BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 23 2010, pages 2961–2968
doi:10.1093/bioinformatics/btq560

Genetics and population analysis Advance Access publication September 30, 2010

Interrogating local population structure for fine mapping in genome-wide association studies

Huaizhen Qin[1], Nathan Morris[1], Sun J. Kang[1], Mingyao Li[2], Bamidele Tayo[3], Helen Lyon[4], Joel Hirschhorn[4,5,6], Richard S. Cooper[3] and Xiaofeng Zhu[1,*]

[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, [2]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, PA, [3]Department of Preventive Medicine and Epidemiology, Loyola University Chicago, Maywood, IL, [4]Department of Genetics, Harvard Medical School, Boston, [5]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge and [6]Program in Genomics and Divisions of Genetics and Endocrinology, Children's Hospital, Boston, MA, USA

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Adjustment for population structure is necessary to avoid bias in genetic association studies of susceptibility variants for complex diseases. Population structure may differ from one genomic region to another due to the variability of individual ancestry associated with migration, random genetic drift or natural selection. Current association methods for correcting population stratification usually involve adjustment of global ancestry between study subjects.

Results: We suggest interrogating local population structure for fine mapping to more accurately locate true casual genes by better adjusting the confounding effect due to local ancestry. By extensive simulations on genome-wide datasets, we show that adjusting global ancestry may lead to false positives when local population structure is an important confounding factor. In contrast, adjusting local ancestry can effectively prevent false positives due to local population structure and thus can improve fine mapping for disease gene localization. We applied the local and global adjustments to the analysis of datasets from three genome-wide association studies, including European Americans, African Americans and Nigerians. Both European Americans and African Americans demonstrate greater variability in local ancestry than Nigerians. Adjusting local ancestry successfully eliminated the known spurious association between SNPs in the LCT gene and height due to the population structure existed in European Americans.

Contact: xiaofeng.zhu@case.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 27, 2010; revised on August 27, 2010; accepted on September 27, 2010

1 INTRODUCTION

Unrecognized population structure can potentially lead to both false positive and false negative findings in genetic association studies (Campbell *et al*., 2005; Lander and Schork, 1994; Zhu *et al*., 2008a, b). Current statistical method used to adjust population stratification in genetic association studies often focus on adjusting global population structure, which is mainly caused by recent migration and genetic random drift. Global population structure can be characterized as an individual's global ancestry calculated as the proportion of the individual's genome inherited from a given ancestral population. In contrast, local population structure can be characterized in a similar way but in local genomic region or at a locus. A popular procedure to eliminate the effect of population structure is the genomic control (GC) approach (Devlin and Roeder, 1999), which assumes that the variance inflation factor of a test statistic is a constant across the genome. However, variance inflation can also be due to locus-specific attributes. For example, the imprint of natural selection has recently been identified in many regions across the genome (Sabeti *et al*., 2007; Voight *et al*., 2006), and can create substantial variation in F_{st} which in turn affects the degree of variance inflation for specific loci (Crow and Kimura, 1970). Thus, the GC approach will eliminate the effect due to the global population structure, but may not be effective in eliminating the effect due to local population structure. An alternative method to correct population structure is based on the principal components (PCs) of the genotypes, in which the PCs of the genotype score matrix of many markers across the entire genome (referred to as global PCs hereafter) are used to represent individual ancestry background and used as covariates in association analysis (Epstein *et al*., 2007; Price *et al*., 2006; Zhu *et al*., 2002, 2008a, b). Various studies have shown that the global PCs can effectively represent human demographic history (Cavalli-Sforza and Bodmer, 1999; Novembre and Stephens, 2008). However, local genomic regions harboring functional variants may be subject to subtle forms of population structure not only as a result of demographic history but natural selection and local random fluctuations of admixture (Genovese *et al*., 2010; Tang *et al*., 2007).

In genome-wide association studies, numerous markers are usually tested one at a time. The test of a non-causal maker locus, which is dependent of a causal gene, may be confounded by a pathway leading through both global and local ancestries (Fig. 1). The dependence between a non-causal marker and a causal gene may due to short range or long-range LD. Short-range LD may exist when two loci are nearby. Long-range LD might exist between two genes in different chromosomes, i.e. due to co-evolution (Rohlfs *et al*., 2010).

*To whom correspondence should be addressed.

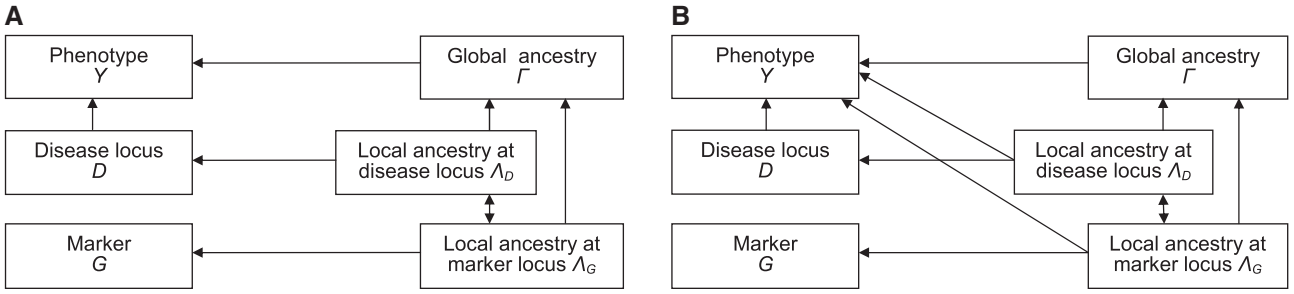

Fig. 1. The pathways which confound the relationship between a non-causative marker G and phenotype Y. A single-arrowed line indicates correlation which is possibly directive (i.e. D to Y), and a double-arrowed line indicates correlation only. Population history determines local ancestries Λ_G, Λ_D and their correlation, i.e. Λ_G and Λ_D can be correlated if markers G and D are close to each other, population admixture or co-evolution occurred on both regions; local ancestries impact the variations of Γ and the allele frequencies of G and D; and D and Γ directly impact phenotypic variation. Therefore, both global and local ancestries may confound $G-Y$ relationship. Λ_G may impact phenotypic variation either indirectly (**A**) due to correlation or directly (**B**) due to natural selection. Adjusting Λ_G can effectively control the type I error due to the confounding by both Λ_G and Γ. Adjusting Γ alone may not well control type I error, especially when Λ_G directly confounds the $G-Y$ association.

We reason that the whole genome is a mosaic consisting of chromosome segments coming from different ancestral populations. Figure 1A illustrates one way of thinking about global and local ancestries. For one individual, we denote by Λ_G the local ancestry of one marker genotype and define the global ancestry (Γ) as the average of Λ_G across the entire genome. Individuals within a population have varying family trees leading to variation in Γ. For example, some African Americans may have 25% European ancestry, whereas some others may have 50% European ancestry. While Γ and Λ_G are correlated as shown in Figure 1A, they are different quantities. The allele frequencies of a marker may be different in different ancestral populations. Hence, the global ancestry (Γ) and genotype (G) are independent conditioning on local ancestry (Λ_G). However, conditioning on global ancestry (Γ), the phenotype (Y) may still be dependent on Λ_G due to the pathway $G-\Lambda_G-\Lambda_D-D-Y$ (see Pearl, 2010 for a further explanation of admissible covariates). Hence, when testing for genotype–phenotype association, type I error may not be adequately controlled in principle by adjusting global ancestry Γ alone. It may be instructive to adjust local ancestry Λ_G. Figure 1B further illustrates why adjusting local ancestry (Λ_G) is important for fine mapping. In this figure, Λ_D is the local ancestry at the causal gene (D), which is associated with the marker (G). There are multiple pathways confounding the genotype–phenotype relationship. The pathways $G-\Lambda_G-Y$ and $G-\Lambda_G-\Lambda_D-Y$ may be blocked by adjusting Λ_G but cannot be blocked by adjusting Γ. When attempting to detect genes, one may not wish to block the pathway $G-\Lambda_G-Y$ because the marker (G) possibly nears a true causal locus. Local ancestry adjustment, however, may serve to refine the location of the true disease locus.

With current high-throughput genotyping technologies and recent methods development on ancestry inference (Patterson *et al*., 2004; Price *et al*., 2009; Pritchard *et al*., 2000; Sankararaman *et al*., 2008; Tang *et al*., 2006; Zhu *et al*., 2006), it is possible to estimate the ancestry of a local genomic region even when the entire genome is a mixture of several ancestral populations. In practice, the real ancestry at a specific locus is usually unknown. For recently admixed populations such as African Americans, the ancestral populations are typically known with good approximation. In this situation, locus-specific ancestry can be inferred accurately using hidden Markov model-based methods (Patterson *et al*., 2004; Price *et al*., 2009; Pritchard *et al*., 2000; Sankararaman *et al*., 2008; Tang *et al*., 2006; Zhu *et al*., 2006). However, it is difficult to accurately infer locus-specific ancestry for a population whose substructure is subtle, due to either the lack of information on the ancestral population or when admixture has occurred within similar populations. For example, among European Americans, where population admixture occurred within populations of similar origin, it is difficult to accurately infer locus-specific ancestry due to the lack of ancestral population information. For such scenarios, we propose using local PCs to present local ancestries. By extensive simulations, we find that the correlation between global PCs and true local ancestries is very weak. Hence, global PCs may not well capture the variations in local ancestry. In contrast, the PCs of the genotype score matrix of the markers within local genomic regions (referred to as local PCs hereafter) are strongly correlated with true local ancestries.

To illustrate the advantages of incorporating local PCs into fine mapping, we evaluated Local Ancestry Principal Components Correction (referred to as LAPCC). Our simulation results illustrated that LAPCC was more effective in eliminating false positives due to local population structure than Global Ancestry Principal Components Correction (referred to as GAPCC) and it retained testing power successfully. We further performed local PC analysis and LAPCC to the datasets from genome-wide association studies (GWAS) in three populations: African-Americans from Maywood, Illinois, Nigerians from Igbo-Ora and Ibadan, Nigeria and European Americans from Framingham, Massachusetts (Kang *et al*., 2010; Levy *et al*., 2009). High degrees of variability in local ancestry of African Americans and European Americans were uncovered whereas very little was found among Nigerians of Yoruba ethnicity. This result is consistent with the demographic histories of the three populations. In addition, LAPCC successfully removed the spurious association between the LCT gene and height, which results from Northern and Southern European population structure in European Americans.

2 METHODS

2.1 Association methods

In LAPCC, we divided each chromosome into 4 Mb adjacent window cores according to the SNP map and constructed an envelope with an 8 Mb-margin to each side of the window core (Supplementary Fig. S1). The left (right) envelopes of the first (last) two windows on each chromosome might be shorter than the inner envelopes. We computed the first 10 PCs of the genotype score matrix of the SNPs within the enveloped 20 Mb window. This window size was chosen based on the linkage disequilibrium due to recent population admixture (Patterson *et al*., 2004; Zhu *et al*., 2006). These local PCs were then used to adjust population structure for the SNPs within the 4 Mb core. At each SNP in the core, we computed genotype score residuals $\tilde{g}_i$ and trait value residuals $\tilde{y}_i$ by regressing genotype scores g_i and trait values y_i on the 10 PCs, respectively. We measured the evidence of trait–SNP association by $s^2=(N-12)r^2/(1-r^2)$, where N was the number of individuals used after excluding individuals with missed genotypes and r was the correlation coefficient between residuals $\tilde{y}_i$ and $\tilde{g}_i$. Asymptotically, s^2 follows χ^2_1 under the null and if population structure is well adjusted. In GAPCC, the first 10 global PCs were computed using the genotypes of all markers across the genome and utilized in the aforementioned way to contrast the association test of each SNP, which is similar to EIGENSTRAT (Price *et al*., 2006). When the true local ancestry is known, we can incorporate the true local ancestry in a regression model and we termed this approach as RegRA. We termed the naïve regression analysis without controlling for population structure as RegUa.

2.2 Local and global PCs

To assess the degree of discrepancies between global PCs and local PCs, we calculated the coefficients of multiple-determination (R^2) and squared coefficients of canonical correlation (λ^2). Let N denote the sample size, $\mathbf{A}=[\mathbf{a}_1,\ldots,\mathbf{a}_K]$ denote the $N\times K$ matrix consisting of the first K global PCs, and $\mathbf{B}=[\mathbf{b}_1,\ldots,\mathbf{b}_K]$ denote the $N\times K$ matrix consisting of the first K local PCs in a local window. The coefficient of multiple-determination R^2_j for $\mathbf{b}_j$ and $\mathbf{A}$ is the R^2 in the linear regression of $\mathbf{b}_j$ on $\mathbf{A}$. The j-th largest squared coefficient of canonical correlation λ^2_j between $\mathbf{A}$ and $\mathbf{B}$ is the j-th largest coefficient of determination between any linear combination of $\mathbf{B}$'s columns and any linear combination of $\mathbf{A}$'s columns. Mathematically, $R^2_1\leq\lambda^2_1$, and $R^2_1+\cdots+R^2_K=\lambda^2_1+\cdots+\lambda^2_K$, R^2_j equals the j-th diagonal element and λ^2_j equals to the j-th largest eigenvalue of $\mathbf{B}'\mathbf{A}\mathbf{A}'\mathbf{B}$, and $R^2_j=r^2_{j1}+\cdots+r^2_{jK}$, where $r^2_{ji}=(\mathbf{b}'_j\mathbf{a}_i)^2$ is the coefficient of determination between $\mathbf{b}_j$ and $\mathbf{a}_i$. Canonical correlation analysis and multiple-determination analysis enable us to evaluate the degree of discrepancies between local and global PCs and facilitate the exploration of population structures. R^2 indicates how much variation of local PCs can be accounted for by global PCs and λ^2 measures the shared variance between local and global PCs.

2.3 GWAS data

In our analysis, we used three genome-wide datasets: Maywood, Nigeria and Framingham (Kang *et al*., 2010; Levy *et al*., 2009). The detailed description and QC for Maywood and Nigeria data can be found in Kang *et al*. (2010). Briefly, the Maywood cohort included 775 unrelated African Americans sampled from Maywood, IL, with 909 622 SNPs genotyped by Affymetrix 6.0 platform. We dropped 74 individuals because of possible DNA contamination, false identity and relatedness. For the 701 retained individuals, we removed 86 800 SNPs whose missing rates >5%, minor allele frequencies <1%. The final analysis dataset included 822 822 SNPs in each of 701 individuals. Similar QCs were applied to Nigeria data, which Affymetrix 6.0 platform was also used for genotyping. After QCs, the Nigeria dataset contained 759 222 SNPs genotyped for 982 individuals. For Framingham dataset, Mendelian errors were checked and the corresponding SNPs with Mendelian inconsistence were set missing. SNPs with HWE $P<10^{-6}$ were dropped. We selected unrelated individuals from each family (i.e. spouses) based on an algorithm that prioritizes individuals with higher genotyping rates, selecting individuals at random when needed. After QCs, the Framingham dataset contained 415 281 SNPs genotyped for 1106 unrelated individuals.

3 SIMULATION STUDIES

3.1 Designs

3.1.1 Admixed populations without selection pressure We first applied the software ADMIXPROGRAM (Zhu *et al*., 2006) to the Maywood dataset to infer the individual's SNP-specific ancestries for the 701 Maywood subjects using 2606 selected ancestry informative SNPs. We observed that the distribution of individual genome-wide CEU allele portions can be well fitted by *Beta*(4.8, 19.2), the beta distribution with mean 0.2 and standard deviation 0.08. Based on this distribution, we simulated an admixed population with average 20% European and 80% African ancestries. Specifically, we simulated N_i individuals with average ancestries w_i (Supplementary Material). For each individual, we simulated the genotypes at the SNPs on 22 autosomes using HapMap dataset, which included 1 969 739 SNPs with complete haplotype information for CEU, YRI and CHB/JPT samples. For each autosome, we simulated the number of crossover points s from the Poisson distribution of mean $\mu=l\times g\times 10^{-7}$, where l was the length of an autosome, g was the generation since the individual began admixture and was randomly sampled from 1 to 10. Then, we uniformly distributed s crossovers across the chromosome. Next, we randomly sampled haplotype segments from CEU or YRI HapMap data between two crossovers independently according to the average ancestry w_i. Using this simulation procedure (denoted by GenoAnceBase0), we simulated the SNP genotypes and ancestries for 2000 unrelated individuals.

3.1.2 Admixed populations with selection pressure We assumed that natural selection will result in non-concordance between local and global ancestry. For each of the 22 autosomes, we randomly picked one locus where selection pressure would be observed. We adopted specific admixture weights for the segments covering the chosen loci. For an odd (even) numbered chromosome, we simulated the segments covering the locus with average 10% (40%) European ancestry and 90% (60%) African ancestry. For the rest of genomic regions, we applied the algorithm GenoAnceBase0. We called the method for generating selection pressure as GenoAnceBase1.

3.1.3 Genetic models To compare different methods, we considered six different genetic models (Supplementary Table S1). Models 1, 2 and 3 were designed to evaluate whether type I errors are under control for different methods. Specifically, in Model 1, only global ancestry contributed to the phenotypic variation; in Model 2, only local ancestries contributed to the phenotypic variation; whereas in Model 3, both global and local ancestries contributed to the phenotypic variation. Models 4, 5 and 6 were designed to evaluate the utilities for fine mapping for different methods. Besides random effect and the effects of 66 QTLs, in Model 4, only global ancestry contributed to the phenotypic variation; in Model 5, only local ancestries contributed to the phenotypic variation; whereas in Model 6, both global and local ancestries contributed to the phenotypic variation. We randomly sampled three SNPs on each chromosome as QTLs for all the three models, but for Models 5

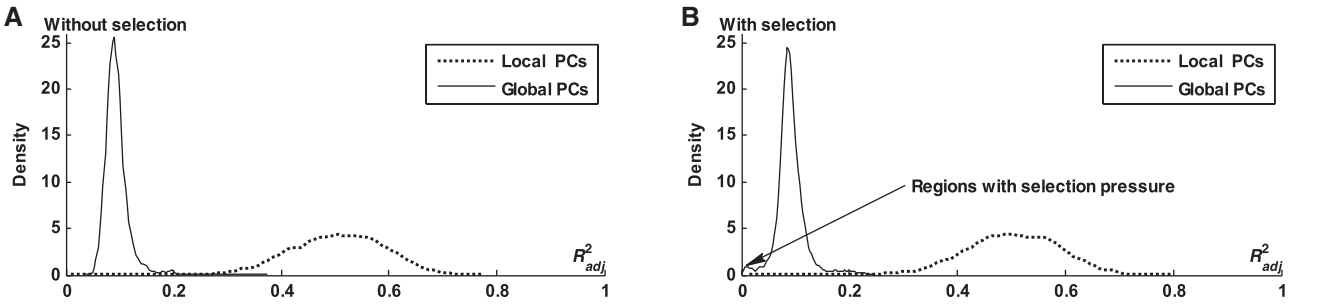

Fig. 2. Distributions of R^2_{adj} values of true ancestry to local PCs and global PCs. (**A**) Data were simulated by GenoAnceBase0 in which no selection pressure was simulated. (**B**) Data were simulated by GenoAnceBase1 in which selection was simulated in 22 local regions. The hump of the global PCs distribution near the origin was caused by the 22 regions where selection was on.

and 6 we let one QTL fall in the region where selection pressure was present. This enabled us to investigate the impact of local ancestry on the power for different methods. Among the 66 QTLs we simulated, one had a strong effect and explained phenotypic variance nine times higher than each of the other 65 QTLs, which explained the same amount of phenotypic variance. The variances expressed by global ancestry, local ancestry and QTLs for all the six models were listed in Supplementary Table S1. For both power and type I error comparisons, we simulated datasets with 2000 individuals each. To mimic the Maywood Affymetrix 6.0 data, we kept 554 297 of the Maywood SNPs which were present in HapMap for CEU, YRI and CHB/JPT samples, and sampled additional 245 703 SNPs from the HapMap via rejection sampling (Supplementary Fig. S2). For Models 4–6, we sampled only 29 QTLs and the others were assumed to have not been genotyped.

3.2 Results

3.2.1 PCs and true local ancestries We used simulations to examine whether the local PCs can represent local ancestries better than global PCs. Using the Maywood and HapMap datasets, we simulated a dataset using the GenoAnceBase0 algorithm to mimic the admixture of two ancestral populations without local population structure, i.e. no selection pressure confined to specific genomic regions. We also simulated a dataset using the GenoAnceBase1 algorithm to ensure the presence of local population structure under the circumstances where selection was acting on specific loci. Since selection may occur in different regions in the genome, in the datasets generated using the GenoAnceBase1 algorithm we randomly selected 1 local region in each of the 22 autosomes with a selection effect. Since the real ancestry at each locus was known, we calculated the adjusted R^2-value (R^2_{adj}) by regressing the real ancestry at each SNP on the first 10 local PCs of the enveloped window. For comparison, we also computed the R^2_{adj} value by regressing the real ancestry at each SNP to the first 10 global PCs. Figure 2 displays the distributions of R^2_{adj} values for local and global PCs in the two simulated datasets. For both simulated datasets, local PCs reflected variation of real ancestry much better than global PCs. For the dataset by GenoAnceBase0, $R^2_{adj} = 0.50 \pm 0.09$ for local PCs wheras $R^2_{adj} = 0.10 \pm 0.03$ for global PCs. For the dataset by GenoAnceBase1, $R^2_{adj} = 0.50 \pm 0.09$ for local PCs, whereas $R^2_{adj} = 0.09 \pm 0.03$ for global PCs. The R^2_{adj} values of the real ancestries for global PCs in the 22 regions under selection pressure were even smaller when compared with other genomic regions, as evidenced by the hump in Figure 2B due to the 22 regions with selection pressure. Our results indicate that global PCs do not capture local ancestry well, especially when significant effects from natural selection are present.

3.2.2 Type I When there are no genetic variants contributing to the phenotypic variation, a quantile-quantile (Q–Q) plot of $-\log_{10}(P\text{-value})$ of a test statistic against a uniform distribution should be close to the diagonal line. Supplementary Figure S3 displays the Q–Q plots of single SNP analysis for the four different methods when applied to the datasets simulated from Models 1 to 3. RegUa performed poorly because there was no adjustment of population structure. When only global ancestry contributed to phenotypic variation (Model 1), both RegRA and LAPCC were within the 95% concentration band and GAPCC was associated with a slightly inflated type I error rate (Supplementary Figs S3Ai and S3Aii). However, when phenotypic variation was influenced by local ancestry (Models 2 and 3), GAPCC yielded notably inflated type I error rates whereas both RegRA and LAPCC still controlled type I error rates close to the nominal level (Supplementary Figs S3B and S3C). The Q–Q plots of GAPCC went beyond the 95% concentration band and the λ-values were substantially larger than 1. GAPCC failed to control type I error rates when the local ancestries contributed to phenotypic variation. In contrast, the Q–Q plots of RegRA and LAPCC were concentrated around the diagonal line and the λ-values were close to 1.

To address the robustness of LAPCC to the change of window-wide SNP density, we dropped the odd-numbered SNPs and re-analyzed the genotype-ancestry data of even-numbered SNPs in the simulated Affymetrix 6.0 datasets and studied the representativeness of the first 10 PCs of 20-Mb local windows for the ancestry structures of their 4-Mb cores. Even though half of SNPs dropped, the patterns of false positive control were essentially unchanged (comparing Supplementary Fig. S4 with S3). This suggested that window-wide local PCs well maintained the representativeness for true local ancestry structures (Supplementary Fig. S5). The denser a window, the better the local PCs capture the true ancestry. However, the improvement is limited in general, suggesting the SNP density of Affymetrix 6.0 or Affymetrix 500K is adequate for controlling population stratification using LAPCC (Supplementary Fig. S6 and Table S2). It should be noted that ideally we would require each

Table 1. Power comparisons of the four methods

Models	Power	GAPCC	RegRA	LAPCC
$\sigma_i^2 = 0.1$				
1	QTL1	0.999	0.996	0.993
	Average[a]	0.310	0.297	0.289
2	QTL1	1.000	0.982	0.984
	Average	0.331	0.312	0.300
3	QTL1	0.998	0.968	0.956
	Average	0.291	0.260	0.250
$\sigma_i^2 = 0.2$				
1	QTL1	1.000	1.000	1.000
	Average	0.534	0.507	0.495
2	QTL1	1.000	1.000	1.000
	Average	0.536	0.508	0.490
3	QTL1	1.000	1.000	1.000
	Average	0.496	0.438	0.420

[a]Under each model, the average power was the mean of the power for the 29 QTLs.

testing SNP is centered with its own 20-Mb window and our approach is only for reducing computational burden when analyzing GWAS data. In this case, the performance should be even better.

3.2.3 Power We evaluated the power of different methods based on 1000 replicate datasets generated under three QTL models (Models 4–6 in Supplementary Table S1). Table 1 shows the power for the three methods, viz, GAPCC, LAPCC and RegRA. Power for LAPCC and RegRA was only slightly lower than GAPCC for all the models we evaluated. However, power of GAPCC could be overestimated when local ancestries contributed to the phenotypic variation because GAPCC might fail to control type I error as noted below. When true ancestry was known, RegRA seemed to have slightly better power than LAPCC.

3.2.4 Fine mapping Since linkage disequilibrium created by admixture may extend over much longer distance than the background LD (Zhu *et al*., 2006), we expect to see association signals for many non-causal markers surrounding a true QTL if local population structure is not properly controlled, even when the distance to the QTL is over several mega base pairs. We examined the performance of LAPCC and GAPCC for the 1351 SNPs within a 5 Mb region of the strongest QTL (Fig. 3). When global ancestry contributed to phenotypic variation, LAPCC and GAPCC performed similarly and we did not observe high significance for the SNPs far away from the QTL (Fig. 3A). However, when local ancestry contributed to the phenotypic variation, we observed that substantially more SNPs were in association with GAPCC than LAPCC, and many of them were several Mbp away from the true QTL (Figs 3B and 3C). Similarly, we also observed more SNPs showing association evidence with GAPCC than LAPCC for QTLs with smaller effects (data not shown). Our results suggest that false positives might be detected if local population structure at non-causal markers is not properly controlled. LAPCC could effectively eliminate such spurious association and thus improve mapping resolution.

4 APPLICATIONS TO GWAS DATASETS

4.1 PCA

We compared local and global PCs from the three GWAS datasets to uncover local ancestry patterns in the three populations. Figure 4 presents the R^2-values of the first local PC in each window and global PCs as well as the largest λ^2-value between local PCs in each window and global PCs. The distribution of the top 10 R^2- and λ^2-values are presented in Supplementary Figure S7. In general, we observed a small degree of correlation between local and global PCs in the samples we studied. However, the R^2 of the first local PC and λ^2 are substantially larger for the Maywood and Framingham samples than for Nigerian sample (Fig. 4), similar for the R^2 values for the rest local PCs (Supplementary Fig. S7). The results of the other λ^2 values for the three GWAS datasets are similar. These results suggest that there is relatively little population structure in the Nigerian sample, whereas Maywood and Framingham display relatively more population structure. Somewhat surprisingly, the Framingham participants demonstrated much more complex local population structure than the other two samples, at odds with the usual assumptions. To alleviate the concern that the finding in the Framingham data was caused by the presence of related individuals we removed 22 participants who shared >15% of their genome identical by descent. The results were essentially the same (data not shown).

4.2 Association analysis

To the three GWAS datasets, we applied RegUa, GAPCC, LAPCC to study the BMI and height association. To remove the potential confounding effect of age and gender, we regressed log(BMI) and height on age and obtained the residuals, within each gender (Kang *et al*., 2010). We then performed linear regression by regressing the residuals on each SNP, which was coded additively. Supplementary Figure S8 shows the *Q–Q* plots of the $-\log_{10}(P\text{-value})$ for association tests for the three methods. We observed that the three methods performed similarly for both the Maywood and the Nigeria datasets, but their performance differs substantially for Framingham (Supplementary Figs S8A and S8B). Interestingly, RegUa had similar GC inflation λ-values as the other two methods for both height and BMI in Maywood, which seems contradictory to the intuition, as it is generally believed that population structure will inflate type I errors. For the Nigeria dataset, the three methods performed equally well, consistent with the aforementioned observation that the Nigerian samples did not have either global or local population structure. For the Framingham data, we observed the largest GC inflation λ-value for RegUa, followed by LAPCC and GAPCC for both BMI and height (Supplementary Fig. S8C). One possible reason for the relatively large λ-value for LAPCC is that the 10 PCs we calculated may not well capture the local population structure if a specific window does not include enough ancestry informative markers (AIMs). It is also possible that there are a large number of height-associated variants across the genome and the collection of these variants will result in departure from the diagonal line for the *Q–Q* plot. To investigate this further, we examined the known spurious association between SNPs in the LCT gene and height, which results from Northern and Southern European population structure in European Americans (Campbell *et al*., 2005). Both LAPCC and

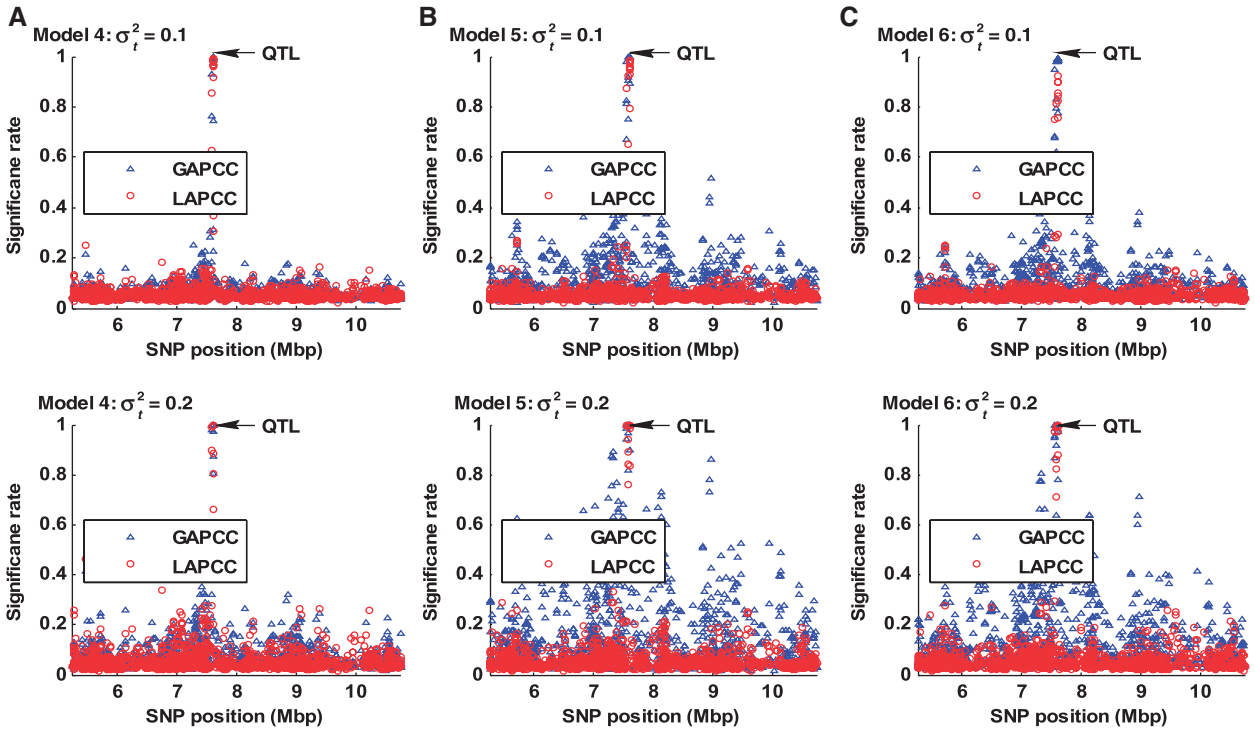

Fig. 3. Significance rates at 1351 SNPs in the 5 Mb region of the strongest QTL. For each method, the significance rate of each SNP was calculated as the proportion of SNPs with a $P < 0.05$ based on 1000 replicates. GAPCC apparently claimed significance more frequently than LAPCC for the SNPs away from the true QTL. (**A**) Global ancestry contributes to phenotypic variation. (**B**) Local ancestries contribute to phenotypic variation. (**C**) Both global and local ancestries contribute to phenotypic variation.

GAPCC eliminated this spurious association but RegUa did not (Table 2).

Many genetic variants have been reported to contribute to variation in human height (Gudbjartsson *et al.*, 2008; Lettre *et al.*, 2008; Weedon *et al.*, 2007, 2008). We searched the published literature and identified 121 SNPs reported to be in strong association with height, 56 out of which were also available in the Framingham dataset. Supplementary Figure S9A displays the *Q–Q* plot obtained from these 56 SNPs. The LAPCC had a slightly larger λ-value than GAPCC. We further examined the 100-Kb regions surrounding the 121 reported SNPs and identified 2879 SNPs genotyped in the Framingham dataset. We then presented the *Q–Q* plot for these SNPs and again LAPCC had a larger λ-value than GAPCC (Supplementary Fig. S9B). In both cases, we observed that LAPCC had larger λ-values than GAPCC, which could be potentially caused by cryptic relatedness among the subjects. To eliminate the impact of relatedness, we further removed the 22 individuals as identified by IBD analysis (Section 4.1) and analyzed the refined data for height association. The *Q–Q* plot patterns were essentially unchanged although the λ-values of RegUa and GAPCC increased slightly (Data not shown). Thus, the observed larger λ-value for LAPCC in Framingham height association might be caused by large amount of true genetic variants contributing to height, although we cautioned that LAPCC could not perform well in some regions without sufficient AIMs. Further simulations (Supplementary Figs S10–S12) indicate that 600–2000 QTLs with small genetic effects together with SNPs in LD with QTLs may lead the genomic inflation factor to be larger than one, further suggested that the observed larger λ-value for LAPCC might be caused by large amount of height QTLs.

5 DISCUSSIONS

In this article, we investigated the effectiveness of local and global ancestry adjustments for fine mapping in genetic association studies. Through extensive simulations, we observed that local PCs were strongly correlated with true local ancestry, whereas global PCs were merely weakly correlated with the true local ancestry. Moreover, our simulations illustrated that adjusting local ancestry (i.e. LAPCC or RegRA) was able to remove spurious association due to both global and local population structure and retain statistical power, whereas global ancestry adjustment methods such as genomic control and global PC-based methods may be inadequte. In our simulations, many significantly associated SNPs surrounding the causal variants in the GAPCC analysis were no longer significant in the LAPCC analysis, whereas the causal variants are essentially not affected. When spurious association is created due to the long-range LD, which is shown in Figure 1 when the correlation between Λ_G and Λ_D is due to coevoluation, our proposed method should also work well. Further, we point out that adjusting local ancestry may in theory reduce type I error equally well as adjusting global ancestry when local ancestry confounds association only through global ancestry. When only global population structure is present, the power of LAPCC is still comparable to the power of GAPCC.

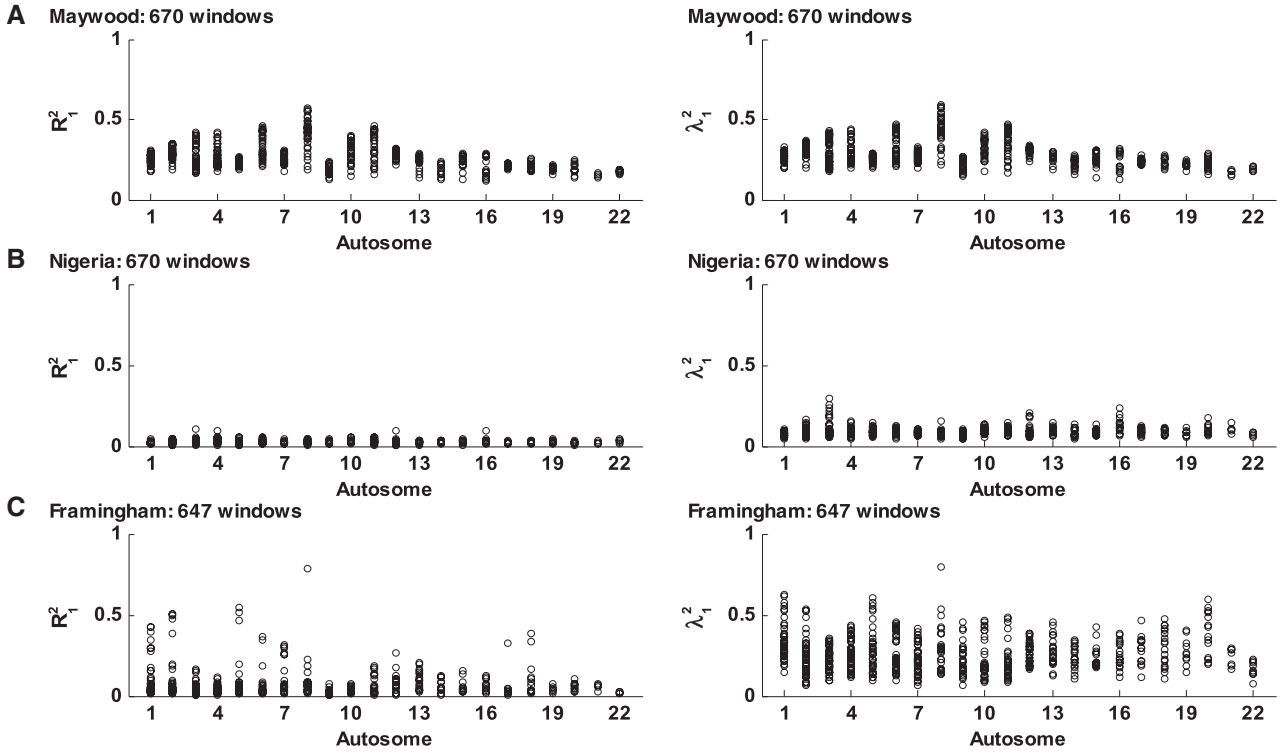

Fig. 4. The distributions of R_1^2 and λ_1^2 values of the local windows in three GWAS datasets. For each window in a given genotype dataset, R_1^2 was the coefficient of multiple determination of the first local PC versus the first 10 global PCs, λ_1^2 was the largest squared coefficient of canonical correlation between the first 10 local and the first 10 global PCs. (**A**) Maywood; (**B**) Nigeria; (**C**) Framingham.

Table 2. The association between polymorphisms in LCT gene and height

Chr.	SNP	*P*-values		
		RegUa	GAPCC	LAPCC
2	rs2322660	0.000062	0.329677	0.625346
2	rs2322659	0.000447	0.303102	0.790376
2	rs2304371	0.004120	0.646870	0.688207
2	rs2015532	0.004254	0.443551	0.962255
2	rs3769013	0.015098	0.813765	0.492297
2	rs1042712	0.016402	0.750703	0.608202
2	rs3769012	0.022521	0.904163	0.500595
2	rs872151	0.040061	0.662543	0.765082
2	rs4954445	0.053337	0.759460	0.765623

Our simulation findings were corroborated by the application of PCA, LAPCC and GAPCC to the three real GWAS datasets. We observed that there could be abundant variation in local ancestry structure across the genome and local ancestry would be weakly correlated to global ancestry for admixed poulations, i.e. African Americans and European Americans, consistent with the demographic histories of these populations. For the three GWAS datasets, the association results produced by LAPCC are roughly comparable to those produced by GAPCC. For the Framingham height data, RegUa had the largest GC inflation λ-value, followed by LAPCC and GAPCC. To investigate whether it was due to the reduced power of GAPCC, we further examined the SNPs surrounding the loci reported to be associated with height. As expected, LAPCC had much larger GC inflation λ-value than GAPCC did, suggesting that the smaller λ-value for GAPCC could be due to the power reduction since global ancestry adjustment might not completely remove the effect of population stratification. Furthermore, we examined the SNPs in the LCT gene which was reported to be spuriously associated with height in European Americans. LAPCC firmly abolished all those SNPs found positive by RegUa, suggesting that local ancestry adjustment would control the effect due to population structure in European populations. For the Maywood data, we observed that RegUa had a similar GC inflation λ-value to LAPCC and GAPCC, fitting a continuous gene flow model under which population structure would not have a large effect in association studies (Rosenberg and Nordborg, 2006). For the Nigerian data, all the three methods performed similarly, and this observation is consistent with our observation that no substantial local or global population structure is present in Nigerians.

Like other approaches, the LAPCC has limitations. The SNP density of a segment may impact its performance. In our analyses, it worked well when a 20-Mb window covered >1000 SNPs in minimum (Supplementary Figs S3–S6 and Table S2). We caution that it may not perform well in the regions without enough AIMs. It is reasonable to run LAPCC on the subset of markers which are found positive by GAPCC. In doing so, we could prevent false positives due to other confounding resources, improve fine mapping and save computing time as well. In addition, it may

be possible to improve LAPCC by introducing some metric of defining a meaningful window size instead of working with a fixed window size.

ACKNOWLEDGEMENTS

We thank the other members in Dr Zhu's lab, Dr Guimin Gao and some other early readers for their helpful comments. We are grateful to the investigators of the Framingham Heart Study who collected and managed the data and to the participants for their invaluable time, patience and dedication to the study.

Funding: National Institutes of Health; National Heart, Lung, and Blood Institute (HL074166, HL086718, HL053353); National Human Genome Research Center (R01HG004517, HG003054).

Conflict of Interest: none declared.

REFERENCES

Campbell,C.D. *et al*. (2005) Demonstrating stratification in a European American population. *Nat. Genet.*, **37**, 868–872.

Cavalli-Sforza,L.L. and Bodmer,W.F. (1999) *The Genetics of Human Populations*. Dover Publications, Mineola, New York.

Crow,J.F. and Kimura,M. (1970) *An Introduction to Population Genetics Theory*. Harper & Row, New York, pp. 469–478.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Epstein,M.P. *et al*. (2007) A simple and improved correction for population stratification in case-control studies. *Am. J. Hum. Genet.*, **80**, 921–930.

Genovese,G. *et al.* (2010) Association of Trypanolytic ApoL1 variants with kidney disease in African-Americans. *Science*, **7**,1–7.

Gudbjartsson,D.F. *et al*. (2008) Many sequence variants affecting diversity of adult human height. *Nat. Genet.*, **40**, 609–615.

Kang,S.J. *et al*. (2010) Genome wide association of anthropometric traits in African and African derived populations. *Hum. Mol. Genet.*, **19**, 2725–2738.

Lander,E.S. and Schork,N.J. (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–2048.

Lettre,G. *et al*. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.*, **40**, 584–591.

Levy,D. *et al*. (2009) Genome-wide association study of blood pressure and hypertension. *Nat. Genet.*, **41**, 677–687.

Novembre,J. and Stephens,M. (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.*, **40**, 646–649.

Patterson,N. *et al*. (2004) Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.*, **74**, 979–1000.

Pearl,J. (2010) An introduction to causal inference. *Int. J. Biostat.*, **62**, 1–59.

Price,A.L. *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Price,A.L. *et al*. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.*, **5**, e1000519.

Pritchard,J.K. *et al*. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rohlfs,R.V. *et al*. (2010) Detecting coevolution through allelic association between physically unlinked loci. *Am. J. Hum. Genet.*, **86**, 674–685.

Rosenberg,N.A. and Nordborg,M. (2006) A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. *Genetics*, **173**, 1665–1678.

Sabeti,P.C. *et al*. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.

Sankararaman,S. *et al.* (2008) Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.*, **82**, 290–303.

Tang,H. *et al*. (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am. J. Hum. Genet.*, **79**, 1–12.

Tang,H. *et al*. (2007) Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am. J. Hum. Genet.*, **81**, 626–633.

Voight,B.F. *et al*. (2006) A map of recent positive selection in the human genome. *PLoS Biol.,* **4**, e72.

Weedon,M.N. *et al*. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.*, **40**, 575–583.

Weedon,M.N. *et al*. (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat. Genet.*, **39**, 1245–1250.

Zhu,X. *et al*. (2002) Association mapping, using a mixture model for complex traits. *Genet. Epidemiol.*, **23**, 181–196.

Zhu,X. *et al*. (2006) A classical likelihood based approach for admixture mapping using EM algorithm. *Hum. Genet.*, **120**, 431–445.

Zhu,X. *et al*. (2008a) A unified association analysis approach for family and unrelated samples correcting for stratification. *Am. J. Hum. Genet.*, **82**, 352–365.

Zhu,X. *et al*. (2008b) Admixture mapping and the role of population structure for localizing disease genes. *Adv. Genet.*, **60**, 547–569.