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.

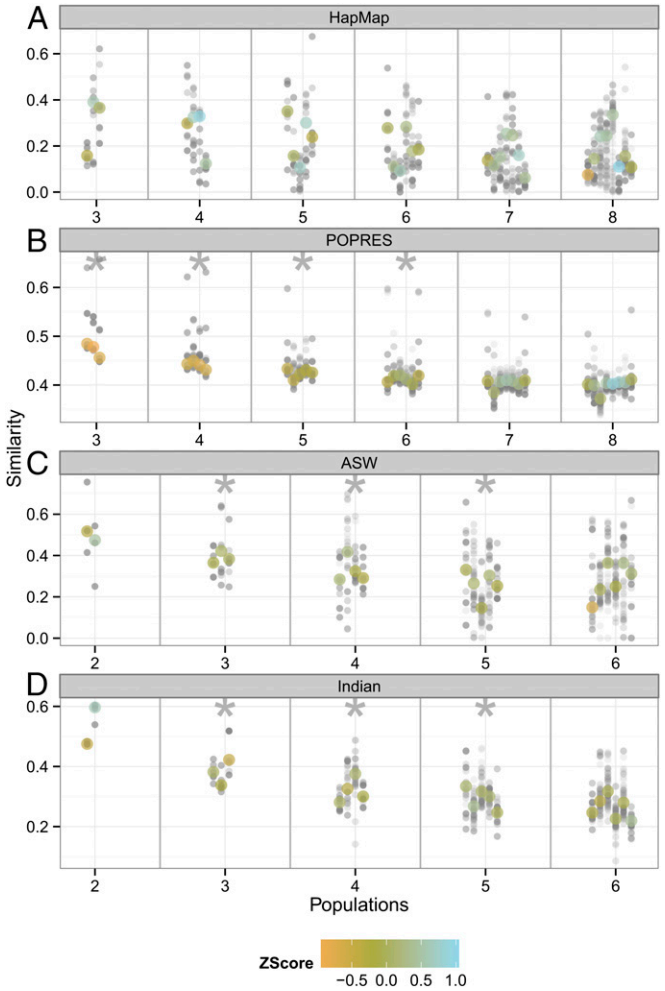

Fig. S1. Average population-specific interindividual similarity, varying *K*, across four studies. Each *x* axis represents the number of ancestral populations in the fitted admixture model; panels represent application to HapMap, POPRES, ASW, and Indian studies, respectively. The *y* axis represents 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 $K = 2,3,4,5,6$.

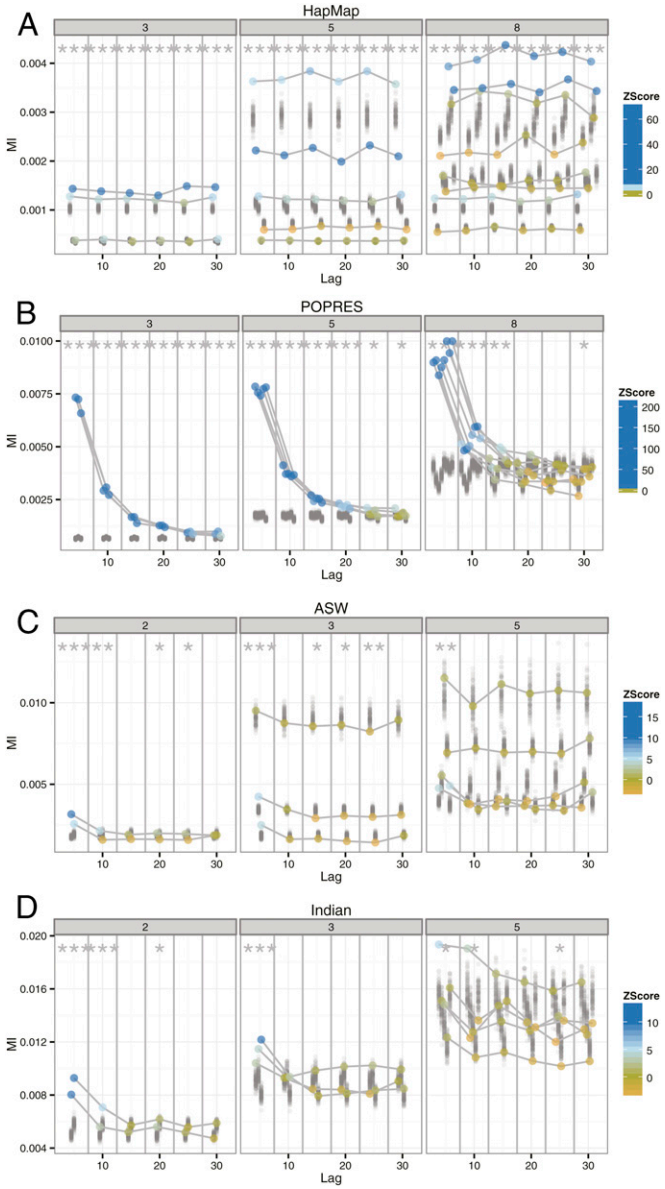

Fig. S2. LD discrepancy function PPC, varying *K*, across four studies. Each *x* axis indicates the number of lag SNPs between pairs of SNPs over which mutual information is averaged. The *y* axis represents the average mutual information between SNPs at varying lags. Each panel shows different numbers 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 $K = 2,3,5$.

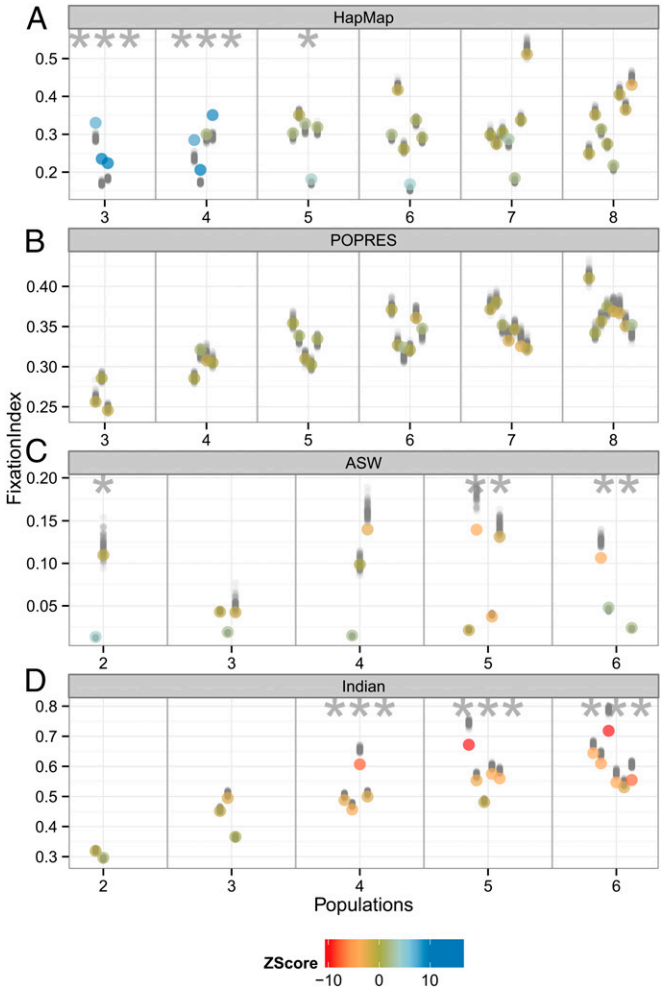

Fig. S3. F_{ST} discrepancy function, varying K, across four studies. The x axis represents the number of ancestral populations K in the fitted admixture model; columns represent application to HapMap, POPRES, ASW, and Indian studies, respectively. The y axis represents the F_{ST} value of a single ancestral 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 $K = 2,3,4,5,6$.

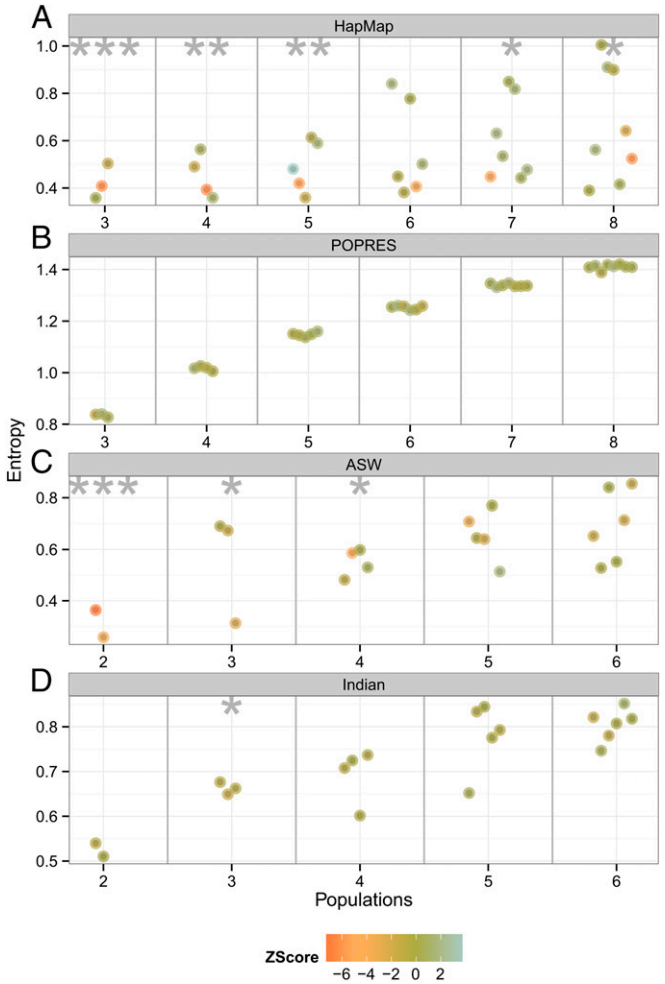

Fig. S4. Average entropy discrepancy function, varying *K*, across four studies. The *x* axis represents the number of ancestral populations in the fitted admixture model; panels represent application to HapMap, POPRES, ASW, and Indian studies. The *y* axis represents the average entropy of the posterior 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 $K = 2,3,4,5,6$.

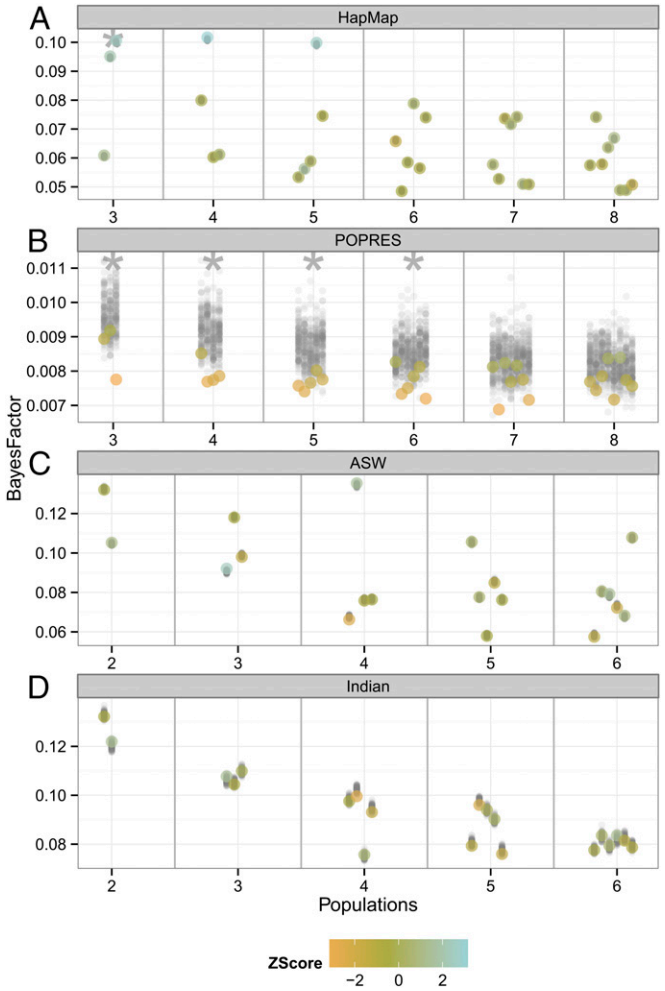

Fig. S5. Association mapping correction PPC, varying *K*, across four studies. The *x* axis represents the number of ancestral populations in the fitted admixture model; panels represent application to HapMap, POPRES, ASW, and Indian studies, respectively. The *y* axis represents maximum $\log_{10}$ 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 $K = 2,3,4,5,6$.