SCIENTIFIC REPORTS

OPEN

Identifying Rare Variant Associations in Admixed Populations

Huaizhen Qin[1,2], Jinying Zhao[1] & Xiaofeng Zhu[3]

Received: 15 March 2018
Accepted: 12 March 2019
Published online: 01 April 2019

An admixed population and its ancestral populations bear different burdens of a complex disease. The ancestral populations may have different haplotypes of deleterious alleles and thus ancestry-gene interaction can influence disease risk in the admixed population. Among admixed individuals, deleterious haplotypes and their ancestries are dependent and can provide non-redundant association information. Herein we propose a local ancestry boosted sum test (LABST) for identifying chromosomal blocks that harbor rare variants but have no ancestry switches. For such a stable ancestral block, our LABST exploits ancestry-gene interaction and the number of rare alleles therein. Under the null of no genetic association, the test statistic asymptotically follows a chi-square distribution with one degree of freedom (1-df). Our LABST properly controlled type I error rates under extensive simulations, suggesting that the asymptotic approximation was accurate for the null distribution of the test statistic. In terms of power for identifying rare variant associations, our LABST uniformly outperformed several famed methods under four important modes of disease genetics over a large range of relative risks. In conclusion, exploiting ancestry-gene interaction can boost statistical power for rare variant association mapping in admixed populations.

Admixture is an omnipresent evolutionary force in complex disease genetics of recently admixed populations. Admixture mapping locates genomic segments that harbor causal alleles with distinct ancestral frequencies through admixture linkage disequilibrium (ALD). It has been successfully applied to locate genetic variants for a range of diseases and traits, e.g., hypertension[1,2], type 2 diabetes[3,4], obesity[5,6] and Alzheimer's dementia[7]. Systematic reviews of admixture mapping approaches can be found in the literature[8–10]. Genome-wide association studies (GWASs) have proven successful in identifying individual common genetic variants associated with common diseases and traits[11]. In the era of GWASs, admixture mapping has become a useful compliment for identifying common variant associations[12]. Several hybrid methods for combining the genotype and allele ancestry at a single-nucleotide polymorphism (SNP) have been developed[13–15]. These methods may claim variants which are in ALD with the true causal variants but not associated with the phenotype in any ancestral population. ALD may extend for substantial distances[16–18]. Qin and Zhu[12] proposed a two-stage fine mapping method to first identify candidate local genomic segments and then identify individual variants responsible for the admixture mapping evidence.

The common variants identified in GWASs merely explain a small proportion of the heritability[19], leading to many explanations of the 'missing' heritability[20–23]. A potential source of the missing heritability is the contribution of rare variants[24,25]. Evidenced by deep sequencing studies[26–28], rare variants may have stronger effects on complex diseases than do common variants. Multiple methods have been developed for identifying rare variant associations. Collapsing methods, e.g., the CAST[29] and the CMC[30], utilize the number of rare alleles in a gene for each individual to enrich association information. The SDWSS[31] scales SNPs in a test set by their minor allele frequencies in unaffected individuals. It utilizes a Wilcoxon type statistic to aggregate information and assesses the significance by permutation. The VT method[32] utilizes the maximum of the test statistics over all allele-frequency thresholds. All these methods implicitly assume that all effects have an identical direction. To combine the effects of opposite directions, the data-adaptive sum test[33] incorporates the signs of the observed effects into the CAST, whereas the C-alpha method[34] and the SKAT[35] test for genetic variance component. In particular, two methods have been proposed to combine the effects of different sizes and opposite directions. The

[1]Department of Epidemiology, College of Public Health and Health Professions and College of Medicine, University of Florida, Gainesville, FL, 32611, USA. [2]Department of Global Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, New Orleans, LA, 70112, USA. [3]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, 10900 Euclid Avenue, Cleveland, Ohio, 44106, USA. Correspondence and requests for materials should be addressed to X.Z. (email: xxz10@case.edu)

ORWSS[36] scales SNP wise numbers of minor alleles by the logarithms of amended odds ratios in the 2×2 tables of disease status by allele states. The EREC method[37] scales SNP wise numbers of minor alleles by the estimated regression coefficients. When families are available, incorporating linkage evidence in rare variants analysis has also been developed[38–41].

However, these existing methods may be underpowered for identifying rare variant associations in admixed populations, because they do not explicitly exploit the association information conveyed by local ancestries, particularly, ancestry-gene interaction. An admixed population and its ancestral populations often bear different burdens of complex diseases, partially due to the ancestral discrepancies in causal alleles, allele frequencies, and effects. Within a chromosomal block harboring causal alleles, an affected admixed individual may have an increased probability of inheriting alleles from the ancestry population of higher disease prevalence[12,42]. Common variants with different ancestral frequencies are correlated to their ancestries[43,44] which provide non-redundant association information[12–15]. We hypothesize that this argument holds for rare variants. Single rare variant association testing has unacceptably limited power since only a small portion of study individuals carry the rare allele.

In this report, we will illustrate the utility of explicitly exploiting local ancestries and genotypes together for rare variant associations. For simplicity, we aim to identify stable ancestral blocks harboring rare variants. For each person, all the SNPs within such a block share an identical ancestry. We propose a heuristic local ancestry boosted sum test (LABST). In a stable ancestral block, our test statistic combines the sum of SNP wise numbers of rare mutations and the ancestry-gene interaction. We mathematically prove that the LABST statistic asymptotically follows a chi-square distribution with 1-df if the test block is not associated with the disease. In extensive simulations, our LABST appropriately controlled type I error rates at preset nominal levels, indicating the ideal accuracy of the asymptotic approximation. Under various multiple rare variant disease modes with a large range of relative risks, our LABST were uniformly more powerful than the benchmark CAST as well as the sophisticated SDWSS and ORWSS. The LABST is a heuristic method designed for unrelated cases and controls. It can be extended to incorporate informative weights, to accommodate covariates and to allow for multiple groups of rare variants.

Methods

In an admixed population of two ancestral populations, let a test chromosomal block contain L rare variants, i.e., the minor allele frequencies (MAFs) $< 2\%$[45]. For n unrelated individuals from the admixed population, let y_i be disease status of individual i ($y_i = 1$, if individual i is affected; $=0$, if unaffected), $G_1 = \{i: y_i = 1\}$ and $G_0 = \{i: y_i = 0\}$ be the index sets of affected and unaffected individuals, respectively. Let g_{ij} denote the number of minor alleles carried by individual i at SNP j, $s_i = \sum_{j=1}^{L} g_{ij}$, and $\boldsymbol{s} = [s_1, \ldots, s_n]$. Let the test block be stable in terms of variant wise ancestries. In other words, we assume that each block wide haplotype of each individual is inherited entirely from one of the two ancestral populations without any ancestry crossover points. Under such an assumption, all the m SNPs within the block share identical ancestry. We define $\boldsymbol{a} = [a_1, \ldots, a_n]$, where a_i denotes the number of ancestries on individual i inherited from the ancestral population of the higher disease prevalence (due to the larger risk haplotype frequency). Let α be the nominal significance level of a test for block-based associations.

The proposed LABST. For each individual i, we define $u_i = (1 + a_i)s_i$ to combine ancestry-gene interaction $a_i s_i$ with s_i. We define a Welch type t statistic

$$W_e = \frac{\bar{u}_1 - \bar{u}_0}{\sqrt{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_0^2/n_0}}, \tag{1}$$

where $\bar{u}_1 = \sum_{i \in G_1} u_i/n_1$, $\hat{\sigma}_1^2 = \sum_{i \in G_1} u_i^2/n_1 - \bar{u}_1^2$, and $\bar{u}_0^2$ and $\hat{\sigma}_0^2$ are likewise defined using the u-scores of n_0 unaffected individuals. We use W_e^2 to measure the association between the test block and the disease status. Let n_1/n_0 converge to a finite positive constant τ when both n_0 and n_1 increase, e.g., $\tau = 1$ if $n_1 = n_0$. If the test block is not associated with the disease status, then the statistic W_e^2 converges in distribution to $\xi \sim \chi_1^2$, the chi-square distribution with 1-df (see Appendix A for a mathematical proof). Thus, we compute the p-value of W_e^2 as $P \stackrel{\text{def}}{=} \Pr(\xi > W_e^2)$ and claim significance when $P < \alpha$, the preset nominal significance level.

Existing methods. Most existing rare variant association methods exploit genotypes without explicitly capitalizing on ancestry-gene interactions. In such methods, the genotypic score of individual i is defined as $x_i = \sum_{j=1}^{L} w_j g_{ij}$, where w_j is a SNP wise weight. The simplest weight is $w_j \equiv 1$ for all the L SNPs as in the CAST[29]. For this universal weight, x_i collapses to s_i, and a benchmark 1-df statistic is constructed by replacing u_i's in our LABST with the s_i's.

The SDWSS[31] weighs a SNP using its minor allele frequency in unaffected individuals among whole-sample individuals. At the j^{th} SNP, $w_j = 1/\sqrt{nq_j(1 - q_j)}$, where $q_j = (1 + m_j)/(2 + 2n_0)$, and m_j is the number of minor alleles at the SNP over the n_0 unaffected individuals. Whole-sample individuals are ranked according to the x_i scores and a rank sum $x = \sum_{i \in G_1} \text{rank}(x_i)$ is defined. Let $x_1^*, \ldots, x_k^*$ be the rank sums based on $k(=1{,}000)$ permutations of disease status, $\hat{\mu}$ and $\hat{\sigma}$, be their mean and standard deviation, respectively. The standardized score is defined as $z = (x - \hat{\mu})/\hat{\sigma}$ and the P value of z is computed according to the standard normal distribution.

The weighting scheme in the SDWSS favors the disease-associated mutations with very low frequencies. As acknowledged by its authors, however, this scheme may reduce the power to detect the disease-associated mutations with higher frequencies. The SDWSS is based on the implicit assumption[32] that $\log(\text{OR}_j) \propto 1/\sqrt{q_{0j}(1 - q_{0j})}$, where OR_j is the odds ratio in the 2×2 table of disease status by the allele at SNP j, and q_{0j} is the MAF in the controls. Thus, Feng *et al.*[36] proposed the ORWSS to jointly analyze rare and common variants. This method keeps

all the steps of the SDWSS except for the weighting scheme. In the ORWSS, a SNP is weighted by the logarithm of the amended odds ratio[46] in the 2×2 table of allele by disease status. The amended odds ratio proves a useful remedy for handling potential empty cells in SNP-wise tables.

Type I error rate inflation factor. Often, a conservative method tends more likely to miss true associations whereas a liberal method tends more likely to claim false positives. A valid powerful method should accurately control the type I error rate at each preset nominal level. Herein, we propose and use type I error rate inflation factor (TIERIF) to measure how accurately a method controls type I error rate. For a given nominal level α, we define the TIERIF of a method as $\gamma_\alpha = \tau_\alpha/\alpha$, where τ_α is the probability that the method rejects the null hypothesis. If $\gamma_\alpha = 1$, then the method is able to controls type I error rate at the given nominal level α. If γ_α is substantially smaller than 1, then the method is overly conservative. If γ_α is substantially larger than 1, then the method is overly liberal.

Usually, it is intractable to mathematically formulate the TIERIF of a sophisticated method. In addition, it is hard to tell what a TIERIF is unacceptably 'small' or 'large'. Herein, we propose an empirical method to estimate this quantity and tell how small (large) is too small (large). Specifically, we define $\hat{\gamma}_\alpha = \hat{\tau}_\alpha/\alpha$ as an estimator of γ_α, where $\hat{\tau}_\alpha$ is the frequency that the method claims significance over R simulation replications generated under the null hypothesis of no association. As R increases, $\hat{\gamma}_\alpha$ converges in probability to γ_α, and $\sqrt{R\alpha}(\hat{\gamma}_\alpha - \gamma_\alpha)/\sqrt{1-\alpha}$ converges in distribution to a standard normal variable (see Appendix B for a mathematical proof). Therefore, if the method properly controls type I error rate at α, then $\hat{\gamma}_\alpha$ concentrates with probability 95% between

$$LB_\alpha = 1 - 1.96\sqrt{(1-\alpha)/(R\alpha)} \tag{2}$$

and

$$UB_\alpha = 1 + 1.96\sqrt{(1-\alpha)/(R\alpha)}. \tag{3}$$

Under the null of no genetic association, the concentration interval $[LB_\alpha, UB_\alpha]$ is the shortest among all the intervals $[LB, UB]$ such that $\lim_{R\to\infty} \Pr_0(LB \le \hat{\gamma}_\alpha \le UB) = 0.95$ (Appendix B). A method is called to be overly conservative if $\hat{\gamma}_\alpha < LB_\alpha$. Likewise, a method is called to be overly liberal if $\hat{\gamma}_\alpha > UB_\alpha$.

Simulation Designs

For method comparisons, we simulated an admixture using the rare variants with the frequency-spectrums of two natural populations. In the simulated admixture, block-wide haplotypes were inherited from the two ancestry populations. Four disease genetic modes were considered, including the dominant, additive, recessive, and multiplicative modes. Under each disease genetic mode, the disease status of an admixed individual was determined by the penetrance conditioning on block-wide risk haplotypes other than individual risk alleles.

Admixture. To simulate a two-way admixture, we downloaded the genotype data of region ENr113.4q26 from the ENCODE project Consortium[47]. Applying the Beagle software[48], we separately inferred 180 CEU (Centre d'Etude du Polymorphisme Humain in Utah, USA) and 180 YRI (Yoruban in Ibadan, Nigeria) haplotypes over the ENr113.4q26 region. Details on the haplotype deconvolution have been described previously[36]. Across the 360 inferred haplotypes, we observed 1,693 SNPs. At each of the region-wide SNPs, we chose the minor allele in the YRI haplotype data ($f_{YRI} \le 0.5$) as the reference allele (Fig. 1a). In the CEU haplotype data, the reference alleles at 1,373 SNPs are of frequencies $f_{CEU} \le 0.5$, whereas at the other 320 SNPs, $f_{CEU} > 0.5$ (Fig. 1b). Based on our previous association study on African Americans[43], we adopted $\omega = 0.8$ vs. $\varpi = 0.2$ as YRI-CEU admixture weights. To 'genotype' one admixed individual in the ENr113.4q26 region, we randomly chose one and another haplotype from the YRI or CEU haplotypes with probabilities ω vs. ϖ. In this simulated admixture, the frequencies of reference alleles at the 1,693 SNPs ($f_{ADX} = \omega f_{YRI} + \varpi f_{CEU}$) range from 0.0011 to 0.5722 (Fig. 1c), where 295 SNPs are of $f_{ADX} < 0.02$, satisfying the conventional criterion of rare variants[45].

Causal haplotypes and ancestries. From the 254 SNPs with $f_{ADX} < 0.015$, we randomly selected 23 SNPs as deleterious allele carriers (Table 1). The minor alleles at these 23 SNPs served as deleterious alleles. The deleterious alleles appear at 32 YRI and 4 CEU haplotypes, respectively. These 36 haplotypes served as risk haplotypes. The proportions of risk haplotypes in YRI and CEU haplotype data sets are $p_{YRI} = \frac{8}{45} \approx 0.1778$ and $p_{CEU} = \frac{1}{45} \approx 0.0222$, respectively. Thus, the proportion of risk haplotypes in the simulated admixture is

$$p_{ADX} = p_{YRI}\omega + p_{CEU}\varpi = \frac{8}{45} \times \frac{4}{5} + \frac{1}{45} \times \frac{1}{5} = \frac{11}{75} \approx 0.1467. \tag{4}$$

For each admixed individual, we let H be number of risk haplotypes and let a be the number of YRI haplotypes. Table 2 presents $\Pr(H, a)$, the joint probability mass of (H, a), where $q_{(\cdot)} = 1 - p_{(\cdot)}$ for YRI, CEU and ADX, respectively. In this simulated admixture, the coefficient of correlation between H and a is

$$\rho = \frac{\sqrt{\varpi\omega}(p_{YRI} - p_{CEU})}{\sqrt{(\omega p_{YRI} + \varpi p_{CEU})(\omega q_{YRI} + \varpi q_{CEU})}} = \frac{7}{12\sqrt{11}} \approx 0.1759. \tag{5}$$

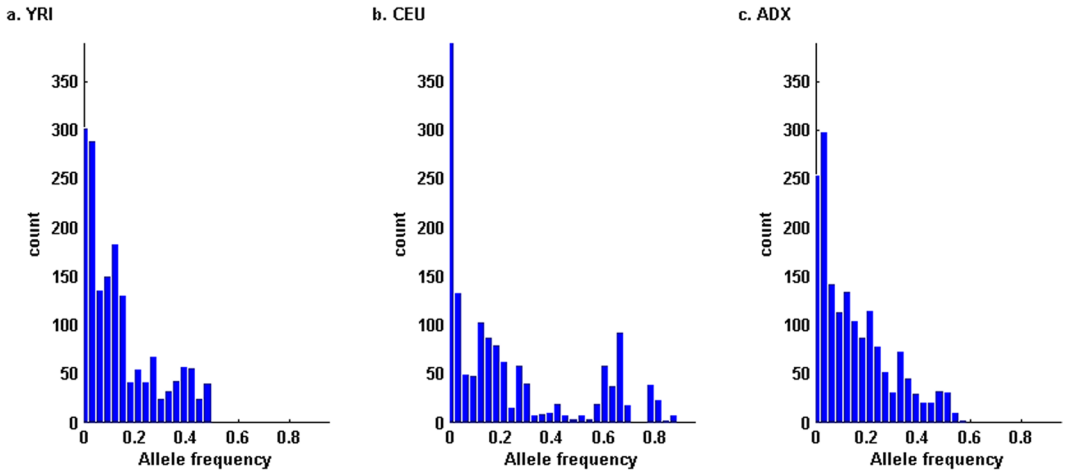

Figure 1. Population wise distributions of the reference alleles at 1,693 SNPs in ENr113.4q26. (**a**) In the YRI haplotype data, all the reference alleles are of frequencies $f_{YRI} \leq 0.5$. (**b**) In the CEU haplotype data, the reference alleles at 1,373 SNPs are of frequencies $f_{CEU} \leq 0.5$. (**c**) In the simulated admixture, the reference alleles at 295 SNPs are rare ($0.001 < f_{ADX} < 0.02$).

SNPs	Base pair positions	Reference alleles	Frequencies		
			f_{YRI}	f_{CEU}	f_{ADX}
rs10020425	118710926	C	1/180	0	0.004444444
rs4834616	118781194	G	1/90	1/180	0.01
rs17866922	118799469	T	1/180	0	0.004444444
rs17866231	118811782	G	1/180	0	0.004444444
rs17869437	118846471	A	1/90	0	0.008888889
rs17866969	118860941	T	1/60	0	0.013333333
rs17866219	118861303	G	1/180	0	0.004444444
rs11931936	118880832	A	1/60	0	0.013333333
rs17869443	118891959	G	1/180	0	0.004444444
rs13138706	118901935	A	1/90	0	0.008888889
rs17865249	118942421	A	1/180	0	0.004444444
rs17867496	118943543	G	1/180	1/60	0.007777778
rs17862037	118948929	C	1/180	1/60	0.007777778
rs17868674	118950202	C	1/180	0	0.004444444
rs17875131	118953676	C	1/180	1/60	0.007777778
rs11562912	118958053	C	1/60	0	0.013333333
rs17867082	118986061	C	1/60	0	0.013333333
rs11945465	119041637	C	1/90	0	0.008888889
rs11929977	119118756	C	1/180	0	0.004444444
rs17867208	119122424	A	1/60	0	0.013333333
rs17869338	119143264	T	1/180	0	0.004444444
rs17866812	119152274	T	1/60	0	0.013333333
rs17867083	119185964	T	1/90	0	0.008888889

Table 1. The distribution of the frequencies of deleterious alleles.

	$a=0$	$a=1$	$a=2$	$\Pr(H)$
$H=0$	$q_{CEU}^2\varpi^2$	$2q_{CEU}q_{YRI}\varpi\omega$	$q_{YRI}^2\omega^2$	q_{ADX}^2
$H=1$	$2p_{CEU}q_{CEU}\varpi^2$	$2\,(q_{CEU}p_{YRI}+p_{CEU}q_{YRI})\varpi\omega$	$2p_{YRI}q_{YRI}\omega^2$	$2p_{ADX}q_{ADX}$
$H=2$	$p_{CEU}^2\varpi^2$	$2p_{CEU}p_{YRI}\varpi\omega$	$p_{YRI}^2\omega^2$	p_{ADX}^2
$\Pr(a)$	ϖ^2	$2\varpi\omega$	ω^2	

Table 2. Generic probability mass function of ($\boldsymbol{H}$, $\boldsymbol{a}$) in the simulated admixture.

	$a=0$	$a=1$	$a=2$	Pr(H)
$H=0$	0.03824198	0.25726420	0.43267160	0.7281778
$H=1$	0.00173827	**0.0614716**	**0.1871012**	0.2503111
$H=2$	0.00000198	**0.0012642**	**0.0202272**	0.0215111
Pr(a)	0.04	0.32	0.64	

Table 3. Specific probabilities of ($\boldsymbol{H}$, $\boldsymbol{a}$) used in the simulation*. *Under this specific joint distribution of (H, a), the variance of $(1+a)H$ is Var$[(1+a)H]=2.019955$.

Modes of disease genetics. Let y be the disease status of an admixed individual (=1, if affected; =0, if unaffected). Let $f_H=\Pr(y=1|H)$ be the penetrance for a given H value (=0, 1, or 2). Then the disease prevalence $\kappa=\Pr(y=1)$ can be formulated as

$$\kappa = \sum_H f_H \Pr(H). \tag{6}$$

Let $RR=f_2/f_0$ be relative risk. Then, $f_1 = f_0 \cdot RR$ for dominant mode, $f_1 = \frac{1}{2}f_0 \cdot (1 + RR)$ for additive mode, $f_1 = f_0 \cdot \sqrt{RR}$ for multiplicative mode, and $f_1=f_0$ for recessive mode. Under each mode, $\Pr(y, a|H) = \Pr(y|H)\Pr(a|H)$, namely, the disease status is independent of local ancestry a given haplotype H. It follows that

$$\Pr(H, a|y) = \frac{\Pr(H, a)\Pr(y|H)}{\Pr(y)} \tag{7}$$

for an arbitrary (H, a, y). Setting $y=0$ in Eq. (7) yielding the joint probability mass of (H, a) in unaffected subpopulation:

$$\Pr(H, a|y = 0) = \frac{\Pr(H, a)(1 - f_H)}{1 - \kappa}. \tag{8}$$

Setting $y=1$ in Eq. (7) yielding the joint probability mass of (H, a) in affected subpopulation:

$$\Pr(H, a|y = 1) = \frac{\Pr(H, a)f_H}{\kappa}. \tag{9}$$

Eqs (8) and (9) and Table 2 are necessary and sufficient to mathematically formulate the Pearson coefficient of the correlation between H and a in the entire affected subpopulation $\rho_1 \stackrel{\text{def}}{=} \text{corr}(H, a|y = 1)$ and that in the entire unaffected subpopulation $\rho_0 \stackrel{\text{def}}{=} \text{corr}(H, a|y = 0)$.

Simulation configurations. Using Table 3, we numerically computed ρ_1 and ρ_0 values for $f_0=0.1$ and each RR under each of the four disease genetic modes (Fig. 2). Under all the four modes, ρ_1 increases and ρ_0 decreases from 0.1759 as RR increases from 1 (no genetic association) to 3. The dominant mode shows the largest ratio ρ_1/ρ_0, followed in turn by the additive mode, the multiplicative mode, and the recessive mode. Of note, $f_0=f_1=f_2=\kappa$ under the null hypothesis of no genetic association. We acknowledge that prevalence (κ) varies for different diseases in an admixed population. For example, about 10% African Americans suffer from lifetime major depressive disorder[49], whereas about 2.7% African American suffer from dementia[50]. In our simulations, we fixed $f_0=0.1$ as a reference value to inspect the type I error rate and power patterns of different association methods with respect to different disease modes, relative risks, sample sizes, and nominal significance levels.

For each RR value under each mode, at each replication we simulated region wide genotypes and ancestries of n_1 affected and n_0 unaffected individuals from the admixed population. For each specific scenario, we adopted sample sizes $n_1=n_0=500$ and then $n_1=n_0=2{,}000$ to inspect the impacts of sample sizes on power levels and type I error rates of the methods under comparison. These sample sizes are realistic in that they reflect the scales of recent deep sequencing studies in African Americans. For example, 489 Alzheimer's cases and 472 controls were sequenced a target sequencing study on Alzheimer's disease[51]. The Jackson Heart Study[52] has deeply sequenced more than 3400 African Americans.

To accurately evaluate the TIERIFs of the methods, we simulated 10^8 replications of genotypes and ancestries by setting $RR=1$ under each of the four disease modes. This number of replications is sufficient and necessary for evaluating type I error rates of gene-based tests at nominal genome-wide significance level (2.5×10^{-6}). For power comparisons, we generated 20,000 replications for each given RR value (>1) under each of the four disease modes. This number of replications would be sufficient for reliably inspecting power patterns.

The ORWSS was designed to accommodate both rare and common variants. Thus, for this method, we used all the region wide variants to compute the weighted-sum of genotypic scores. The SDWSS have been observed to reduce statistical power when more neutral common variants are included into the test statistic. Intuitively, the other two methods will also reduce statistical power if common neutral variants are included. In our power comparisons, therefore, we used the conventional threshold 0.02 to choose rare variants[36] to perform the other three methods. In the simulated admixture, the reference alleles at 295 SNPs proved rare ($f_{ADX}<0.02$). Hence, we used the numbers of minor alleles at these 295 SNPs to compute the sum scores in the CAST and our LABST as well as the weighted-sum score in the SDWSS.

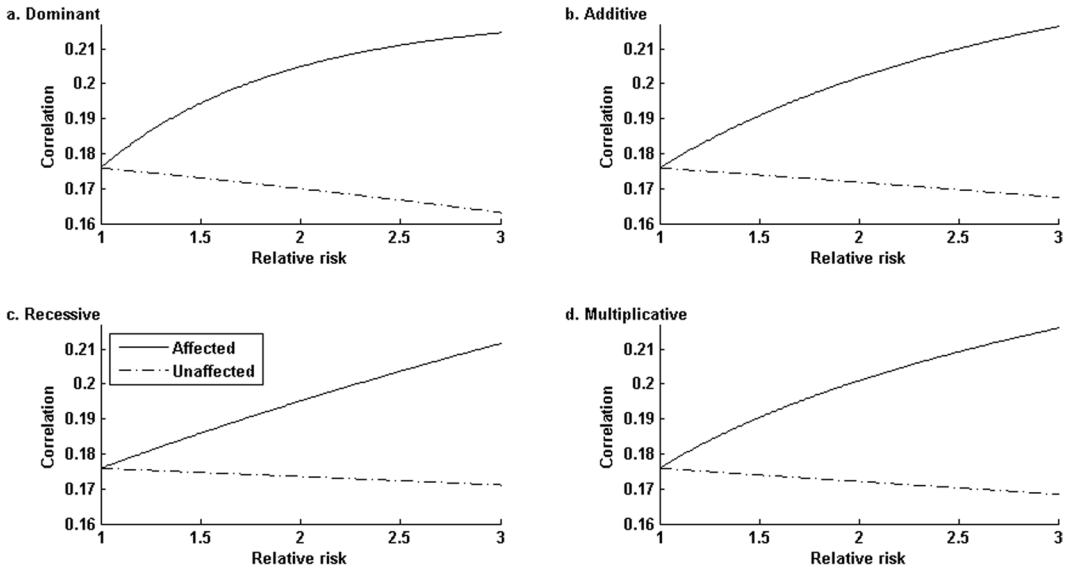

Figure 2. The trends of correlation between the numbers of causal haplotypes and ancestries under four modes of disease genetics. The correlation curves in each panel were generated by fixing $f_0 = 0.1$ and varying (f_1, f_2) according to the underlying genetic modes. Generically, as relative risk increases from 1 to 3, the coefficient of correlation between H and a in affected group corr$(H, a|D = 1)$ increases from 0.1759, whereas that in unaffected group decreases from 0.1759. The dominant mode shows the largest ratio of corr$(H, a|D = 1)$ to corr$(H, a|D = 0)$, followed in turn by the additive, multiplicative and recessive modes.

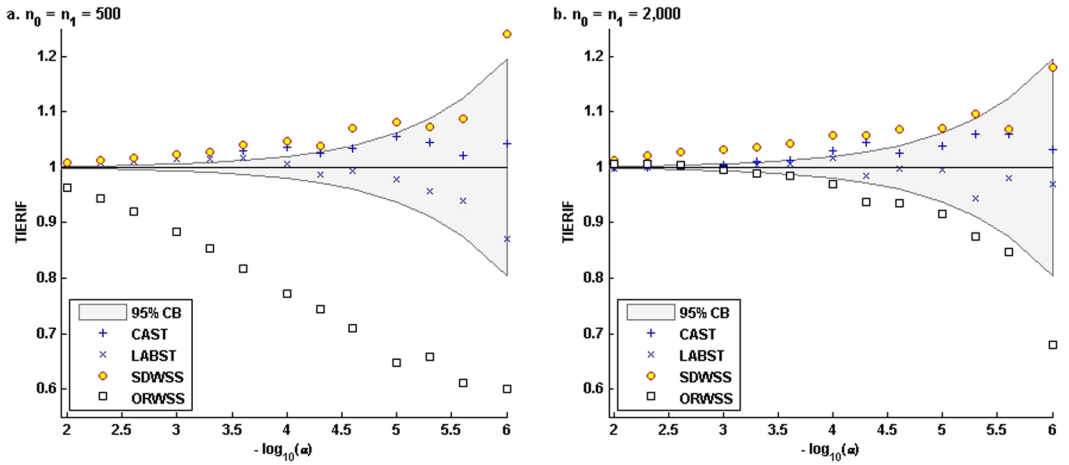

Figure 3. Empirical TIERIFs of the four methods based on different sample sizes and various nominal levels. In each panel, we generated 10^8 replications of region wide genotypes and ancestries of the specified numbers of affected and unaffected admixed to evaluate the TIERIF of each method at each nominal level. In the ORWSS, we used all the 1,693 variants to compute the individual weighted-sums of genotypic scores. In the other methods, we used the 295 variants of $f_{ADX} < 0.02$ to compute the individual sums/weighted-sums of genotypic scores. The LABST and the CAST accurately controlled the type I error rates. The SDWSS appeared overly liberal. The ORWSS appeared over conservative, particularly for the smaller samples.

Results

Type I error rates. Figure 3 presents the TIERIFs of the four methods. For both sets of sample sizes, the CAST and our LABST well control type I error rates for various nominal levels across interval $[10^{-6}, 0.05]$. They do not inflate or deflate type I error rates. Their TIERIF curves consistently concentrate around 1 and within the 95% concentration band (CB). The SDWSS appears overly liberal and the type I error rate inflation is quite robust to the increase in sample size. Its TIERIF curves clearly break the upper bound of the 95% CB for both the smaller and the larger sample size settings. These results would suggest that the SDWSS suffers a systematic bias in calibrating the tail probability of its test statistic. In contrast, the ORWSS appears overly conservative. Its TIERIF curves clearly break the lower bound of the 95% CB, especially for the smaller sample sizes. It becomes essentially less conservative for the larger sample sizes. These results would suggest that the ORWSS better calibrates the tail probability of its test statistic for larger sample sizes.

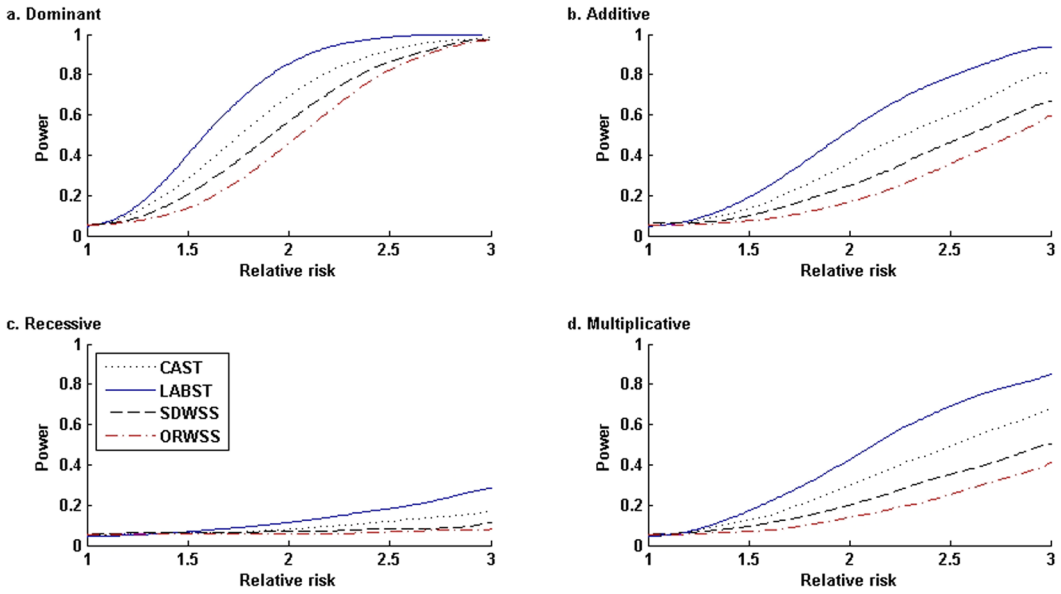

Figure 4. Power comparisons under four disease modes with various relative risks: $n_0 = n_1 = 500$, nominal level $\alpha = 0.05$. In the simulated admixture, all the 23 risk allele frequencies are less than 0.015, and the cumulative risk haplotype frequency is 0.1467. Under each mode, at each relative risk, we evaluated the power of each method based on 20,000 simulated replications. In the ORWSS, we used all the 1,693 variants to compute the individual weighted-sums of genotypic scores. In all the other methods, we used the 295 variants of $f_{ADX} < 0.02$ to compute the individual sums/weighted-sums of genotypic scores.

Power comparisons. Figure 4 presents the power comparisons under four disease genetic modes with sample sizes $n_0 = n_1 = 500$ and nominal level $\alpha = 0.05$. Overall, each method performs the best at the dominant mode, followed in turn by the additive mode, multiplicative mode, and lastly, recessive mode. Under each disease genetic mode, our LABST performs the best for all relative risks, followed in turn by the CAST, the SDWSS, and the ORWSS. Exploiting an identical set of rare variants, the CAST uniformly outperforms the SDWSS. Since the SDWSS is robustly liberal, its power inferiority would be caused by the transformation of the weighted-sums of genotypic scores to ranks, which would lose information. For the moderate sample sizes, the ORWSS appears unacceptably conservative and lacks ability to effectively separate the true causal variants from the other variants.

Figure 5 presents the power comparisons when increasing sample size to $n_0 = n_1 = 2{,}000$ but keeping all the other parameters used in Fig. 4. All the methods show increased power across all the different disease genetic modes. The LABST keeps its uniform preference, followed by the CAST, which outperforms the SDWSS and the ORWSS. However, the ORWSS now outperforms the SDWSS for a wide range of relative risks under the dominant, additive, and multiplicative modes. Under the recessive mode, the SDWSS slightly outperforms the ORWSS for all the relative risks. When increasing the sample size, the ORWSS becomes much less conservative and better scales the causal SNPs especially when RR becomes relatively large.

Figure 6 presents the power comparisons when reducing the nominal level to $\alpha = 10^{-6}$ but keeping all the other parameters in Fig. 5. At this significance level, the LABST still outperforms the other methods across all the scenarios. Under the recessive mode, all methods have very low or no power with respect to various relative risks. Under the other three modes, the CAST is more powerful than the SDWSS but becomes less powerful than the ORWSS, whereas the ORWSS become the second best among our compared methods for a wide range of relative risks.

Discussion

The primary objective of this report is to illustrate the utility of leveraging local ancestry for rare variant association analysis. We present the LABST to combine local ancestry of a test block with the sum of genotypic scores of block-wide rare variants. Under the null of no genetic association, we mathematically prove that the LABST statistic asymptotically follows the chi-square distribution of one degree of freedom. This explicit asymptotic null distribution enables us analytically compute the significance of each ancestry block. Under our extensive simulations, the LABST properly controls type I error rates at various preset nominal levels. These results indicate that the null distribution of the LABST statistic can be accurately approximated by the 1-df chi-square distribution. Based on our results, the permutation-based evaluations of significance in the SDWSS and ORWSS are not accurate enough for genome-wide scans for samples from admixed populations. The SDWSS tends to inflate type I error rates and the inflation appears robust to the changes of sample sizes. In other words, the SDWSS would suffer a systematic bias in calibrating the tail probability of its test statistic. The ORWSS appears severely conservative when the numbers of affected and unaffected individuals are moderate. Its conservativeness becomes less significant when the sample sizes are essentially increased. The conservativeness of the ORWSS would stem from the unideal stability and effectiveness of its weighting scheme.

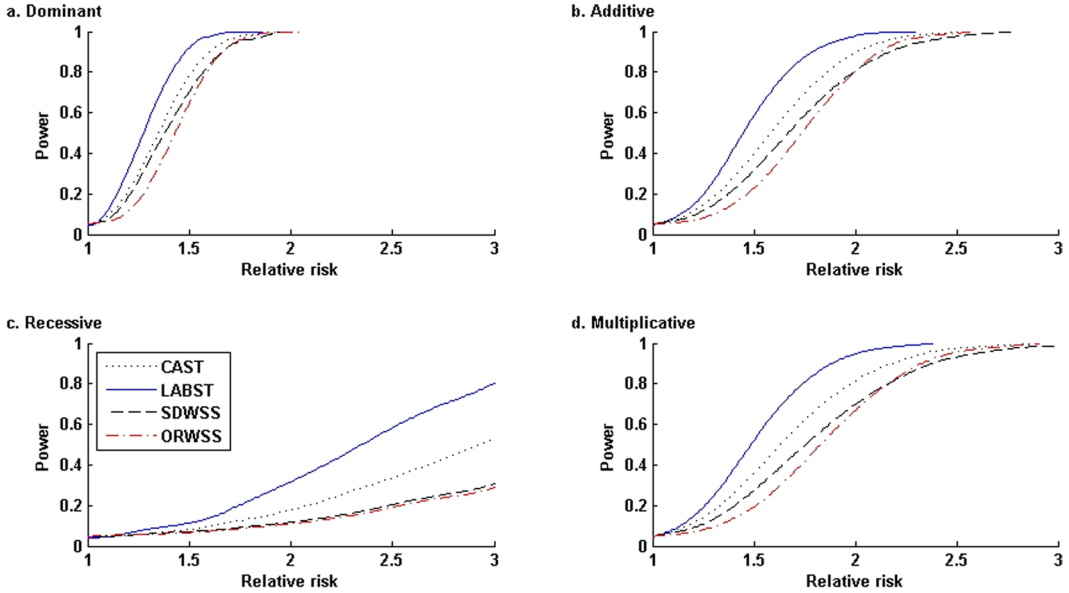

Figure 5. Power comparisons under four disease modes with various relative risks: $n_0 = n_1 = 2{,}000$, nominal level $\alpha = 0.05$. In the simulated admixture, all the 23 risk allele frequencies are less than 0.015, and the cumulative risk haplotype frequency is 0.1467. Under each mode, at each relative risk, we evaluated the power of each method based on 20,000 simulated replications. In the ORWSS, we used all the 1,693 variants to compute the individual weighted-sums of genotypic scores. In all the other methods, we used the 295 variants of $f_{ADX} < 0.02$ to compute the individual sums/weighted-sums of genotypic scores.

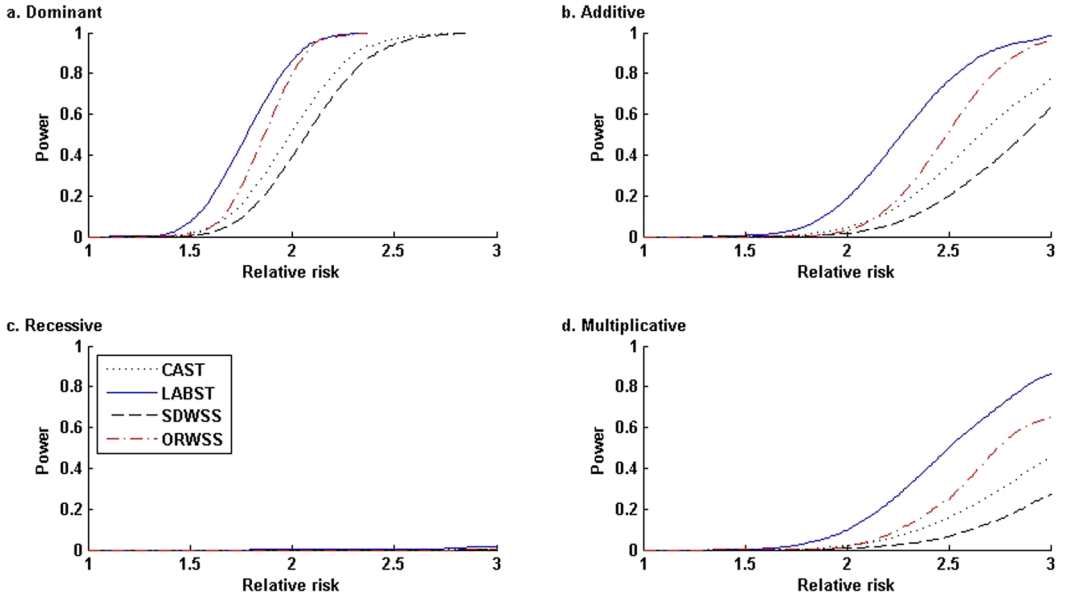

Figure 6. Power comparisons under four disease modes with various relative risks: $n_0 = n_1 = 2{,}000$, nominal level $\alpha = 10^{-6}$. In the simulated admixture, all the 23 risk allele frequencies are less than 0.015, and the cumulative risk haplotype frequency is 0.1467. Under each mode, at each relative risk, we evaluated the power of each method based on 20,000 simulated replications. In the ORWSS, we used all the 1,693 variants to compute the individual weighted-sums of genotypic scores. In all the other methods, we used the 295 variants of $f_{ADX} < 0.02$ to compute the individual sums/weighted-sums of genotypic scores.

In our simulations for the power comparisons, we hypothesize that certain haplotypes of some rare alleles are direct causal factors. This assumption allows for diverse SNP wise MAFs but does not necessarily mean that alleles with smaller MAFs have larger effect sizes. Under four disease genetic modes, the LABST are uniformly more powerful than the CAST, SDWSS, and ORWSS across all relative risks, sample sizes, and nominal levels investigated. Its power gain stems from explicit incorporation of the interaction between a gene and the local ancestry. The CAST is more powerful than the SDWSS uniformly across all our simulations, even though the SDWSS

is robustly liberal. The SDWSS loses portion of information when transforming the original weighted-sums of numerical genotypic scores to Wilcoxon rank-sum statistic. As pointed out by Wilcoxon himself, ranks are not sufficient statistics[53] and hence rank-sum test would not be the most powerful test. The superiority of the ORWSS to the CAST and the SDWSS varied across different disease genetic modes, relative risks, sample sizes, and nominal significance levels. Based on our results, the ORWSS would have limited utility for studies of small to moderate samples, whereas it would be useful for studies with large samples from a homogeneous population. Based on our results, a liberal method is not necessarily more powerful uniformly than a conservative method. The preference of a method depends on how effectively it can aggregate the association information of rare variants. Although derived from simulations on a particular region, all the conclusions are generalizable for an arbitrary admixed population with different ancestry-haplotype correlations between cases and controls. Such differences often stem from the different ancestral frequencies of the risk haplotypes, disease modes, and/or relative risks.

Our LABST can be extended to Hotelling's two-sample T-squared test[54] to jointly analyze multiple groups of variants when desired. Following the LABST, we can define u_{ij} as the integrative score of individual i in group j. For d groups, we write $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})'$ as the $d \times 1$ vector of integrative scores, $\overline{\boldsymbol{u}}_0 = \sum_{i \in G_0} \boldsymbol{u}_i / n_0$, $\overline{\boldsymbol{u}}_1 = \sum_{i \in G_1} \boldsymbol{u}_i / n_1$, and $\boldsymbol{V} = (n-2)^{-1} [\sum_{i \in G_1} (\boldsymbol{u}_i - \overline{\boldsymbol{u}}_1)(\boldsymbol{u}_i - \overline{\boldsymbol{u}}_1)' + \sum_{i \in G_0} (\boldsymbol{u}_i - \overline{\boldsymbol{u}}_0)(\boldsymbol{u}_i - \overline{\boldsymbol{u}}_0)']$. We define Hotelling's statistic as $T^2 = (n_0 n_1 / n)(\overline{\boldsymbol{u}}_1 - \overline{\boldsymbol{u}}_0) \boldsymbol{V}^{-1} (\overline{\boldsymbol{u}}_1 - \overline{\boldsymbol{u}}_0)$. The covariance matrix $\boldsymbol{V}$ converges in probability to a positively definite matrix as long as the integrative scores are not in co-linearity. The statistic T^2 converges in distribution to the chi-square distribution with d degrees of freedom if group set is not associated with the disease. In addition, informative group wise weights, if available, can be readily incorporated into the Hotelling's T^2 test.

In this investigation, individual local ancestries were assumed to be known. In practice, local ancestries can be inferred from available genomic data. Several software packages, such as SABER[18], HAPAA[55], HAPMIX[56], MULTIMIX[57], CSVs[58], and ELAI[59], have been established for inferring local ancestries. These packages utilize available marker-wise genotypes of a target individual and the haplotypes/genotypes from certain ancestral panels. When dense SNPs are genotyped across the genome, the local ancestries can be highly accurately inferred. For each admixed individual, our LABST assumes that within a short block there is no ancestry crossover. This assumption is reasonable for haplotype blocks in ancestral populations[60,61]. Such haplotype blocks are of little evidence for historical recombination and much shorter than ALD regions. Gene based rare variant associations often fall in such blocks. In practice, it would be important to accommodate covariates (e.g., population structure variables, environmental factors). Let $\boldsymbol{z}_i = [1, z_{i1}, \ldots, z_{ic}]'$ contain the covariates of individual i and let $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]'$ be the whole sample covariates matrix. To adjust for the covariates, let $e_i = u_i - z_i'(Z'Z)^{-1}Z'u$ for each individual i, where $\boldsymbol{u} = [u_1, \ldots, u_n]'$. It is clear that the vector of residuals $\boldsymbol{e} = [e_1, \ldots, e_n]'$ is orthogonal to $\boldsymbol{Z}$, that is, $\boldsymbol{e}'\boldsymbol{Z} = 0$. Replacing u_i's in the statistic W_e (Eq. 1) with e_i's is one way to adjust for covariates.

Like many existing methods, our LABST assumes that individuals are randomly recruited from a target admixed population. It will be instructive to develop particular integrative methods for other sampling schemes that enrich rare variants. For example, the individuals with extreme values of a quantitative trait are often recruited for sequencing studies. Under such a trait-oriented sampling scheme, the LABST is valid but its power would be improved by combining local ancestry with a direct quantitative association analysis that incorporates the sampling scheme. In addition, individuals can be selected according to a secondary sampling trait, which is conveniently and economically measured. Only for the recruited individuals, the values of the primary study trait are measured. For such a sampling scheme, we will develop novel effective methods to combine block wise ancestries and genotypes with multiple phenotypes for identifying pleiotropic genes. Currently, the LABST only works for a recent (several-generations) admixture of two ancestral populations with different genetic architectures, i.e., distinct causal allele frequencies and/or effects. One typical example is the current African American population, which suffers from disproportionately heavier burdens of multiple diseases[1–7] than European Americans. The LABST can be extended to allow for multi-way admixtures such as Hispanic and Latino Americans. For example, it can be extended to a $(d+1)$-way admixture by using Hotelling's two-sample T-squared test with d degrees of freedom, which is similar to the above extension to combine multiple groups of variants.

References

1. Zhu, X. *et al.* Admixture mapping for hypertension loci with genome-scan markers. *Nature genetics* **37**, 177 (2005).
2. Zhu, X. *et al.* Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. *Human molecular genetics* **20**, 2285–2295 (2011).
3. Kao, W. L. *et al.* MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nature genetics* **40**, 1185 (2008).
4. Elbein, S. C., Das, S. K., Hallman, D. M., Hanis, C. L. & Hasstedt, S. J. Genome-wide linkage and admixture mapping of type 2 diabetes in African American families from the American Diabetes Association GENNID (Genetics of NIDDM) Study Cohort. *Diabetes* **58**, 268–274 (2009).
5. Basu, A. *et al.* Admixture mapping of quantitative trait loci for BMI in African Americans: evidence for loci on chromosomes 3q, 5q, and 15q. *Obesity* **17**, 1226–1231 (2009).
6. Cheng, C.-Y. *et al.* Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X. *PLoS genetics* **5**, e1000490 (2009).
7. Hohman, T. J. *et al.* Global and local ancestry in African-Americans: Implications for Alzheimer's disease risk. *Alzheimer's & Dementia* **12**, 233–243 (2016).
8. Zhu, X., Tang, H. & Risch, N. Admixture mapping and the role of population structure for localizing disease genes. *Advances in genetics* **60**, 547–569 (2008).
9. Winkler, C. A., Nelson, G. W. & Smith, M. W. Admixture mapping comes of age. *Annual review of genomics and human genetics* **11**, 65–89 (2010).
10. Seldin, M. F., Pasaniuc, B. & Price, A. L. New approaches to disease mapping in admixed populations. *Nature Reviews Genetics* **12**, 523 (2011).
11. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367 (2009).

12. Qin, H. & Zhu, X. Power Comparison of Admixture Mapping and Direct Association Analysis in Genome-Wide Association Studies. *Genetic epidemiology* **36**, 235–243 (2012).
13. Tang, H., Siegmund, D. O., Johnson, N. A., Romieu, I. & London, S. J. Joint testing of genotype and ancestry association in admixed families. *Genetic epidemiology* **34**, 783–791 (2010).
14. Lettre, G. *et al*. Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS genetics* **7**, e1001300 (2011).
15. Pasaniuc, B. *et al*. Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS genetics* **7**, e1001371 (2011).
16. Parra, E. J. *et al*. Estimating African American admixture proportions by use of population-specific alleles. *The American Journal of Human Genetics* **63**, 1839–1851 (1998).
17. Patterson, N. *et al*. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics* **74**, 979–1000 (2004).
18. Tang, H., Coram, M., Wang, P., Zhu, X. & Risch, N. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics* **79**, 1–12 (2006).
19. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187 (2011).
20. Manolio, T. A. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747 (2009).
21. Turkheimer, E. Still missing. *Research in Human Development* **8**, 227–241 (2011).
22. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109**, 1193–1198 (2012).
23. Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13**, 135 (2012).
24. Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. & Amos, C. I. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *The American Journal of Human Genetics* **82**, 100–112 (2008).
25. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics* **11**, 415 (2010).
26. Cohen, J. C. *et al*. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
27. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
28. Ahituv, N. *et al*. Medical sequencing at the extremes of human body mass. *The American Journal of Human Genetics* **80**, 779–791 (2007).
29. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **615**, 28–56 (2007).
30. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311–321 (2008).
31. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics* **5**, e1000384 (2009).
32. Price, A. L. *et al*. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics* **86**, 832–838 (2010).
33. Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Human heredity* **70**, 42–54 (2010).
34. Neale, B. M. *et al*. Testing for an unusual distribution of rare variants. *PLoS genetics* **7**, e1001322 (2011).
35. Wu, M. C. *et al*. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93 (2011).
36. Feng, T., Elston, R. C. & Zhu, X. Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genetic epidemiology* **35**, 398–409 (2011).
37. Lin, D.-Y. & Tang, Z.-Z. A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics* **89**, 354–367 (2011).
38. Wang, H. *et al*. Variants in angiopoietin-2 (ANGPT2) contribute to variation in nocturnal oxyhaemoglobin saturation level. *Hum Mol Genet*, https://doi.org/10.1093/hmg/ddw324 (2016).
39. Wang, H. *et al*. Combined linkage and association analysis identifies rare and low frequency variants for blood pressure at 1q31. *Eur J Hum Genet*, https://doi.org/10.1038/s41431-018-0277-1 (2018).
40. He, K. Y. *et al*. Leveraging linkage evidence to identify low-frequency and rare variants on 16p13 associated with blood pressure using TOPMed whole genome sequencing data. *Hum Genet*, https://doi.org/10.1007/s00439-019-01975-0 (2019).
41. He, K. Y. *et al*. Rare variants in fox-1 homolog A (RBFOX1) are associated with lower blood pressure. *PLoS Genet* **13**, e1006678, https://doi.org/10.1371/journal.pgen.1006678 (2017).
42. Halder, I. & Shriver, M. D. Measuring and using admixture to study the genetics of complex diseases. *Human genomics* **1**, 52 (2003).
43. Qin, H. *et al*. Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics* **26**, 2961–2968 (2010).
44. Wang, X. *et al*. Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* **27**, 670–677 (2010).
45. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics* **40**, 695 (2008).
46. Agresti, A. *Categorical data analysis*. Vol. 482 (John Wiley & Sons, 2003).
47. Consortium, E. P. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *nature* **447**, 799 (2007).
48. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**, 1084–1097 (2007).
49. Williams, D. R. *et al*. Prevalence and distribution of major depressive disorder in African Americans, Caribbean blacks, and non-Hispanic whites: results from the National Survey of American Life. *Archives of general psychiatry* **64**, 305–315 (2007).
50. Mayeda, E. R., Glymour, M. M., Quesenberry, C. P. & Whitmer, R. A. Inequalities in dementia incidence between six racial and ethnic groups over 14 years. *Alzheimer's & Dementia* **12**, 216–224 (2016).
51. Logue, M. W. *et al*. Targeted Sequencing of Alzheimer Disease Genes in African Americans Implicates Novel Risk Variants. *Frontiers in neuroscience* **12**, 592, https://doi.org/10.3389/fnins.2018.00592 (2018).
52. Zekavat, S. M. *et al*. Deep coverage whole genome sequences and plasma lipoprotein (a) in individuals of European and African ancestries. *Nature communications* **9**, 2606 (2018).
53. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics bulletin* **1**, 80–83 (1945).
54. Hotelling, H. In *Breakthroughs in statistics* 54–65 (Springer, 1992).
55. Sundquist, A., Fratkin, E., Do, C. B. & Batzoglou, S. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome research* **18**, 676–682 (2008).
56. Price, A. L. *et al*. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics* **5**, e1000519 (2009).

57. Churchhouse, C. & Marchini, J. Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic epidemiology* **37**, 1–12 (2013).
58. Brown, R. & Pasaniuc, B. Enhanced methods for local ancestry assignment in sequenced admixed individuals. *PLoS computational biology* **10**, e1003555 (2014).
59. Guan, Y. Detecting structure of haplotypes and local ancestry. *Genetics* **196**, 625–642 (2014).
60. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature genetics* **29**, 229 (2001).
61. Gabriel, S. B. *et al*. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).

Acknowledgements

This work was partially funded by the National Institutes of Health grants HG003054, HL113338, R01DK091369, R01MH097018, RF1AG052476, Carol Lavin Bernick Faculty Grant (632119), Tulane's Committee on Research fellowship (600890), and Tulane Innovative Programs Hub grant (632037). The funders had no role in study design, data analysis, preparation of the manuscript, or decision to publish.

Author Contributions

H.Q. and X.Z. conceived the research. H.Q. performed the analysis and wrote the first manuscript. X.Z. and J.Z. criticized, revised, and finalized the manuscript. All the authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-41845-3.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.