Strong selection at MHC in Mexicans since admixture

Quan Zhou\textsuperscript{1,2,3}, Liang Zhao\textsuperscript{1,2}, Yongtao Guan\textsuperscript{1,2,3,4,5}

July 8, 2015

1 USDA/ARS Children’s Nutrition Research Center
2 Department of Pediatrics, Baylor College of Medicine
3 Program of Structure and Computational Biology and Molecular Biophysics, Baylor College of Medicine
4 Department of Molecular and Human Genetics, Baylor College of Medicine
5 Corresponding author. Address: 1100 Bates, Room 2070, Houston TX 77030.
   Tel: 713-798-0362 Email: \texttt{yongtaog@bcm.edu}

Running head: Strong selection at MHC in Mexicans
Keywords: Admixture, ELAI, Local ancestry, Mexicans, MHC, Selection

Abstract

Mexicans are a recent admixture of Amerindians, Europeans, and Africans. We performed local ancestry analysis of Mexican samples from two genome-wide association studies obtained from dbGaP, and discovered that at the MHC region Mexicans have excessive African ancestral alleles compared to the rest of the genome, which is the hallmark of recent selection for admixed samples. The estimated selection coefficients are 0.05 and 0.07 for two datasets, which put our finding among the strongest known selections observed in humans, namely, lactase selection in northern Europeans and sickle-cell trait in Africans. Inaccurate Amerindian training samples were a major concern for the credibility of previously reported selection signals in Latinos. Taking advantage of the flexibility of our statistical model, we devised a model fitting technique that can learn Amerindian ancestral haplotype from the admixed samples, which allows us to infer local ancestries for Mexicans using only European and African training samples. The strong selection signal at the MHC remains without Amerindian training samples. Finally, we note that medical history studies suggest such a strong selection at MHC is plausible in Mexicans.
1 Introduction

In 1492 Columbus discovered America. Europeans, led by the Spaniards, and armed with horses, wheels, germs, and steel, rapidly conquered the New World [1], and promptly Africans were brought there as slave labor. During the past 500 or so years, three populations — Amerindians, Europeans, and Africans — have occupied the same space and time, albeit asymmetrically, and were genetically admixing. Twenty generations later, the majority of the people inhabiting Central America, Caribbean Islands, and South America, such as Mexicans, Puerto Ricans, and Columbians have become an admixture of the three continental ancestral populations. These recently admixed populations are of great interest for modern genetic studies [2].

In 2007, Tang and colleagues analyzed a small cohort of Puerto Rican samples and reported three regions that are under strong recent selection [3]. Using their then state-of-the-art local ancestry inference software Saber [4], Tang and colleagues discovered in Puerto Rican samples genomic regions whose mean local ancestries (averaged over individuals) significantly deviated from the genome-wide average — a hallmark of recent selection for admixed samples. Price and colleagues cautioned that the strong selection discovered by Tang and colleagues might be artifacts and they provided three arguments [5]. First, Saber only models linkage disequilibrium (LD), the non-independence of genetic markers in a population, between adjacently markers and thus may produce unreliable local ancestry estimates in regions that harbor long-range LD. It was noted that all three loci under selection that Tang and colleagues reported are within the long-range LD regions. Second, the Amerindian training samples used by Tang and colleagues, which are Maya and Pima samples from human genetic diversity panel (HGDP) [6], is an inaccurate ancestral population for Puerto Ricans, which might produce unreliable local ancestry inference. Third, Price and colleagues analyzed a larger sample using their software AncestryMap [7] and did not discover the deviation of local ancestry reported by Tang and colleagues.

This paper does not aim to resolve the dispute between two parties on the issue — we do not yet have access to the genetic data of Puerto Ricans beyond the 55 individuals in the 1000 Genomes project [8]. Nevertheless, we would like to make following comments. First, the AncestryMap uses the so called ancestry informative markers (AIMs) to infer local ancestry; because that AIMs are sparse and that ancestry informative haplotypes may not contain sufficient amount of AIMs, the statistical method underlying AncestryMap is evidently under-powered in detecting local ancestry compared to those that attempt to model haplotypes, particularly more recent model-based method such as HapMix [9] and ELAI [10]. Therefore, negative results from AncestryMap cannot convincingly refute positive findings by Tang and colleagues. Second, the other two concerns raised by Price and colleagues — long-range LD and inaccurate Amerindian training samples — are potential sources of false positives. We would like to note, however, that ELAI models the full-range LD — both short and long — and the long-range LD does not undermine the local ancestry inference but
improve it. In addition, ELAI has the capacity to use only European and African training samples — discarding the Amerindian training samples — to infer local ancestry for Mexicans (more on this point later).

Recently Bhatia and colleagues analyzed a large number of African American samples and reported that no selection was detected using local ancestry inference [11]. Although they only analyzed African American samples, in an effort to generalize their findings, in the discussion they questioned the plausibility of positive findings in Mexicans reported by ELAI [10], which discovered signature of recent selection in HapMap3 [12] Mexican samples based on a departure of local ancestry from the global average. On a side note, reviewers of the ELAI paper asked the author to tone down the claim about the selection findings in the ELAI paper [10], and the author abided and cautioned readers to wait for confirmation from analysis of larger datasets. Under these circumstances, we analyzed local ancestry of two genome-wide association study (GWAS) datasets whose subjects are of Mexican descent, which we obtained from the database of genotype and phenotype (dbGaP), and confirmed the strong selection in the MHC region in Mexicans.

2 Results

We applied for access and downloaded two GWAS datasets of Mexican descent from the dbGaP. One is the Viva La Familial obesity-diabetes familial risk study (henceforth Viva) that contain 815 individuals from 263 families [13]. The other is the Mexican hypertriglyceridemia study (henceforth Lipid) that contains 2229 unrelated individuals [14]. After stringent QC (see Methods), we applied ELAI to infer local ancestry of each individual. ELAI outperforms other competing state-of-the-art methods in local ancestry inference [10]. It is also convenient to use as it does not require phasing for either training samples or cohort samples, nor does it require recombination map or global admixture proportions as inputs. Public resources, such as HapMap and 1000 Genomes projects, contain high quality European and African haplotypes, but not Amerindian haplotypes; this makes ELAI even more attractive than others in analyzing Mexican samples.

2.1 Patterns of global and local ancestry

In VIVA the global ancestry proportions (that is, the admixture proportions) for Amerindian, European, and African components are 0.484, 0.452, and 0.064 respectively. In Lipid the numbers are 0.552, 0.409, and 0.039. Compared to Viva, Lipid has a higher Amerindian ancestry proportion and lower European and African ancestry proportions. The sampling location is likely to account for the difference: participants in Lipid were recruited in Mexico City, Mexico, whereas participants in Viva were recruited in Houston, Texas. For each ancestry component, there are substantial variations among individuals (see two triangular plots in Figure 1). For both datasets, the topological resemblance between the triangular plot and the principal component (PC) plot is remarkable. The
relative positions of the Mexican outlier individuals are well matched, and an African American individual accidentally recruited in Viva is rather obvious. This suggests that ELAI estimates are sensible, and that using PC to derive admixture proportions has some merits [15]. There is a folklore that including East Asians as additional Amerindian training samples might improve ancestry inference of Latinos. Our experience suggests that this practice has little impact, and the PC plots, in which Chinese separate from Amerindians inconsistently in two datasets, seem to corroborate our experience.

We computed at each marker the average dosages separately for each ancestral component by averaging that component over all individuals. Figure 2 shows variation of African average dosages along each autosome. (Figure S1 has averages dosages for all ancestries.) The spikes on chromosome 6 in both datasets are rather striking. For Viva, the sample standard deviation (ssd) of average dosages for Amerindian, European, and African components are 0.046, 0.043, and 0.024 respectively. The largest deviations, measured by the ssd of average dosages for each ancestry, are 5.4, 4.8, and 9.9. The locus whose African average dosage is 9.9 ssd above the mean is inside the MHC region, and under the normal approximation, a 9.9 ssd corresponds to a p-value of $2 \times 10^{-23}$, which surpasses any reasonable significant threshold for a genome-wide analysis (in GWAS such a significant threshold is $5 \times 10^{-8}$). The same region inside MHC was again identified as significant in Lipid; the largest deviation of African average dosages is 14.8 ssd above the mean, which corresponds to a p-value of $3 \times 10^{-49}$. The region identified in MHC is the same region identified by analyzing HapMap3 Mexican samples [10]. In that study, a region on chromosome 8 was also identified as border-line significant in Amerindian average dosages. In both Viva and Lipid, however, this region is not replicated.

2.2 Effects of different European and African training samples

We used HapMap3 Utah Residents with Northern and Western European Ancestry (CEU) as European training samples; Yoruba in Ibadan, Nigeria (YRI), from west Africa, as African training samples; and Maya and Pima from HGDP [6] (MAYA) as Amerindian training samples. To test the robustness of our results against different choices of training samples, we first investigated European and African training samples as they both have alternative choices in HapMap3. We used Tuscani in Italia (TSI), from south Europe, as an alternative to CEU, and Maasai in Kinyawa, Kenya (M KK), from east Africa, as an alternative to YRI, and these produced four combinations: CEU–YRI–MAYA, CEU–M KK–MAYA, TSI–YRI–MAYA, and TSI–M KK–MAYA. We also combined all training samples to perform inference (CEU+TSI–YRI+MKK–MAYA), and together we had five comparisons shown in Figure 3(a) and (b). We made the following observations: 1) Using TSI to replace CEU produced a less significant deviation at the MHC region. 2) Using MKK to replace YRI produced a more significant deviation at MHC. 3) Combining all training samples produced a significant deviation at MHC, and the significant level is intermediate among other combinations.
4) Outside the MHC region, different combinations of training samples produced congruent results.

Figure 3(c) shows the difference in inferred European average dosages between two European training samples (average difference between TSI-YRI-MAYA vs CEU-YRI-MAYA and TSI-MKK-MAYA vs CEU-MKK-MAYA). Interestingly, the highest peak contains HLA-B and HLA-C loci. We naturally suspect that TSI has more genetic diversity than CEU at the MHC, because more genetically diverse European training samples tend to produce higher estimates of European ancestry dosages. Amerindian average dosages are congruent between choices of CEU and TSI training samples (Figure S2), and the deficiency in African average dosages when using TSI as training samples are compensated by sufficiency of European average dosages. We extracted 8679 SNPs in the extended MHC region, 25–35Mb on chromosome 6, from European and African training samples, and ran ELAI using two upper clusters without specifying the population label, which is essentially haplotype-based structure analysis [10]. One admixture component was arbitrarily chosen to make comparison, and the admixture component was averaged over 10 EM runs (after adjusting for label-switching across EM runs). The violin plots in Figure 3(d) show that TSI is indeed more diverse than CEU at the MHC, MKK is more diverse than YRI, and MKK is the most diverse among four non-admixed populations, which agrees with the theory of east African origin of modern humans [16]. Recently admixed African Americans (ASW) were included for sanity check of the haplotype-based structure inference.

2.3 Effects of with and without Amerindian training samples

Next we turn to Amerindian training samples. The 1000 Genomes admixture analysis group used a collection of Amerindian samples [17] different from the Maya and Pima from HGDP that we used, but we had difficulty in obtaining that data. Moreover, a practical concern is that any specific choice of Amerindian training samples will be subject to suspicion of inaccuracy. To test the robustness of our inference against different Amerindian training samples, we elected to remove Amerindian training samples and used only European and African training samples to perform inference — but of course we kept the setting of three ancestral populations. ELAI can function with the absence of one training population as long as there are enough genetic components of that ancestry in the cohort samples. Because Mexicans have a large Amerindian ancestry proportion, when Amerindian training samples are missing, ELAI is still able to learn Amerindian ancestral haplotypes relatively easily from Mexican samples as long as the sample size is large. The same is true for European training samples, but it becomes more difficult if African training samples are missing. To borrow an analogy from next-generation sequencing, a large number of Mexican samples and a high ancestry proportion to local ancestry inference is analogous to a high coverage of sequencing reads to variant call.

The recommended practice in an early version of ELAI is to split a large dataset into small subsets. Doing so not only improves computational efficiency on a computer cluster, but also allows
ELAI to jointly fit training and cohort datasets. It is evident — at least among specialists — that a cluster model becomes less fit to the training samples in the presence of an overwhelmingly large number of cohort samples, which undermines the performance of local ancestry inference (or imputation). Recall that removing Amerindian training samples requires a large number of cohort samples jointly fitting the model with training samples — we are seemingly in a quandary. The solution is rather simple. During parameter estimation of the two-layer model underlying ELAI [10], we can arbitrarily adjust relative weights between cohort and training samples without changing the expected ancestral allele (haplotype) frequency estimates. In other words, we can take an arbitrarily large number of cohort samples and down weight their contribution to parameter estimation. When the training samples are available, the weighting ensures the model fits to training samples sufficiently; otherwise, the ancestral alleles are estimated exclusively by cohort samples, and the weight cancels out in the parameter estimation as long as we assign equal weight to all cohort samples. Thus, the weighting allows us to take the extreme measure of removing Amerindian training samples.

We implemented the weighting scheme and applied it to both datasets. We combined CEU and TSI as European training samples and YRI and MKK as African training samples. Figure 4 demonstrates, using both Viva and Lipid datasets, the difference, or lack of it, in the estimated African average dosages with and without Amerindian training samples. Comparing the Amerindian average dosages, however, the estimates without Amerindian training samples are higher than with. The mean differences are 0.09 for Viva and 0.08 for Lipid. This is not too surprising considering 1) Maya and Pima samples have some European ancestral components (PC plots in Figure 1); and 2) Maya and Pima samples may be imperfect representatives of the Amerindian source populations for Mexicans, and learning Amerindian ancestry components from a large number of cohort samples may provide a better fit. Our results shall eliminate concerns of possible artifacts caused by inaccurate Amerindian training samples.

2.4 Strong selection at the MHC region

If purely by chance, it is very unlikely that Amerindians share more alleles with Africans at MHC than the rest of the genome at such a significant level; that the pathogens from the Old world are often lethal to the native inhabitants of the New World seems to argue against such a peculiar sharing. The effect of the population bottleneck and the drift do not distinguish the MHC from the rest of the genome [18]. If selection happened in Africans before admixture, one would expect to see such selection signals in African Americans, which are not there [11]. Therefore, it is safe to assume that the African average dosages in Mexicans rose from the genome-wide mean \( p_0 \) to the inferred value of \( p_1 \) at MHC in the past 20 generations, and it is selection at work. A selection coefficient \( s \) can be computed via a formula \( p_0 \times (1 + s)^{20} = p_1 \). For Viva \( s = 0.05 \) and for Lipid \( s = 0.07 \). Both estimates indicate very strong selection, on par with lactase selection in northern
To understand how many SNPs have contributed to the selection signal in MHC, we assigned a phenotypic value to each individual based on their African ancestry dosage at the identified region in MHC (detailed in Methods), regressed out six leading principal component and admixture proportions, and performed the single-SNP association test using BIMBAM \[21\]. At a very liberal threshold of \(\log_{10}\) Bayes factor > 10, we discovered 1700 SNPs in the extended MHC region to be genome-wide significant (Figure S3). Considering the high correlation among SNPs in the region, we next performed multi-SNP analysis using a Bayesian variable selection regression procedure implemented in the software pi MASS \[22\]. pi MASS implements a Markov chain Monte Carlo (MCMC) procedure to sample the posterior distribution of model space (SNP sets) under sparse and shrinkage priors. The output contains posterior probability of association (PPA) for each SNP, which roughly reflects how often the SNP is being selected in an additive model. We ran pi MASS using all markers from chromosome 6 of Lipid with 10,000 burn-in steps and 1 million sampling steps. Two independent runs were conducted. In both runs, the proportion of variation explained (the narrow sense heritability) estimates had the same posterior mean of 0.88, with ssd of 0.015 and 0.017 respectively. The posterior mean model sizes (the number of SNPs in the model sampled) were 93 ± 10.7 and 83 ± 7.1 respectively (mean ± ssd). The two runs had 126 and 116 SNPs with PPA > 0.1; among them, 60 SNPs overlapped, and the union contained 182 SNPs. We removed these 182 SNPs and reran local ancestry inference of chromosome 6. The pattern of the local ancestry was essentially unaffected. These exercises shall eliminate concerns that a few questionable SNPs drive the selection signal.

3 Discussion

In this paper we analyzed two existing GWAS datasets of Mexican subjects and demonstrated that the MHC region is under strong recent selection in Mexicans. Because Viva contains related individuals, we split individuals into non-overlapping subsets, each containing 40-50 unrelated individuals; performed local ancestry inference separately for each subset; and aggregated the inferences. This practice produced congruent results as our combined analysis. In Lipid, samples were assigned case-control labels according to their triglyceride levels. The results presented in the paper ignored the case-control status. We analyzed cases and controls separately, and the results were highly congruent to that of the combined analysis. A few gene regions outside of MHC, however, showed significant differences in Amerindian average dosages between cases and controls (the largest difference is 0.17). Considering that Mexicans have a higher prevalence of dyslipidemia \[23\], these regions may harbor candidate genes responsible for high triglyceride \[24\]. We may report relevant findings elsewhere. We also analyzed African American samples in HapMap3 and did not find any region under selection, which agrees with a recent study \[11\]. This serves as a negative control.
for ELAI. We devised a model fitting technique to introduce weighting into parameter estimation, which makes it possible to infer local ancestry of Mexicans using only European and African training samples. This rids the concern that the detected selection signals in Mexicans are artifacts produced by inaccurate Amerindian training samples.

Bhatia and colleagues made a rather plain argument which essentially states “if it’s there, why haven’t others found it already?” [11] We nevertheless took this opportunity to address this argument using simulations. We simulated genotypes using a demographic model that mimic the out-of-Africa migration events [25], performed forward simulations to mimic admixture and selection at three linked loci (details in Methods), and then inferred local ancestry using other competing methods as well as ELAI. The 1000 Genomes used consensus call from four programs: HapMix [9], LAMP-LD [26], RFMix [27], and MultiMix [28]. The publicly available version of HapMix was designed exclusively for two-way admixture, and the extended version used to analyzing the 1000 Genomes data was not available to us [8]. Thus it was excluded from our analysis. MultiMix performed poorly in our easiest scenario compared to LAMP-LD and RFMix and was excluded from the comparison as well. Both LAMP-LD and RFMix require phased training samples, and RFMix requires phased cohort samples as well. We first supplied true phasing (both training and cohort samples) to RFMix and LAMP-LD. Then, in an attempt to investigate how phasing accuracy affects their local ancestry inference, we introduced 2% switch-errors into cohort haplotypes and training haplotypes that mimic Amerindians, 1% switch-errors into European and African training samples, and redid the analysis. ELAI ignored the phasing. For both LAMP-LD and RFMix we used the same parameter settings as those used in the 1000 Genomes admixture analysis group [8, Supplementary p.68]. Using true phasing, both LAMP-LD and RFMix performs on par with ELAI; RFMix under-performs when switch-errors are introduced (Figure S4). It is worthwhile to note that MHC is notoriously hard to phase, and phasing for admixed samples at MHC is even more challenging as it requires the phasing algorithm to correctly identify local ancestry — a catch-22 for RFMix. We were surprised at the worse-than-the-expected performance of RFMix in the presence of switch-errors (Figure S5), and after further investigation, we discovered that its window size parameter has a sweet-spot (Figure S6), which is smaller than 0.1 centi-Morgan (cM) or roughly 100 SNPs, a window size used by the 1000 Genomes analysis group, and is much smaller than 0.2 cM, a suggested default value. When using the best window size RFMix performed on par with ELAI (Figure S7). LAMP-LD is robust to switch-errors and performs as well as ELAI. But previously we identified that LAMP-LD tends to make more mistakes in small ancestry track lengths in a simulation study where admixed samples were created from HapMap haplotypes [10]. Going back to “if it’s there, why haven’t others found it already?”, in addition to the obvious “ELAI is more effective,” our simulation studies suggest that 1) the democratic strategy adopted by the 1000 Genomes admixture analysis group, which used consensus calls from four methods to identify local ancestry, is perhaps not optimal; and 2) parameter settings in individual software
A natural question is “what genes or alleles are under selection?” Answering this satisfactorily requires sequencing the MHC region of Mexican samples and training samples. We only have limited data on the HLA allele for Mexicans and Africans, little for Amerindians, but plenty for Europeans. Recall that, everything else being equal, substituting CEU with TSI produced reduced estimates of African average dosages at MHC in Mexicans, this reduction was compensated for by increased estimates of European average dosages (Figure 3(b)), and Amerindian average dosages remain largely unchanged at MHC (Figure S2). Therefore, by studying the allele frequencies of CEU and TSI and comparing them against allele frequencies of Mexicans, we gleaned some insights regarding which alleles have been under selection. We imputed allele types for each CEU and TSI sample at HLA-B and HLA-C using HLA*IMP [29]. These two loci were chosen because they are located at places where exchanging CEU and TSI produced the largest differences in estimated European average dosages (Figure 3(c)) and because their Mexican allele frequencies are readily available [30]. We identified alleles whose frequencies differ by at least 0.05 between CEU and TSI, filtered out those whose frequencies in Mexicans are less than 0.02, and retained those whose frequencies are closer to those of TSI than CEU. These alleles are candidate selection targets because they are more likely explained by TSI, which, compared to CEU, increases Europeans ancestry estimates and decreases African ancestry estimates in Mexicans. We identified 3 such HLA alleles (Table S1). In the future we may obtain allele frequencies for Amerindians and examine whether these alleles are genuine selection targets.

The MHC region influences susceptibility and resistance to a broad range of infectious agents such as viruses, bacteria, and parasites. It is sensible to observe more alleles of African ancestry at MHC in Mexicans if those alleles confer selective advantages in the presence of certain infectious agents. The European conquerors brought to America European and African diseases such as smallpox, measles, and typhus. Spaniards imposed an urbanized life style and farming practice on native people. A sudden increase in local population concentration, displacement, social upheaval, food shortages, and stress made them much vulnerable to infectious diseases. An estimated 5–8 million native people perished in a smallpox epidemic alone in early 1500s [31]. Nevertheless, after “difficult struggles of the formative period,” the acceptance and enthusiasm of the new life emerged from the persistence of the old; for a brief period a “fusion of European and Mesoamerican cultures seemed ready to emerge” [32]. But severe drought hit and lethal pandemic broke out [31, 32]. The epidemic, called “huey cocoliztli,” was symptomatically different from those imported from the Old World; some medical historians suspect it was a hemorrhagic fever caused by arenavirus carried by rodents [32]. It first broke out in 1545 and lingered until 1815 [32, 33]. The epidemic selectively targeted native people, and 90% of the population perished in a few generations [31, 33]. This sustained epidemic harbors plenty of opportunities for strong selection at MHC, which fits our analysis. Once again history left its mark in genomes for posterity [34].
4 Methods

4.1 Datasets

The first dataset, Viva La Familia obesity-diabetes familial risk study (dbGaP Study Accession: phs000616.v1.p1), contains 858 genotyped individuals [13]. Among them, 815 Mexicans children from 263 families were genotyped with Illumina HumanOmni 1-v1.0 BeadChips, and the remaining 43 children were genotyped on HumanOmni 2.5-8v1 BeadChips. We chose to analyze the 815 samples that were typed on the same chip. Study participants in Viva La Familia study were recruited in Houston, Texas. The second dataset, Mexican hypertriglyceridemia study (dbGaP Study Accession: phs000618.v1.p1), contains 2229 samples with 1117 cases and 1112 controls, where the case-control status was ascertained based on an individual’s serum triglyceride level [14]. Note that although there were 4350 study samples reported in the paper, the dbGaP contains only 2229 that were genotyped with Illumina Human610-Quad BeadChips — stage 1 of the GWAS. The rest samples were only typed on selected 1200 SNPs — stage 2. Study participants in this study were recruited in Mexico City. We call the first dataset Viva and the second Lipid.

4.2 Data quality control

We removed all A/T, C/G SNPs whose potential allele flipping between different datasets cannot be identified without additional information. A SNP was removed if it was missing in one of the datasets, either training or cohort. We also removed SNPs whose missing proportion was larger than 5%. Although we realized that the Hardy-Weinburg disequilibrium test is not appropriate for admixed samples, we used it anyway to remove SNPs whose HWD test p-values < $10^{-6}$. It is understood that this practice errs toward the safe side by eliminating possibly good SNPs. Finally, we obtained the cluster plots for each SNP, devised a simple algorithm to assign quality scores to each SNP cluster plot, and visually inspected those SNPs whose score indicated low quality. We removed those SNPs that contained a fourth cluster, or whose clusters were not distinct (examples of such cluster plots can be found in [35]). We were particularly stringent to conduct such SNP quality control at the MHC region. Of the two GWAS datasets we obtained from dbGaP, Viva contains SNP cluster information, Lipid does not. In the end, we had 352,754 SNPs from Viva and 479,757 SNPs from Lipid. The low number of SNPs in Viva reflected small number of overlapping SNPs between the Illumina HumanOmni 1-v1.0 and the Illumina 650Y arrays, the latter of which was used by the HGDP study that generates the Maya and Pima genotypes used as Amerindian training samples.
4.3 Infer local ancestry

We used ELAI \cite{10} for local ancestry inference, which has been demonstrated to outperform competing methods such as HapMix \cite{9} and LAMP-LD \cite{26}. ELAI implements a two-layer cluster model and the model is fitted via the EM algorithm. The upper-layer clusters are parameterized to represent haplotypes from ancestral populations, and the lower-layer clusters contemporary haplotypes. The two-layer model was motivated by approximating the coalescent with recombination. It directly applies to diplotypes and automatically integrates out phase uncertainty. It can also estimate the recombination rates between markers, and hence doesn’t require recombination map as an input. Thus, the requirement to run ELAI is minimal — just genotypes and marker positions. To run ELAI, one needs to provide training samples. We used European and African samples from HapMap3 and Maya and Pima samples from HGDP as default training samples (or reference panels, or source populations). ELAI is a cluster-based model and we wanted to specify numbers of clusters. The number of upper-layer clusters represents the number of source populations and we set it as 3; the number of lower-layer clusters was set as 15. Extensive simulations demonstrated that this parameter setting performs well \cite{10}. Lastly, we needed to specify number of admixing generations and we used 20. All ELAI results were averaged over 10 independent EM runs of 20 steps each, unless otherwise noted.

4.4 Define phenotype for association test

We defined a marker set A that contained markers whose African average dosages were greater than 0.30. This threshold was 13 sample standard deviations away from the mean (in Lipid dataset), and the resulting markers formed a consecutive region within MHC. We assigned each individual a phenotypic value obtained by averaging African ancestry dosages over markers in A.

4.5 Simulation study

We used a population genetics model that mimics the out-of-Africa migration events to simulate a 3 Mb region of three source populations that mimic Amerindian, European, and African \cite{25}. After setting aside 200 haplotypes from each source population as training haplotypes, we used the remaining haplotypes to simulate three-way admixed individuals by a one-pulse model \cite{36}. Specifically, we randomly selected 50,000 haplotypes from the three source populations using proportions of 50%, 45%, and 5%, mimicking the admixture proportion of Mexicans. We split 3 Mb into three segments, and assigned at two splitting points recombination hotspots. At each hotspot, we assumed equal recombination probability of 0.1, 0.2, and 0.5 per generation. We sampled two haplotypes with replacement and introduced possible crossover events at hotspots to produce two new haplotypes. We repeated the pairing and crossover 25,000 times to produce 50,000 haplotypes for the next generation. The admixture simulation was done for 20 generations. To simulate selec-
tion, we designated the mid-section as the locus under selection, and assumed selection coefficients of 0.02, 0.05, and 0.10 per generation. The alleles under positive selection were those 5% from the source population that mimicked Africans. After 20 generations, we randomly chose 2,000 admixed haplotypes, pairing them to form 1,000 diplotypes as cohort samples. We used two sizes of mid-section: 0.5 Mb and 1 Mb. A small mid-section produces a more challenging problem. To investigate how switch-errors affected local ancestry inference for different methods, in addition to perfect phasing situation, we also introduce 2% phasing errors into Amerindian training samples and the cohort samples, and 1% phasing errors to European and African training samples. To do so, at randomly selected heterozygous marker, from left to right we crossed-over two haplotypes.
References


Figure 1: Global ancestry proportions and principal components. The PC plots shown are mirror images of the original as indicated by “−” sign in labels. To produce a triangular plot, note that each individual associates a vector of $x, y,$ and $z$ with $x + y + z = 1,$ and a unique point can be determined such that within an equilateral triangle its distances to three edges are $x, y,$ and $z.$
Figure 2: African average dosages. Plot shows all 22 autosomes for two GWAS datasets. The spike at MHC region on chromosome 6 is rather striking in both datasets. The blue lines are the genome-wide mean of average dosages; the gray lines are $mean \pm 4ssd$ (ssd stands for sample standard deviation).
Figure 3: Comparison between different European and African training samples. The comparison was performed with chromosome 6 of Lipid dataset. African (a) and European (b) average dosages for five sets of training samples shown in legend, where ALL means CEU+TSI–YRI+MKK–MAYA. (c) The difference of estimated European average dosages of Mexicans between two European training samples (see main text for explanation). (d) The violin plots of structure analysis of five HapMap3 populations, where ASW denotes Americans from the Southwest, an African American population. On each violin plot, gray dot denotes the median and black dot the mean.
Figure 4: Comparison with and without Amerindian training samples. (a) African average dosages of Viva. (b) Amerindian average dosages of Viva. (c) African average dosages of Lipid. (d) Amerindian dosages of Lipid. We combined CEU and TSI as European training samples, and YRI and MKK as African training samples.
### Table S1: Candidate alleles that are under selection in Mexicans.

Columns under the headings CEU and TSI are imputed allele frequencies, and the column under the heading Mexican was obtained from [30].

<table>
<thead>
<tr>
<th>HLA-Allele</th>
<th>CEU</th>
<th>TSI</th>
<th>Mexican</th>
</tr>
</thead>
<tbody>
<tr>
<td>B*15:01</td>
<td>0.0860</td>
<td>0.0326</td>
<td>0.0214</td>
</tr>
<tr>
<td>B*07:02</td>
<td>0.1674</td>
<td>0.0803</td>
<td>0.0406</td>
</tr>
<tr>
<td>C*04:01</td>
<td>0.0910</td>
<td>0.1475</td>
<td>0.1859</td>
</tr>
</tbody>
</table>
Figure S1: Average dosages for Amerindian (blue), European (red), and African (green) ancestries for Viva and Lipid datasets. Plot shows all 22 autosomes. The previous identified signal on chromosome 8, although still there, fails to be statistically significant.
Figure S2: Comparison of Amerindian ancestry dosages for five sets of training samples shown in the legend, where ALL means CEU+TSI–YRI+MKK–MAYA. The comparison was performed with chromosome 6 of Lipid dataset.
Figure S3: SNP associated with enrichment of African local ancestry in Mexican lipid study. Bayes factors (BF) were computed using BIMBAM. The horizontal blue line is $\log_{10} BF = 10$. 
Figure S4: Comparison between LAMP-LD, RFMix, and ELAI under different simulation conditions. There are 9 combinations of crossover probability (0.1, 0.2, and 0.5) and selection coefficients (0.02, 0.05, and 0.10) for two sizes of the mid-section: 1 Mb (a) and 0.5 Mb (b). The mid-section harbors alleles under selection, and a smaller size produces a more challenging problem. Plots also compare effects of phasing errors (2% for cohort and the Amerindian training sample and 1% for the other two training samples). RFMix performed significantly worse when phasing errors were introduced in (b). Compared to RFMix, LAMP-LD was less sensitive to phasing error. ELAI was unaffected by phasing errors. Parameters for LAMP-LD: window size = 100, number of HMM states = 25; for RFMix: window size = 0.1 cM, which approximately contains 100 SNPs. Both parameter settings were used by the 1000 Genomes analysis group.
Figure S5: RFMix performance with different switch-errors. This is the same dataset as used in Figure S4(b). In the legends, the number before the plus sign is the switch-error for cohort and the Amerindian training sample, and the number after is for the European and African training samples.
Figure S6: RFMix’s performance with different choices of window size. The mean absolute deviation was computed from the same dataset that was used in Figure S4(b) and averaged over 9 simulation parameter (recombination probability and selection strength) settings. The switch-errors were 2% for cohort and the Amerindian training sample and 1% for the other two training samples. We used centi-Morgan (cM) to measure the window size, which is also the unit required by RFMix. In our simulations, 1 cM contains roughly 1000 SNPs. Different ancestry components have slightly different optimal window sizes.
Figure S7: RFMix performs well with the optimal window size. This is the same dataset as used in Figure S4(b). When the optimal window size is used, RFMix is no longer sensitive to switch-errors introduced (2% for cohort and the Amerindian training sample and 1% for the other two training samples).