ARTICLE

Genome-wide Significance Thresholds for Admixture Mapping Studies

Kelsey E. Grinde,[1,*] Lisa A. Brown,[1,2] Alexander P. Reiner,[3,4] Timothy A. Thornton,[1] and Sharon R. Browning[1]

Admixture mapping studies have become more common in recent years, due in part to technological advances and growing international efforts to increase the diversity of genetic studies. However, many open questions remain about appropriate implementation of admixture mapping studies, including how best to control for multiple testing, particularly in the presence of population structure. In this study, we develop a theoretical framework to characterize the correlation of local ancestry and admixture mapping test statistics in admixed populations with contributions from any number of ancestral populations and arbitrary population structure. Based on this framework, we develop an analytical approach for obtaining genome-wide significance thresholds for admixture mapping studies. We validate our approach via analysis of simulated traits with real genotype data for 8,064 unrelated African American and 3,425 Hispanic/Latina women from the Women's Health Initiative SNP Health Association Resource (WHI SHARe). In an application to these WHI SHARe data, our approach yields genome-wide significant p value thresholds of 2.1×10^{-5} and 4.5×10^{-6} for admixture mapping studies in the African American and Hispanic/Latina cohorts, respectively. Compared to other commonly used multiple testing correction procedures, our method is fast, easy to implement (using our publicly available R package), and controls the family-wise error rate even in structured populations. Importantly, we note that the appropriate admixture mapping significance threshold depends on the number of ancestral populations, generations since admixture, and population structure of the sample; as a result, significance thresholds are not, in general, transferable across studies.

Introduction

Understanding the genetic causes of human diseases and traits has long been of interest in the scientific community. However, the large majority of the research in this area has been conducted in populations of European descent.[1–3] Admixed populations, such as African Americans and Hispanics/Latinos, are historically underrepresented in genetic studies, yet their mixed and diverse ancestry presents unique opportunities for detecting genetic variants associated with complex traits and diseases.

Due to the processes involved in the inheritance of genetic material, the genomes of admixed individuals are a mosaic of segments with different ancestral origins (Figure 1). This mosaic pattern of locus-specific ancestry, or *local ancestry*, varies considerably across individuals within an admixed population and proves useful for identifying causal genetic variants via admixture mapping. Admixture mapping studies scan the genomes of admixed individuals for associations between local ancestry and a trait of interest.[4–6] Disease prevalence and trait values often differ across ancestral groups (e.g., asthma,[7] prostate cancer,[8] blood pressure[9]), due to a combination of genetic and environmental causes. By looking for associations between a trait and local ancestry, admixture mapping seeks to identify the genetic variants that differ in frequency across these ancestral groups and drive the observed phenotypic differences. In recent years, admixture mapping has become more widely used and has proven to be a powerful approach for localizing causal genetic variants.[8,10–30]

A single genome-wide admixture mapping study will typically involve hundreds of thousands or millions of hypothesis tests, and multiple testing correction procedures are needed to control the overall type I error rate. Perhaps the best-known multiple testing correction procedure involves a Bonferroni correction on the total number of hypothesis tests. Although easy to implement, this approach is widely criticized for yielding conservative significance thresholds in the presence of correlated tests. A related approach involves a Bonferroni correction on the estimated effective number of independent tests;[13,19,31,32] however, a number of authors[33–35] have suggested that this approach does not always guarantee family-wise error rate control in genome-wide association studies. Permutation-based[22,24,36,37] and simulation-based[20,38,39] multiple testing correction procedures are often considered to be the gold standard for genetic association studies but can be very computationally intensive. Alternatives to these procedures, based on the multivariate normal distribution, have been suggested to speed up computation time.[35,40,41]

A promising alternative to the above-mentioned approaches involves an analytic multiple testing correction.[36,42,43] In particular, Siegmund and Yakir[43] derived the correlation of admixture mapping test statistics and used that theoretical result to provide an analytic approximation to the appropriate significance threshold for admixture mapping studies in admixed populations

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA; [2]Seattle Genetics, Bothell, WA 98021, USA; [3]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; [4]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
*Correspondence: grindek@uw.edu
https://doi.org/10.1016/j.ajhg.2019.01.008.
© 2019 American Society of Human Genetics.

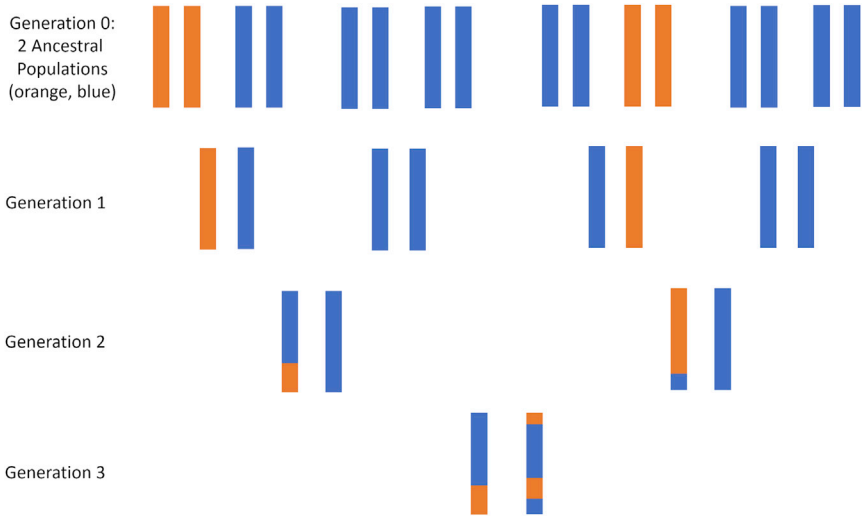

Figure 1. Inheritance of Genetic Material in an Admixed Population with Two Ancestral Populations
Colors indicate the ancestral origin of genetic material across a single chromosome, with two ancestral populations (orange, blue). In each generation, genetic material is passed from parents to offspring, with crossover events leading to chromosomes with a mixture of genetic material from the parent's maternal and paternal chromosomes. Over time, blocks of ancestry are broken up by additional crossover events, resulting in a mosaic of segments of different ancestral origins. We refer to the locus-specific ancestral origin of genetic material as *local ancestry*, while *admixture proportion* or *global ancestry* refers to the overall (genome-wide) proportion of genetic material inherited from each ancestral population.

with two ancestral populations and equal admixture proportions across individuals. However, many admixed populations have more than two ancestral populations (e.g., Hispanics/Latinos) and/or unequal admixture proportions across individuals in the population,[44–48] the latter being a consequence of population structure.[49]

In this paper, we develop a theoretical framework which applies to admixed populations with any number of ancestral populations or distribution of admixture proportions, and then use that theoretical framework to develop multiple testing correction procedures for admixture mapping studies in admixed populations with population structure. We apply our proposed procedures to genotype data for samples of African American and Hispanic/Latina ancestry from the Women's Health Initiative SNP Health Association Resource (WHI SHARe). We also perform a simulation study using these WHI SHARe genotype data and simulated traits to validate our theoretical work and evaluate the performance of our approach relative to other commonly used multiple testing correction procedures.

Material and Methods

Admixture Mapping Model

Following previous studies,[6,39,50] we use marginal regression to perform admixture mapping in samples with unrelated individuals, regressing the trait of interest on inferred local ancestry at each observed locus across the genome. At each locus, we quantify local ancestry as the number of alleles (0, 1, or 2) inherited from each ancestral population at that locus. In an admixed population with K ancestral populations, we characterize the local ancestry for admixed individual i at locus j via the vector $\boldsymbol{a}_{ij} = \begin{pmatrix} a_{ij1} & \dots & a_{ijK} \end{pmatrix}^T$, where $\sum_{k=1}^{K} a_{ijk} = 2$ and the k$^{\text{th}}$ component of this vector, a_{ijk}, denotes the number of alleles inherited by individual i from ancestral population k at locus j. Similarly, we represent the admixture proportions for each individual via the vector $\boldsymbol{\pi}_i = \begin{pmatrix} \pi_{i1} & \dots & \pi_{iK} \end{pmatrix}^T$, where $\sum_{k=1}^{K} \pi_{ik} = 1$ and the components of this vector represent the overall (genome-wide) proportion of genetic material inherited by individual i from each ancestral population. To perform admixture mapping, we regress the trait of interest, $\boldsymbol{y}$, on each component of the local ancestry vector $(k = 1, \dots, K)$ at each locus $(j = 1, \dots, m)$ using the marginal regression model

$$E\left[y_i \middle| a_{ijk}, \boldsymbol{\pi}_i\right] = \alpha + \beta_{jk} a_{ijk} + \boldsymbol{\gamma} \boldsymbol{\pi}_{i,-K}, \qquad \text{(Equation 1)}$$

where $\boldsymbol{\pi}_{i,-K} = \begin{pmatrix} \pi_{i,1} & \dots & \pi_{i,K-1} \end{pmatrix}^T$ includes the first $K-1$ components of the vector of admixture proportions. We fit separate regression models for each ancestral group in order to investigate which ancestral population(s) drive the association between the trait and local ancestry at each locus, and we adjust for estimated admixture proportions in all models to account for potential population structure.[5,50] We test for association between the trait and local ancestry using a Wald test, where the test statistic is the ratio of the estimated regression coefficient for the local ancestry term and its standard error $(Z_{jk} = \widehat{\beta}_{jk} / \widehat{se}(\widehat{\beta}_{jk}))$, with one test statistic per locus and ancestral component.

Theoretical Framework: Joint Distribution of Admixture Mapping Test Statistics

Our goal is to derive a significance threshold that controls the family-wise error rate, or the probability of making at least one type I error, for a genome-wide admixture mapping study. In other words, we wish to find the genome-wide test statistic threshold Z^* such that

$$\Pr\left(\max_{1 \le j \le m,\ 1 \le k \le K} |Z_{jk}| > Z^* \mid \beta_{jk} = 0 \ \forall j, k \right) = \alpha^*,$$

for some pre-specified level α^* (e.g., 0.05). To derive this threshold, we must understand the asymptotic joint distribution of our admixture mapping test statistics $Z_{11}, \dots, Z_{mK}$.

The first step is to characterize the correlation of local ancestry vectors at pairs of loci across the genome. For an admixed population with any number of ancestral populations, generations since admixture, or distribution of admixture proportions across the population, we can show that the correlation of local ancestry vectors $\left(\boldsymbol{a}_j, \boldsymbol{a}_{j'}\right)$ at two loci (j, j') depends on the recombination fraction between the loci (θ), the number of generations since admixture (g), and the population mean (E), variance (V), and covariance (Cov) of the admixture proportions:

We refer to this result as *Lemma 1*, and a proof is available in Appendix A. Note that if all individuals in the population have the same admixture proportions (i.e., $\boldsymbol{\pi}_i = \boldsymbol{\pi}\ \forall i$), then the local ancestry correlation is simply $(1-\theta)^g$ when $k = k'$, as had been shown previously in the context of admixed populations with two ancestral populations.[43]

$$Corr\left(a_{jk},\ a_{j'k'}\right) = \begin{cases} (1-\theta)^g + [1-(1-\theta)^g]\dfrac{2V(\pi_k)}{E(\pi_k) - E^2(\pi_k) + V(\pi_k)} & if\ \ k = k' \\ \dfrac{2Cov(\pi_k,\ \pi_{k'}) - (1-\theta)^g[Cov(\pi_k, \pi_{k'}) + E(\pi_k)E(\pi_{k'})]}{\sqrt{[E(\pi_k) - E^2(\pi_k) + V(\pi_k)]\,[E(\pi_{k'}) - E^2(\pi_{k'}) + V(\pi_{k'})]}} & if\ k \neq k'. \end{cases}$$

Using *Lemma 1*, it is straightforward to derive the asymptotic joint distribution of our collection of admixture mapping test statistics $\boldsymbol{Z} = (Z_{11} \quad \ldots \quad Z_{mK})^T$. For an admixed population with any number of ancestral populations, generations since admixture, or distribution of admixture proportions across the population, we can show that the asymptotic joint distribution of the test statistics $\boldsymbol{Z}$ can be approximated by a mean zero Gaussian process with covariance (and correlation) given by

$$Cov\left(Z_{jk},\ Z_{j'k'}\right) = \begin{cases} (1-\theta)^g\ if\ k = k' \\ -(1-\theta)^g \dfrac{E(\pi_k)E(\pi_{k'})}{\sqrt{E(\pi_k)[1-E(\pi_k)]E(\pi_{k'})[1-E(\pi_{k'})]}}\ if\ k \neq k', \end{cases}$$

where the recombination fraction (θ), generations since admixture (g), and population mean admixture proportions ($E(\pi_k)$) are defined as above. We refer to this result as *Theorem 1*, and a proof is available in Appendix A. Note that the covariance of test statistics simplifies conveniently when the admixed population has only two ancestral populations ($K = 2$):

$$Cov\left(Z_{jk},\ Z_{j'k'}\right) = \begin{cases} (1-\theta)^g \approx e^{-0.01g\delta}\ if\ k = k' \\ -(1-\theta)^g \approx -e^{-0.01g\delta}\ if\ k \neq k', \end{cases}$$

where δ is the genetic distance, in centimorgans (cM), between loci j, j'; it follows that the distribution of admixture mapping test statistics can then be approximated by an Ornstein-Uhlenbeck process.

Multiple Testing Correction Procedures

We propose two multiple testing correction procedures which use the asymptotic joint distribution of admixture mapping test statistics provided by Theorem 1 to derive a genome-wide significance threshold that will control the family-wise error rate in admixture mapping studies. Both approaches are implemented in our R package *STEAM* (Significance Threshold Estimation for Admixture Mapping).

Simulation-Based Approach

To estimate the appropriate genome-wide test statistic threshold for an admixture mapping study, we simulate test statistics from their asymptotic joint distribution (Theorem 1) and choose the threshold that controls the empirical family-wise error rate at the desired level. This approach differs from traditional simulation-based multiple testing approaches in that we simulate test statistics directly, rather than simulating traits and re-calculating test statistics at each locus for each simulation replicate. By simulating test statistics directly, computation time for our multiple testing correction procedure is considerably reduced and, importantly, is independent of sample size. To simulate admixture mapping test statistics from this distribution, we have developed a fast algorithm that requires only the genetic distances between loci, the estimated admixture proportions for individuals in the sample, and an estimate of the parameter g (see Appendix B for details).

Analytic Approximation Approach

An alternative approach for deriving genome-wide significance thresholds in the special case of admixed populations with two ancestral populations ($K = 2$) was developed previously.[43] Siegmund and Yakir[43] showed that, under some assumptions, the asymptotic joint distribution of admixture mapping test statistics can be approximated by an Ornstein-Uhlenbeck process, and then used that result to provide an analytic approximation to the family-wise error rate:[43]

$$\Pr\left(\max_{1\leq j\leq m,\ k}|Z_{jk}| > z\right) \approx 1 - \exp\left\{-2C[1-\Phi(z)] - 2\beta L z\phi(z)\nu\left(z\sqrt{2\beta\Delta}\right)\right\}, \quad \text{(Equation 2)}$$

where C is the number of chromosomes analyzed, having total genetic length L cM; Δ is the marker density; Φ and ϕ are the cumulative distribution and density functions, respectively, of the standard normal distribution; $\beta = 0.01g$; and the function ν is an infinite sum which can be approximated by $\nu(y) \approx (2/y)[\Phi(y/2) - 0.5]/[(y/2)\Phi(y/2) + \phi(y/2)]$. Although this analytic approximation was initially proposed for admixture mapping studies in populations with equal admixture proportions across individuals,[43] our work (i.e., Theorem 1) shows that it is also applicable to populations with heterogeneous admixture proportions, provided that the admixture proportions are included as covariates in the regression analysis. As a result, we can use this analytic approximation to find the admixture mapping test statistic threshold that will control the family-wise error rate at the desired level (α^*) in an admixed population with two ancestral populations and any distribution of admixture proportions: we simply find the

value z that sets the right hand side of Equation 2 equal to α^*. This involves an optimization step that can be quickly solved using existing tools (e.g., *uniroot* in R[51]). Simulation is not required for this approach, so the significance threshold can be derived in a matter of seconds.

Estimating the Number of Generations since Admixture

Both the analytic approximation and simulation-based multiple testing correction approaches rely on the number of generations since admixture. We can estimate the number of generations since admixture (g) using the observed pattern of local ancestry correlation in our sample, since g determines the rate of decay of this correlation (see Lemma 1). We propose an approach similar to that of Hellenthal et al.,[52] where we use non-linear least-squares to find the value of g that provides the best fit to the observed local ancestry correlation curves. We implement this approach in our R package *STEAM*, along with our multiple testing correction procedures.

Analysis of WHI SHARe Data

We applied our multiple testing correction procedures to two cohorts of admixed individuals with African American and Hispanic/Latina ancestry from the Women's Health Initiative SNP Health Association Resource (WHI SHARe), and also used these data to perform simulation studies comparing the performance of our proposed multiple testing correction procedures to competing approaches.

The Data

The WHI is a long-term health study of postmenopausal women in the United States. A total of 161,808 postmenopausal women aged 50–79 years old were recruited, including 12,151 self-identified African Americans (AA) and 5,469 self-identified Hispanic Americans (HA) who had consented to genetic research. Study design details and cohort characteristics are described elsewhere.[53] A subsample of these women were selected for genotyping, using the Affymetrix Genome-Wide Human SNP Array 6.0 that contains 906,000 single-nucleotide polymorphisms (SNPs) and more than 946,000 probes for the detection of copy-number variants. In these analyses, we focus only on the SNPs. The genotype data were processed for quality control, including call rate, concordance rates for blinded and unblinded duplicates, and sex discrepancy, leaving 871,309 unflagged SNPs with a genotyping rate of 99.8% and 12,008 (8,421 AA and 3,587 HA) women used in the current analysis.[14]

Local Ancestry Inference

To implement and evaluate our proposed multiple testing correction procedures in the WHI SHARe data, we first needed to infer local ancestry. First, we formed reference panels for local ancestry inference using individuals of European, African, and Native American descent from the International HapMap Project[54] (HapMap) and the Human Genome Diversity Project[55] (HGDP). In particular, the reference panels for both the AA and HA cohorts included 165 individuals of European descent (HapMap CEU, Utah residents with Northern and Western European ancestry) and 203 individuals of African descent (HapMap YRI, Yoruba in Nigeria), and the HA reference panel additionally included 63 individuals of Native American descent from HGDP. We identified a set of 551,025 and 536,374 SNPs common to the reference panels and the WHI AA and HA samples, respectively. Second, we used an iterative procedure suggested by Conomos et al.[56] to identify sets of 8,064 and 3,425 mutually unrelated AA and HA individuals, respectively. Third, we performed phasing and imputation of sporadic missing genotypes using Beagle[57] version 3. Genetic distances were estimated using the publicly available HapMap genetic map.[58] After these pre-processing steps, we performed local ancestry inference using RFMix[59] to estimate the number of alleles inherited from each ancestral population at each locus across the genome.

Application of Multiple Testing Correction Procedures

We implemented the analytic approximation approach in the AA cohort and our test statistic simulation-based approach (with 10,000 replications) in both the AA and HA cohorts. Both approaches require the number of generations since admixture, which we estimated from the observed pattern of local ancestry correlation in these samples using our non-linear least-squares approach described above. Our simulation-based approach additionally requires admixture proportions, which we estimated for each individual using the genome-wide average inferred local ancestry.

Simulation Study Using WHI SHARe Genotypes

To evaluate the performance of our proposed methods, we simulated 10,000 sets of traits for each individual according to the model $y_i \sim_{iid} N(0,1)$. We used PLINK v.1.9[60] to perform admixture mapping in each cohort, adjusting for estimated admixture proportions. We calculated the observed correlation of these tests across simulation replicates to compare to our theoretical result (Theorem 1) and evaluated the empirical family-wise error rate of our methods across the 10,000 simulation replicates. Finally, we compared our approaches to two competing methods: a Bonferroni correction on the total number of hypothesis tests and the trait simulation approach (with 10,000 replicates).

Results

Population Structure and Validation of Theoretical Results in WHI SHARe

The WHI SHARe African American (AA) and Hispanic American (HA) cohorts exhibit considerable heterogeneity in estimated admixture proportions (Figure 2), indicating that the theoretical work of previous authors[43] would not be applicable to these samples, even in the case of the AA cohort with just two ancestral populations. However, we do observe that the patterns of local ancestry and test statistic correlation in the WHI SHARe samples are consistent with our new theoretical results (Figure 3). Furthermore, a non-linear least-squares regression on the observed local ancestry curves yields estimates of the generations since admixture for each cohort ($\hat{g}_{AA} = 5.9$, $\hat{g}_{HA} = 9.6$) that are consistent with previously published studies.[49,61–64]

Comparison of Multiple Testing Correction Procedures in WHI SHARe

In the African American cohort, our multiple testing correction procedures yield genome-wide p value thresholds of 2.1×10^{-5} and 2.0×10^{-5} for the test statistic simulation and analytic approximation approaches, respectively. Both thresholds are consistent with the threshold given by the trait simulation approach (see Table 1) and are three orders of magnitude less stringent than the Bonferroni threshold. The empirical family-wise error rate for

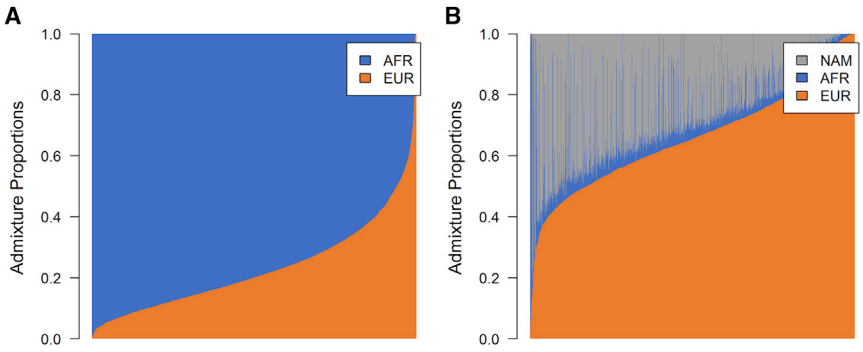

Figure 2. Estimated Admixture Proportions in WHI SHARe
(A) Estimated proportions of genetic material inherited from European (EUR) and African (AFR) ancestral populations for the African American samples.
(B) Proportions of European (EUR), African (AFR), and Native American (NAM) ancestry for the Hispanic/Latina samples.

each approach from a simulation study using simulated traits is reported in Table 2. As expected, the trait simulation approach controls the empirical family-wise error rate exactly at the nominal level 0.05. Our proposed correction procedures also control family-wise error rate at the nominal level, while the Bonferroni correction, as expected, is very conservative.

The derived significance thresholds for the Hispanic American cohort are more stringent than those in the African American cohort, reflecting the differences between the two cohorts in terms of the number of ancestral populations, number of generations since admixture, and distribution of admixture proportions. Our test statistic simulation procedure yields a p value threshold of 4.5×10^{-6}, which is again consistent with the trait simulation threshold and controls the empirical family-wise error rate at the nominal level (see Tables 1 and 2). As in the African American cohort, the Bonferroni correction yields a significance threshold that is orders of magnitude too conservative.

Computation Time

Computation time differs considerably across the four approaches. The Bonferroni correction can be used to compute the significance threshold nearly instantaneously. The analytic approximation approach is also very quick, taking under half a second on a 12-core 2.4 GHz computer with Intel Xeon E5-2630Lv2 processors and 128 GB of memory. The slowest is the trait simulation approach: for our WHI SHARe analyses, each replicate (which involved running a genome-wide admixture mapping study) took approximately 5 min on the same computer, for a total of more than 800 h of computation time to run all 10,000 replicates. In comparison, our test statistic simulation approach took only a fraction of a second per replicate, amounting to less than 10 min to run all 10,000 replicates in the African American and Hispanic American cohorts.

Discussion

We have developed a theoretical framework to characterize the correlation of local ancestry vectors and admixture mapping test statistics in admixed populations with any number of ancestral populations and distribution of admixture proportions. Our application to data from the Women's Health Initiative SNP Health Association Resource highlights the importance of this extension, as both the African American and Hispanic American samples display considerable heterogeneity in admixture proportions (Figure 2). Based on these new theoretical results, we show that an existing analytic approximation[43] can be used to derive significance thresholds for admixture mapping studies in admixed populations with two ancestral populations, even in the presence of population structure, as long as the admixture mapping model adjusts for admixture proportions. For admixed populations with any number of ancestral populations, we propose an approach that simulates test statistics directly from their asymptotic joint distribution, saving considerable computation time relative to the trait simulation approach, while still yielding an appropriate significance threshold that controls the family-wise error rate.

Our multiple testing correction procedures are based on theoretical work that explicitly models the correlation of admixture mapping test statistics, so are not conservative like the commonly used Bonferroni correction; this will translate to gains in power in genome-wide admixture mapping studies. Compared to the trait simulation approach, our correction procedures yield comparably appropriate significance thresholds but are far less computationally intensive, and we provide an R package for easy implementation. Furthermore, by simulating test statistics directly from their asymptotic distribution, the computation time of our simulation-based multiple testing does not increase with sample size, which will prove useful as future studies are able to recruit larger and larger numbers of individuals. We believe that our approaches provide an attractive alternative for researchers looking to control for multiple testing in genome-wide admixture mapping studies, particularly in admixed populations with population structure.

In this paper, our theoretical work and data analyses have focused on genome-wide admixture mapping studies with quantitative traits and unrelated individuals. However, preliminary analyses indicate that our theoretical work extends easily to binary traits (see

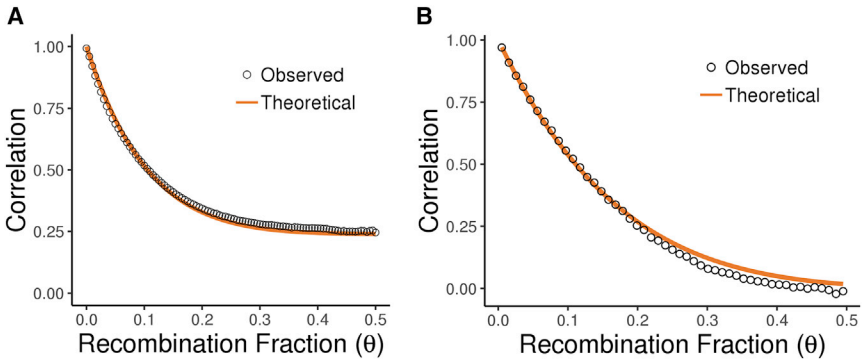

Figure 3. Correlation of Local Ancestry and Test Statistics in WHI SHARe
(A) Comparison of the observed and expected (theoretical) correlation of the European component of local ancestry vectors in the Hispanic/Latina cohort, averaging across pairs of markers falling into bins defined by their distance apart. The expected correlation comes from Lemma 1, with $g = 9.6$.
(B) Comparison of the observed and expected (theoretical) correlation of admixture mapping test statistics in the African American cohort, with expected correlation corresponding to Theorem 1, using $g = 5.9$.

Figure S1). In the case of quantitative traits that are heavily skewed (or otherwise depart considerably from normality) larger sample sizes may be needed for asymptotic normality of the test statistics to be achieved; to address this problem, transformations such as rank normalization[65,66] could be considered. The presence of relatedness, accounted for by use of a mixed model,[67] should not change the marginal distribution of admixture mapping test statistics, but would likely change their correlation structure. We expect that this will not have a large impact on the appropriate significance threshold, but further investigation is needed to confirm this hypothesis.

Our multiple testing correction procedures require estimates of the admixture proportions for each admixed individual, the number of generations since admixture, and the genetic distance between consecutive loci. To assess sensitivity to the choice of genetic map used to produce these pairwise genetic distances, we implemented *STEAM* in the WHI African American cohort using both the HapMap genetic map and an African American-specific genetic map.[68] Although these maps are quite different in some regions of the genome, we found that they still produce similar estimates of the number of generations since admixture (HapMap: 5.9, African American map: 5.7) and the genome-wide p value threshold (HapMap: 2.1×10^{-5} [95% CI: 1.9×10^{-5}, 2.2×10^{-5}]; African American map: 2.0×10^{-5} [1.8×10^{-5}, 2.3×10^{-5}]). Our estimates of the number of generations since admixture (g) may be sensitive to assortative mating or departures from the assumption of a single instantaneous admixture event. Assortative mating can lead to increased variability in admixture proportions across a population,[49,69] which our approach accounts for by allowing these proportions to vary, and may additionally change the pattern of local ancestry correlation in the sample,[69] which will impact our estimate of the number of generations since admixture. However, in application to real admixed populations (e.g., WHI SHARe) where departures from the assumption of a single instantaneous admixture event and/or random mating (e.g., due to geographic constraints) are likely, we find that our approach still works well. In estimating the parameter g from observed data using our proposed method, we are able to appropriately capture the correlation structure of admixture mapping test statistics in the sample, which is what is important for estimating an appropriate genome-wide significance threshold.

The p value threshold 5×10^{-8} has become quite widely adopted as a control for multiple testing in genome-wide association studies,[38,70–72] but there is no such "established" threshold for admixture mapping studies. Even in the specific context of the WHI SHARe genotype data, at least four different genome-wide p value thresholds have been used in published admixture mapping analyses in the African American cohort (including 7×10^{-6},[14,17] 1×10^{-5},[15] 1.5×10^{-5},[16] and 1.82×10^{-5}[22]), demonstrating the lack of consensus up to this point (even across analyses of the same dataset) on how best to derive significance thresholds for genome-wide admixture mapping studies. In practice, many admixture mapping studies cite the work of other studies (e.g., Tang et al.[36]) as the basis for their chosen significance threshold. However, our theoretical work and analysis of the African American and Hispanic WHI SHARe cohorts demonstrate that admixture mapping significance thresholds are not necessarily transferable across studies. In particular, the appropriate significance threshold depends on the number of ancestral populations, generations since admixture (to which it is particularly sensitive), population structure (through the distribution of admixture proportions), and density of markers tested, all of which often differ from one study to another. We encourage investigators to take this important point into consideration when choosing a significance threshold for their own genome-wide admixture mapping study.

Appendix A: Proofs of Theoretical Results

Lemma 1: Local Ancestry Correlation

Consider an admixed population with K ancestral populations, g generations since admixture, and admixture proportions distributed according to $\boldsymbol{\pi} \sim F$, where F has finite first and second moments. Then, the correlation of local

Table 1. Comparison of p Value Thresholds from Four Multiple Testing Correction Procedures in WHI SHARe African American (AA) and Hispanic American (HA) Samples

		Simulation (10,000 reps)		
	Bonferroni (# Tests)	Traits	Test Statistics	Analytic Approximation
AA	9.1×10^{-8}	2.1×10^{-5}	2.1×10^{-5}	2.0×10^{-5}
(95% CI)		$(1.9, 2.3) \times 10^{-5}$	$(1.9, 2.2) \times 10^{-5}$	
HA	3.1×10^{-8}	4.3×10^{-6}	4.5×10^{-6}	n/a
(95% CI)		$(4.0, 4.6) \times 10^{-6}$	$(3.9, 4.9) \times 10^{-6}$	

For simulation-based approaches, we also provide a 95% bootstrap confidence interval. Both simulation-based approaches used 10,000 replications. The nominal genome-wide type I error rate (α^*) is 0.05.

ancestry vectors at two loci j, j' separated by recombination fraction θ is given by:

$$Corr\left(a_{jk},\ a_{j'k'}\right) = \begin{cases} (1-\theta)^g + [1-(1-\theta)^g]\dfrac{2V_F(\pi_k)}{E_F(\pi_k) - E_F^2(\pi_k) + V_F(\pi_k)} & \text{if } k = k' \\ \dfrac{2Cov_F(\pi_k,\ \pi_{k'}) - (1-\theta)^g[Cov_F(\pi_k, \pi_{k'}) + E_F(\pi_k)E_F(\pi_{k'})]}{\sqrt{\left[E_F(\pi_k) - E_F^2(\pi_k) + V_F(\pi_k)\right]\left[E_F(\pi_{k'}) - E_F^2(\pi_{k'}) + V_F(\pi_{k'})\right]}} & \text{if } k \neq k'. \end{cases}$$

Proof

Consider an admixed population with K ancestral populations and g generations since admixture. Denote the admixture proportions by the vector $\boldsymbol{\pi} = (\pi_1 \ldots \pi_K)^T, \sum_{k=1}^K \pi_k = 1$, and let that vector be drawn from a distribution F with finite first and second moments. Let $\boldsymbol{a}_j = \left(a_{j1} \ldots a_{jK}\right)^T$ be the local ancestry vector, where $\sum_{k=1}^K a_{jk} = 2$ and a_{jk} denotes the number of alleles inherited from ancestral population k at locus j. Similarly, let $\boldsymbol{a}_j^P = \left(a_{j1}^P \ldots a_{jK}^P\right)^T$ be the parental local ancestry vector, where now $\sum_{k=1}^K a_{jk}^P = 1$ and a_{jk}^P denotes the number of alleles inherited from parent P ($P = M, F$ for mother and father, respectively) that are derived from ancestral population k at locus j. Consider two loci j, j' separated by recombination fraction θ or, equivalently, genetic distance δ cM. We wish to derive the correlation of local ancestry vectors $\boldsymbol{a}_j,\ \boldsymbol{a}_{j'}$ at these loci, but first we will consider the correlation of the parental local ancestry vectors $\boldsymbol{a}_j^P,\ \boldsymbol{a}_{j'}^P$.

Note that for the parental local ancestry vector $\boldsymbol{a}_j^P$, exactly one of the components takes the value 1 and the other $K-1$ components must take the value 0. Then, conditional on the admixture proportions $\boldsymbol{\pi}$, the correlation of components k, k' of the parental local ancestry vectors at loci j, j' is:

$$Corr\left(a_{jk}^P,\ a_{j'k'}^P \mid \boldsymbol{\pi}\right) = \frac{Cov\left(a_{jk}^P,\ a_{jk}^P \mid \boldsymbol{\pi}\right)}{\sqrt{V\left(a_{jk}^P \mid \boldsymbol{\pi}\right)V\left(a_{j'k'}^P \mid \boldsymbol{\pi}\right)}} = \frac{\Pr\left(a_{jk}^P = 1,\ a_{j'k'}^P = 1 \mid \boldsymbol{\pi}\right) - \pi_k\pi_{k'}}{\sqrt{\pi_k(1-\pi_k)\pi_{k'}(1-\pi_{k'})}}$$

To reduce this further, we must consider two cases: $k = k'$ and $k \neq k'$. First, suppose that $k = k'$. Then, $\Pr(a_{jk}^P = 1, a_{j'k'}^P = 1 \mid \boldsymbol{\pi}) = (1-\theta)^g \pi_k + [1-(1-\theta)^g]\pi_k^2$. Second, suppose that $k \neq k'$. Now, $\Pr(a_{jk}^P = 1,\ a_{j'k'}^P = 1 \mid \boldsymbol{\pi}) = (1-\theta)^g \times 0 + [1-(1-\theta)^g]\pi_k\pi_{k'}$. After simplifying, it follows that

$$Corr\left(a_{jk}^P, a_{j'k'}^P \mid \boldsymbol{\pi}\right) = \begin{cases} (1-\theta)^g \ \text{ if } k = k' \\ (1-\theta)^g \dfrac{-\pi_k\pi_{k'}}{\sqrt{\pi_k(1-\pi_k)\pi_{k'}(1-\pi_{k'})}} \text{ if } k \neq k'. \end{cases}$$

At each locus, we can separate the local ancestry vector $\boldsymbol{a}_j$ into the sum of the parental local ancestry vectors $\boldsymbol{a}_j^P$, such that $\boldsymbol{a}_j = \boldsymbol{a}_j^M + \boldsymbol{a}_j^F$. The parental local ancestry vectors are conditionally independent (conditional on $\boldsymbol{\pi}$). Thus, $Cov(a_{jk},\ a_{j'k'} \mid \boldsymbol{\pi}) = Cov(a_{jk}^M,\ a_{j'k'}^M \mid \boldsymbol{\pi}) + Cov(a_{jk}^F,\ a_{j'k'}^F \mid \boldsymbol{\pi}) = 2Cov(a_{jk}^P,\ a_{j'k'}^P \mid \boldsymbol{\pi})$, and the conditional correlation of components of the local ancestry vectors $\boldsymbol{a}_j$ is the same as that of the parental vectors $\boldsymbol{a}_j^P$:

$$Corr\left(a_{jk},\ a_{j'k'} \mid \boldsymbol{\pi}\right) = \frac{2Cov\left(a_{jk}^P,\ a_{jk}^P \mid \boldsymbol{\pi}\right)}{\sqrt{2V\left(a_{jk}^P \mid \boldsymbol{\pi}\right)2V\left(a_{j'k'}^P \mid \boldsymbol{\pi}\right)}} = Corr\left(a_{jk}^P,\ a_{j'k'}^P \mid \boldsymbol{\pi}\right).$$

We use the laws of total expectation, variance, and covariance to find the marginal correlation of local ancestry vectors:

Table 2. Empirical Family-wise Error Rate of Four Multiple Testing Correction Procedures in Simulation Studies using WHI SHARe African American (AA) and Hispanic American (HA) Genotype Data

		Simulation (10,000 reps)		
	Bonferroni (# Tests)	Traits	Test Statistics	Analytic Approximation
AA	5×10^{-4}	0.050	0.050	0.048
HA	8×10^{-4}	0.050	0.052	n/a

Empirical family-wise error rate was calculated across 10,000 replications of a simulated null trait. The nominal genome-wide type I error rate (α^*) is 0.05.

$$Corr\left(a_{jk},\ a_{j'k'}\right) = \frac{Cov\left(a_{jk},\ a_{j'k'}\right)}{\sqrt{V(a_{jk})V\left(a_{j'k'}\right)}} = \frac{E\left[Cov\left(a_{jk},\ a_{j'k'} \mid \boldsymbol{\pi}\right)\right] + Cov\left[E(a_{jk} \mid \boldsymbol{\pi}),\ E\left(a_{j'k'} \mid \boldsymbol{\pi}\right)\right]}{\sqrt{\{E\left[V(a_{jk} \mid \boldsymbol{\pi})\right] + V\left[E(a_{jk} \mid \boldsymbol{\pi})\right]\}\left\{E\left[V\left(a_{j'k'} \mid \boldsymbol{\pi}\right)\right] + V\left[E\left(a_{j'k'} \mid \boldsymbol{\pi}\right)\right]\right\}}}$$
$$= \frac{E\left[2Cov\left(a^P_{jk},\ a^P_{j'k'} \mid \boldsymbol{\pi}\right)\right] + Cov[2\pi_k, 2\pi_{k'}]}{\sqrt{\{E[2\pi_k(1-\pi_k)] + V[2\pi_k]\}\{E[2\pi_{k'}(1-2\pi_{k'})] + V[2\pi_{k'}]\}}}$$

With a bit of algebra (excluded from this proof for the sake of brevity), it follows that the marginal correlation of components k, k' of the local ancestry vectors at loci j, j' is

Proof

Consider an admixed population with K ancestral populations and g generations since admixture. Denote

$$Corr\left(a_{jk},\ a_{j'k'}\right) = \begin{cases} (1-\theta)^g + [1-(1-\theta)^g]\dfrac{2V_F(\pi_k)}{E_F(\pi_k) - E_F^2(\pi_k) + V_F(\pi_k)} & if\ \ k = k' \\ \dfrac{2Cov_F(\pi_k,\ \pi_{k'}) - (1-\theta)^g[Cov_F(\pi_k, \pi_{k'}) + E_F(\pi_k)E_F(\pi_{k'})]}{\sqrt{\left[E_F(\pi_k) - E_F^2(\pi_k) + V_F(\pi_k)\right]\left[E_F(\pi_{k'}) - E_F^2(\pi_{k'}) + V_F(\pi_{k'})\right]}} & if\ \ k \neq k', \end{cases}$$

as desired.

Theorem 1: Test Statistic Correlation

Consider an admixed population with K ancestral populations, g generations since admixture, and admixture proportions distributed according to $\pi \sim F$, where F has finite first and second moments. For loci $j \in \{1, \ldots, m\}$ and ancestry components $k \in \{1, \ldots, K\}$, define test statistics $Z_{jk} = \widehat{\beta}_{jk}/\widehat{se}(\widehat{\beta}_{jk})$ based on the regression model in Equation 1. Then, under the universal null hypothesis ($\beta_{jk} = 0\ \forall j,k$), the collection of test statistics $\boldsymbol{Z} = (Z_{11} \ldots Z_{mK})^T$ has an asymptotic multivariate normal distribution with mean $\mathbf{0}$ and covariance (and correlation) given by

the admixture proportions by the vector $\boldsymbol{\pi} = (\pi_1 \ldots \pi_K)^T$, $\sum_{k=1}^K \pi_k = 1$, and let that vector be drawn from a distribution F with finite first and second moments. Let $\boldsymbol{a}_j = (a_{j1} \ldots a_{jK})^T$ be the local ancestry vector, where $\sum_{k=1}^K a_{jk} = 2$ and a_{jk} denotes the number of alleles inherited from ancestral population k at locus j. Define Wald test statistics $Z_{jk} = \widehat{\beta}_{jk}/\widehat{se}(\widehat{\beta}_{jk})$ based on the marginal linear regression model $E[\boldsymbol{y}|a_{jk}, \boldsymbol{\pi}_{-K}] = \alpha + \beta_{jk}a_{jk} + \boldsymbol{\gamma}\boldsymbol{\pi}_{-K}$ (Model 1). Suppose that the universal null hypothesis holds, such that $\beta_{jk} = 0\ \forall j \in \{1, \ldots, m\}, k \in \{1, .., K\}$. We must show that the collection of test statistics $\boldsymbol{Z}$ is asymptotically multivariate normal with mean $\mathbf{0}$ and covariance as defined above.

$$Cov\left(Z_{jk},\ Z_{j'k'}\right) = \begin{cases} (1-\theta)^g \approx e^{-0.01g\delta}\ if\ \ k = k' \\ -(1-\theta)^g h(E_F(\pi_k),\ E_F(\pi_{k'})) \approx -e^{-0.01g\delta} h(E_F(\pi_k),\ E_F(\pi_{k'}))\ if\ \ k \neq k', \end{cases}$$

where $h(x,y) = xy/\sqrt{x(1-x)y(1-y)}$, θ is the recombination fraction between loci j, j', and δ is the genetic distance (cM) between those loci.

It is straightforward to show (e.g., by using the asymptotic equivalence between Wald tests and score tests, combined with existing results about the asymptotic

distribution of score tests from such a model[35,40]) that the test statistics $\boldsymbol{Z}$ are asymptotically multivariate normal with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}$ with elements $\boldsymbol{\Sigma}_{jk,j'k'} = Cov\left(Z_{jk},\ Z_{j'k'}\right)$. Furthermore, we can show (e.g., as in Joo et al.[73]) that test statistics $\boldsymbol{Z}$ from the unadjusted admixture mapping model (Model 1 without $\boldsymbol{\pi}_{-K}$) have covariance $Cov(Z_{jk}^{UNADJ},\ Z_{j'k'}^{UNADJ}) = Corr(a_{jk},\ a_{j'k'})$. From the adjusted model (Model 1 *with* $\boldsymbol{\pi}_{-K}$), the covariance of test statistics is simply the correlation of local ancestry conditioned on the covariates $\boldsymbol{\pi}_{-K}$:[35] $Cov\left(Z_{jk},\ Z_{j'k'}\right) = Corr\left(a_{jk}, a_{j'k'}|\boldsymbol{\pi}\right)$. Combining these results with *Lemma 1* and the approximation $(1-\theta)^g \approx \exp(-0.01g\delta)$ from Siegmund and Yakir,[43] we see that asymptotically the test statistics $\boldsymbol{Z}$ will have covariance defined by $Cov(Z_{jk},\ Z_{j'k'}) = Corr(a_{jk},\ a_{j'k'}|\boldsymbol{\pi} = E[\boldsymbol{\pi}])$, where

$$Corr\left(a_{jk}, a_{j'k'} \mid \boldsymbol{\pi} = E[\boldsymbol{\pi}]\right) = (1-\theta)^g \approx e^{-0.01g\delta} \text{ if } k = k',$$

$$\begin{aligned} &Corr\left(a_{jk}, a_{j'k'} \mid \boldsymbol{\pi} = E[\boldsymbol{\pi}]\right) \\ &= (1-\theta)^g \frac{-E[\pi_k]E[\pi_{k'}]}{\sqrt{E[\pi_k](1-E[\pi_k])E[\pi_{k'}](1-E[\pi_{k'}])}} \approx e^{-0.01g\delta} \\ &\quad \times \frac{-E[\pi_k]E[\pi_{k'}]}{\sqrt{E[\pi_k](1-E[\pi_k])E[\pi_{k'}](1-E[\pi_{k'}])}} \text{ if } k \neq k', \end{aligned}$$

as desired.

Corollary

When $K = 2$, $h(E_F(\pi_k),\ E_F(\pi_{k'})) = 1$ since $\sum_{k=1}^{K} E_F(\pi_k) = 1$, so the covariance of test statistics simplifies nicely:

$$Cov\left(Z_{jk},\ Z_{j'k'}\right) = \begin{cases} (1-\theta)^g \approx e^{-0.01g\delta} \text{ if } k = k' \\ -(1-\theta)^g \approx -e^{-0.01g\delta} \text{ if } k \neq k', \end{cases}$$

and the distribution of test statistics can then be approximated by an Ornstein-Uhlenbeck process, as shown by Siegmund and Yakir.[43]

Appendix B: Recursive Simulation Algorithm

We propose a simulation-based multiple testing correction approach that simulates admixture mapping test statistics $\boldsymbol{Z} = (Z_{11}\ Z_{12} \ldots Z_{mK})^T$ from their asymptotic joint distribution (Theorem 1). To do so, we use the following recursive algorithm:

1. Set $\boldsymbol{Z}_1 = (Z_{11} \ldots Z_{1K})^T = \boldsymbol{L}\boldsymbol{X}_1$, *where* $\boldsymbol{X}_1 \sim N_{K-1}(0,\ \boldsymbol{I}_{K-1})$.
2. Set $\boldsymbol{Z}_2 = a_{12}\boldsymbol{Z}_1 + b_{12}\boldsymbol{L}\boldsymbol{X}_2$, *where* $\boldsymbol{X}_2 \sim N_{K-1}(0,\ \boldsymbol{I}_{K-1})$.
3. Set $\boldsymbol{Z}_3 = a_{23}\boldsymbol{Z}_2 + b_{23}\boldsymbol{L}\boldsymbol{X}_3$, where $\boldsymbol{X}_3 \sim N_{K-1}(0,\ \boldsymbol{I}_{K-1})$.

…

m. Set $\boldsymbol{Z}_m = a_{m-1,m}\boldsymbol{Z}_{m-1} + b_{m-1,m}\boldsymbol{L}\boldsymbol{X}_m$, where $\boldsymbol{X}_m \sim N_{K-1}(0,\ \boldsymbol{I}_{K-1})$.

Here, K is the number of ancestral populations, m is the number of loci, $\boldsymbol{L}$ is a $K \times (K-1)$ matrix which depends on the first and second moments (which we can estimate using their sample equivalents) of the distribution of admixture proportions in the population, and a_{ij}, b_{ij} are scalars which depend on the number of generations since admixture (g) and the genetic distance (in cM) between consecutive loci i and j.

Others[35,73] have proposed multiple-testing correction procedures that similarly utilize knowledge of the asymptotic distribution of test statistics; however, our approach takes advantage of the specific, convenient form of this distribution for admixture mapping studies to speed up computation time. Note that this algorithm scales linearly with the number of loci m, but run time does not depend on the number of samples n (except through calculation of the first and second moments of the sample admixture proportions). Computation time can be drastically reduced by considering just a single locus per unique ancestry block (if applicable; not all local ancestry inference programs perform calling within windows). Run time does increase slightly, but not drastically, with increasing number of ancestral populations K. Running 10,000 replicates on the WHI SHARe data took approximately 8 and 9 min for the African American ($K = 2$) and Hispanic American ($K = 3$) samples, respectively. We have implemented this algorithm in our R package *STEAM* (Significance Threshold Estimation for Admixture Mapping), which is available on GitHub.

Supplemental Data

Supplemental Data include one figure and can be found with this article online at https://doi.org/10.1016/j.ajhg.2019.01.008.

Acknowledgments

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. Funding for WHI SNP Health Association Resource (WHI-SHARe) genotyping was provided by NHLBI contract N02-HL-64278. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A listing of WHI investigators can be found at http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf. K.E.G. was supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1256082. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Declaration of Interests

The authors declare no competing interests.

Received: November 5, 2018
Accepted: January 17, 2019
Published: February 14, 2019

Web Resources

STEAM, https://github.com/kegrinde/STEAM

References

1. Need, A.C., and Goldstein, D.B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends Genet. *25*, 489–494.
2. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163–165.
3. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
4. Rife, D.C. (1954). Populations of hybrid origin as source material for the detection of linkage. Am. J. Hum. Genet. *6*, 26–33.
5. McKeigue, P.M. (1998). Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am. J. Hum. Genet. *63*, 241–251.
6. Shriner, D. (2013). Overview of admixture mapping. Curr Protoc Hum Genet *76*, 1.23.1–1.23.8.
7. Forno, E., and Celedon, J.C. (2009). Asthma and ethnic minorities: socioeconomic status and beyond. Curr. Opin. Allergy Clin. Immunol. *9*, 154–160.
8. Freedman, M.L., Haiman, C.A., Patterson, N., McDonald, G.J., Tandon, A., Waliszewska, A., Penney, K., Steen, R.G., Ardlie, K., John, E.M., et al. (2006). Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc. Natl. Acad. Sci. USA *103*, 14068–14073.
9. Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo, J.L., Jr., Jones, D.W., Materson, B.J., Oparil, S., Wright, J.T., Jr., Roccella, E.J.; Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. National Heart, Lung, and Blood Institute; and National High Blood Pressure Education Program Coordinating Committee (2003). Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension *42*, 1206–1252.
10. Zhu, X., Luke, A., Cooper, R.S., Quertermous, T., Hanis, C., Mosley, T., Gu, C.C., Tang, H., Rao, D.C., Risch, N., and Weder, A. (2005). Admixture mapping for hypertension loci with genome-scan markers. Nat. Genet. *37*, 177–181.
11. Reich, D., Patterson, N., Ramesh, V., De Jager, P.L., McDonald, G.J., Tandon, A., Choy, E., Hu, D., Tamraz, B., Pawlikowska, L., et al.; Health, Aging and Body Composition (Health ABC) Study (2007). Admixture mapping of an allele affecting interleukin 6 soluble receptor and interleukin 6 levels. Am. J. Hum. Genet. *80*, 716–726.
12. Winkler, C.A., Nelson, G.W., and Smith, M.W. (2010). Admixture mapping comes of age. Annu. Rev. Genomics Hum. Genet. *11*, 65–89.
13. Zhu, X., Young, J.H., Fox, E., Keating, B.J., Franceschini, N., Kang, S., Tayo, B., Adeyemo, A., Sun, Y.V., Li, Y., et al. (2011). Combined admixture mapping and association analysis identifies a novel blood pressure genetic locus on 5p13: contributions from the CARe consortium. Hum. Mol. Genet. *20*, 2285–2295.
14. Reiner, A.P., Beleza, S., Franceschini, N., Auer, P.L., Robinson, J.G., Kooperberg, C., Peters, U., and Tang, H. (2012). Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women. Am. J. Hum. Genet. *91*, 502–512.
15. Carty, C.L., Johnson, N.A., Hutter, C.M., Reiner, A.P., Peters, U., Tang, H., and Kooperberg, C. (2012). Genome-wide association study of body height in African Americans: the Women's Health Initiative SNP Health Association Resource (SHARe). Hum. Mol. Genet. *21*, 711–720.
16. Ochs-Balcom, H.M., Preus, L., Wactawski-Wende, J., Nie, J., Johnson, N.A., Zakharia, F., Tang, H., Carlson, C., Carty, C., Chen, Z., et al. (2013). Association of DXA-derived bone mineral density and fat mass with African ancestry. J. Clin. Endocrinol. Metab. *98*, E713–E717.
17. Coram, M.A., Duan, Q., Hoffmann, T.J., Thornton, T., Knowles, J.W., Johnson, N.A., Ochs-Balcom, H.M., Donlon, T.A., Martin, L.W., Eaton, C.B., et al. (2013). Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations. Am. J. Hum. Genet. *92*, 904–916.
18. Galanter, J.M., Gignoux, C.R., Torgerson, D.G., Roth, L.A., Eng, C., Oh, S.S., Nguyen, E.A., Drake, K.A., Huntsman, S., Hu, D., et al. (2014). Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. J. Allergy Clin. Immunol. *134*, 295–305.
19. Gomez, F., Wang, L., Abel, H., Zhang, Q., Province, M.A., and Borecki, I.B. (2015). Admixture mapping of coronary artery calcification in African Americans from the NHLBI family heart study. BMC Genet. *16*, 42.
20. Schick, U.M., Jain, D., Hodonsky, C.J., Morrison, J.V., Davis, J.P., Brown, L., Sofer, T., Conomos, M.P., Schurmann, C., McHugh, C.P., et al. (2016). Genome-wide association study of platelet count identifies ancestry-specific loci in Hispanic/Latino Americans. Am. J. Hum. Genet. *98*, 229–242.
21. Brown, L.A., Sofer, T., Stilp, A.M., Baier, L.J., Kramer, H.J., Masindova, I., Levy, D., Hanson, R.L., Moncrieft, A.E., Redline, S., et al. (2017). Admixture mapping identifies an Amerindian ancestry locus associated with albuminuria in Hispanics in the United States. J. Am. Soc. Nephrol. *28*, 2211–2220.
22. Giri, A., Hartmann, K.E., Aldrich, M.C., Ward, R.M., Wu, J.M., Park, A.J., Graff, M., Qi, L., Nassir, R., Wallace, R.B., et al. (2017). Admixture mapping of pelvic organ prolapse in African Americans from the Women's Health Initiative Hormone Therapy trial. PLoS ONE *12*, e0178839.
23. Sofer, T., Baier, L.J., Browning, S.R., Thornton, T.A., Talavera, G.A., Wassertheil-Smoller, S., Daviglus, M.L., Hanson, R., Kobes, S., Cooper, R.S., et al. (2017). Admixture mapping in the Hispanic Community Health Study/Study of Latinos reveals regions of genetic associations with blood pressure traits. PLoS ONE *12*, e0188400.
24. Giri, A., Edwards, T.L., Hartmann, K.E., Torstenson, E.S., Wellons, M., Schreiner, P.J., and Velez Edwards, D.R. (2017). African genetic ancestry interacts with body mass index to modify risk for uterine fibroids. PLoS Genet. *13*, e1006871.

25. Cyr, D.D., Allen, A.S., Du, G.J., Ruffin, F., Adams, C., Thaden, J.T., Maskarinec, S.A., Souli, M., Guo, S., Dykxhoorn, D.M., et al. (2017). Evaluating genetic susceptibility to *Staphylococcus aureus* bacteremia in African Americans using admixture mapping. Genes Immun. *18*, 95–99.
26. Parra, E.J., Mazurek, A., Gignoux, C.R., Sockell, A., Agostino, M., Morris, A.P., Petty, L.E., Hanis, C.L., Cox, N.J., Valladares-Salgado, A., et al. (2017). Admixture mapping in two Mexican samples identifies significant associations of locus ancestry with triglyceride levels in the BUD13/ZNF259/APOA5 region and fine mapping points to rs964184 as the main driver of the association signal. PLoS ONE *12*, e0172880.
27. Uribe-Salazar, J.M., Palmer, J.R., Haddad, S.A., Rosenberg, L., and Ruiz-Narváez, E.A. (2018). Admixture mapping and fine-mapping of type 2 diabetes susceptibility loci in African American women. J. Hum. Genet. *63*, 1109–1117.
28. Wang, H., Cade, B.E., Sofer, T., Sands, S.A., Chen, H., Browning, S., Stilp, A.M., Louie, T.L., Thornton, T.A., Craig Johnson, W., et al. (2018). Admixture mapping identifies novel loci for obstructive sleep apnea in Hispanic/Latino Americans. Hum. Mol. Genet. Published online November 7, 2018. https://doi.org/10.1093/hmg/ddy387.
29. Gignoux, C.R., Torgerson, D.G., Pino-Yanes, M., Uricchio, L.H., Galanter, J., Roth, L.A., Eng, C., Hu, D., Nguyen, E.A., Huntsman, S., et al. (2018). An admixture mapping meta-analysis implicates genetic variation at 18q21 with asthma susceptibility in Latinos. J. Allergy Clin. Immunol., S0091-6749(18)31274-0.
30. Spear, M.L., Hu, D., Pino-Yanes, M., Huntsman, S., Eng, C., Levin, A.M., Ortega, V.E., White, M.J., McGarry, M.E., Thakur, N., et al. (2018). A genome-wide association and admixture mapping study of bronchodilator drug response in African Americans with asthma. Pharmacogenomics J. Published online September 12, 2018. https://doi.org/10.1038/s41397-018-0042-4.
31. Li, J., and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb) *95*, 221–227.
32. Shriner, D., Adeyemo, A., and Rotimi, C.N. (2011). Joint ancestry and association testing in admixed individuals. PLoS Comput. Biol. *7*, e1002325.
33. Dudbridge, F., and Koeleman, B.P. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am. J. Hum. Genet. *75*, 424–435.
34. Salyakina, D., Seaman, S.R., Browning, B.L., Dudbridge, F., and Muller-Myhsok, B. (2005). Evaluation of Nyholt's procedure for multiple testing correction. Hum. Hered. *60*, 19–25, discussion 61–62.
35. Conneely, K.N., and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. Am. J. Hum. Genet. *81*, 1158–1168.
36. Tang, H., Siegmund, D.O., Johnson, N.A., Romieu, I., and London, S.J. (2010). Joint testing of genotype and ancestry association in admixed families. Genet. Epidemiol. *34*, 783–791.
37. Qin, H., and Zhu, X. (2012). Power comparison of admixture mapping and direct association analysis in genome-wide association studies. Genet. Epidemiol. *36*, 235–243.
38. Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M.J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet. Epidemiol. *32*, 381–385.
39. Brown, L. (2016). Statistical Methods in Admixture Mapping: Mixed Model Based Testing and Genome-wide Significance Thresholds. PhD Thesis (University of Washington).
40. Seaman, S.R., and Müller-Myhsok, B. (2005). Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. Am. J. Hum. Genet. *76*, 399–408.
41. Han, B., Kang, H.M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genet. *5*, e1000456.
42. Sha, Q., Zhang, X., Zhu, X., and Zhang, S. (2006). Analytical correction for multiple testing in admixture mapping. Hum. Hered. *62*, 55–63.
43. Siegmund, D., and Yakir, B. (2007). The Statistics of Gene Mapping (New York: Springer).
44. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.M., Wambebe, C., Tishkoff, S.A., and Bustamante, C.D. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc. Natl. Acad. Sci. USA *107*, 786–791.
45. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044.
46. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos. Am. J. Hum. Genet. *98*, 165–184.
47. Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E., and Shriver, M.D. (1998). Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. *63*, 1839–1851.
48. Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., and Ostrer, H. (2010). Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc. Natl. Acad. Sci. USA *107* (*Suppl 2*), 8954–8961.
49. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. *96*, 37–53.
50. Redden, D.T., Divers, J., Vaughan, L.K., Tiwari, H.K., Beasley, T.M., Fernández, J.R., Kimberly, R.P., Feng, R., Padilla, M.A., Liu, N., et al. (2006). Regional admixture mapping and structured association testing: conceptual unification and an extensible general linear model. PLoS Genet. *2*, e137.
51. R Core Team (2018). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
52. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. Science *343*, 747–751.
53. Hays, J., Hunt, J.R., Hubbell, F.A., Anderson, G.L., Limacher, M., Allen, C., and Rossouw, J.E. (2003). The Women's Health Initiative recruitment methods and results. Ann. Epidemiol. *13* (9, Suppl), S18–S77.
54. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu,

F., Peltonen, L., et al.; International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
55. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.
56. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. *98*, 127–148.
57. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.
58. International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.
59. Maples, B.K., Gravel, S., Kenny, E.E., and Bustamante, C.D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am. J. Hum. Genet. *93*, 278–288.
60. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
61. Smith, M.W., Patterson, N., Lautenberger, J.A., Truelove, A.L., McDonald, G.J., Waliszewska, A., Kessing, B.D., Malasky, M.J., Scafe, C., Le, E., et al. (2004). A high-density admixture map for disease gene discovery in african americans. Am. J. Hum. Genet. *74*, 1001–1013.
62. Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G., and McKeigue, P.M. (2004). Design and analysis of admixture mapping studies. Am. J. Hum. Genet. *74*, 965–978.
63. Price, A.L., Patterson, N., Yu, F., Cox, D.R., Waliszewska, A., McDonald, G.J., Tandon, A., Schirmer, C., Neubauer, J., Bedoya, G., et al. (2007). A genomewide admixture map for Latino populations. Am. J. Hum. Genet. *80*, 1024–1036.
64. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.
65. Beasley, T.M., Erickson, S., and Allison, D.B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? Behav. Genet. *39*, 580–595.
66. Sofer, T., Zheng, X., Gogarten, S.M., Laurie, C.A., Grinde, K., Shaffer, J.R., Shungin, D., O'Connell, J.R., Durazo-Arvizo, R.A., Raffield, L., et al. (2019). A fully-adjusted two-stage procedure for rank normalization in genetic association studies. Genet. Epidemiol., 1–13.
67. Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. *38*, 203–208.
68. Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170–175.
69. Zaitlen, N., Huntsman, S., Hu, D., Spear, M., Eng, C., Oh, S.S., White, M.J., Mak, A., Davis, A., Meade, K., et al. (2017). The effects of migration and assortative mating on admixture linkage disequilibrium. Genetics *205*, 375–383.
70. Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.
71. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.
72. Jannot, A.S., Ehret, G., and Perneger, T. (2015). $P < 5 \times 10(-8)$ has emerged as a standard of statistical significance for genome-wide association studies. J. Clin. Epidemiol. *68*, 460–465.
73. Joo, J.W., Hormozdiari, F., Han, B., and Eskin, E. (2016). Multiple testing correction in linear mixed models. Genome Biol. *17*, 62.

The American Journal of Human Genetics, Volume *104*

Supplemental Data

Genome-wide Significance Thresholds for Admixture Mapping Studies

Kelsey E. Grinde, Lisa A. Brown, Alexander P. Reiner, Timothy A. Thornton, and Sharon R. Browning

Supplemental Figures

Figure S1. Correlation of admixture mapping test statistics with binary versus quantitative traits. To confirm the validity of our theoretical work for binary traits, we simulated traits and local ancestry at pairs of loci for admixed individuals in a variety of populations. For each admixed individual $i = 1, \dots, n$, we first drew global ancestry proportions from a pre-specified distribution F representing different population structure scenarios: no structure ($\boldsymbol{\pi_i} = \boldsymbol{\pi}\ \forall i$), subpopulations ($\boldsymbol{\pi_i} = \{\boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \dots, \boldsymbol{\pi_P}\}$ with probability $\{p_1, p_2, \dots, p_P\}$, where $\sum_{l=1}^{P} p_l = 1$), or Dirichlet global ancestry ($\boldsymbol{\pi_i} \sim_{iid} Dirichlet(\boldsymbol{\alpha})$). We considered various choices of number of individuals (n), number of ancestral populations (K), and hyperparameters for the distribution of global ancestry proportions F. For each individual's two haplotypes, we independently simulated crossover events between two loci separated by recombination fraction θ across g generations according to a Poisson process. We simulated ancestry at the first locus according to a Multinoulli (categorical) distribution with probabilities equal to the global ancestry vector $\boldsymbol{\pi_i}$. Using the simulated crossover history, we determined whether any recombination had occurred between the two loci; if so, we independently simulated ancestry at the second locus according to the same Multinoulli distribution; if not, we set ancestry at the second locus equal to ancestry at the first. We simulated binary traits for each individual according to the model $y_i \sim_{iid} Bernoulli(0.2)$, and quantitative traits according to $y_i \sim_{iid} N(0,1)$. Finally, we paired the individuals' haplotypes, recorded the local ancestry vectors for each individual, and calculated admixture mapping test statistics at each locus using the simulated traits. We repeated this process 10,000 times, calculated the correlation of admixture mapping test statistics at the two loci across simulation replicates, then compared the observed patterns of correlation to the expected correlation given by our theoretical results (Theorem 1). Panels (A) and (B) present the observed versus theoretical correlation of admixture mapping test statistics (testing the first

ancestry component at two loci separated by recombination fraction θ) in an admixed population with $K = 3$, $n = 10{,}000$, $\boldsymbol{\pi_i} \sim_{iid} Dirichlet([1,1,1])$, and either binary (Panel A) or quantitative (Panel B) traits.