A previous study has shown that genomic control loses nearly all power and EIGENSTRAT suffers a partial power loss in GWAS if causal SNPs confound with highly differentiated SNPs between substructures,[22] However, this problem could be avoided in CGAS using AIMs. For example, as we provided in this study a list of top 1000 highly informative AIMs, a good way of choosing AIMs from the list is to avoide using those AIMs within or nearby candidate gene regions for controlling population stratification.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

These studies were supported by the National Science Foundation of China (NSFC) grants (31171218; 30971577), by the Shanghai Rising-Star Program 11QA1407600, and by the Science Foundation of the Chinese Academy of Sciences (KSCX2-EW-Q-1-11; KSCX2-EW-R-01-05; KSCX2-EW-J-15-05). SX is Max Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. SX also gratefully acknowledges the support of the National Program for Top-notch Young Innovative Talents and the support of KCWong Education Foundation, Hong Kong. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

1 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220–228.
2 Freedman ML, Reich D, Penney KL *et al*: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36**: 388–393.
3 Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512–517.
4 Campbell CD, Ogburn EL, Lunetta KL *et al*: Demonstrating stratification in a European American population. *Nat Genet* 2005; **37**: 868–872.
5 Price AL, Butler J, Patterson N *et al*: Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet* 2008; **4**: e236.
6 Tian C, Plenge RM, Ransom M *et al*: Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet* 2008; **4**: e4.
7 Paschou P, Drineas P, Lewis J *et al*: Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet* 2008; **4**: e1000114.
8 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.
9 Chen J, Zheng H, Bei J-X *et al*: Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am J Hum Genet* 2009; **85**: 775–785.
10 Suo C, Xu H, Khor C-C *et al*: Natural positive selection and north-south genetic diversity in East Asia. *Eur J Hum Genet* 2012; **20**: 102–110.
11 Abecasis GR, Auton A, Brooks LD *et al*: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**: 56–65.
12 Altshuler DM, Ra Gibbs, Peltonen L *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.
13 Korn JM, Kuruvilla FG, McCarroll SA *et al*: Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008; **40**: 1253–1260.
14 Tang H, Peng J, Wang P, Risch NJ: Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol* 2005; **28**: 289–301.
15 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
16 Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**: 1358–1370.
17 Ding L, Wiener H, Abebe T *et al*: Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC genomics* 2011; **12**: 622.
18 Rosenberg NA, Li LM, Ward R, Pritchard JK: Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 2003; **73**: 1402–1422.
19 Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955; **11**: 375–386.
20 Benjamini Y, Hochberg Y: Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Statist Soc B* 1995; **57**: 289–300.
21 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
22 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.
23 Su B, Xiao J, Underhill P *et al*: Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. *Am J Hum Genet* 1999; **65**: 1718–1724.
24 Lao O, Lu TT, Nothnagel M *et al*: Correlation between genetic and geographic structure in Europe. *Curr Biol* 2008; **18**: 1241–1248.
25 Qu HQ, Li Q, Xu S *et al*: Ancestry informative marker set for han chinese population. *G3* 2012; **2**: 339–341.
26 Seldin MF, Price AL: Application of ancestry informative markers to association studies in European Americans. *PLoS Genet* 2008; **4**: e5.

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