PNAS PLUS

Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure

David Mimno[a], David M. Blei[b], and Barbara E. Engelhardt[c,1]

[a]Department of Information Science, Cornell University, Ithaca, NY 14853; [b]Department of Statistics, Department of Computer Science, Columbia University, New York, NY 10027; and [c]Department of Computer Science, Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08540

Edited by Simon Tavare, University of Southern California, Los Angeles, CA, and accepted by the Editorial Board April 23, 2015 (received for review July 3, 2014)

Admixture models are a ubiquitous approach to capture latent population structure in genetic samples. Despite the widespread application of admixture models, little thought has been devoted to the quality of the model fit or the accuracy of the estimates of parameters of interest for a particular study. Here we develop methods for validating admixture models based on posterior predictive checks (PPCs), a Bayesian method for assessing the quality of fit of a statistical model to a specific dataset. We develop PPCs for five population-level statistics of interest: within-population genetic variation, background linkage disequilibrium, number of ancestral populations, between-population genetic variation, and the downstream use of admixture parameters to correct for population structure in association studies. Using PPCs, we evaluate the quality of the admixture model fit to four qualitatively different population genetic datasets: the population reference sample (POPRES) European individuals, the HapMap phase 3 individuals, continental Indians, and African American individuals. We found that the same model fitted to different genomic studies resulted in highly study-specific results when evaluated using PPCs, illustrating the utility of PPCs for model-based analyses in large genomic studies.

posterior predictive checks | admixture models | population structure | model checking | genomic data

One essential problem for population genetics is to characterize latent population structure in genetic samples. Inferred population structure is used to control for confounding effects in both genome-wide association studies (GWAS) and quantitative trait mapping (1, 2), and to explore genetic relationships when studying population ancestry and history (3–6).

An important and influential approach to characterizing latent population structure is the admixture model, first implemented in the Structure program (7). The admixture model is a Bayesian model of a collection of genomes. It represents each genome as a convex combination of K ancestral populations and each ancestral population as population-specific allele frequencies for each genetic locus. Given genetic data, admixture models recover both the genetic variation within ancestral populations and proportions of each ancestral population within a genome. Because of their descriptive power, admixture models have become essential for exploratory analyses of genomic studies (8); they have transformed modern research in population genetics.

The admixture model, like all statistical models, makes simplifying assumptions about the data. These assumptions enable complex data to be modeled in a way that is both analytically useful and computationally tractable. For example, the admixture model assumes that individuals in the sample are distantly related, that all locus-specific allele frequencies are equally likely, and that genetic loci are independent. Population genetics tells us, however, that genomic data violate these assumptions (7). Paraphrasing the quip by statistician George Box, our question is not whether the model is true—we know that it is not—but whether it is useful. Do the fitted parameters help with the analytic task, or do the simplifying assumptions direct us to unsupported conclusions?

Here, we show that the effect of the admixture assumptions on inference of population structure depends on the data at hand; thus diagnosing admixture model misspecification should be regular practice in their application. A misspecified model may indicate spurious associations between diseases and genetic variants after correcting for latent structure, or lead to blind alleys of ancestral history while exploring inferred structure that is an artifact of the simplifying assumptions. When we fit the admixture model to genetic data—whether we are exploring latent structure or using inferred population structure in downstream analyses—we rely on the recovered representation from a statistical procedure to be meaningfully connected to the true genetic structure that has emerged from a complex evolutionary process. It is essential that we assess the strength of this connection.

To this end, we develop a general statistical procedure for checking the goodness-of-fit of an admixture model to genomic data. Our procedures are based on posterior predictive checks (PPCs), a technique from Bayesian statistics used to quantify the effect of Bayesian model misspecification (9–13). A PPC works as follows. We first fit a model using observed data, estimating the posterior distribution of latent parameters. The fitted model induces a distribution of future data conditioned on the observations; this distribution is called the posterior predictive distribution (PPD). We use the PPD to generate synthetic datasets. Finally, we check whether the simulated datasets are close to the observed dataset when summarized through a statistic of interest, called the discrepancy function.

Significance

Bayesian models, including admixture models, are a powerful framework for articulating complex assumptions about large-scale genetic data; such models are widely used to explore data or to study population-level statistics of interest. However, we assume that a Bayesian model does not oversimplify the complexities in the data, to the point of invalidating our analyses. Here, we develop and study procedures for quantitatively evaluating admixture models of genetic data. Using four large genetic studies, we demonstrate that model checking should be an important part of the modern genetic data analysis pipeline. Our methods help to support inferences drawn from recovered population structure, to protect scientists from being misled by a misspecified model class, and to point scientists toward useful model extensions.

Author contributions: D.M., D.M.B., and B.E.E. designed research; D.M. and B.E.E. performed research; D.M. and B.E.E. analyzed data; and D.M., D.M.B., and B.E.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. S.T. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

Software to to run the admixture model and the PPCs we developed on large genomic datasets is available at github.com/mimno/admixture-ppc.

[1]To whom correspondence should be addressed. Email: bee@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1412301112/-/DCSupplemental.

The idea behind PPCs is that, if the model assumptions are appropriate, then data generated from the PPD will look like the observed data. The discrepancy measures a relevant property of the data that we hope to capture. If the model is well specified for a specific dataset, then the observed data, viewed through the discrepancy, will be a likely draw from the estimated PPD. If the model is not well specified, then the observed discrepancy will look like an outlier.

Given observed genotype data, checking admixture models with a PPC works as follows (Fig. 1). We first fit an admixture model to the data, estimating ancestral populations and individual-specific population proportions. Most analyses end here, e.g., with illustrations of the population proportions as in Fig. 2. We then simulate genomes from the posterior predictive distribution, using posterior estimates of the latent parameters to draw synthetic genetic data that share the same structure as the observed data; we repeat this process many times to create a collection of replicated data. Finally, we compare discrepancies computed on the observed data to the empirical distribution of the discrepancies computed on the replicated data. When an observed discrepancy is not likely relative to the empirical distribution of the replicated discrepancies, the PPC suggests that the model is misspecified with respect to the discrepancy. We maintain that PPC assessments are best made visually (12, 13).

We used PPCs to check for misspecification in four genomic datasets: HapMap phase 3, Europeans, African Americans, and continental Indians. These data have been previously characterized using an admixture model and have qualitatively different types of latent population structure (Fig. 2). We developed five discrepancy functions to check for important types of model misspecification in admixture model analyses. The discrepancy may be a function of both observed and latent variables (11, 12). We based these discrepancies on common measures in population genetics:

- Identity by state (IBS): We test for the impact of long-range single nucleotide polymorphism (SNP) correlations on within-population variance estimates by quantifying the genomic variation among pairs of individuals within alleles from the same ancestral population;
- Background linkage disequilibrium (LD): We test for the impact of short-range SNP correlations on allele frequency estimates by computing autocorrelation between SNPs, or background LD;
- F_{ST}: We check the number of ancestral populations by computing F_{ST} among labeled and inferred ancestry;
- Assignment uncertainty: We test how distinct the ancestral populations are from one another by quantifying uncertainty in ancestral population assignment; and
- Association tests: We test whether or not the inferred population structure adequately controls for confounding latent population structure in association mapping studies by quantifying the difference in statistical significance of corrected associations versus uncorrected associations under the null hypothesis of no association.

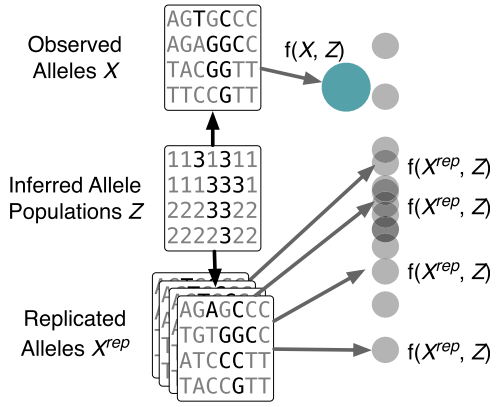

Fig. 1. Schematic diagram showing PPCs applied to genetic studies. The admixture model is fitted to a collection of genomic data. A distribution of the discrepancy function $f(\cdot,\cdot)$ for each of K inferred populations to the replicated genomic data X^{rep}, given the estimated ancestral population parameters Z, approximates the posterior predictive distribution. The probability of observing the discrepancy function applied to the observed data X with respect to this approximate posterior predictive distribution quantifies the goodness-of-fit.

Using and comparing PPCs with several discrepancies allows scientists to innovate the model in the directions for which it is most important for the analysis task (*Discussion*).

With five discrepancies and four datasets, PPCs reveal that each application of the admixture model meets and diverges from its assumptions in different ways. Each PPC indicates when we might extend the model to better match the complexities of the data at hand, and in the *Discussion* we point to some specific extensions that address each direction of misspecification. Although we focus here on the simple admixture model, we emphasize that assessing model fitness is important in any application of statistical models to data. We will demonstrate that PPCs give an intuitive framework for visually and quantitatively understanding when inferred hidden structure can or cannot be safely interpreted and used in downstream analyses.

Results

We used PPCs to assess the fit of an admixture model to data from four genomic studies. We developed five discrepancies, each measuring the degree to which the fitted model captures one aspect of the data. Here we report our findings.

We first describe the admixture model and outline our procedure for using a PPC to check for misspecification. The admixture model uses allele frequencies across genomes to recover individual-specific ancestry proportions and population-specific allele frequencies. The observed data include $i = 1, \ldots, n$ individuals, each with $\ell \in 1, \ldots, L$ SNPs. Each SNP ℓ for individual i is represented as two binary variables, $x_{i,\ell,1}, x_{i,\ell,2} \in \{0,1\}$, where $x_{i,\ell,1} + x_{i,\ell,2}$ is the number of copies of the minor (less frequent) allele (genetic variant); here we assume diploid organisms (two complete sets of chromosomes) and biallelic SNPs. The admixture model assumes that there are $k = 1, \ldots, K$ ancestral populations that describe these data, where K is specified a priori, and that each population is associated with location-specific allele frequencies $\beta_{k,\ell}$. Individual i's genotype data are generated as follows: (*i*) draw individual-specific ancestry proportions θ_i from a uniform Dirichlet distribution; (*ii*) for each SNP ℓ, draw two categorical variables $z_{i,\ell,1}$ and $z_{i,\ell,2}$ from a multinomial distribution with parameter θ_i; these latent variables indicate the ancestral populations assigned to the two copies of that SNP; and (*iii*) conditioned on the assigned ancestral populations, draw two alleles for SNP ℓ using the corresponding allele frequencies; that is, draw each $x_{i,\ell,\cdot}$ from a Bernoulli with parameter $\beta_{z_{i,\ell,\cdot},\ell}$.

Conditioned on data, the admixture model estimates the posterior distribution of latent population structure. This structure is encoded in individual specific population proportions θ_i (illustrated in Fig. 2), the population-specific allele frequencies β_k, and the assigned ancestral populations z. We use expectation maximization (EM) to estimate model parameters. See *Methods* for complete model specification and description of EM.

To estimate the posterior predictive distribution for this fitted model, we generated replicated data from parameter estimates. For each observed genomic data matrix x and K ancestral populations, we first estimated the latent parameters: β, θ, and z. With these estimates, we generated a collection of replicated datasets. We drew L SNPs for the same n individuals, holding the parameters z and β fixed. This resulted in a set of simulated individuals x^{rep}. In particular, for all i and ℓ, we drew $x^{rep}_{i,\ell,1} | z_{i,\ell,1}, \beta_\ell \sim Bernoulli(\beta_{z_{i,\ell,1},\ell})$ and similarly for $x^{rep}_{i,\ell,2}$. The replicated

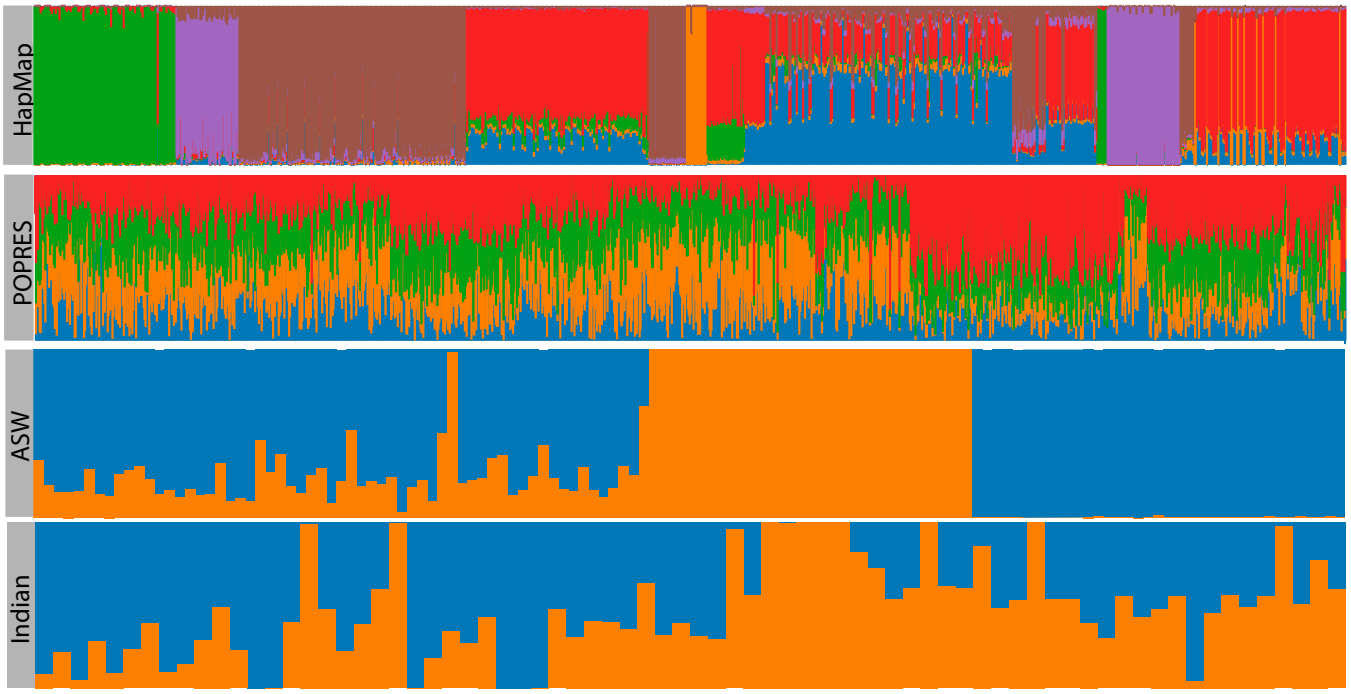

Fig. 2. Illustration of the variation in individual-specific ancestry proportions across the four genomic studies. The *x* axis represents the individuals in each study, and the *y* axis is the proportion of the genome with maximum likelihood assignment in each ancestral populations (colors distinguish ancestral populations). (*First Row*) HapMap phase 3 individuals, clustered by geographic origin of sample, fitted to six ancestral populations (14). Individuals, for the most part, have ancestry in one of the six continental ancestral populations, and are not substantially admixed. (*Second Row*) POPRES individuals, clustered by geographic location in Europe, fitted with four ancestral populations (15, 16). Because of the genetic proximity of the four populations, representing four corners of Europe, each individual has a proportion of ancestry in each population. (*Third Row*) ASW individuals, clustered by reported African ancestry, fitted to two ancestral populations (17). Because we include African and European individuals, we see individuals have either ~80% African and ~20% European ancestry, or all African or all European ancestry. (*Fourth Row*) Indian individuals, fitted to two ancestral populations (18). Most individuals have some proportion of ancestry in the two ancestral populations, because of an ancient admixture event.

data were generated conditional on the inferred latent variables; we do not need to reestimate them at any point in our analysis. For each dataset x and fitted admixture model we generated 100 replicated datasets x^{rep}.

For each PPC, we developed a discrepancy function $f(x,z,\theta)$, which is a function of the data and inferred latent structure. In our PPCs, each discrepancy partitions the alleles by assigned population and produces K scalar values. We computed the observed discrepancy $f(x,z,\theta)$ and the replicated discrepancy $f(x^{rep},z,\theta)$ for each replicated dataset. The empirical distribution of $f(x^{rep},z,\theta)$ is an estimate of the PPD of the discrepancy. Thus, we checked model fitness by locating the observed discrepancy in this distribution. If the observed discrepancy was an outlier with respect to this estimated PPD, then we conclude that the model is not a good fit to our data with respect to the discrepancy.

For each PPC, we used visualizations and assessments of significance to summarize the results (19). The PPC plots visualize the observed discrepancy against its PPD. We plotted the value of the replicated discrepancies $f(x^{rep},z=k,\theta)$ with gray circles and the observed discrepancy $f(x,z=k,\theta)$ with an offset solid circle. We colored the observed discrepancy to encode its z score, the number of SDs from the mean of the replicated discrepancy. Finally, we quantified the likelihood that the K z scores were jointly generated from a standard normal distribution. The number of gray stars at the top of each figure corresponds to the level of deviation from standard normal (*Methods*), which quantifies the magnitude of model misspecification with respect to a discrepancy.

Our results include evaluations of four genomic studies (see *Methods* for details). We set the number of ancestral populations based on prior work; we extend the results to diverse K in the *Supporting Information*. We study the following datasets:

- The HapMap phase 3 (*HapMap*) data include 1,043 individuals genotyped at 468,167 SNPs from populations worldwide; these individuals descend from distinct (and, in a few cases, admixtures of distinct) ancestral populations (20). Each individual is labeled with the location of the sample collection. Following prior work, we set $K=6$ (21).
- The population reference sample (POPRES) (*POPRES*) data include 1,387 individuals genotyped at 197,146 SNPs from continental Europe (16, 22). We expect the ancestry of these individuals to be an admixture of populations defined on a continuous geographic gradient instead of distinct ancestral populations (15). Each individual has a label describing the geographic birthplace of their grandparents, and we included individuals only if all four grandparents shared the same birthplace. Following prior work, we set $K=4$ (15).
- The African American (*ASW*) data include 61 African American individuals from the southwest United States, 32 individuals from Utah of Northern and Western European ancestry (CEU), and 37 Yoruban individuals from Ibadan, Nigeria (YRI) in the 1000 Genomes Project sequenced at 88,885 SNPs (17). We expect the ancestry of the genotypes from the ASW individuals are admixed from two distinct ancestral populations: European and West African. Individuals are labeled as ASW, CEU, or YRI. We set $K=2$ because all individuals are thought to have ancestry from only European and Yoruban populations.
- The Indian data (*Indian*) include 74 individuals genotyped at 196,375 SNPs from 15 distinct groups across India (18). The ancestry of each of these individuals' genotypes is admixed between two distinct but closely related ancestral groups: Ancestral North Indians and Ancestral South Indians (18).

Group labels are associated with each individual. Following prior work, we set $K = 2$ (18).

We number and describe the estimated ancestral populations in Table 1 (see also Fig. 2). We are using a fixed number of ancestral populations for each study, and we keep this numbering consistent throughout the figures in the *Results*.

Below, we describe each discrepancy and how each study fared under its lens. Then, we shift the focus to the genomic studies, summarizing how well the model fits the data across the collection of discrepancies.

Discrepancy in Within-Population Genomic Variation. Population admixture models are often used to explain similarities between individuals' genotypes: if two individuals have portions of their chromosomes that are descended from the same ancestral population, they will, probabilistically, have more alleles in common than two individuals descended from distinct populations according to the admixture model. How many more alleles the individuals share depends on the separation between ancestral populations, the proportion of the genomes with shared ancestry, and within-population genetic variation.

We developed a discrepancy to compute the IBS between pairs of individuals, which measures conditional similarity of alleles between individuals (23). For a pair of individuals, we averaged the number of alleles in common across the genome for SNPs that share ancestry assignments in the fitted model (24). For this discrepancy, a pair of individuals with a discrepancy of 1.0 for population k indicates that 100% of the alleles assigned to population k in both individuals are identical, whereas 0.0 indicates 100% differing alleles. A misspecified model with respect to this IBS discrepancy may imply that we cannot safely ignore admixture LD, or (possibly long distance) genetic correlations induced by recent admixture of distinct populations, in our admixture model (25).

We performed the PPC and found that, for HapMap, ASW, and Indian, the within-population variation is well captured by the admixture model (Fig. 3). In contrast, the POPRES data showed underestimated average interindividual similarity, indicating model misspecification. Note that the two studies with distinct populations (HapMap, ASW) tended to have z scores greater than zero indicating greater than expected interindividual similarity on the observed data; the two studies with less well-separated populations (POPRES, Indian) have z scores below zero, indicating less similarity than expected.

We also looked at the observed discrepancy without the context of the replicated discrepancy. In the POPRES data and, to a lesser degree, the Indian data, the average similarity of individuals is similar across ancestral populations. This might be expected when modeling data with continuous population structure. In contrast, HapMap and, to a lesser extent, ASW exhibit more variability across ancestral populations. This may be a function of variable heterozygosity within the distinct ancestral populations (26). The higher observed discrepancy in ASW relative to the same recovered ancestral populations in HapMap may suggest that, within this ASW population of African Americans, there is less variability within ancestral populations than in estimates of these populations from nonadmixed individuals (European and Yoruban HapMap individuals, for example). This is interesting in light of recent estimates of the effective population size of the ASW population, which is an order of magnitude greater than either the European or Yoruban effective population sizes (27).

A mismatch in observed and replicated interindividual similarity leading to a failure of this PPC may be due to admixture linkage disequilibrium. Admixture LD arises as a result of admixture across two distant populations; because of the difference in allele frequencies between these two populations, there will be large regions across chromosomes that are well correlated because of shared ancestry (25). More generations of admixture reduces the effect of admixture LD for genomic regions that are inherited independently, leaving only correlations among neighboring genomic loci (background LD). Thus, for recently admixed individuals such as in ASW, we might expect to see evidence of admixture LD that leads to higher-than-expected similarity among individuals. Lower than expected similarity is likely due to the proximity of the ancestral populations: if the ancestral populations are difficult to distinguish with respect to allele frequencies, then SNP-specific ancestral assignments will be arbitrary, and within-population allele variation will be higher in the observed data than is modeled by the Bernoulli and found in the replicated data (Fig. S1).

Discrepancy in Background LD. Linkage disequilibrium LD is the nonrandom assortment of alleles across the genome. LD occurs when alleles are not inherited independently: the alleles at one genomic locus provide information about the alleles at another locus. The process of recombination in diploid chromosomes implies that alleles that are nearby on a chromosome are inherited together unless a recombination event occurs, creating dependencies in local genotypes populationwide. Recombination is an infrequent event across a genome: for a single offspring, there are on average a small number of recombination events per chromosome, with high variance (28). Recombination events within a population lead to a blocklike correlation structure of genotypes across the genome, where polymorphisms adjacent in the chromosome will be well correlated (referred to as background LD, in contrast to possibly long-distance dependencies induced by admixture LD). This correlation decays as the chromosomal distance increases—but not uniformly. Although each study we analyzed contains local correlation patterns, the admixture model assumes independence of every SNP conditional on population ancestry;

Table 1. Numbering and composition of each of the estimated ancestral populations for each study

	HapMap	POPRES	ASW	Indian
1	Pakistan (0.6); N Europe (0.2); Russia (0.2); Israel (0.2)	Ukraine (0.5); Slovenia (0.5); Croatia (0.4); Bosn. Herz. (0.4)	YRI (1.0); ASW (0.8)	Kurumba (0.8); Bhil (0.8); Tharu (0.7); Vaish (0.7)
2	New Guinea (1.0); S Europe (0.2)	Norway (0.6); Latvia (0.5); Finland (0.5); Denmark (0.5)	CEU (1.0); ASW (0.2)	Vysya (1.0); Velama (0.8); Srivastava (0.8); Kamsali (0.7)
3	S Africa (1.0); C Africa (1.0); N Africa (0.3)	Portugal (0.3); Spain (0.3); Bulgaria (0.3); Switz. (0.3)		
4	N Europe (0.7); Israel (0.7); N Africa (0.7); S Europe (0.7)	Slovakia (0.6); Turkey (0.6); Greece (0.5); Italy (0.5)		
5	S America (1.0); C America (0.9)			
6	Japan (0.9); China (0.9); SE Asia (0.8); Russia (0.3)			

Proportions in parentheses are the estimated proportions for individuals with those geographic labels that have ancestry in that estimated ancestral population.

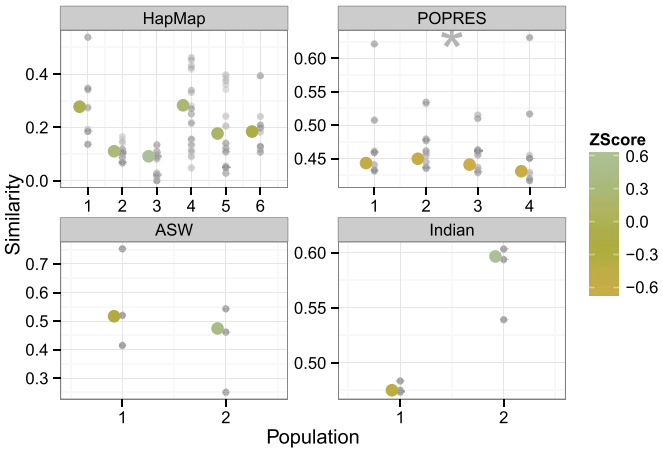

Fig. 3. Average population-specific interindividual similarity across four studies. Each *x* axis represents the study for which the admixture model was fitted. Each *y* axis represents mean interindividual similarity across individuals conditioned on one ancestral population. Gray points represent values from replicated data. Larger points represent values from observed data, colored by their *z* scores relative to the empirical distribution of the replicated values. Stars indicate significant divergence in *z* scores from a standard normal. POPRES data show underestimated average interindividual similarity relative to the data replicates.

this PPC check fails when this independence assumption does not hold in the posterior predictive distribution.

We assessed the fitted admixture model for local correlation structure among SNPs. To do this, we built a discrepancy function (LD discrepancy) that measures local dependencies among SNPs using mutual information (MI). The history of statistics to quantify LD is rich (29), and there are many possibilities for this discrepancy. Mutual information quantifies the reduction in uncertainty about random variable $x_{i,\ell,j}$ given knowledge of the state of $x_{i,\ell',j'}$ (or vice versa; MI is symmetric) for a pair of discrete random variables, here, the observed alleles at two loci, $x_{i,\ell,j}$ and $x_{i,\ell',j'}$ (30) (Eq. 1). If two SNPs are independent (i.e., in linkage equilibrium, or inherited independently), MI will, theoretically, be zero. However, finite samples imply that, even when two SNPs are in linkage equilibrium, the MI may be nonzero by chance. Our discrepancy uses MI to measure dependence between adjacent pairs of assayed SNPs. We checked lags between pairs of SNPs varying between 1 and 30, indicating 0 to 29 intervening SNPs between the pairs of tested SNPs along the chromosome.

When we applied this LD PPC to our data, we found that, across studies, the models were generally misspecified at small lags (Fig. 4). For all studies but HapMap, the *z* scores show that, as the lag grows, the observed MI tend toward underestimates of MI in the posterior predictive distribution. The *z* scores for POPRES deviate substantially from a standard normal for all lags indicating more observed background LD than expected. The model's independence assumption across SNPs manifests in nearly identical replicated discrepancies across lags, regardless of the number of ancestral populations (Fig. S2). Failure of this PPC may suggest using a subset of SNPs with low pairwise LD; however, admixture models rely on greater SNP densities to enable the separation of similar ancestral populations.

We also looked at the observed discrepancies alone. In POPRES and Indian data, background LD decayed rapidly with observed SNP lag (31). In the POPRES data, the measurements of background LD were consistent across inferred populations. In the HapMap data, background LD differs among distinct populations (32); this is expected because LD is impacted by population-specific effects, including migration, deviations from random mating, and selection (33). In HapMap and, less so, ASW, background LD is fairly uniform across different lags. This difference may be a function of the density of SNPs and SNP ascertainment biases, which differs across studies, but we found that these results persist for lags up to 1,000 adjacent SNPs. Note also that the HapMap data include almost 2.5 times more SNPs than the POPRES data.

Discrepancy in Reported Ancestry. It is often assumed that self-reported ancestries provide no additional information above fitted individual-specific ancestry parameters. To test misspecification with respect to this assumption, we developed a PPC to compare inferred individual-specific ancestry proportions to geography labels using the fixation index F_{ST}. Using the fitted admixture model to partition alleles into one of K inferred ancestral populations, F_{ST} measures whether or not subdividing these alleles by individual-specific geography labels reduces the variance of those allele frequencies. The F_{ST} is zero if no additional population structure exists in the geography labels beyond what is already recovered in the inferred ancestral populations. Conversely, large values of F_{ST} indicate that the geography labels capture additional structure not found in the estimated populations. For a fitted admixture model with K populations, we evaluated this population-specific F_{ST} discrepancy function by computing K F_{ST} values (Eq. **2**).

We computed the PPC for the F_{ST} discrepancy. We found that, except in ASW, the estimated genetic ancestry often reflects all of the population structure in the reported ancestry (Fig. 5). If we perform this PPC over many different values of ancestral populations K, we found that this PPC tends to fail for values of K that are inadequate to explain variation in the observed data (Fig. S3). This discrepancy appears to be, in these four studies, an effective approach to evaluate the range of well-specified numbers of ancestral

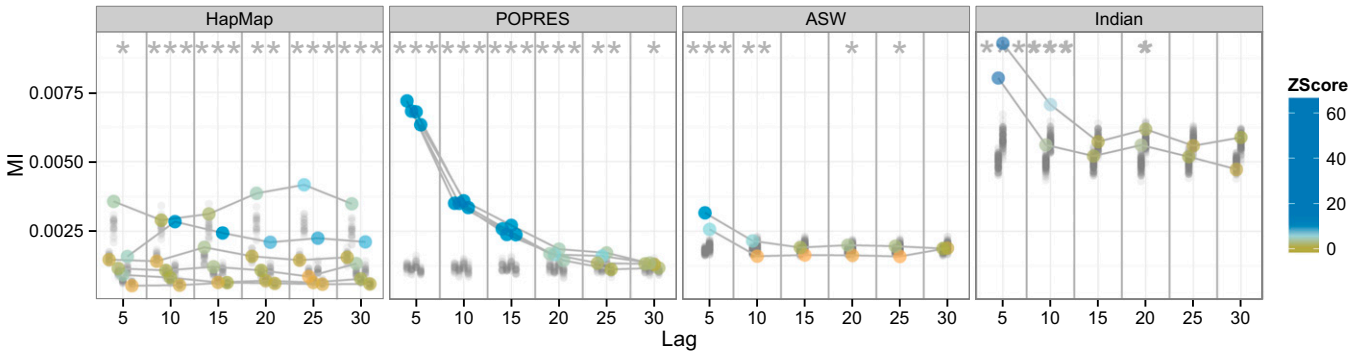

Fig. 4. LD discrepancy PPC across four studies. The *x* axis indicates the number of lag SNPs between pairs of SNPs over which mutual information is averaged. The *y* axis represents the average MI between SNPs at varying lags. Observed values for each population are connected by lines across lags and colored by *z* score. None of these studies are well fit across all SNP lags, highlighting the impact of the assumption of independence between SNPs.

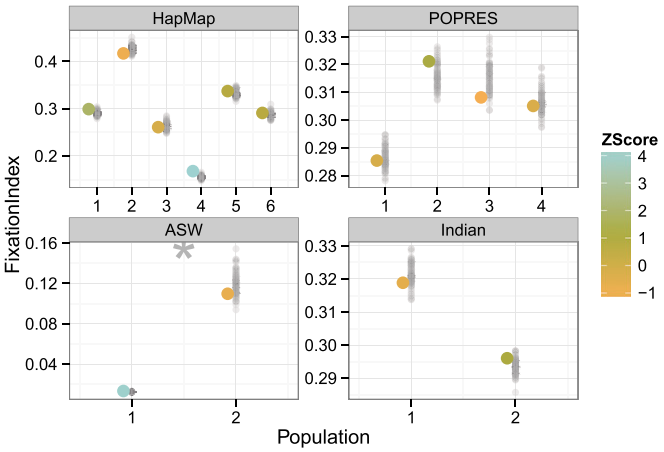

Fig. 5. F_{ST} discrepancy function across four studies. The *x* axis columns represent application to HapMap, POPRES, ASW, and Indian studies, respectively. The *y* axis represents the F_{ST} value of a single ancestral population *k* with respect to the geography labels. Smaller values indicate a closer match between the labeled and estimated populations. In ASW, the number of ancestral populations $K = 2$ is outside of an acceptable range.

populations in the admixture model. Based on recent observations in these 1000 Genomes data, we hypothesize that these ASW individuals have an uncharacteristically large proportion of Native American ancestry relative to prior genomic studies from African American individuals, and that the correct number of populations here is indeed $K = 3$ (34); this is confirmed in the success of this PPC for $K = 3$ as described in the *Supporting Information*.

The observed F_{ST} varied considerably across studies (Fig. 5). As with other discrepancy functions, HapMap and ASW show greater variability of observed F_{ST} values across populations than POPRES and Indian due to greater heterogeneity of ancestral populations. The two ASW populations have consistently lower observed F_{ST} values, suggesting that the fitted admixture model captures most of the information in the reported ancestries (ASW, CEU, and YRI). Indeed, F_{ST} is undefined for several populations within the ASW models because all alleles assigned to those populations were from individuals with a single reported ancestry. In the context of the replicated data, however, this conclusion is shown to be incorrect: there is latent structure in the data that is not captured in this model, specifically, the presence of Native American ancestry. The Indian data, with 14 reported ancestries, have higher overall F_{ST} values: individuals within a reported ancestry have a range of admixture between Ancestral North Indian and Ancestral South Indian (18), and geographic labels are, unsurprisingly, a poor indicator of admixture proportions across these individuals.

Discrepancy in Uncertainty in Ancestral Population Assignments. We chose these four studies because they represent distinct patterns of admixture (Fig. 2): in HapMap, most individuals have ancestry explained by one or two populations; in POPRES, although regional variation in ancestry proportions is clear, most individuals have substantial ancestry contributions from all four populations. We measured the degree of uncertainty of individual-specific population assignments to determine whether this uncertainty is characteristic of true latent structure or evidence of model misspecification.

The discrepancy function for this task quantifies the average entropy, or uncertainty, of the ancestral population assignment for alleles in the fitted model. We used the fitted admixture model to compute estimates of the conditional probability of the assignment of each SNP allele to each ancestral population $k \in \{1, \ldots, K\}$. We then computed, for each population, the average entropy of this conditional probability across individuals and SNPs (Eq. **3**).

We performed PPCs with this discrepancy and found that this model is misspecified for ASW with respect to uncertainty in allele-specific ancestry assignments (Fig. 6); the other three studies did not indicate problems. We found the small variation in average entropy among the replicated data surprising; indeed, the replicated points appear near identical at the scale of the visualization. Again we hypothesize that the failure of this PPC on ASW is due to additional Native American ancestry that is not captured with $K = 2$ populations (Fig. S4).

The success of this PPC across three studies is also unexpected when considering the observed discrepancies alone. The POPRES and Indian studies showed high entropy, indicating a high degree of uncertainty in population assignments (Fig. 6). This uncertainty does not appear to be caused by lack of EM convergence: this same uncertainty is observed for models fit with 10 times as many EM iterations. In contrast, the HapMap and ASW studies, which have distinct ancestral populations, have greater variance of average entropy across populations within *K*, indicating greater uncertainty in some population assignments relative to others.

Discrepancy in Correcting for Population Structure in Genome-Wide Association Studies. The final discrepancy function we considered is the value of using estimates of individual-specific ancestry proportions to correct for population structure in association studies (1, 2, 35–37). Association mapping uncovers associations between SNPs and traits (e.g., height, disease status, cholesterol levels). Correcting for population structure in association mapping is essential because latent structure leads to false positive associations when alleles that have different frequencies across populations are mapped to traits with differential rates across populations (38, 39). We note that, in the original manuscript describing the use of admixture model parameters to correct for population stratification in association studies (1), they effectively performed the same PPC as we applied here without describing it as a posterior predictive check.

For this task, we quantified the effect of the fitted admixture parameters on the $\log_{10}$ Bayes factors (BFs) comparing the null hypothesis of no association between a SNP and the trait controlling for population assignment, and the alternative hypothesis of an association between a SNP and the trait, controlling for population assignment (following prior work; ref. 1). We randomly

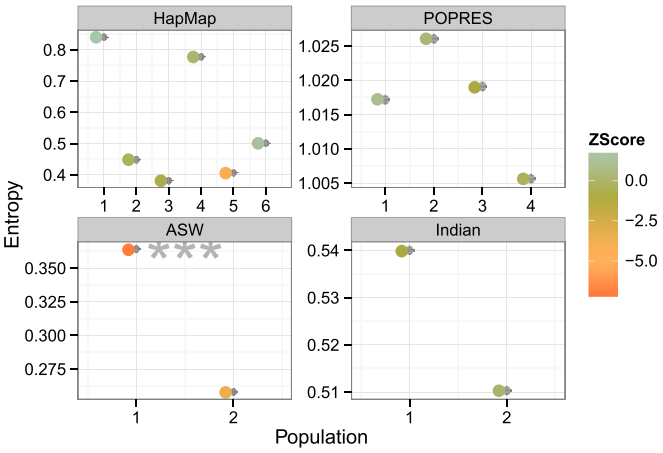

Fig. 6. Average entropy discrepancy function across four studies. The *x* axis represents the number of populations in the fitted admixture model; panels represent application to HapMap, POPRES, ASW, and Indian studies. The *y* axis represents the average entropy of the posterior distribution over populations for each allele. Larger absolute value entropy represents greater uncertainty over population assignments. This discrepancy function shows small variance among replicates and finds a misspecified model for ASW.

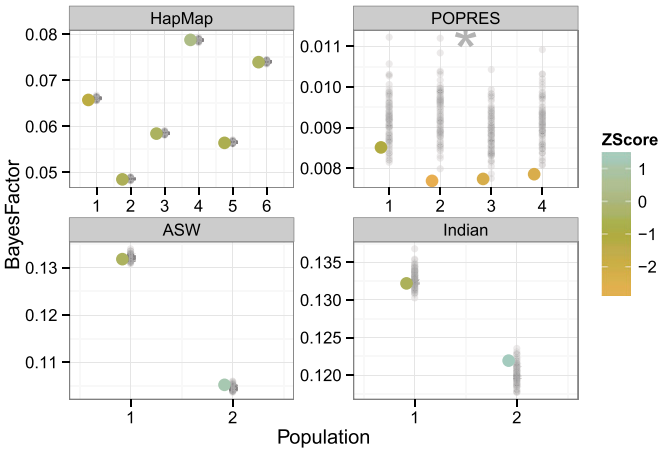

Fig. 7. Association mapping correction PPC across four studies. The *x* axis represents application to HapMap, POPRES, ASW, and Indian studies, respectively. The *y* axis represents maximum $\log_{10}$ BFs. Maximum $\log_{10}$ BFs from replicated genomes are shown with gray points. Colors of the observed data discrepancy values represent the *z* score with respect to the replicated samples from fitted admixture models. Stars indicate deviation from normality in *z* scores. The POPRES data appears to deviate from the estimated posterior predictive distribution for association mapping.

generated binary traits for each study using a population k in the model to create phenotypes with different rates within our estimated populations but with no explicit association with any SNP. Thus, controlling for population structure, any significant association will be a false positive. The value of the discrepancy for a specific population is computed as the maximum $\log_{10}$ BF over all SNPs. We then computed the *z* score of the $\log_{10}$ BFs of the observed data associations with respect to the $\log_{10}$ BFs from the replicated data.

We performed this mapping PPC and found that the observed maximum $\log_{10}$ BFs are generally within the expected range when sampling alleles from the fitted model, which provides additional confidence in our ability to reject false positives (Fig. 7). The variation in the POPRES replicates highlights why the PPC fails: controlling for fine levels of population structure with noisy discrete estimates is not effective control for structure. It is counterintuitive, although, that the four populations in POPRES have lower than expected observed corrected $\log_{10}$ BFs (Fig. S5).

In the observed discrepancies alone, we found that the maximum $\log_{10}$ BF across studies was small (<0.14), indicating that this approach to correcting for population structure is broadly effective at avoiding false positive associations. As in previous discrepancies, the variability in maximum $\log_{10}$ BF between populations was greater for HapMap and ASW data than POPRES and Indian data, reflecting greater distinction between populations in the first two studies. The maximum $\log_{10}$ BFs for POPRES are smaller than the other studies, reflecting less significance in tests for association between SNPs and populations; this is unsurprising given the homogeneity of the inferred populations and the corresponding randomly generated phenotypes.

Summarizing PPC Results Within Study. We turn our attention to summarizing the results from the five PPCs for each of our genomic studies. Our results across PPCs tell a complex story for each study that indicates specific misspecified assumptions.

HapMap Phase 3. In our application of PPCs to the HapMap phase 3 data, we found that the background LD discrepancy PPC indicated a gross model misspecification, highlighting the negative impact of the assumption of independent SNPs. Other PPCs did not find misspecification with $K = 6$ on these data. Two ways to address this model misspecification with respect to background LD in these well-separated ancestral populations would be to (*i*) prune the SNP data drastically to select a near-independent set of SNPs from which ancestry may be estimated, or (*ii*) model background LD explicitly (*Discussion*).

European Samples. In our application of PPCs to the POPRES data, we found that the admixture model with $K = 4$ ancestral populations generally indicated similar variation in discrepancy within and across populations. Many of the failures of the PPCs on these data could be used together to highlight model misspecification with respect to the continuous structure of the ancestral populations, which is poorly captured by discrete structure assumed by an admixture model. We might model latent structure for these data with a continuous population model (e.g., principal components based; refs. 40, 41).

African Americans. PPCs on the ASW data with $K = 2$ showed that the ancestral population corresponding to European ancestry was well captured in the observed data with respect to the replicated data, but the population corresponding to Yoruban ancestry was often misspecified for the observed data with respect to the replicated data; the F_{ST} PPC and the entropy PPC show this differential fit. For these data, explicitly modeling Native American ancestry with $K = 3$ addresses these specific model misspecifications (see *Supporting Information* for results).

Continental Indians. With PPCs on the Indian data with $K = 2$, we found that the failure of the entropy PPC indicates that the underlying estimates of the two ancestral populations have substantial uncertainty. Relying on these estimates to characterize ancestral population allele frequencies or determine admixture proportions for each individual is unjustifiable. These data may also benefit from using a continuous model of ancestry because of the difficulty of differentiating these two fairly proximal ancestral populations many generations after the admixture events (18); alternatively, denser biomarkers may facilitate separation of the two ancestral populations.

Discussion

We have developed posterior predictive checks for analyzing genomic datasets with an admixture model. We have demonstrated that the PPC gives a valuable perspective on genetic data beyond statistical inference of model parameters. Research on fitted admixture models is often accompanied by a "just so" story to explain the inferred parameters and how they are reflective of ancestral truth (14). The model may suggest specific hypotheses, but only conditioned on the model being a good fit for the observed data. PPCs check this assumption of good fit, giving support to hypotheses by confirming that the underlying assumptions do not oversimplify the existing structure in the observed data. In this paper, we developed PPCs for the admixture model. We designed biological discrepancy functions to quantify the effect of the model assumptions on interpreting and using the estimated parameters for downstream analyses.

Statistical modeling of genetic data requires us to balance the complexity of the model with its capacity to capture the data at hand. We are often limited, for example, by insufficient data to support an overly complex model, or by computational constraints on the class of model we wish to fit. Thus, we support the iterative practice of fitting the simplest model (i.e., the one we fit here), checking whether a higher resolution model is needed, and then improving the model only in the ways that result in more reliable interpretations of the results. PPCs drive this process of targeted model development, pointing us toward enriched Bayesian models to quantifiably improve their performance for the exploratory tasks at hand. With this practice in mind, we revisit the PPCs described above. We discuss how we extend the admixture

model, or choose a variant from the research literature, when we detect a misspecified assumption.

Many population studies have applied admixture models to explore and quantify genetic variation between individuals within and across ancestral populations (14, 42, 43); these analyses may benefit from the interindividual PPC. For studies where this PPC indicates misfit, prior work has adapted the admixture model to control admixture LD by explicitly modeling haplotype blocks for each ancestral population (44). In particular, the SNP-specific ancestry assignment z variables for each individual are modeled by a Markov chain, where the probability of transitioning to a different ancestral population from one position to the next has an exponential distribution. This specifies a Poisson process describing haplotype block lengths across the genome, with global rate parameter r.

Many studies have noted that background LD may lead to phantom ancestral populations (45); applying admixture models to genomic data that contain background LD may find the SNP autocorrelation PPC useful. After identifying model misspecification using the background LD discrepancy, we could extend the admixture model to explicitly capture background LD. Above we described a Markov model on the z variables, where genotypes remain independent conditional on ancestral population assignment. Extending this idea, Saber (46) implements a Markov hidden Markov model (MHMM) to capture both haplotype blocks and background LD by adding a Markov chain across the observed SNPs x. Others have further extended this model in various ways, including estimating recombination events explicitly in the MHMM (47), and incorporating sophisticated models of haplotypes and LD (48).

The F_{ST} discrepancy function effectively checks for a misspecified number of ancestral populations K. The ubiquitous problem of selecting a number of ancestral populations is, arguably, the most substantial hurdle to overcome in applying admixture models (or more general latent factor models) to data (2, 7, 14). Methods and statistics have been proposed to evaluate the proper number of latent ancestral populations, often motivated by F_{ST} (7, 49); additionally, Bayesian nonparametric models estimate the posterior distribution of K (50, 51). We propose a PPC with the F_{ST} discrepancy for general use in evaluating appropriate ranges of the number of ancestral populations for a specific study. A simple adaptation of the model to correct for a failure of this PPC is to change the number of ancestral populations K (Fig. S3).

There are explicit model extensions that affect the F_{ST} of the inferred ancestral populations. For example, one can build hierarchical models that allow the sharing of allele frequencies across populations for some SNPs; this was implemented in the Structure 2.0 model, which includes a hierarchical component to allow similar allele frequencies across ancestral populations (the F model) (44). A second example is from the topic model literature (similar models applied to text documents), where the ancestral populations are captured in a tree-structured hierarchy (52, 53). In the corresponding admixture model, the root node would include SNPs that have shared allele frequencies across all ancestral populations; the leaves would include SNPs that have a frequency in a single ancestral population that is different from the frequency in all other ancestral populations (referred to as ancestry informative markers, ref. 21).

Previous population studies have explored and interpreted the population-specific SNP frequencies estimated by admixture models (54–56); almost all applications of the admixture model have used point estimates of ancestry assignments to determine the proportion of admixture in individuals (15, 21). The average entropy PPC will check uncertainty in ancestry assignment, and has implications for interpreting estimates of SNP frequencies. To adapt the model to this misspecification, the hyperparameters for the Dirichlet-distributed allele-specific ancestry assignments may be changed. We and others set $\alpha = 1$ (7), giving equal weight to all possible contribution across ancestries for each SNP. In particular, we might give higher weight to admixture proportions near 0 and 1 by setting $\alpha < 1$ for studies where we expect low levels of admixture (e.g., the HapMap data). The equivalent change for the hyperparameters in the population-specific allele frequency parameters would encourage allele frequency distributions that more closely match observed allele frequency spectrums (57).

Another relevant model adaptation would be to modify the distribution of a SNP to be not Bernoulli but instead Poisson (58), Gaussian (59), or something more sophisticated (60, 61). Although these extensions seem reasonable, the PPC with this discrepancy found little need to modify the admixture model assumptions in our current studies. The exception to this point is the ASW study, although we note that setting $K = 3$ as suggested above will, in part, address this misspecification (Fig. S4).

We believe that methods to control for population stratification in association mapping will benefit from application of the mapping PPC, including linear mixed models and nongenerative methods such as EIGENSTRAT (2, 62). Failure of the association mapping PPC indicates that the estimates of population structure are insufficient to correct for the confounding latent structure in the individuals. There are many directions to consider for mitigating this type of model misspecification. As examples, one may use larger numbers of estimated principal components or populations, use alternative approaches to specifying the latent structure variables, or correct for structure estimated on local regions of the genome.

Applied statisticians develop models to capture the complexity of data. To form hypotheses from these models, however, we need assurances that the data can support them. PPCs provide a simple mechanism to quantify when a model is sufficient or when it needs additional structure to support downstream analysis for specific data. Although we have focused on the admixture model, the PPC procedure applies to any probabilistic model. For example, we believe there could be a substantial role for PPCs in evaluating demographic models. As we continue to collect genomic data, we continue to develop complex models to explain them. Equally important to building our repertoire of statistical models for analyzing genomic data are to build our repertoire of ways to check these analyses.

Materials and Methods

Software for all methods described here is publicly available at github.com/mimno/admixture-ppc. Note that this PPC software to generate replicated data from a fitted model may be used to evaluate any admixture model where the alleles are independent Bernoulli variables from one of K populations indicated by some fitted latent variable z, which is fixed in the data replicates. This class of admixture model includes Structure 2.0 (44). The discrepancies, however, will work for all admixture models, although they will not be meaningful for some that encode different assumptions.

Genomic Study Data. We downloaded the HapMap phase 3 release 3 genotype data from the National Center for Biotechnology Information (NCBI) website (20). We downloaded the POPRES data from dbGaP (accession no. phs000145) and filtered these genotype data as described in prior work (15, 22). We downloaded the ASW genotype data from 61 individuals, 32 CEU individuals, and 37 YRI individuals from the 1000 Genomes Project Phase 1 (17) from the 1000 Genomes FTP server (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results) and pruned these data by randomly sampling 1% of SNPs with minor allele frequency (MAF) ≥5%. We downloaded the Indian data from the Reich Lab FTP server (18) and processed it as in previous work (15).

Admixture Model and Parameter Estimation. We used a standard admixture model in this work, as implemented in the Structure software (7), although we built our own software to fit this model to large-scale data as described below. We represented ancestral population $k \in \{1, \ldots, K\}$ with a Dirichlet-distributed random variable $\phi_{\ell,k}$, for each SNP $\ell \in \{1, \ldots, L\}$, over all possible alleles; we assumed exactly two alleles for each SNP, so this simplified to a beta distribution. We represented individual-specific ancestry proportions for individual $i \in \{1, \ldots, n\}$ as a Dirichlet-distributed variable θ_i over K populations. To generate a diploid genotype i from this model, for each SNP ℓ we sampled two multinomial variables with parameter θ_i, $z_{i,\ell,1}$ and $z_{i,\ell,2}$, one for each allele copy. We then sampled alleles $x_{i,\ell,1}$ and $x_{i,\ell,2}$ from the distributions $\phi_{\ell,z_{i,\ell,1}}$ and $\phi_{\ell,z_{i,\ell,2}}$, respectively. The generative model has the following form:

$$\begin{aligned}\theta_i &\sim Dir_K(\alpha)\\ \phi_{\ell,z_{i,\ell,j}} &\sim Beta(\gamma,\gamma)\\ z_{i,\ell,j} &\sim Mult(\theta_i) \text{ for } j\in\{1,2\}\\ x_{i,\ell,j} &\sim Mult\left(\phi_{\ell,z_{i,\ell,j}}\right) \text{ for } j\in\{1,2\},\end{aligned}$$

where $j\in\{1,2\}$ represents the two allele copies. For simplicity, we set $\alpha=1$ and $\gamma=1$, indicating a uniform prior on the ancestry proportions for each individual and the site-specific allele frequencies for each population (63). The likelihood for this model has the form:

$$P(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta},\phi)=\prod_i^n\prod_\ell^L\prod_{j\in\{1,2\}}\phi_{\ell,z_{i,\ell,j}}^{x_{i,\ell,j}}\left(1-\phi_{\ell,z_{i,\ell,j}}\right)^{1-x_{i,\ell,j}}\theta_{i,z_{i,\ell,j}}.$$

Without loss of generality, we represent a heterozygous SNP (encoded as 1) as $x_{i,\ell,1}=1$ and $x_{i,\ell,2}=0$, where 1 and 0 are the two alleles at that site (7).

We estimated parameters and latent variables with EM, alternating between (*i*) (E step) estimating the posterior mode for population assignments of alleles Z given estimates for ϕ and $\boldsymbol{\theta}$, and (*ii*) (M step) maximizing the individual- and population-level parameters given posterior modes of Z (63, 64). The update equations for the E step, estimating the posterior mode for population assignment for an allele, are

$$\begin{aligned}q(z_{i,\ell,j}=k|x_{i,\ell,j}=1,\phi_{\ell,k},\theta_{i,k})&\propto\theta_{i,k}\phi_{\ell,k}\\ q(z_{i,\ell,j}=k|x_{i,\ell,j}=0,\phi_{\ell,k},\theta_{i,k})&\propto\theta_{i,k}(1-\phi_{\ell,k}).\end{aligned}$$

The update equations for the M step, estimating the model parameters for n individuals and L SNPs given the posterior mode for $\boldsymbol{Z}$, are:

$$\begin{aligned}\phi_{\ell,k}=\frac{1}{2n}\sum_{i=1}^n\sum_{j\in\{1,2\}}&q(z_{i,\ell,j}=k)^{\mathbb{1}[x_{i,\ell,j}=1]}\\ &(1-q(z_{i,\ell,j}=k))^{\mathbb{1}[x_{i,\ell,j}=0]}\\ \theta_{i,k}=\frac{1}{2L}\sum_{\ell=1}^L\sum_{j\in\{1,2\}}&q(z_{i,\ell,j}=k).\end{aligned}$$

We initialized the population-specific minor allele frequency parameters $\phi_{\ell,k}$ to the empirical proportion of the minor allele in the training data, plus a uniform random variable $u\sim\mathcal{U}(0,0.1)$, truncating extreme values so that $0.05<\phi_{\ell,k}<0.95$. To initialize population proportions for each individual, we drew K uniform random variables $u_k\sim\mathcal{U}(0,1)$ and set $\theta_{i,k}=u_k/\sum_{k'}^K u_{k'}$. We iterated between these E and M steps for 1,000 iterations. We also ran a number of models to 10,000 iterations, but found no substantial differences in the fitted models, supporting convergence of the parameter estimates in 1,000 iterations.

Discrepancy: Interindividual Similarity for IBS. We measured similarity between individuals (identity by state) with respect to population k by counting the number of alleles that are shared between individuals with population assignment $z_{i,\ell,j}=k$ (i.e., the Manhattan distance) and dividing by the total number of alleles with population assignment $z_{i,\ell,j}=k$ (24). The value of the discrepancy function is the average proportion of shared alleles over all pairs of individuals. We ignored pairs of genomes that shared fewer than 500 alleles assigned to a population k.

Discrepancy: Mutual Information for Background LD. We calculated population-specific MI for each pair of adjacent SNPs within a lag of all integers $m\in\{2,\ldots,30\}$. This window is with respect to the ordering of the SNPs in our filtered dataset based on chromosomal position, although the actual distance in base pairs between SNPs varies dramatically across studies and SNP pairs.

Mutual information—quantified in bits—is computed between alleles at two adjacent loci separated by $m-1$ SNPs, $x_{i,\ell,j}$ and $x_{i,\ell',j'}$. For haploid genotypes, both generated from the same population k, we computed:

$$I(x_{i,\ell,j};x_{i,\ell',j'})=\sum_{x_{i,\ell,j}\in\{0,1\}}\sum_{x_{i,\ell',j'}\in\{0,1\}}p(x_{i,\ell,j},x_{i,\ell',j'})\log_2\left(\frac{p(x_{i,\ell,j},x_{i,\ell',j'})}{p(x_{i,\ell,j})p(x_{i,\ell',j'})}\right). \quad [1]$$

As calculating this statistic is computationally intensive, we computed MI for this 30 SNP window only over the first 10,000 SNPs in each study.

Discrepancy: Distance Between Estimated Ancestral Populations and Geographic Labels. We computed the F_{ST} statistic on alleles assigned to one ancestral population relative to the geographic labels using Wright's estimator of F_{ST}, which considers single alleles at each genomic locus (65). In our data, each individual has exactly one geographic label g, but individual's alleles are assigned to possibly many ancestral populations. The probability of an allele for SNP ℓ and population k is computed as $\overline{p}_{k,\ell}=\frac{1}{N_{k,\ell}}\sum_i\sum_{j\in\{1,2\}}x_{i,\ell,j}\mathbb{1}[z_{i,\ell,j}=k]$, where $N_{k,\ell}$ is the total number of alleles assigned to population k, and $\mathbb{1}[\cdot]$ is an indicator function. We further partition these population-specific probabilities by geographic label g: $\overline{p}_{k,g,\ell}=\frac{1}{N_{k,g,\ell}}\sum_{i\in g}\sum_{j\in\{1,2\}}x_{i,\ell,j}\mathbb{1}[z_{i,\ell,j}=k]$, where $N_{k,g,\ell}$ is the number of alleles assigned to population k for individuals with geographic label g, which may be zero, in which case this probability is set to zero. With these probabilities in hand, and assuming Bernoulli distribution of alleles (which fixes the variance given the expectation), we calculate fixation index F_{ST} as

$$F_{ST}=\frac{\frac{1}{R_k}\sum_{g=1}^{|g|}\left(\overline{p}_{k,\ell}-\overline{p}_{k,g,\ell}\right)^2}{\overline{p}_{k,\ell}\left(1-\overline{p}_{k,\ell}\right)}, \quad [2]$$

where R_k is the number of geographic labels with nonzero alleles in population k.

Discrepancy: Average Entropy. We computed the average entropy discrepancy function as the average entropy for each estimated ancestral population over all alleles assigned to a population. Given estimates of the distribution of ancestral populations for an individual i, $\hat{\theta}_{i,\cdot}$, and the probability of the minor allele for a SNP across populations, $\phi_{\ell,\cdot}$, we calculated the posterior probability of each population k for the observed alleles $x_{i,\ell,\cdot}$ in that individual's genome. The entropy is

$$\mathcal{H}(z_{i,\ell,\cdot}|x_{i,\ell,\cdot})=-\sum_{k=1}^K p(z_{i,\ell,\cdot}=k|x_{i,\ell,\cdot})\log_2 p(z=k|x_{i,\ell,\cdot}). \quad [3]$$

Then we computed the discrepancy as the average entropy for each ancestral population k over all alleles assigned to a population.

Discrepancy: Correcting Latent Structure in Association Mapping. For each model and each ancestral population k, we generated a binary phenotype vector of length n, simulating a condition that is associated with that population. In particular, for the kth phenotype vector, we sampled a Bernoulli random variable c_i for each individual with probability $0.5\theta_{k,i}+0.1(1-\theta k,i)$, so individuals with ancestry in population k are more likely to exhibit the phenotype. To compute the discrepancy function for the GWAS association controlling for ancestry, we computed, for each SNP, the 2 ln BF for each SNP as the ratio of the likelihood of the genotype given the phenotype and the population assignment of the SNP ($p(x_i|c_i,z_i)$), and the likelihood of the genotype given the population assignment ($p(x_i|z_i)$) (1, 66). We used a beta-binomial model with symmetric smoothing parameter 0.1 for the generalized linear model. The value of the discrepancy for population k is the maximum 2 ln BF over all SNPs. For each population k, we sampled random phenotypes 10 times using that population as the risk factor and averaged the result of the discrepancy function over the 10 samples.

Diploid Data. Diploid data complicates some discrepancy functions, because, conditionally, each of the two copies of an allele may be assigned to different ancestral populations. As in the original specification of the admixture model, we divided single ternary observations $x\in\{0,1,2\}$ into the sum of two binary observations $x_1,x_2\in\{0,1\}$. Each binary observation has its own population assignment z_1,z_2. Two of our discrepancy functions (i.e., IBS, MI) compared pairs of SNPs and were modified to account for as many as two population assignments per SNP.

Let $z_{a1},z_{a2},z_{b1},z_{b2}$ be the first and second assignment variables at SNPs a and b, and let $x_{a1},x_{a2},x_{b1},x_{b2}$ be the alleles associated with those hidden variables. There are four possible comparisons, depending on how many of the population assignment variables are set to k:

1. No match. If one SNP is not assigned to k, there are no comparisons.
2. Single match. If exactly one allele in both SNPs is assigned to k, those alleles are comparable; e.g., if $z_{a1}=k$ and $z_{b2}=k$, but $z_{a2}\neq k$ and $z_{b1}\neq k$, then we compare x_{a1} with x_{b2}.
3. Single-double match. If one SNP has one allele assigned to k, and the other SNP has both alleles assigned to k, there are two comparisons; e.g., if $z_{a1}=k,z_{a2}=z_{b2}=k$, we compare x_{a1} with x_{b2} and x_{a2} with x_{b2}.
4. Double-double match. When all four alleles are assigned to k, there are four comparisons: x_{a1} with x_{b1}, x_{a1} with x_{b2}, x_{a2} with x_{b1}, x_{a2} with x_{b2}.

Posterior Predictive Checks. We generated replications of the observed data by sampling from the fitted model distributions:

$$x_{i,\ell,j}^{rep} \sim \mathcal{M}ult\left(\hat{\phi}_{\ell,\hat{z}_{i,\ell}}\right). \quad [4]$$

For a discrepancy function of alleles and their population assignment variables, $f(\boldsymbol{x}, \boldsymbol{z})$, we calculated $f(\boldsymbol{x}^1, \boldsymbol{z}), \ldots, f(\boldsymbol{x}^R, \boldsymbol{z})$, holding the latent variables fixed. For each of our PPCs, we set $R = 100$, except for the IBS discrepancy that considers every pair of individuals, where we used $R = 30$. We calculated the mean μ_k and SD σ_k of the replicated density over all R replications for each of K ancestral population. We then computed an empirical z score to compare the observed discrepancy $\boldsymbol{x}$ to the mean and SD of the replicated density: $\sigma_k^{-1}(f(\boldsymbol{x}, \boldsymbol{z}) - \mu_k)$. Given these population-specific z scores, we estimated the significance of their deviation from a standard normal distribution by computing a likelihood ratio (LR) as the ratio of density of the z scores under a normal distribution with maximum likelihood estimates of parameters over the likelihood of the z scores under the standard normal. We report this LR in figures—three stars for $2 \log LR > 10$, two stars for $2 \log LR > 6$, and one star for $2 \log LR > 2$—where a larger LR indicates greater distance from the standard normal for the z scores.

ACKNOWLEDGMENTS. The authors would like to acknowledge the groups that made these data publicly available: HapMap Consortium, GlaxoSmithKline, 1000 Genomes Project, and David Reich. B.E.E. was funded through National Institutes of Health (NIH) R00 HG006265 and NIH R01 MH101822.

1. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. *Am J Hum Genet* 67(1):170–181.
2. Price AL, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38(8):904–909.
3. Reich D, et al. (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89(4):516–528.
4. Moorjani P, et al. (2011) The history of African gene flow into southern Europeans, Levantines, and Jews. *PLoS Genet* 7(4):e1001373.
5. Wang S, Lachance J, Tishkoff SA, Hey J, Xing J (2013) Apparent variation in Neanderthal admixture among African populations is consistent with gene flow from Non-African populations. *Genome Biol Evol* 5(11):2075–2081.
6. Hellenthal G, et al. (2014) A genetic atlas of human admixture history. *Science* 343(6172):747–751.
7. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155(2):945–959.
8. Gilbert KJ, et al. (2012) Recommendations for utilizing and reporting population genetic analyses: The reproducibility of genetic clustering using the program STRUCTURE. *Mol Ecol* 21(20):4925–4930.
9. Box GEP (1980) Sampling and Bayes' inference in scientific modelling and robustness. *J R Stat Soc [Ser A]* 143:383–430.
10. Rubin D (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann Stat* 12:1151–1172.
11. Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80:267–278.
12. Gelman A, Meng X, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Stat Sin* 6:733–807.
13. Gelman A, Shalizi CR (2013) Philosophy and the practice of Bayesian statistics. *Br J Math Stat Psychol* 66(1):8–38.
14. Rosenberg NA, et al. (2002) Genetic structure of human populations. *Science* 298(5602):2381–2385.
15. Engelhardt BE, Stephens M (2010) Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* 6(9):e1001117.
16. Nelson MR, et al. (2008) The Population Reference Sample, POPRES: A resource for population, disease, and pharmacological genetics research. *Am J Hum Genet* 83(3):347–358.
17. Abecasis GR, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.
18. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461(7263):489–494.
19. Gelman A (2004) Exploratory data analysis for complex models. *J Comput Graph Stat* 13:755–779.
20. Altshuler DM, et al.; International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52–58.
21. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* 73(6):1402–1422.
22. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456(7218):98–101.
23. Bishop DT, Williamson JA (1990) The power of identity-by-state methods for linkage analysis. *Am J Hum Genet* 46(2):254–265.
24. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38(6):1358–1370.
25. Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: Limits and guidelines. *Am J Hum Genet* 55(4):809–824.
26. Conrad DF, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 38(11):1251–1260.
27. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genet* 9(10):e1003905.
28. Fledel-Alon A, et al. (2011) Variation in human recombination rates and its genetic determinants. *PLoS ONE* 6(6):e20321.
29. Slatkin M (2008) Linkage disequilibrium—Understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9(6):477–485.
30. Cover TM, Thomas JA (1991) *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, Hoboken, NJ), pp 576.
31. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437(7063):1299–1320.
32. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A (2003) Linkage disequilibrium patterns of the human genome across populations. *Hum Mol Genet* 12(7):771–776.
33. McEvoy BP, Powell JE, Goddard ME, Visscher PM (2011) Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* 21(6):821–829.
34. Kenny EE, et al. (2014) Inferring patterns of demography and assortative mating in the thousand genomes project admixed populations from the Americas. American Society of Human Genetics.
35. Hoggart CJ, et al. (2003) Control of confounding of genetic associations in stratified populations. *Am J Hum Genet* 72(6):1492–1504.
36. Satten GA, Flanders WD, Yang Q (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 68(2):466–477.
37. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55(4):997–1004.
38. Patterson N, et al. (2004) Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* 74(5):979–1000.
39. Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. *Nat Genet* 36(5):512–517.
40. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2(12):e190.
41. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 11(7):459–463.
42. Jorde LB, Wooding SP (2004) Genetic variation, classification and 'race'. *Nat Genet* 36(11, Suppl):S28–S33.
43. Serre D, Pääbo S (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Res* 14(9):1679–1685.
44. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164(4):1567–1587.
45. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145):661–678.
46. Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* 79(1):1–12.
47. Sankararaman S, Kimmel G, Halperin E, Jordan MI (2008) On the inference of ancestries in admixed populations. *Genome Res* 18(4):668–675.
48. Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genet* 8(1):e1002453.
49. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol Ecol* 14(8):2611–2620.
50. Huelsenbeck JP, Andolfatto P (2007) Inference of population structure under a Dirichlet process model. *Genetics* 175(4):1787–1802.
51. Teh YW, Jordan MI, Beal MJ, Blei DM (2006) Hierarchical Dirichlet processes. *J Am Stat Assoc* 101(476):1566–1581.
52. Blei D, Griffiths T, Jordan M, Tenenbaum J (2004) Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference* (MIT Press, Cambridge, MA).
53. Adams R, Ghahramani Z, Jordan M (2010) Tree-structured stick breaking for hierarchical data. *Advances in Neural Information Processing Systems 23*, eds Lafferty JD, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A (Curran Associates, Inc., San Francisco), pp 19–27.
54. Ardlie KG, Lunetta KL, Seielstad M (2002) Testing for population subdivision and association in four case-control studies. *Am J Hum Genet* 71(2):304–311.
55. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11):e1002967.
56. Gautier M, Vitalis R (2013) Inferring population histories using genome-wide allele frequency data. *Mol Biol Evol* 30(3):654–668.
57. Nelson MR, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337(6090):100–104.
58. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.
59. Rasmussen CE (2000) The infinite Gaussian mixture model. *In Advances in Neural Information Processing Systems 12*, eds Solla SA, Leen TK, Muller KR (MIT Press, Cambridge, MA), pp 554–560.
60. Yang WY, Novembre J, Eskin E, Halperin E (2012) A model-based approach for analysis of spatial structure in genetic data. *Nat Genet* 44(6):725–731.
61. Paisley J, Wang C, Blei DM (2012) The discrete infinite logistic normal distribution. *Bayesian Anal* 7:997–1034.
62. Kang HM, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42(4):348–354.
63. David H, Alexander JN, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655–1664.
64. Dempster A, Laird N, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc, B* 39:1–38.
65. Wright S (1969) *Evolution and the Genetics of Populations*, Vol. II. The Theory of Gene Frequencies (Univ of Chicago Press).
66. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 10(10):681–690.

Supporting Information

Mimno et al. 10.1073/pnas.1412301112

SI Results

In the main manuscript, we chose to present results for four different study data with five separate discrepancies and a single number of ancestral populations per study. The results from this analysis are simple to interpret and present an elegant motivating example for the use of PPCs for generative models. In reality, we fit a finite admixture model to one of four genomic datasets across a range of numbers of latent ancestral populations. Although these additional results include ideas that require additional explanation beyond the results presented in the paper for single values of K, we include them in the supplemental information to show a more thorough but complex analysis of model fit to genomic data.

Discrepancy in Genomic Similarity. When we apply the discrepancy for interindividual similarity to the four studies across multiple numbers of ancestral populations K, we observed a general decreasing trend in similarity as we increased K. Taken by itself, this trend might be seen as indicating that larger numbers of populations lead to overfitting, but when we compare the observed values to replications from the fitted model, we see the same trend. Similarities for replicated genomes are consistently lower, and more variable, as we increase K, indicating that lower similarity values are an expected feature of more fine-grained models. We can compare observed values to the distribution of replicated values using the z scores for each population, which are shown with a color scale. The POPRES data shows the strongest pattern, with similarities that are consistently lower than expected for $K = 3$ and 4, but values closer to the mean replicated similarity around $K = 7$.

Individuals in the POPRES, ASW, and Indian data are, on average, more similar to each other overall than individuals in HapMap, as we might expect considering the low relative heterozygosity in these regional studies relative to HapMap.

Across these studies, interindividual similarity tends to decrease as the number of ancestral populations increases. In the Indian data, the average similarity across individuals in all three populations in the $K = 3$ model is greater than the average similarity across individuals in all six populations at $K = 6$, suggesting that the estimates of the allele specific distributions associated with each ancestral populations have greater uncertainty (i.e., minor allele frequency [MAF] estimates closer to 0.5 than 0 or 1) as the number of populations increases.

HapMap z scores show that, regardless of the number of ancestral populations, the within-population variation is well captured by the admixture model. For POPRES, ASW, and Indian data, however, there is a preference for certain numbers of ancestral populations. In ASW and Indian data we found that two populations captured within-population variation well, and for ASW, Indian, and POPRES, larger numbers of populations (seven for POPRES) were necessary to capture within-population genetic variance.

Discrepancy in Background LD. When we apply the discrepancy for interindividual similarity to the four studies across multiple numbers of ancestral populations K, we found that fitting an admixture model with additional ancestral populations K generally increases the background LD observed within each population: for a fixed lag (say, 20 SNPs) there is generally an increase in the average MI within ancestral populations as the number of ancestral populations increases, indicating greater LD at farther distances as population structure is modeled at finer resolutions.

Next, we applied the LD discrepancy to our replicated data (Fig. S2). The z scores for the POPRES data deviate substantially from a standard normal for all lags at $K = 3$ indicating more observed background LD than expected, but are better captured by a standard normal for $K = 5$ at lags 25 and 30, and are below the replicated MI values at the same lags for $K = 8$, indicating less observed background LD than expected with respect to the sample replicates. Thus, despite having larger absolute MI values, admixture models with larger numbers of ancestral populations capture less background LD than expected for larger lags. This may be due to smaller population-specific sample sizes: as K increases, MI values on the observed data are estimated from fewer alleles, leading to greater variance.

Discrepancy in Reported Ancestry. We applied this F_{ST} discrepancy to the observed data from the four studies across different numbers of ancestral populations. We found that, in general, the average F_{ST} across the K ancestral populations increased as we increased the number of populations (Fig. S3), suggesting that, as we divide genomes more finely between larger numbers of ancestral populations, within each inferred population less information is captured about reported ancestries.

The naïve interpretation of the F_{ST} discrepancy applied to the observed data are that increasing the number of estimated populations leads to lower-quality models that fail to capture meaningful population structure, including structure information available in reported ancestries. Another interpretation is that the increase in F_{ST} as K increases is because the number of alleles from which variance is estimated shrinks as we partition genomes both by inferred ancestry and again by reported ancestry. Consider an extreme case where every individual within an estimated population k has the same reported ancestry A except for one with ancestry B. Because the allele frequencies for population k and ancestry B are estimated from only one genome, the variances are drastically underestimated.

We applied the F_{ST} discrepancy function to replicates from each of the fitted models, conditional on the inferred ancestry assignments for each SNP, to compute the z scores for this discrepancy. HapMap requires six or more ancestral populations to properly model the structure in the ancestry labels (there are 14). In particular, the z scores for the POPRES data are well captured by a standard normal across all numbers of inferred populations, in agreement with previous results that inferred admixture populations capture the same information as the 32 reported geographic labels at approximately $K \geq 4$ (1, 2). In contrast, the Indian data ancestry labels best capture the underlying heterogeneity in the data with two populations, which is believed to be the truth (3), but poorly fit a standard normal distribution for $K > 3$ despite the large number of ancestral assignments (there are 15). We hypothesize that the ancestry labels for the POPRES and Indian data do not reflect underlying population structure, but instead split each inferred population into partitions that do not have a strong genomic signature, but instead reflect geographic or cultural basis. On the HapMap data, the z scores indicate a poor model fit to the population labels for $K < 6$. As prior work suggests, the "best" number of ancestral populations in these data were six (4).

Discrepancy in Uncertainty in Ancestral Population Assignments. We applied the population assignment discrepancy to the observed

data from the four studies across different numbers of ancestral populations. We found that, in general, the average entropy across ancestral populations increased as we increased the number of ancestral populations (Fig. S4), illustrating that, as the number of ancestral populations grows, the uncertainty in the population assignments of alleles increases. This trend is stronger in the two studies that have poorly separated ancestral populations (POPRES, Indian).

The HapMap data are notable for this PPC: at every value of K there is at least one population with average entropy lower than expected by at least three SDs (z score < -3) (Fig. S4). These populations with greater certainty than expected are enriched for individuals with reported ancestry in South and Central America; for $K = 3$, this population with lower-than-expected entropy combines the Americas and East Asia. In absolute terms, these populations with greater than expected certainty have average entropy within range across k. It is only when we compared observed entropy to replication entropy that misspecification with respect to these populations emerged.

Discrepancy in Correcting for Population Structure in Genome-Wide Association Studies. We applied the association mapping discrepancy function to the observed data and found variable results across studies and numbers of ancestral populations (Fig. S5). We found that the maximum $\log_{10}$ BF across studies and K was small (<0.15) indicating that the model is effective at avoiding false positives. This maximum $\log_{10}$ BF tended to decrease as we increased the number of populations, indicating that overfitting the admixture model is advantageous when using the parameters for downstream structure corrections.

We then performed the PPC with this discrepancy function on replicated data, and, as for our four other discrepancy functions, we found that the PPC supports different conclusions than the observed discrepancy. The largest deviations from normality in z scores are at low numbers of populations for HapMap and POPRES, but these studies violate normality of z scores in opposite directions. In HapMap at $K = 3$, $\log_{10}$ BFs are significantly greater than expected under the model, showing that controlling for biased estimates of latent structure limits the ability to reject false positives in association testing. In POPRES at $K \leq 7$, not only are maximum $\log_{10}$ BFs small, they are significantly smaller than expected according to the replicated data.

Summarizing PPC Results Within Study. Our results across PPCs do not show a succinct picture of how admixture models are misspecified for genomic data, but instead tell a complex story for each study. Although these results are written for the case of multiple values of K, they are meant to supplement and detail the summarized results in the main manuscript.

HapMap phase 3. Across our application of PPCs to the HapMap phase 3 data, we found, not surprisingly, that there is substantial allelic heterogeneity within individuals in ancestral populations, illustrated in both interindividual PPC and the entropy PPC. Moreover, we found substantial variability in allelic heterogeneity across ancestral populations: admixture LD is badly misspecified in the admixture model for these data. These data did not show position specific background LD patterns we found in other studies, but background LD was also misspecified for these data across all tested values for K. The appropriate number of ancestral populations is in the range $K \geq 6$. Below this range the model does not fully account for information present in reported ancestries, and cannot effectively filter false-positive gene associations. For exploratory analyses relating to contrasting within- and across-population heterogeneity, these PPCs would suggest using more admixture models that capture background LD and admixture LD with $K \geq 6$ for these data, such as SABER (5).

European samples. Across our application of PPCs to the POPRES data, we found that there was strong allelic homogeneity among individuals within ancestral populations, and we also found that there were strikingly similar levels of interindividual homogeneity across populations from the interindividual PPC, background LD PPC, and entropy PPC. Admixture LD and background LD were misspecified in this application, but background LD was fit well when $K = 5$ and lag was greater than 15. This indicates a possible correction may be to subsample the genomic data every fifteenth SNP in the data, although this would remove a large number of SNPs that may be essential to discriminating these relatively similar ancestral populations. It appears, across PPCs, that a good range of K is around $4 \leq K \leq 7$; when correcting for population structure, fitted parameters seem to perform well when $K = 7$ for these data. For exploratory analyses related to European population substructure, these PPCs would suggest using admixture models that capture admixture LD, such as the Structure 2.0 method (6); another indication would be to model latent structure with a continuous population model (e.g., principal components based).

African Americans. Across our application of PPCs to the ASW data, we found strong allelic homogeneity within some ancestral populations, and a large variance across populations for within-population homogeneity. Unlike the first two applications, most of the PPCs appear well fit for $K = 4$ for many downstream analyses; the one exception is for the background LD, where all models appeared misspecified across most lags in SNP adjacency. For these data, it may be useful to include a more descriptive model of background LD (5), which may change the appropriate number of ancestral populations (7). Nonetheless, this application, with recent admixture between two well-separated ancestral populations, appears well suited for analysis using the admixture model.

Continental Indians. Across our application of PPCs to the Indian data, we found substantial allelic homogeneity among individuals within a population, and little variation among the estimated ancestral populations in the homogeneity of the individuals, although there is more variation in this application than in the POPRES application. Because we see so little variation among estimated ancestral population, but we believe the true ancestral populations may be well separated, it is possible that the more ancient admixture events and the absence of an individual with ancestry in only one of the ancestral populations imply poor estimation of the ancestral populations. Despite this, it appears that, across the PPCs, $K = 2$ with a base-pair lag of greater than 10 is well-specified across many downstream analyses.

1. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456(7218): 98–101.
2. Engelhardt BE, Stephens M (2010) Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genet* 6(9):e1001117.
3. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461(7263):489–494.
4. Rosenberg NA, et al. (2002) Genetic structure of human populations. *Science* 298(5602): 2381–2385.
5. Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet* 79(1):1–12.
6. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164(4):1567–1587.
7. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447(7145): 661–678.

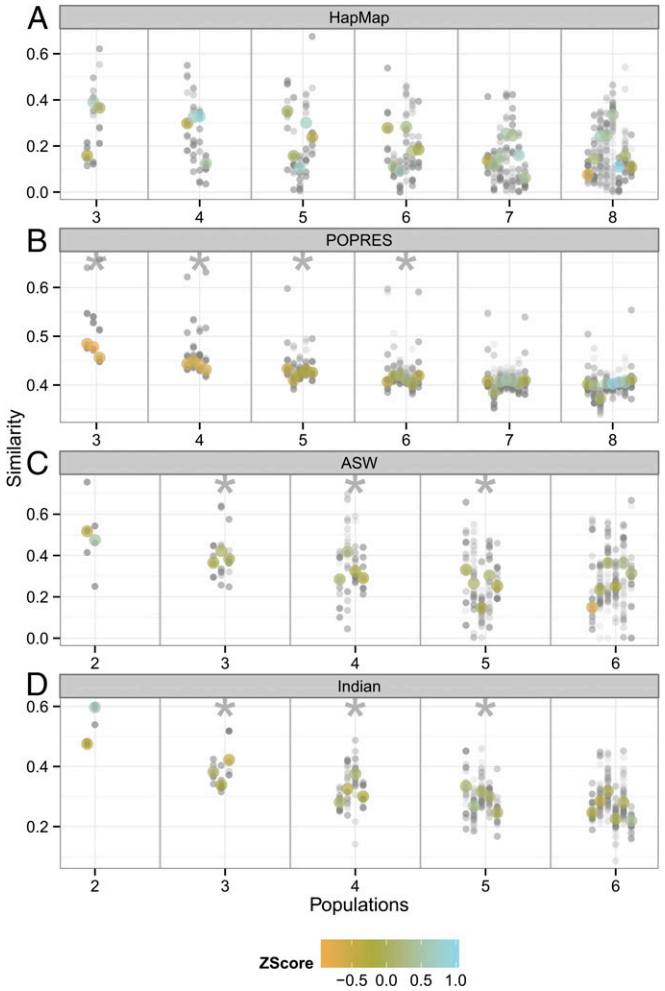

Fig. S1. Average population-specific interindividual similarity, varying *K*, across four studies. Each *x* axis represents the number of ancestral populations in the fitted admixture model; panels represent application to HapMap, POPRES, ASW, and Indian studies, respectively. The *y* axis represents the mean interindividual similarity across individuals conditioned on each ancestral population. Small semitransparent points represent values from replicated data. Larger points represent values from real data, colored by their *z* scores relative to the empirical distributions of the replicated values. Stars indicate significant divergence in *z* scores from a standard normal. (*A*) HapMap data; (*B*) POPRES data; (*C*) ASW data; and (*D*) Indian data, with $K = 2,3,4,5,6$.

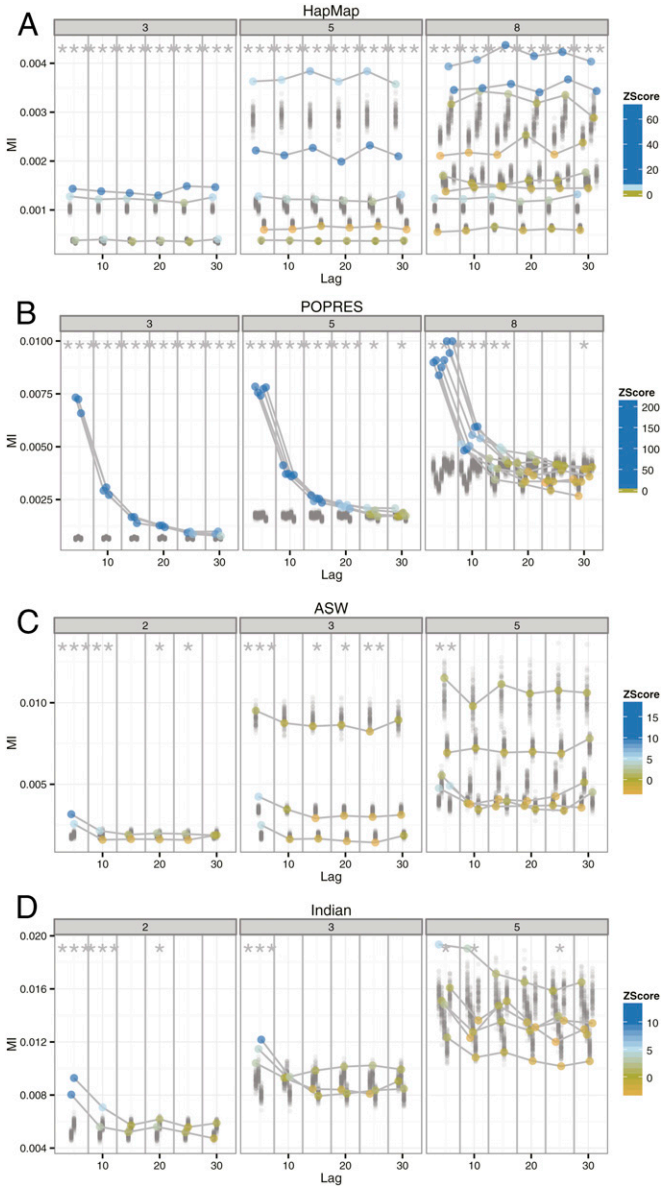

Fig. S2. LD discrepancy function PPC, varying *K*, across four studies. Each *x* axis indicates the number of lag SNPs between pairs of SNPs over which mutual information is averaged. The *y* axis represents the average mutual information between SNPs at varying lags. Each panel shows different numbers of ancestral populations in the fitted admixture models. Observed values for each population are connected by lines across lags and colored by *z* score. Stars indicate significant divergence in *z* scores from a standard normal. (*A*) HapMap data, with $K = 3,5,8$; (*B*) POPRES data, with $K = 3,5,8$; (*C*) ASW data, with $K = 2,3,5$; and (*D*) Indian data, with $K = 2,3,5$.

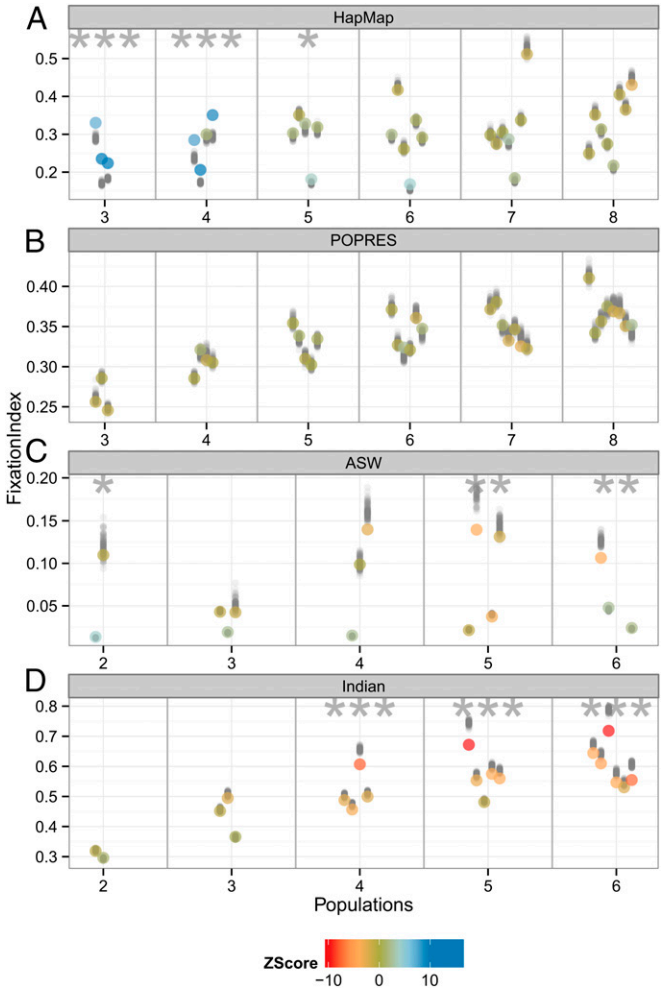

Fig. S3. F_{ST} discrepancy function, varying K, across four studies. The x axis represents the number of ancestral populations K in the fitted admixture model; columns represent application to HapMap, POPRES, ASW, and Indian studies, respectively. The y axis represents the F_{ST} value of a single ancestral population k with respect to the self-reported ancestry information. Smaller values indicate a closer match between the reported and estimated ancestral populations. POPRES values are within the expected range for all K, indicating that the increasing observed values are an artifact of increasing subdivision of the data. For HapMap $K = 6$ is the smallest model for which reported ancestries are no longer informative. With the smaller datasets (ASW, Indian), models with more populations show consistently lower association between alleles and reported labels than expected. (*A*) HapMap data; (*B*) POPRES data; (*C*) ASW data; and (*D*) Indian data, with $K = 2,3,4,5,6$.

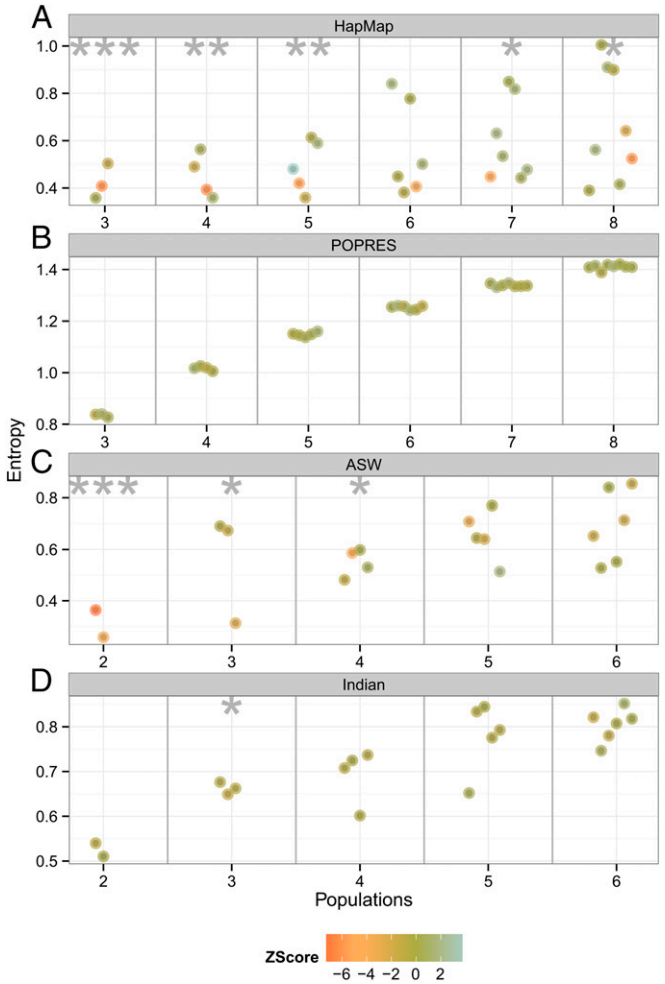

Fig. S4. Average entropy discrepancy function, varying *K*, across four studies. The *x* axis represents the number of ancestral populations in the fitted admixture model; panels represent application to HapMap, POPRES, ASW, and Indian studies. The *y* axis represents the average entropy of the posterior distribution over populations for each allele. Larger absolute value entropy represents greater uncertainty over population assignments. Replicated values are shown as smaller gray circles. (*A*) HapMap data; (*B*) POPRES data; (*C*) ASW data; and (*D*) Indian data, with $K = 2,3,4,5,6$.

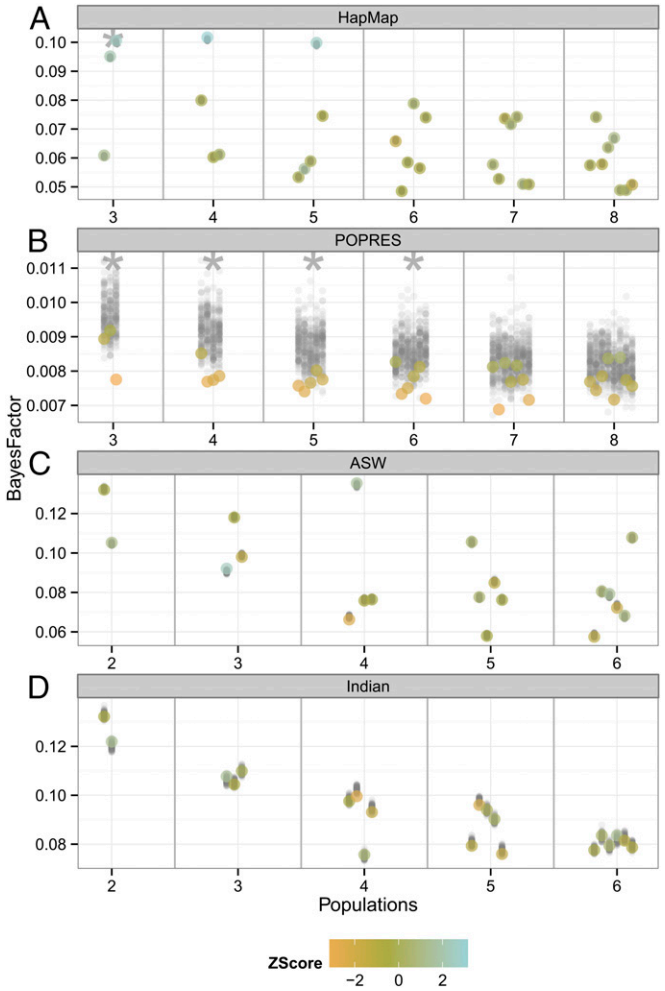

Fig. S5. Association mapping correction PPC, varying *K*, across four studies. The *x* axis represents the number of ancestral populations in the fitted admixture model; panels represent application to HapMap, POPRES, ASW, and Indian studies, respectively. The *y* axis represents maximum $\log_{10}$ BFs. Maximum $\log_{10}$ BFs from replicated genomes are shown with smaller gray points. Colors of the observed data discrepancy values represent the *z* score with respect to the replicated samples from fitted admixture models. Stars indicate deviation from normality in *z* scores, suggesting that larger numbers of populations are more effective in controlling for latent population structure in POPRES and, to a lesser degree, HapMap. (*A*) HapMap data; (*B*) POPRES data; (*C*) ASW data; and (*D*) Indian data, with $K = 2,3,4,5,6$.