RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference

Brian K. Maples,[1,2] Simon Gravel,[1,3] Eimear E. Kenny,[1,4,5,6,7,8] and Carlos D. Bustamante[1,8,*]

Local-ancestry inference is an important step in the genetic analysis of fully sequenced human genomes. Current methods can only detect continental-level ancestry (i.e., European versus African versus Asian) accurately even when using millions of markers. Here, we present RFMix, a powerful discriminative modeling approach that is faster (~30×) and more accurate than existing methods. We accomplish this by using a conditional random field parameterized by random forests trained on reference panels. RFMix is capable of learning from the admixed samples themselves to boost performance and autocorrect phasing errors. RFMix shows high sensitivity and specificity in simulated Hispanics/Latinos and African Americans and admixed Europeans, Africans, and Asians. Finally, we demonstrate that African Americans in HapMap contain modest (but nonzero) levels of Native American ancestry (~0.4%).

Introduction

Nonrandom mating and genetic drift have led to discernible allele-frequency differences among many human populations.[1–4] Coupled with recent advances in computational and high-throughput genomics, these allele-frequency differences afford high-resolution ancestry inference across individual human genomes. Local-ancestry inference (LAI), or ancestry deconvolution, is critical for the analysis of admixed genomes and is a standard part of genetic analysis in a wide range of fields, ranging from pharmacogenomics to human demographic history.[5–9] Although previous studies have focused on continental ancestry (e.g., European versus East Asian versus sub-Saharan African ancestry), it has become evident that subcontinental ancestries must also be considered.[10] For example, European populations are genetically heterogeneous, and many biomedical traits (including height, blood pressure, and cholesterol levels) show gradients that mirror genetic clines.[11] Likewise, despite the fact that both groups are classified as Latino, Puerto Ricans and Mexicans living in the United States have the highest and lowest incidence, morbidity, and mortality of asthma, respectively, in the country.[12] A final example is the South Africa Colored population, which derives its ancestry from an admixture of multiple African populations, as well as European and Asian populations, and exhibits large variation in (and high incidence of) susceptibility to tuberculosis.[13] The ability to uncover patterns of subcontinental ancestry in such populations is critical for disentangling the role of ancestry versus environment versus individual genetic markers on these and other complex traits.

Recently, the 1000 Genomes Project Phase I released low-pass sequencing, exome, and dense genotyping data for 1,092 individuals from 14 popualtions.[1] A key finding of this work and others[14,15] was that the vast majority of genetic variants in the human genome are rare in frequency and are population specific. We hypothesize that these features of whole-genome sequence data will allow the differentiation of even closely related populations so that most individuals in the world will trace their ancestry to multiple genetically discernible ancestral populations. Given that human populations have expanded dramatically from less than 100,000,000 people 10,000 years ago to 7,000,000,000 people today, the model of multiple finite and genetically discernible ancestral populations is a testable one.

Numerous computational approaches to LAI have been developed. Early approaches, such as STRUCTURE, were designed for unlinked markers[16–18] and modeled local-ancestry correlations due to common ancestry by using Hidden Markov Models (HMMs) instead of explicitly modeling linkage disequilibrium (LD). Although useful for inferring highly diverged populations, these approaches do not fully exploit the potentially rich information in haplotypes (particularly for differentiating closely related populations). Most approaches that do incorporate LD explicitly (such as HAPMIX) can only consider two ancestral populations at a time because of computational limitations.[19–21] Among the state-of-the-art approaches is the LAMP algorithm, which is able to draw inference accurately across more than two ancestral populations and does so in a significantly shorter time than HAPMIX.[22] A potential limitation of many current methods is that they require large reference panels that are good proxies for the true ancestries of the admixed samples. Despite the continuing contributions of organized efforts such as HapMap[2] and the 1000 Genomes

[1]Department of Genetics, Stanford University, Stanford, CA 94305, USA; [2]Biomedical Informatics Training Program, Stanford University, Stanford, CA 94305, USA; [3]Department of Human Genetics, McGill University, Montreal, QC H3A 1B1, Canada; [4]Department of Genetics and Genome Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [5]Center of Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [6]The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [7]Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sanai, New York, NY 10029, USA
[8]These authors contributed equally to this work
*Correspondence: cdbustam@stanford.edu
http://dx.doi.org/10.1016/j.ajhg.2013.06.020.

Table 1. Comparison of Accuracies and Speeds between Methods

	Perfectly Phased Data		Beagle-Phased Data	
	Diploid Accuracy	Run Time (s)	Diploid Accuracy	Run Time (s)
RFMix	95.6%	26	93.2%	108
LAMP-HAP or LAMP-LD[a]	93.7%	878	90.7%	914
SupportMix	91.9%	44	77.3%	45

The table shows diploid accuracy and run time when local ancestry was inferred in simulated Latino samples under different phasing conditions.
[a]LAMP-HAP and LAMP-LD were used for LAI on the perfectly phased data and Beagle-phased data, respectively.

Project,[1] publically available population-scale data sets remain sparse, most notably for individuals from Native American groups;[22] thus, the accuracy of all methods that rely solely on ancestry panels is limited by the available samples. This motivates the development of methods that can utilize the ancestry information contained within the admixed samples themselves and that are also fast enough to analyze the tens of millions of SNPs recorded in whole-genome sequence panels.

To address these issues, we depart from the *generative* approach taken by all commonly used LAI algorithms, wherein an explicit probabilistic model for the observed variables (the alleles) and unobserved variables (the ancestry) is fitted to the data via a HMM or an extension thereof.[19–23] We developed RFMix, a *discriminative* approach that models ancestry along an admixed chromosome given observed haplotype sequences of known or inferred ancestry. Consider a system that contains an unobserved variable of interest Y and observed variables X that we have measured to help us infer Y. Discriminative approaches model P(Y | X), the dependence of Y on X, directly, whereas generative approaches first estimate P(Y, X), the joint dependence between all the variables in the system, before using Bayes' rule to estimate P(Y | X). Examples of discriminative approaches include regular and logistic regression. Whereas generative models offer advantages when data are sparse, discriminative models have lower asymptotic error.[24] Because the amount of available human genome data will keep growing over the coming years, we expect that a well-designed discriminative approach to LAI will outperform its generative counterpart.

In addition to providing increased accuracy, RFMix allows for considerable gains in speed (Table 1). This allows us to improve performance by running the method iteratively and using inferred ancestry assignments to augment the training set. Incorporating the ancestral tracts inferred in the admixed samples into the reference panels is advantageous for at least three reasons. First, haplotypes in the admixed populations are direct descendants from the actual ancestral populations rather than of a proxy population and thus should be able to better resolve ancestral haplotype patterns. Second, by augmenting the reference panel with chromosomes from admixed individuals, we increase the total number of observed haplotypes and, thus, the training panel size. Third, identity-by-descent (IBD) information across individuals in the admixed population can be directly leveraged for ancestry inference. These advantages are particularly beneficial in situations where reference samples from close proxy populations are not available or the number of reference samples collected is low. This is often the case when panels must be chosen from pre-existing publically available data sets.

We finally provide a RFMix generalization that jointly models phasing errors and local ancestry. This is particularly important when phasing in the admixed population is performed statistically with the use of population data; in such cases, long-range phasing is often very inaccurate, and we hypothesize that modeling ancestry and phase jointly could lead to improved inference of both.

We use simulated continental-scale admixtures of Native American, African, and European ancestries to demonstrate that RFMix is faster and more accurate than the state-of-the-art methods and is capable of utilizing ancestry information from admixed samples to substantially increase performance across a number of realistic scenarios. We also show that RFMix scales to whole-genome sequence data to achieve high accuracy across a number of simulated subcontinental admixtures. In addition, we show that the phase-correction strategy not only improves phasing but also allows accurate LAI in admixed haplotypes containing phase errors. Finally, we apply RFMix to HapMap African Americans to study the existence of Native American ancestry within this group.

Material and Methods

Theory

In brief, our discriminative modeling approach works by dividing each chromosome into windows and inferring local ancestry within each window by using a conditional random field (CRF) parameterized by random forests trained on reference panels (Figure 1).[25,26] Once ancestries have been assigned to the windows within admixed chromosomes, they are used for refining our knowledge of haplotype patterns in the ancestral populations and improving inference accuracy with an expectation-maximization (EM) step. For simplicity, we first explain the initial iteration of the calling strategy with no phase-error correction. Also, because local ancestry can be inferred on each chromosome in the genome independently, we describe the analysis of one chromosome in the genome.

Inputs and Windowing

RFMix uses the genetic location of SNPs to divide the chromosome into W contiguous disjoint windows such that the maximum distance between all SNPs in any window is d cM. The N phased chromosomes in the admixed and reference panels are read, whereby one reference panel is supplied for each of the R ancestries. The haplotypes of these chromosomes across windows can be represented by a random N × W matrix H, where the value of the $(i,j)^{th}$ element $H_{i,j}$ is the sequence of alleles $H_{i,j}^{(1)}$, $H_{i,j}^{(2)}$, …, $H_{i,j}^{(s\_j)}$ of the i^{th} haplotype in the j^{th} window, where s_j is the number of SNPs in the j^{th} window. Similarly, the local ancestry of these chromosomes can be represented by a random N × W matrix A,

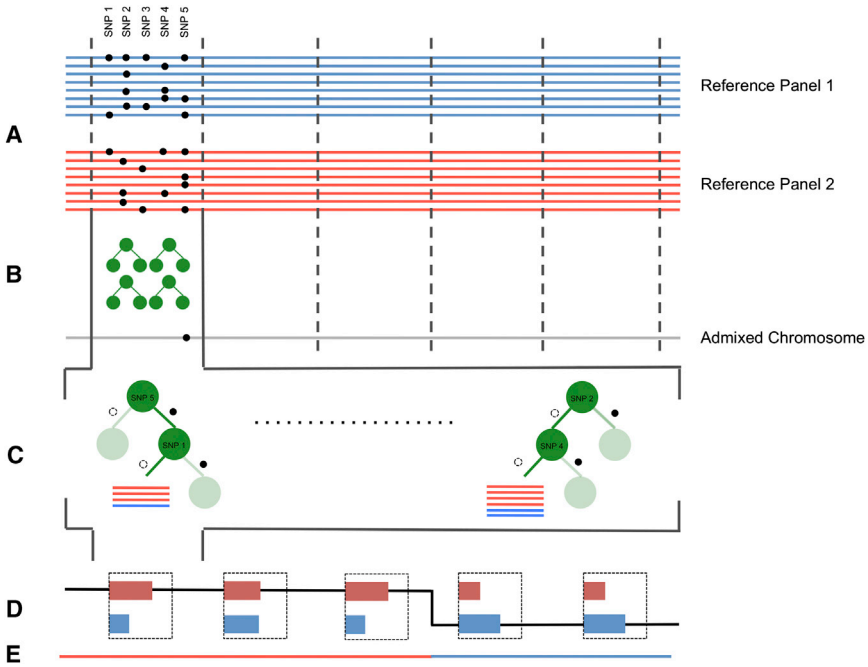

Figure 1. The LAI Algorithm
To illustrate the working of RFMix, we consider a single admixed chromosome from an individual with ancestry from two diverged populations.
(A) For building reference panels, samples are collected from proxy populations related to the ancestral populations. Phased chromosomes are divided into windows of equal size on the basis of genetic distance.
(B) For each window, a random forest is trained to distinguish ancestry by using the reference panels.
(C) Considering the admixed chromosome, each tree in the random forest generates a fractional vote for each ancestry by following the path through the tree corresponding to the admixed sequence.
(D) These votes are summed, producing posterior ancestry probabilities within each window. These posterior probabilities are used for determining the most likely sequence of ancestry across windows via MAP inference (black line) or via max marginalization of the forward-backward posterior probabilities (not shown).
(E) The local ancestries inferred by MAP across the admixed chromosome.

where the $(i,j)^{th}$ element $A_{i,j}$ is the local ancestry of the i^{th} chromosome in the j^{th} window. Although all elements of H are observed, only the elements of A in rows corresponding to chromosomes designated as references are initially observed. For notational purposes, $H_{i,*}$ and $A_{i,*}$ represent the haplotype structure and local ancestry, respectively, along the entire i^{th} haploid chromosome.

LAI

A CRF framework is used for LAI. We use a linear-chain CRF to model the conditional distribution $P(A_{i,*}|H_{i,*}:\Theta)$. The CRF can be represented in log-linear form:

$$P(A_{i,*}\,|\,H_{i,*}:\Theta)=\frac{1}{Z(H_{i,*})}\exp\left\{\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{h\epsilon\mathcal{H}_w}\theta^{A}_{w,r,h}1_{\{A_{i,w}=r\}}1_{\{H_{i,w}=h\}}+\sum_{p=1}^{W-1}\sum_{j=1}^{R}\sum_{k=1}^{R}\theta^{T}_{p,j,k}1_{\{A_{i,p}=j\}}1_{\{A_{i,p+1}=k\}}\right\},$$

where

$\mathcal{H}_w$ is the set of all possible haplotypes in window w
$1_{\{x\,=\,x'\}}$ is an indicator function that equals 1 when x equals x′ and 0 otherwise

$$Z(H_{i,*})=\sum_{A_{i,*}}\exp\left\{\sum_{w=1}^{W}\sum_{r=1}^{R}\sum_{h\epsilon\mathcal{H}_w}\theta^{A}_{w,r,h}1_{\{A_{i,w}=r\}}1_{\{H_{i,w}=h\}}+\sum_{p=1}^{W-1}\sum_{j=1}^{R}\sum_{k=1}^{R}\theta^{T}_{p,j,k}1_{\{A_{i,p}=j\}}1_{\{A_{i,p+1}=k\}}\right\}$$

$$\theta^{A}_{w,r,h}=\ln(P(A_{i,w}=r\,|\,H_{i,w}=h))$$

$$\theta^{T}_{p,j,k}=\ln\left(P\left(A_{i,p}=j,\,A_{i,p+1}=k\right)\right)$$

θ^{A} and θ^{T} are the two sets of model parameters. The former set is learned by the training of a random forest on the reference panels for each window, and the latter is set with the admixture model described by Falush et al.[16] (see below). Inference can then be performed with maximum-a-posteriori (MAP) estimation or smoothing, analogous to the Viterbi and forward-backward inference approaches used in HMMs.

Learning Model Parameters

Learning the θ^{A} Parameters. For each window, a random forest is trained with segments of the reference haplotypes within that window and then used for estimating the posterior probability of each ancestry given the segment of the admixed haplotype within that window. The predictor variables in each window are the alleles observed at the biallelic SNPs within that window, and the response variable is the local ancestry in that window. Although any discriminative classifier could in theory be used, random forests have the advantage in that they can perform classification with any number of ancestral classes, have a direct probabilistic interpretation, and work optimally with binary predictor variables, which is the case when biallelic SNPs are used. In addition, they are computationally fast and able to find high-dimensional interactions between subsets of variables even in the presence of many uninformative variables. This is ideal for characterizing haplotype structure in data with many SNPs, such as whole-genome sequence data.

The random-forest algorithm that we use is similar to the one originally described by Leo Breiman,[26] but it has two modifications. The first changes the bootstrapping subalgorithm from one step to two. Instead of sampling each haplotype from all reference panels with uniform probability, it first randomly samples an ancestry with uniform probability and then randomly chooses a haplotype from that ancestry with uniform probability. This is to address any potential class-imbalance problem, where, for example, one ancestry might happen to have many more samples collected than another. This is especially important when ancestral tracts inferred from admixed individuals are used because it is likely that one ancestry is significantly more represented than another.

The second modification replaces the per-tree majority unit vote with a fractional vote that depends on the composition of the node that an admixed haplotype maps to. For example, if an admixed haplotype maps to a node with e_1 haplotypes from ancestry 1 and e_2 haplotypes from ancestry 2, the fractional votes cast by this tree for ancestries 1 and 2 would be $e_1 / (e_1 + e_2)$ and $e_2 / (e_1 + e_2)$, respectively. This strategy has been found to improve the accuracy of posterior-class-probability estimates from bagged classifiers.[27]

Learning the θ^T Parameters. The joint probability of local ancestries in adjacent windows depends on the global proportion of each ancestry and the probability of recombination between the two windows. For the former, we simply assume a uniform distribution of ancestry, although we could modify it to take advantage of demographic knowledge or iteratively update it in the EM step. In calculating the probability of recombination between two loci, we assume the admixture model described by Falush et al.[16] Thus, the joint probability distribution is

$$P(A_{i,p} = j,\ A_{i,p+1} = k) = \begin{cases} q_j(\exp(-d_pG) + (1 - \exp(-d_pG))q_k) \text{ if } j = k \\ q_j(1 - \exp(-d_pG))q_k \text{ otherwise,} \end{cases}$$

where q_j is the proportion of ancestry j in the admixed population, G is the number of generations since admixture, and d_p is the distance between the middle of windows p and p + 1.

Incorporating Information from Admixed Individuals

Above, we described how to model $P(A_{i,*}|H_{i,*} : \Theta)$ for each admixed chromosome independently. Ideally, we would model $P(A \mid H : \Theta)$, the joint ancestry across all admixed and reference panel chromosomes, so as to incorporate information from the admixed panel and discover latent admixture in the reference panels. To accomplish this in a computationally tractable manner, we take an EM approach.

First, we initialize the local-ancestry assignments of the admixed chromosomes independently by using the approach described above. For the M step, because we assume a uniform distribution of global ancestry, the θ^T parameters do not need to be updated because they do not depend on the local-ancestry-state assignments. Otherwise, we could use the estimated global-ancestry proportions of the admixed individual or the admixed population as a whole to modify these parameters. To update the θ^A parameters, we train random forests in each window by using the local-ancestry assignments for chromosomes in that window. Ideally, for each chromosome in a window, we would train a random forest on all other chromosomes and use that to infer the local-ancestry distribution in that window for that chromosome. Although this would avoid the problem of using a classifier trained on the data we want to analyze, it would significantly slow down our approach. Instead, in each window we divide the set of chromosomes randomly into b bins such that each bin has as close to the same number of chromosomes from each ancestry in it as possible. Then for each bin, we train a random forest on the remaining $b - 1$ bins and use it to infer the probability distribution of local ancestry for each chromosome in that bin. This underscores the importance of speed in the central approach taken because this increases the runtime by a factor of b times the number of iterations of EM. For the E step, we use the updated parameters to infer local ancestry in each chromosome via MAP or max marginalization as above.

Accounting for Phase Errors

We now model $P(A_{i,*}, A_{i_c,*}, H_{i,*}, H_{i_c,*}|O_{i,*}, O_{i_c,*} : \Theta)$, where i and i_c are the indices of both copies of the chromosome being analyzed for a particular admixed individual and $O_{i,}*$ is the observed phased sequence for chromosome i obtained from some phasing algorithm. We assume that, at most, one strand-flip error occurs per window per individual and let $F(o_{i,w}, o_{i_c,w})$ map a given ordered pair of phased haplotypes in window w to the set of all possible ordered pairs of phased haplotypes that can be achieved by the addition of one strand flip or less to the input. Thus, with a log-linear representation, the CRF is

$$\begin{aligned} P(A_{i,*}, A_{i_c,*}, H_{i,*}, H_{i_c,*} \mid O_{i,*}, O_{i_c,*} : \Theta) = \\ \frac{1}{Z(O_{i,*}, O_{i_c,*})} \exp\Bigg\{ \sum_{w=1}^{W} \sum_{r=1}^{R} \sum_{r_c=1}^{R} \sum_{o,o_c \in \vartheta_{i,w}} \sum_{h,h_c \in F(o,o_c)} \Big(\theta^A_{w,r,h} + \theta^A_{w,r_c,h_c} \\ + \theta^F_{i,w,h,h_c,o,o_c} \Big) 1_{\{A_{i,w}=r\}} 1_{\{A_{i_c,w}=r_c\}} 1_{\{H_{i,w}=h\}} 1_{\{H_{i_c,w}=h_c\}} 1_{\{O_{i,w}=o\}} \\ \times 1_{\{O_{i_c,w}=o_c\}} + \sum_{p=1}^{W-1} \sum_{j=1}^{R} \sum_{k=1}^{R} \theta^T_{p,j,k} 1_{\{A_{i,p}=j\}} 1_{\{A_{i,p+1}=k\}} \\ + \sum_{p_c=1}^{W-1} \sum_{j_c=1}^{R} \sum_{k_c=1}^{R} \theta^T_{p_c,j_c,k} 1_{\{A_{i_c,p_c}=j_c\}} 1_{\{A_{i_c,p_c+1}=k_c\}} \Bigg\}, \end{aligned}$$

where

$\vartheta_{i,w}$ is the set of all possible haplotypes that could be constructed from the genotypes in window w for the sample comprising chromosomes i and i_c

$\theta^F_{i,w,h,h_c,o,o_c} = \ln(P(H_{i,w} = h, H_{i_c,w} = h_c|O_{i,w} = o, O_{i_c,w} = o_c))$

$Z(O_{i,*}, O_{i_c,*})$ is the normalizing factor.

To calculate θ^F_{w,h,h_c,o,o_c}, we assumed that the probability of a strand-flip error at any heterozygous site would be 0.07. Thus, if there were n heterozygous sites in window w for individual i,

$$\theta^F_{i,w,h,h_c,o,o_c} = \begin{cases} \ln\left((0.07) * (0.93)^{n-1}\right) \text{ if one switch} \\ \ln((0.93)^n) \text{ otherwise.} \end{cases}$$

MAP inference results in a new phasing for the haplotypes of each individual, as well as local-ancestry calls along each haplotype.

Simulations

Processing HapMap and Native American Samples

HapMap3 trio-phased samples were obtained, and individuals who had a pairwise IBD proportion greater than 0.05 were removed. We used LiftOver[28] to get the build 37 genetic locations of the SNPs and removed any SNPs that were unable to be mapped. We also obtained Affymetrix (Affy) 6.0 genotype data for 43 Native American individuals[29] who had been determined to have insignificant European admixture by ADMIXTURE and phased them with Beagle. We removed all instances of duplicate SNPs in the Native American data and intersected the remaining SNPs with the HapMap data. Finally, we removed all A/T and G/C SNPs.

Using SNP Array Data for LAI on Simulated Latinos with Three-Way Continental Admixture

We generated reference panels and simulated ten Latino genomes with 45% Native American, 50% European, and 5% African ancestry by using a two-step process. In the first step, we used a Wright-Fisher simulation of 400 diploid individuals to construct ten individuals with local-ancestry assignments sampled 12 generations after admixture. In the second step, we generated genotype

assignments for each individual by using processed NAT (Native American)[29] and HapMap CEU (Utah residents with ancestry from northern and western Europe from the CEPH collection) and YRI (Yoruba in Ibadan, Nigeria) samples.[2] Samples not used for constructing simulated genomes were used for building reference panels composed of 30 samples from each population.

LAI was then performed on chromosome 1 of each simulated admixed sample with the use of the three ancestral population samples of 30 individuals each as reference panels. For all RFMix runs in this paper, we used input parameters of 0.2 cM window sizes, eight generations of admixture, and 100 trees per random forest. We also performed LAI by using LAMP-HAP with the preliminary phasing step removed, as well as SupportMix.

Incorporating Information from the Admixed Panel with Small Reference Panels

We simulated an additional 30 Latino samples in the same manner as above. We performed LAI on the combined 40 simulated Latino chromosome 1's by using ideal CEU, YRI, and NAT reference panels of three individuals each. We then performed the EM step for five iterations and used MAP inference to set ancestries at each iteration. For all EM steps in this paper, we used ten bins per EM iteration. For comparison, we also performed LAI by using LAMP-HAP with the preliminary phasing step removed.

Incorporating Information from the Admixed Panel with Proxy Reference Panels

We used 30 MKK (Maasai in Kinyawa, Kenya), 30 JPT (Japanese in Tokyo, Japan) and CHB (Han Chinese in Beijing, China), and 30 TSI (Toscani in Italy) samples from HapMap as proxy reference panels for the African, Native American, and European ancestries, respectively. We used these reference panels to perform LAI on the 40 simulated Latino samples. We then performed the EM step for five iterations and used MAP inference to set ancestries at each iteration. We also performed LAI for this scenario by using LAMP-HAP with the preliminary phasing step removed.

Incorporating Information from the Admixed Panel with Small, Proxy Reference Panels

We used three individuals from each proxy reference panel as references to infer ancestry in the 40 simulated Latino samples. We then performed the EM step for five iterations and used MAP inference to set ancestries at each iteration. We also performed LAI for this scenario by using LAMP-HAP with the preliminary phasing step removed.

Accounting for Strand-Flip Errors

We used Beagle to phase the initially simulated Latino chromosomes after removing all nonvariant and singleton sites and used the phased chromosomes of the ideal reference panels as a reference. We then inferred local ancestry by using the phase-correcting model with the ideal reference panels.[30] Because the NAT data did not contain trios and thus did not provide a highly accurate phasing against which we could compare our phase corrections, we acquired Affy 6.0 data for ten Native American trios[31] and simulated ten additional Beagle-phased Latino individuals with these data. We used the original ideal reference panels to perform LAI. We also simulated ten African American individuals with 82% African and 18% European ancestry by using a similar approach to the above and phased them with Beagle. We inferred local ancestry in these individuals as well by using 30 YRI and 30 CEU individuals as ideal references. For comparison, we also performed LAI with LAMP-LD and SupportMix.

LAI on Subcontinental Admixtures with and without Sequence Data

We used an approach similar to the above to construct admixed genomes sampled 12 generations after a 50/50 admixture of Japanese and Han Chinese South (CHS) (JPT/CHS), British from England and Scotland (GBR) and Tuscans (GBR/TSI), Finnish from Finland (FIN) and Tuscans (FIN/TSI), British and Finnish (GBR/FIN), and Yoruba and Luhya in Webuye, Kenya (LWK) (YRI/LWK). We used phased, consensus data from the 1000 Genomes Project Phase I to create reference panels and fill the genotypes of the admixed individuals.

We performed LAI on chromosome 11 for each admixed individual by using the samples in the two ancestral populations as a reference. Variant, nonsingleton sites from the integrated call set were used for inference. This inference was repeated for both the Affy 6.0 subset and the OMNI 2.5M subset of these sites. To determine the effect of reference-panel size on accuracy, we used the integrated call set and repeatedly halved the sizes of the reference panels (we rounded down when necessary) and used each size to infer local ancestry. For all analyses, we made ancestry calls at each SNP by doing max marginalization on the smoothed posteriors. Accuracy was determined for different confidence thresholds (50%, 90%, 99%, 99.9%, and 99.99%) on these maximum posteriors. The number of sites where no call was made was also recorded for each threshold.

Native American Ancestry in African Americans

We first used HapMap data to simulate ten African American genomes resulting from an admixture of YRI, CEU, and NAT populations eight generations in the past and used proportions of 82%, 17.5%, and 0.5%, respectively, based on previous estimates.[31] We then removed four individuals who had a global Native American ancestry proportion greater than 1%, resulting in a mean global Native American ancestry component of 0.54% across the remaining six samples. We inferred local ancestry across all autosomes in each sample by using the max-marginalization approach with thresholds of 50%, 90%, 99%, 99.9%, and 99.99%. We then inferred local ancestry by using ten simulated African American genomes generated from an admixture of 82% YRI and 18% CEU populations. We used a proxy European reference panel of TSI samples, whereas we used ideal panels of YRI and NAT individuals for the other populations. Each panel was composed of 30 individuals. Finally, 20 trio-phased ASW (African Ancestry in Southwest US) samples were obtained from HapMap, and local ancestry was inferred as above with panels composed of 85 CEU, 97 YRI, and 43 NAT samples.

Results

Using SNP Array Data for Fast Inference of Local Continental Ancestries

To evaluate the power and speed of RFMix for LAI in Hispanic/Latino populations, we simulated Latino individuals sampled from a three-way admixed population composed of 45% Native American, 50% European, and 5% African ancestry with admixture occurring 12 generations in the past. We simulated these individuals and built reference panels by using Affy 6.0 data from HapMap CEU and YRI samples, as well as Affy 6.0 NAT data.[29] For comparison, we also inferred local ancestry by using LAMP-HAP,[22] the state-of-the-art LAI method, and SupportMix,[23] a recently developed machine-learning method that trains Support Vector Machines in a

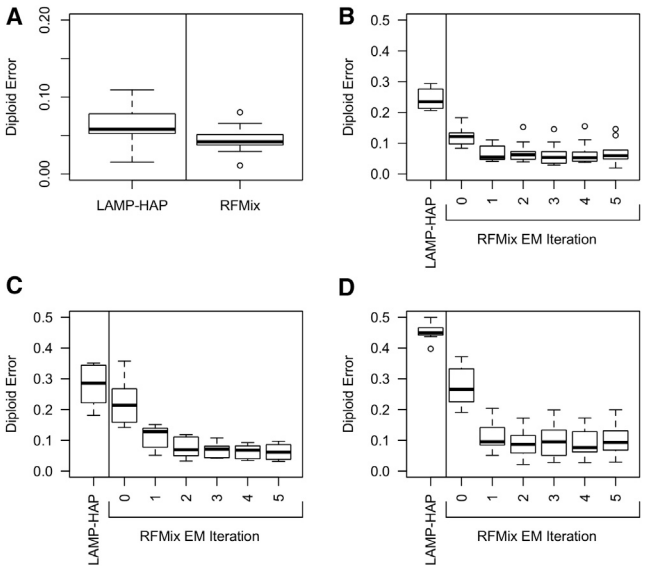

Figure 2. Comparison of Diploid Ancestry Error between LAMP-HAP and RFMix Inferences for Simulated Latinos across a Range of Real-World Scenarios

We simulated ten Latino individuals as described in the text and used reference panels composed of 30 ideal samples (A). We then simulated an additional 30 Latino individuals and used reference panels composed of three ideal samples (B), 30 proxy samples (C), or three proxy samples (D). The 0^{th} EM iteration of RFMix refers to the initial round of learning and inference shown in Figure 1.

sliding-window HMM framework (Table 1). We began by considering the case where exact phasings are known for the admixed samples. The performance criterion that we compared was diploid ancestry accuracy because we later wanted to assess performance when phasing of the admixed genomes was imperfect. RFMix had a mean diploid ancestry accuracy of 95.6% (0.68 SEM), which was significantly more accurate than the mean diploid ancestry accuracy of 93.7% (0.82 SEM) for LAMP-HAP (one-tailed paired-sample Wilcoxon signed-rank test p value = 0.005), although both methods performed well on this data set across all samples (Figure 2A). The difference was accentuated when small reference panels were used (Figures 2B and 2D). The observed average accuracy for LAMP-HAP was within 1 SE of the average accuracy observed by Baran et al. on a similar data set with the use of LAMP-HAP.[22] SupportMix had a lower mean diploid accuracy of 91.9% (0.57 SEM).

Combined learning and inference across ten simulated samples took 26 s for RFMix, over 33-fold faster than LAMP-HAP and 1.7-fold faster than SupportMix (Table 1) (all methods ran on an Intel Xeon 3.0 GHz processor with 24 GB RAM). Further, the discrete-window approach employed in RFMix allows for further speed optimization via parallelization. For example, during multithreading across two processors, the time required for RFMix dropped to nearly half of the nonparallelized time. As additional parallelization was added, this trend continued.

Incorporation of Ancestry Information from Admixed Individuals

The speed of RFMix allowed us to integrate it into an EM framework for incorporating the ancestry information contained within admixed individuals. We hypothesized that integrating ancestry information from admixed samples would most likely significantly improve performance in several practical scenarios, including (1) when reference populations closely related to the ancestral populations are unavailable and (2) when only a few samples of the reference populations can be collected. In order to gauge the effectiveness of the EM approach, we simulated an additional 30 Latino individuals and constructed three scenarios where EM is predicted to improve inference.

The first scenario featured small reference panels of three CEU, three NAT, and three YRI individuals. We used RFMix to infer local ancestry in the 40 simulated admixed individuals and tracked performance through five iterations of the EM. For comparison, we also inferred local ancestry with LAMP-HAP. To compare this scenario to the original scenario with larger reference panels, we only calculated diploid ancestry accuracy in the ten initially simulated Latinos. Interestingly, RFMix without an EM step had an accuracy of 87.8% (0.99 SEM), compared to 75.5% (1.1 SEM) for LAMP-HAP. After one iteration of EM, the average accuracy of RFMix increased to 93.2% (0.87 SEM), and further iterations did not significantly change this accuracy (Figure 2B).

In the second scenario, ancestry panels were different from those used for generating the simulated individuals. We refer to these reference panels as "proxy" panels, which contrast with the "ideal" reference panels discussed above. We used HapMap MKK, TSI, and combined JPT and CHB individuals to construct proxy references for African, European, and Native American ancestry, respectively. Each panel contained 30 individuals. RFMix-run LAI with these reference panels resulted in an average accuracy of 78.2% (2.1 SEM) before the EM step and an average accuracy of 93.2% (0.75 SEM) after three iterations of EM; no significant change was observed over subsequent iterations (Figure 2C). LAMP-HAP produced inferences with average accuracies of 72.2% (2.0 SEM).

The third scenario was a combination of the first two—we used three individuals from each proxy reference panel for reference. Before the EM step, RFMix had an average accuracy of 73.0% (2.0 SEM), and after two iterations of EM, it had an average accuracy of 91.3% (1.4 SEM); no significant change in accuracy was observed over subsequent iterations (Figure 2D). By comparison, LAMP-HAP had an average accuracy of 54.8% (0.83 SEM).

To test what effect incorrectly inferred latent admixture in the reference panels would have on performance, we repeated the above experiments by using an RFMix option that discards the original reference panels after the initial inference step so that only the (imperfect) inferred ancestries within the admixed samples are used as a reference in the subsequent EM stage. We found that the EM stage

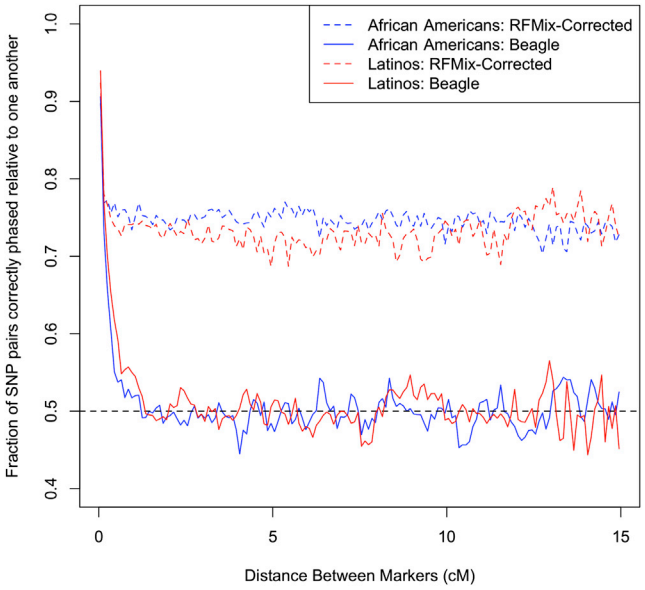

Figure 3. Phase Correction with Local-Ancestry Information
Fraction of SNP pairs that are within contiguous heterozygous ancestry regions and that are phased correctly with respect to each other as a function of the genetic distance between them. Blue and red lines correspond to simulated Latino and African American samples, respectively. Dashed and solid lines correspond to phasings that have and have not been corrected with local ancestry, respectively. The horizontal line at 0.5 marks the expected performance of random phasing.

still improved accuracies despite the occasionally high error rate in the initial ancestry estimates (Figure S1, available online). In addition, although discarding the panels resulted in lower accuracies, the effect was not strong and accuracies remained high overall and did not diverge over 30 iterations of EM. We also looked at the fraction of inferred SNP ancestries that changed between each iteration of EM. We found that this number converged to approximately 1% in less than six iterations for all scenarios considered whether reference panels were kept after the initial inference step or were discarded after this step (Figure S2).

Autocorrecting Phase Errors

We extended our local-ancestry approach to simultaneously model potential phase errors along with local ancestry (see Material and Methods). We used this extended approach to perform LAI on simulated unrelated African American and Latino samples that had been phased with Beagle. The average diploid accuracy of inference with RFMix for the ten African Americans was 98.9% (0.21 SEM). When we used RFMix and the original NAT panel to infer local ancestry in the Latino samples, we observed an average diploid accuracy of 93.2% (0.49 SEM), which was significantly greater than the 90.7% (1.1 SEM) accuracy with LAMP-LD (one-tailed paired-sample Wilcoxon signed rank test p value = 0.0049) but not statistically different from the 93.7% (0.82 SEM) accuracy observed with LAMP-HAP on the perfectly phased data (two-tailed paired-sample Wilcoxon signed rank test p value = 0.492) (Table 1). For the Latino individuals simulated with the trio-phased Native American data, we observed an average diploid accuracy of 93.4% (0.48 SEM). To illustrate the importance of modeling strand-flip errors when inferring local ancestry on haplotypes rather than diplotypes, we also performed LAI with SupportMix, which does not account for phase errors when inferring haplotype local ancestry and was more severely impacted by the presence of strand-flip errors in the Beagle-phased data (Table 1).

We hypothesized that using local-ancestry information to correct phasing would reduce the occurrence of strand flips in heterozygous ancestry regions, and so for each pair of heterozygous sites in these regions, we examined the probability that they would be phased in the correct orientation relative to each other on the basis of the distance between them. We grouped SNP pairs into bins on the basis of their distance and calculated this metric for the original Beagle-phased admixed chromosomes and for the new phasings generated by our approach. With both the simulated African American and Latino samples, we found that utilizing local-ancestry information improved the long-range phasing within heterozygous ancestry regions from statistically random to approximately 75% (Figure 3).

Using Whole-Genome Sequence Data for Inference of Local Subcontinental Ancestries

The 1000 Genomes Project Phase I has made population-scale combined SNP-chip and sequence data sets publically available for the first time.[1] To assess whether improved resolution can be obtained with the use of sequence data in addition to SNP-chip data, we simulated five two-way subcontinental admixtures and used RFMix to infer local ancestries by using the phased integrated call sets. We used combinations of the TSI, YRI, JPT, LWK, CHS, GBR, and FIN panels. In all cases, increasing the marginal probability threshold on whether to call a site increased the average accuracy, suggesting that the calculation of marginal probabilities is consistent (Figure 4). A large proportion of loci (43%–91%) in all subcontinental admixtures had their local ancestry inferred with >90% accuracy (Figure 4). To investigate the benefit of adding exome and low-coverage whole-genome sequence data to SNP array data, we also performed inference by using only the Affy 6.0 and Illumina OMNI 2.5M subsets of sites. Interestingly, performance in some simulated admixtures was significantly improved by the additional data, whereas others showed no improvement (Figure 5).

To determine the effect of sample size on accuracy, we repeatedly downsampled reference panels by half for each admixture. Reference-panel size had a significant impact on inference performance (Figure 6). For the JPT/CHS admixture, doubling the reference panel sizes from 39 JPT and 49 CHS to 79 JPT and 90 CHS resulted in approximately the same gain in accuracy as did adding

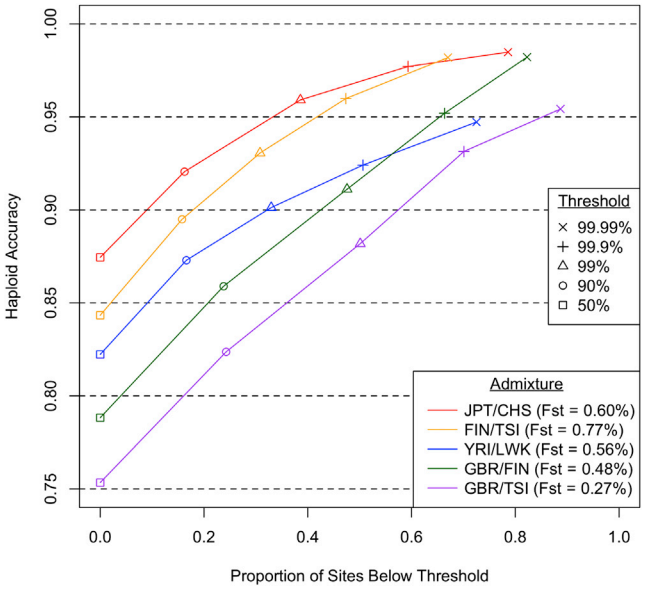

Figure 4. Accuracy of LAI of Subcontinental Admixtures with a Data Set Integrating SNP Array Data and Exome and Whole-Genome Sequence Data

We inferred ancestry for simulated admixed individuals. Low- and high-confidence call sets were generated with posterior-probability thresholds of 50%, 90%, 99%, 99.9%, or 99.99%. We show the accuracy in the resulting call sets as a function of the proportion of the genome that did not meet the threshold.

sequence data to OMNI 2.5M data. The results also suggest that doubling the largest reference panel size could result in >90% accuracy across all sites for the JPT/CHS, FIN/TSI, and YRI/LWK admixtures.

Inference on these whole-genome sequence data sets remained fast with RFMix. Combined learning and inference on the ten FIN/TSI samples with 519,937 SNPs in one chromosome took 4 min and 30 s without parallelization.

Native American Ancestry in African Americans

A potential application of RFMix is identifying low-occurrence ancestry. A previous study estimated the proportion of Native American ancestry in African Americans as 0.5%.[31] To gauge RFMix's ability to detect low levels of Native American admixture in African Americans, we used HapMap data to simulate six African American genomes with mean Native American ancestry of 0.56%. Because there was concern that Native American tracts inferred in African Americans would actually be rare Eurasian haplotypes, we used a proxy European reference panel of TSI samples, whereas we used ideal panels of YRI and NAT individuals for the other populations. We inferred local ancestry on these simulated genomes and calculated both the true-positive and the false-positive rates of Native American ancestry. To gauge the amount of Native American ancestry inferred when none is present, we also simulated ten African American genomes with no Native American ancestry. At a 99.9% confidence threshold on inferred ancestry, the average proportion of Native American ancestry was close to the true amount in the simulated samples containing true Native American ancestry (Figure 7). Also, at this threshold, the amount of false-positive Native American ancestry was nearly zero when none was present in the samples. In addition, when Native American ancestry was present, the positive predictive value for this ancestry was 83.1%. Thus, we are confident that (1) we did not falsely infer the presence of Native American ancestry in the real samples, (2) the estimated global proportion of Native American ancestry was accurate, and (3) the positive predictive value for loci inferred as Native American was high at this threshold. Using the 99.9% confidence threshold, we inferred that Native American ancestry comprises slightly over 0.44% of the total ancestry of African Americans, validating the previous estimate.[31]

Discussion

We have described a discriminative approach for LAI and have demonstrated (1) its improved performance compared to that of the state-of-the-art method with three-way continental admixtures, (2) its ability to use the ancestry information within admixed samples to improve performance in several real-world scenarios, (3) its ability to rapidly and accurately infer ancestry in subcontinental admixtures with the use of both SNP array data and large sequencing data sets, and (4) its ability to improve long-range phasing by using local ancestry. Obtaining good proxy reference panels for admixture deconvolution remains a challenge for many researchers despite the growing availability of publically available population-scale data provided by international efforts such as HapMap and the 1000 Genomes Project. Thus, our approach's ability to utilize the ancestry information within the admixed samples represents a significant advance in the field. The fact that accuracy with subcontinental admixtures significantly increased when the reference-panel sizes were increased also lends additional motivation for expanding the publically available data sets. The gain in accuracy observed from adding information from sequencing data also further motivates population-scale sequencing and public data release of properly consented samples for method development.

We have also demonstrated that local-ancestry information can improve long-range phasing. Because a large number of people are admixtures of at least two subcontinental populations, combining subcontinental admixture deconvolution with local-ancestry phase correction could allow significantly improved long-range phasing in individuals not traditionally thought of as admixed. IBD analysis will also benefit from this work, given that phase accuracy significantly affects the power to detect IBD segments.[32] Other future work includes using RFMix in situations where no proxy reference panel for one or more of the ancestral populations exists. One potential way in which to do this is to use a global-ancestry-inference algorithm such as ADMIXTURE to determine which

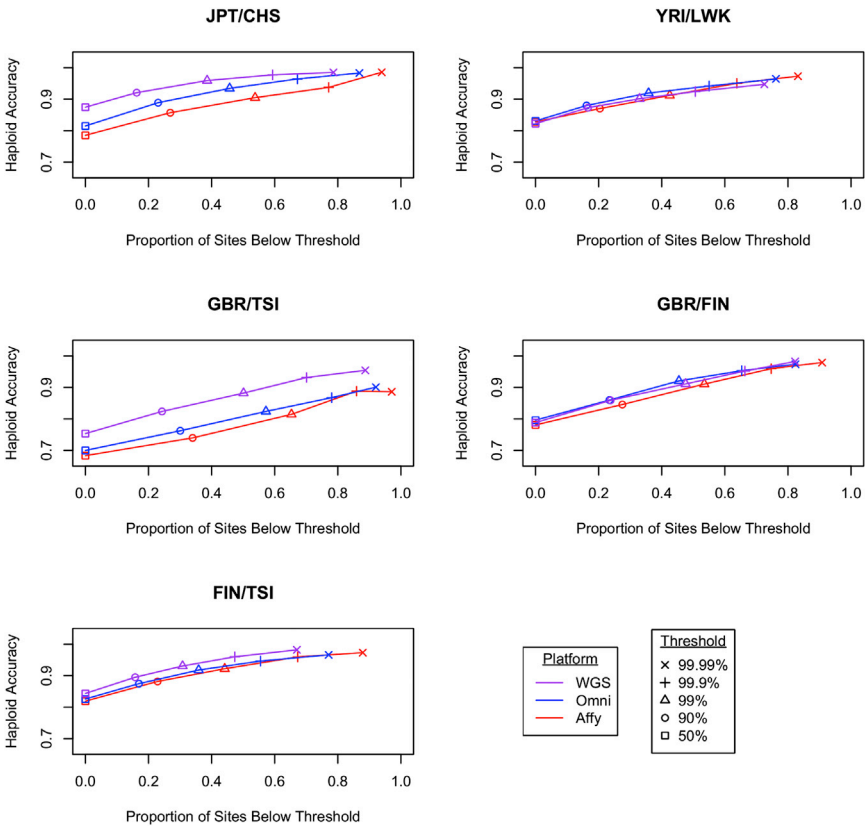

Figure 5. Integrating Exome and Whole-Genome Sequence Data with SNP Array Data

We generated haploid ancestry call sets by using posterior-probability thresholds of 50%, 90%, 99%, 99.9%, or 99.99% and simulated individuals, as in Figure 4, for different ancestry pairs and data sets. The following abbreviations are used: WGS, 1000 Genomes integrated data set; OMNI, 1000 Genomes OMNI 2.5M data set; Affy, the data set composed of the subset of WGS sites present on the Affy 6.0 SNP array.

admixed individuals have the greatest proportion of each ancestry and then use these individuals as references for those ancestries in an initial inference step followed by several iterations of the EM step. Another example of future work involves slightly modifying the algorithm to take advantage of its speed. For example, one approach we have tried is performing inference through a majority vote from multiple overlapping windows on each SNP. We achieved this by running RFMix multiple times on a sample with a range of different window sizes. However, this approach resulted in only modest gains in accuracy (Table S1).

The improvement in accuracy from adding information from sequence data also motivates future work for determining the best way in which this type of data can be utilized. One challenge will be in dealing with the higher rate of sequencing and phasing errors due to rare variants. Fortunately, these errors are most prominent for singleton variants, which provide very little information about local ancestry and can be discarded from the analysis. Because of imputation and joint calling, common variants can be called accurately with the use of whole-genome sequence data, even where coverage is low. Because random-forest classifiers are somewhat robust to training errors, we speculate that high-coverage, whole-genome data will lead to higher accuracy than will genotyping chip data.

Finally, we used RFMix to infer tracts of Native American ancestry in African Americans, thus confirming previous observations.[31,33] Future work will include determining the subcontinental Native American populations of these tracts, as well as applying this analysis to uncover Native American admixture in European Americans. As the amount of data available for reference continues to grow, we expect that it will become possible to predict subcontinental ancestry across the entire genome with high accuracy.

Supplemental Data

Supplemental Data include two figures and one table and can be found with this article online at http://www.cell.com/AJHG.

Acknowledgments

We thank Andres Moreno for providing data and Fouad Zakharia and Suyash Shringarpure for helpful comments. This work was supported by National Science Foundation (NSF) Graduate Research Fellowship grant DGE-1147470, National Library of Medicine training grant LM007033, National Human Genome Research Institute grant 2R01HG003229, and NSF Division of Mathematical Sciences grant 1201234. C.D.B. consults for Personalis, Inc., Ancestry.com, Invitae (formerly Locus Development), and the 23andMe.com project "Roots into the Future." None of these entities played any role in the design of the research or interpretation of results presented here.

Received: March 21, 2013
Revised: May 13, 2013
Accepted: June 21, 2013
Published: August 1, 2013

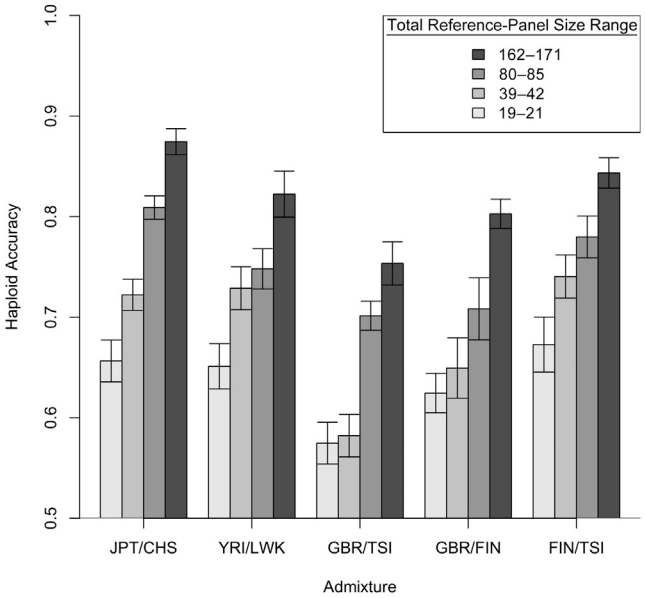

Figure 6. Effect of Reference-Panel Size on Performance of Subcontinental Admixture Deconvolution with the 1000 Genomes Integrated Whole-Genome Call Set
We simulated admixed individuals as we did for Figures 4 and 5 and inferred local ancestry by using different reference-panel sizes that correspond to roughly $^1/_2$, $^1/_4$, and $^1/_8$ of the original panel. Error bars represent the SEM. Because the same admixed individuals are used in simulations with different panel sizes, errors for different sample sizes are correlated.

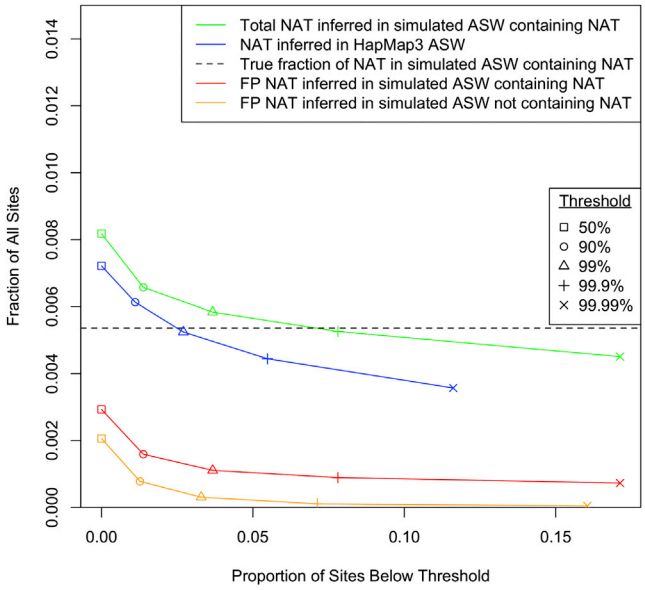

Figure 7. Native American Ancestry in African Americans
For the HapMap African American genome, the proportion inferred to have Native American ancestry (blue) is compared to the proportion inferred in a simulated population with 0.5% of Native American ancestry (green). For comparison, we estimated false-positive ("FP") rates on the basis of simulation with and without Native American ancestry. To ensure that the false-positive rates correspond to a realistic situation, we simulated individuals by using segments of CEU, YRI, and NAT ancestry and performed inference by using TSI, YRI, and NAT reference panels.

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes Project, http://www.1000genomes.org/
BEAGLE, http://faculty.washington.edu/browning/beagle/beagle.html
International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/
RFMix, http://med.stanford.edu/bustamantelab/
UCSC Genome Browser, http://genome.ucsc.edu/

References

1. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
2. International HapMap Consortium. (2003). The International HapMap Project. Nature *426*, 789–796.
3. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.
4. Nelson, M.R., Bryc, K., King, K.S., Indap, A., Boyko, A.R., Novembre, J., Briley, L.P., Maruyama, Y., Waterworth, D.M., Waeber, G., et al. (2008). The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am. J. Hum. Genet. *83*, 347–358.
5. Yang, J.J., Cheng, C., Devidas, M., Cao, X., Fan, Y., Campana, D., Yang, W., Neale, G., Cox, N.J., Scheet, P., et al. (2011). Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nat. Genet. *43*, 237–241.
6. Pool, J.E., and Nielsen, R. (2009). Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics *181*, 711–719.
7. Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G.K., Tandon, A., Kao, W.H.L., Ruczinski, I., Fornage, M., Siscovick, D.S., Zhu, X., et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a Breast Cancer Consortium. PLoS Genet. *7*, e1001371.
8. Wang, X., Zhu, X., Qin, H., Cooper, R.S., Ewens, W.J., Li, C., and Li, M. (2011). Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics *27*, 670–677.
9. Gravel, S. (2012). Population genetics models of local ancestry. Genetics *191*, 607–619.
10. Winkler, C.A., Nelson, G.W., and Smith, M.W. (2010). Admixture mapping comes of age. Annu. Rev. Genomics Hum. Genet. *11*, 65–89.
11. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
12. Torgerson, D.G., Gignoux, C.R., Galanter, J.M., Drake, K.A., Roth, L.A., Eng, C., Huntsman, S., Torres, R., Avila, P.C., Chapela, R., et al. (2012). Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. J. Allergy Clin. Immunol. *130*, 76–82, e12.
13. de Wit, E., Delport, W., Rugamika, C.E., Meintjes, A., Möller, M., van Helden, P.D., Seoighe, C., and Hoal, E.G. (2010).

Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum. Genet. *128*, 145–153.
14. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., and Bustamante, C.D.; 1000 Genomes Project. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA *108*, 11983–11988.
15. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al.; NHLBI Exome Sequencing Project. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.
16. Falush, D., Stephens, M., and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics *164*, 1567–1587.
17. Hoggart, C.J., Shriver, M.D., Kittles, R.A., Clayton, D.G., and McKeigue, P.M. (2004). Design and analysis of admixture mapping studies. Am. J. Hum. Genet. *74*, 965–978.
18. Patterson, N., Hattangadi, N., Lane, B., Lohmueller, K.E., Hafler, D.A., Oksenberg, J.R., Hauser, S.L., Smith, M.W., O'Brien, S.J., Altshuler, D., et al. (2004). Methods for high-density admixture mapping of disease genes. Am. J. Hum. Genet. *74*, 979–1000.
19. Tang, H., Coram, M., Wang, P., Zhu, X., and Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. Am. J. Hum. Genet. *79*, 1–12.
20. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. *5*, e1000519.
21. Sundquist, A., Fratkin, E., Do, C.B., and Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Res. *18*, 676–682.
22. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics *28*, 1359–1367.
23. Omberg, L., Salit, J., Hackett, N., Fuller, J., Matthew, R., Chouchane, L., Rodriguez-Flores, J.L., Bustamante, C., Crystal, R.G., and Mezey, J.G. (2012). Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. BMC Genet. *13*, 49.
24. Ng, A.Y., and Jordan, M.I. (2001). On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. In Advances in Neural Information Processing Systems 14, T.G. Dietterich, S. Becker, and Z. Ghahramani, eds. (Cambridge: MIT Press), pp. 841–848.
25. Lafferty, J., McCallum, A., and Pereira, F.C.N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning, 282–289.
26. Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32.
27. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (New York: Springer).
28. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*(Database issue), D590–D598.
29. Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., López Herráez, D., et al. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. *6*, e1001116.
30. Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.
31. Kidd, J.M., Gravel, S., Byrnes, J., Moreno-Estrada, A., Musharoff, S., Bryc, K., Degenhardt, J.D., Brisbin, A., Sheth, V., Chen, R., et al. (2012). Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. Am. J. Hum. Genet. *91*, 660–671.
32. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.
33. Gonçalves, V.F., Prosdocimi, F., Santos, L.S., Ortega, J.M., and Pena, S.D.J. (2007). Sex-biased gene flow in African Americans but not in American Caucasians. Genet. Mol. Res. *6*, 256–261.