PLOS ONE

RESEARCH ARTICLE

A Continuous Correlated Beta Process Model for Genetic Ancestry in Admixed Populations

Zachariah Gompert*

Department of Biology, Utah State University, Logan, UT, United States of America

* zach.gompert@usu.edu

Abstract

Admixture and recombination create populations and genomes with genetic ancestry from multiple source populations. Analyses of genetic ancestry in admixed populations are relevant for trait and disease mapping, studies of speciation, and conservation efforts. Consequently, many methods have been developed to infer genome-average ancestry and to deconvolute ancestry into continuous local ancestry blocks or tracts within individuals. Current methods for local ancestry inference perform well when admixture occurred recently or hybridization is ongoing, or when admixture occurred in the distant past such that local ancestry blocks have fixed in the admixed population. However, methods to infer local ancestry frequencies in isolated admixed populations still segregating for ancestry do not exist. In the current paper, I develop and test a continuous correlated beta process model to fill this analytical gap. The method explicitly models autocorrelations in ancestry frequencies at the population-level and uses discriminant analysis of SNP windows to take advantage of ancestry blocks within individuals. Analyses of simulated data sets show that the method is generally accurate such that ancestry frequency estimates exhibited low root-mean-square error and were highly correlated with the true values, particularly when large (±10 or ±20) SNP windows were used. Along these lines, the proposed method outperformed *post hoc* inference of ancestry frequencies from a traditional hidden Markov model (i.e., the linkage model in `structure`), particularly when admixture occurred more distantly in the past with little on-going gene flow or was followed by natural selection. The reliability and utility of the method was further assessed by analyzing genetic ancestry in an admixed human population (Uyghur) and three populations from a hybrid zone between *Mus domesticus* and *M. musculus*. Considerable variation in ancestry frequencies was detected within and among chromosomes in the Uyghur, with a large region of excess French ancestry harboring a gene with a known disease association. Similar variation was detected in the mouse hybrid zone, with notable constancy in regions of excess ancestry among admixed populations. By filling what has been an analytical gap, the proposed method should be a useful tool for many biologists. A computer program (`popanc`), written in `C++`, has been developed based on the proposed method and is available on-line at http://sourceforge.net/projects/popanc/.

Citation: Gompert Z (2016) A Continuous Correlated Beta Process Model for Genetic Ancestry in Admixed Populations. PLoS ONE 11(3): e0151047. doi:10.1371/journal.pone.0151047

Editor: Francesc Calafell, Universitat Pompeu Fabra, SPAIN

Received: August 27, 2015

Accepted: February 23, 2016

Published: March 11, 2016

Data Availability Statement: Computer code, compiled binaries, and example files are available from http://sourceforge.net/projects/popanc/.

Funding: The author has no support or funding to report.

Competing Interests: The author has declared that no competing interests exist.

Introduction

Genetic admixture between differentiated populations or species is common in plants and animals [1–6], including humans [7–10]. Admixture and recombination result in individuals whose genomes comprise a mosaic of chromosome segments with different genetic ancestry, that is to say, chromosome segments that have been inherited from different source populations. Given the prevalence of admixture, analyses of genetic ancestry are relevant in many areas of biology [11, 12]. For example, patterns of admixture and introgression in the wild show that species boundaries are often porous, and have been used to characterize the genetic basis of adaptation and reproductive isolation [13–19]. Patterns of genetic variation in admixed populations also provide information about past demographic events [11, 20–22]. Moreover, an accurate characterization of genetic ancestry is required when using genome-wide association or admixture mapping to identify genetic variants associated with trait variation or disease [11, 23–27]. Finally, admixture can be catalyzed by anthropogenic habitat alteration or species introductions, and can cause species collapse or extinction [28–31]. Thus, an understanding of admixture can be important for biodiversity conservation and wildlife management.

Numerous statistical methods have been developed to infer genetic ancestry from molecular data (reviewed in [12]). Early methods considered unlinked genetic markers and were primarily concerned with inference of genome-average ancestry, that is, the proportion of an individual's genome inherited from each of K potential source populations (as in the admixture model in `structure` [23]). More recently, a variety of methods have been proposed to resolve genetic ancestry into a series of continuous blocks of DNA inherited from different source populations, and thereby infer local or locus-specific ancestry along chromosomes [32–37]. Local ancestry inference can be based on population allele frequencies or haplotypes. Hidden Markov models (HMMs) are commonly used to model correlations in local ancestry along chromosome (as in the linkage HMM in `structure`), and in some cases, background linkage disequilibrium (this includes Markov-HMMs and infinite-HMMs as in `sabre` and `mspectrum`) [32, 33, 36]. Local ancestry inference can be very accurate, particularly when samples from well-defined source populations and phased DNA sequence data are available [35–37]. Methods that summarize genetic ancestry for a population or lineage also exist. These include tree-based methods used to infer population admixture proportions [7, 38] and genomic cline models, which can be used to quantify differential introgression in hybrid zones [39, 40].

Different ancestry inference methods are better suited for different tasks or under different conditions. For example, estimates of genome-average ancestry from `structure` can be used to identify recent hybrids [41], whereas tree-based methods are better able to detect ancient introgression [7, 38]. My primary focus in this paper is on local ancestry inference at the population-level. Because of recombination, genetic drift and selection in admixed populations, population local ancestry frequencies can vary across the genome [15, 39, 42–44]. In other words, local ancestry from a given source population can be more common in some regions of the genome than others. Such variation in local ancestry frequencies precedes genome stabilization during hybrid speciation [11], and has been associated with adaptation in several systems, including maize [19], humans [45], and butterflies [17, 18]. Several approaches have been used to quantify variation in local ancestry frequencies. In particular, ancestry frequencies can be inferred *post hoc* from resolved local ancestry blocks [19], or in the case of very ancient admixture, using tree-based methods [17]. Similarly, genomic cline methods can provide derived summaries of local ancestry frequencies when hybridization is an ongoing process [39, 40].

Herein, I propose and evaluate a new statistical method to estimate local ancestry frequencies. The primary motivation for this method is a desire to infer ancestry frequencies in admixed populations when ongoing gene flow from source populations is rare or absent, but

before genome stabilization is complete. Such situations exists in nature [6, 46], but do not meet the assumptions of existing methods (local ancestry inference generally assumes ancestry frequencies do not vary across the genome and tree-based methods ignore segregating variation within lineages). Moreover, analyses of ancestry frequencies in isolated admixed populations should provide novel insights about the relative roles of selection, drift and recombination in shaping genomes (e.g., [47]), and on the genetic basis of trait variation (e.g., [27]). I first describe the proposed method, which combines discriminant analysis with a continuous correlated beta process model to jointly estimate local ancestry within individuals and local ancestry frequencies at the population-level. I then assess the accuracy of the method by applying it and a traditional HMM approach to simulated data sets. The reliability and utility of the method is further demonstrated by using it to analyze genetic ancestry in an admixed human population (Uyghur), and three admixed populations from a house mouse hybrid zone. These analyses show that the method is both accurate and useful. Computer software implementing these methods (`popanc`) is available on-line at http://sourceforge.net/projects/popanc/.

Methods

Model

As a basis for the proposed statistical method, consider a model where admixture between two populations, *A* and *B*, occurs *t* generations in the past. The resulting admixed population then evolves until the present by recombination, drift, and selection, but with little or no ongoing gene flow from the source populations (as in [48, 49]). Under this model, genome-average ancestry should initially vary among individuals because of variation in the number of migrant ancestors (genealogy variance) and variation in the contribution of genetic material from each ancestor (assortment variance) (Fig 1, [21]). However, these sources of variation should decay rapidly, and be replaced by genome-wide variation in local ancestry frequencies among chromosome segments [44]. Eventually, chromosome segments will fix for local ancestry from population *A* or *B*. This process has been referred to as genome stabilization, particularly in the context of homoploid hybrid speciation [11, 50]. My current focus is on inference during the intermediate stages of this processes, that is, once variation in genome-average ancestry has mostly been removed, but while variation for local ancestry is still segregating in the admixed population (i.e., before genome stabilization; Fig 1). Once genome stabilization is complete or nearly complete, tree-based methods can be used to analyze local ancestry at a population or species-level, as intra-population variation in genetic ancestry can be ignored. However, during this intermediate period, methods are needed that allow for variable ancestry frequencies and that can account for genetic divergence in the admixed population (via drift or selection).

Most traditional methods for local ancestry inference use homogeneous HMMs or extensions of these [32, 33, 36]. HMMs are parameterized by a transition probability matrix that gives the probability of switching from one ancestry state (or ancestral haplotype) to another as one moves along a chromosome. Homogeneous HMMs assume that the transition probability matrix is constant and independent of the position in the genome. In essence, this assumes that local ancestry frequencies are the same everywhere in the genome. Although this assumption is reasonable when admixture occurred recently or when hybridization is ongoing, it becomes less tenable as progress towards genome stabilization occurs (Fig 1). This problem could be circumvented by defining a non-homogeneous HMM where the transition probability matrix varies along chromosomes to reflect variation in local ancestry frequencies, but computational methods for non-homogeneous HMMs are not well developed. Instead, my proposed method uses a continuous correlated beta process model (CCBPM; [51]) to co-estimate local ancestry

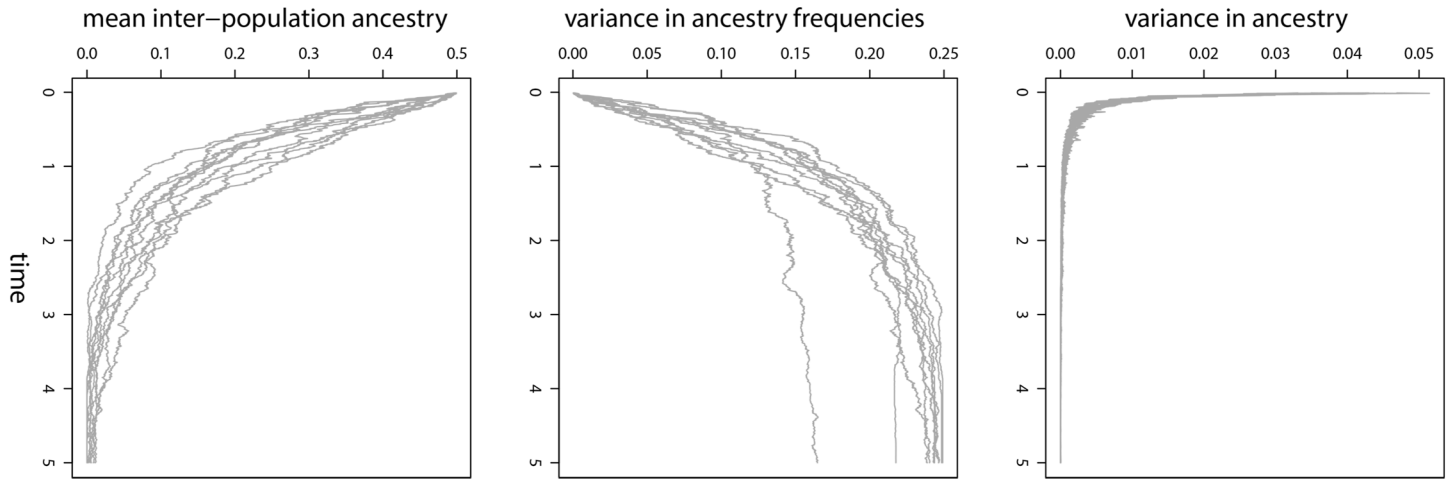

Fig 1. Variance in ancestry. Plots depict summaries of genetic ancestry from 10 replicate simulations (gray lines). Each simulation followed a Wright-Fisher model starting with an admixed population composed entirely of F1s. The population then evolved by recombination and genetic drift. Population size was constant ($2N = 200$), and simulations ran for $t = 1000$ generation (time is reported relative to populations size as $t/2N$). Ancestry was followed at 100 genetic loci that were equally spaced on a 1 Morgan chromosome. Plots show the variance in genome-average ancestry among individuals (top pane), the variance in local ancestry frequencies among loci, and the proportion of loci where individuals are expected to have one gene copy from each source population (i.e., inter-population ancestry [6]). Note that variation in ancestry frequencies persists long-after variation in genome-average ancestry has decayed.

doi:10.1371/journal.pone.0151047.g001

within individuals and population-level local ancestry frequencies while explicitly modeling the genomically autocorrelated variation in the latter. Key model parameters can be estimated in a computationally efficient way via Gibbs sampling. An additional advantage of this method relative to many HMMs is that inference does not depend on phased data. While such data might be available for humans and some model systems, sequencing strategies commonly used in non-model organisms, such as genotyping-by-sequencing (GBS) methods [18, 52, 53], generate sparse, un-phased SNP data. My goal is to develop a method that can be used in these situations, as long as a draft reference genome is available.

First, I begin with a general description of a CCBPM. Consider a series of binomial experiments where the probability of success (θ) varies from experiment to experiment. In a Bayesian framework, it would be natural to place a conjugate beta prior on the probability of success for each experiment (θ_x), and thereby obtain the posterior probability distribution for θ_x (also a beta distribution), which would be $\Pr(\theta_x|y_x, n_x, \alpha_0, \beta_0) \propto \theta_x^{y_x+\alpha_0-1}(1-\theta_x)^{n_x-y_x+\beta_0-1}$, where y_x denotes the number of successes out of n_x trials, and α_0 and β_0 are shape parameters for the beta prior. Note that the posterior is a beta distribution with shape parameters $\alpha = y_x + \alpha_0$ and $\beta = n_x - y_x + \beta_0$. Now assume that experiments are conducted one after another and that the probability of success is autocorrelated in time (or space) such that successive experiments have similar values of θ (i.e., θ_x and θ_{x+1} tend to be similar). While the model described above could still be used for inference, it would be sub-optimal as it does not provide a means to share information about θ among experiments. An alternative solution is to estimate θ using a CCBPM [51], which is a graphical model that generalizes the Bayesian model above by allowing for information sharing among experiments. A kernel function $K(x, x')$ dictates the extent that information is shared among experiments. Different kernel functions are possible, but the kernel should be a decreasing function of the time or distance between a pair of experiments. Under this model an approximate posterior distribution for each θ_x can be generated by drawing samples from $\Pr(\theta_x|\mathbf{y}, \mathbf{n}, \alpha_0, \beta_0) = \text{beta}(\alpha = (\Sigma_i\, y_i\, k(x, i)) + \alpha_0, \beta = (\Sigma_i(n_i - y_i)k(x, i)) + \beta_0)$.

As the general description above makes clear, a CCBPM can naturally be used to model local ancestry frequencies in an admixed population. In particular, at each locus (x) the number of gene copies with ancestry from source population A (z_x) in a sample of n_x diploid individuals can be modeled as the outcome of a binomial experiment with the probability of success given by the population ancestry frequencies at that locus (q_x and $1 - q_x$). Because of recombination and the chromosomal nature of inheritance, ancestry frequencies should be autocorrelated along chromosomes. The CCBPM is used to share information across linked loci when estimating each q_x. Thus, the approximate conditional posterior distribution for each q_x is,

$$\begin{aligned}\Pr(q_x|\mathbf{z}, \mathbf{2n}, \alpha_0, \beta_0) &= \text{beta}\left(\alpha = \left(\sum_i z_i k(x,i)\right) + \alpha_0, \beta = \left(\sum_i (2n_i - z_i)k(x,i)\right) + \beta_0\right) \\ &\propto q_x^{\alpha_0 - 1 + \sum_i z_i k(x,i)} (1 - q_x)^{\beta_0 - 1 + \sum_i (2n_i - z_i)k(x,i)}.\end{aligned} \tag{1}$$

Here I use the squared exponential kernel $k(x, i) = \exp\left(\frac{-(x-i)^2}{\sigma}\right)$, where σ is a scale parameter. While the scale parameter could be fixed, I instead propose to place an uniform prior on this parameter and estimate it from the data (though my main aim is to integrate over uncertainty in this nuisance parameter rather than make inferences about it).

In this model, local ancestry ($\mathbf{z}$) is not an observed quantity, but instead represents a latent variable that must be inferred from the data. z_x can be decomposed as $z_x = \sum_j z_{xj}$ where the sum is over individuals and $z_{xj} \in \{0, 1, 2\}$ denotes the number of gene copies at locus x that individual j inherited from source population A. Following Bayes' rule, a posterior distribution for z_{xj} can be specified as,

$$\Pr(z_{xj}|s_{xj}, q_x) \propto \Pr(s_{xj}|z_{xj})\Pr(z_{xj}|q_x) \tag{2}$$

where s_{xj} is the DNA sequence data (discussed more below). The first term on the right side of Eq (2) represents the probability of the observed sequence data for individual j conditional on that individual having 0, 1 or 2 gene copies derived from source population A, and the last term is the prior probability of inheriting z_{xj} gene copies from source population A. The latter is clearly given by,

$$\Pr(z_{xj}|q_x) \sim \text{binomial}(q_x, 2), \tag{3}$$

but specifying a probability distribution for the first term (i.e., the likelihood of z_{xj} given the data) is more complicated. Thus, I will describe my approach for specifying $\Pr(s_{xj}|z_{xj})$ in detail. Because of segregating allelic variation in the source populations and drift (or selection) in the admixed and source populations, sequence data from any particular nucleotide variant (i.e., SNP) can be rather uninformative about local ancestry. To overcome this limitation and to model expected autocorrelations in local ancestry within individuals, I propose to approximate $\Pr(s_{xj}|z_{xj})$ using discriminant analysis (DA; a related approach was used by [54] for global ancestry inference).

Here DA is used to provide different weights to different SNPs such that they are maximally informative about local ancestry. The analysis proceeds one SNP or locus at a time. First, a window is defined around each genetic locus; a window includes a specific number of neighboring SNPs (this could be constrained by physical or recombination distance). Choosing a reasonable window size can be important (this is discussed more in the Discussion). In particular, larger windows will contain more information about local ancestry, but if windows become too large they will frequently span ancestry breakpoints, which is undesirable (see e.g., [34, 55]). Also, the window cannot include more SNPs than reference individuals (i.e., the number of

observations must exceed the number of variables). DA is then used to generate a discriminant function and thereby distinguish between individuals with 0, 1 or 2 gene copies from source population *A* for each window (i.e., to infer z_{xj}). A set of reference samples from each source population (i.e., populations *A* and *B*) is required to generate the discriminant function. These represent $z_x = 0$ (source *B*) and $z_x = 2$ (source *A*). Reference samples with one gene copy from each source population can then be simulated to represent $z_x = 1$. My implementation of DA then proceeds as follows. Let S_x be a $N \times P$ matrix with reference individuals as rows and the genotypic data from the set of genetic variants within a window as columns. Here the genotypic data are centered counts of one of the two alleles for each SNP. The within group covariance matrix (S_w) is then calculated as,

$$S_w = \frac{\sum_g (n_g - 1) s_{xg}^T s_{xg}}{\sum_g n_g - 3} \tag{4}$$

where the summation is over the three groups (i.e., samples with 0, 1 or 2 gene copies from source population *A*), n_g is the sample size for group *g*, and s_{xg} is the sub-matrix containing only individuals from group *g*. Next, the between group scatter matrix (S_B) is obtained as,

$$S_v = \frac{1}{3} \sum_g (\mu_g - \mu)(\mu_g - \mu)^T \tag{5}$$

where μ_g and μ are the group and grand means of s_{xg} and S_x, respectively. Eigenvalue decomposition of the canonical matrix $S_w^{-1} S_b$ can then be used to obtain the discriminant function. Specifically, the eigenvector associated with the largest eigenvalue of the canonical matrix contains the discriminant coefficients. These coefficients can be used to calculate a discriminant score for each reference sample and thereby transform the reference samples onto a new space that maximizes the genetic differences among the three groups relative to within group variation. Note that only the first discriminant function is required to separate the three groups because the $z_x = 1$ reference group is intermediate between the two other groups.

The mean and variance of the discriminant scores for the reference individuals from each group are used to define $\Pr(s_{xj}|z_{xj})$ such that,

$$\Pr(s_{xj}|z_{xj}) = \Pr\Big(d_{xj} = f(s_{xj})|z_{xj}\Big) = \text{normal}\Big(\mu = \bar{d}_{xg=z_x}, \sigma^2 = \text{var}(d_{xg=z_x})\Big), \tag{6}$$

where $\bar{d}_{xg=z_x}$ and $var(d_{xg\,=\,z_z})$ are the mean and variance of the discriminant scores for the set of reference samples with ancestry z_x, $f(s_{xj})$ is the discriminant function, and d_{xj} is the discriminant score for an individual with unknown ancestry. Thus, after transforming the admixed individuals onto the new sample space with the discriminant function developed from the reference set, Eq (6) can be used to calculate the probability of the sequence data (or more precisely the probability of the discriminant score based on the sequence data) if the individual has 0, 1 or 2 gene copies from source population *A*.

A computer program (`popanc`), written in `C++`, has been developed to generate parameter estimates from the model described above. The program first performs the DA using linear algebra functions provided by the GNU Scientific Library [56]. Markov chain Monte Carlo (MCMC) is then used to obtain samples from the approximate posterior distributions for each of the model parameters (this is an approximation because Eq (1) represents a process-model generalization of Bayes' rule [51]). MCMC includes Gibbs samplers for population ancestry frequencies (**q**) based on Eq (1) and individual local ancestry (**z**) based on Eqs (3) and (6). A Metropolis update is performed for the scale parameter σ. HDF5 is used for efficient storage and processing of the MCMC samples [57].

Simulations and data analysis

Multiple data sets were simulated and analyzed to assess the efficacy of the proposed method. Simulations were individual-based, and tracked ancestry segments rather than genetic markers (as in [11, 39, 40]). In each simulation an admixed population of N diploid individuals was initiated via hybridization between two source populations, A and B. This was followed by t discrete generations, where mating occurred within the admixed population with or without ongoing gene flow (m). Genomes consisted of 2, 1 Morgan chromosomes (i.e., 1 recombination event occurred per chromosome each generation). All simulations included genetic drift and some also included natural selection (details below). At the end of each simulation, 50 individuals were sampled from each source population and the admixed population, and genetic marker (SNP) data were generated for these individuals. Allele frequencies in the source populations were drawn from independent uniform distributions bounded by 0.05 and 0.95. Genotypes were then obtained by randomly sampling from the population allele frequencies. A similar procedure was used to generate genotypes for the admixed individuals, except alleles were sampled based on the source population allele frequencies and local ancestry. Moreover, allele frequencies within ancestry segments were modified to account for genetic drift (genetic drift affects ancestry frequencies, and also allele frequencies within ancestry segments). Specifically, for each genetic marker and ancestry type (source A and B) a new allele frequency was sampled from beta($\alpha = p\gamma$, $(1-p)\gamma$), where $\gamma = -1\frac{F-1}{F}$ and $F = -\exp\left(-\frac{t}{N}\right)\left(\exp\left(\frac{t}{N}\right) - 1\right)$ [58, 59]. Evolutionary dynamics depend on the ratio of the time since admixture and the population size (i.e., $\frac{t}{N}$). Allowing drift to affect ancestry and allele frequencies (rather than just ancestry frequencies) is realistic and important as it makes local ancestry inference considerably more difficult.

A series of data sets was simulated to determine how time since admixture affects the accuracy of ancestry frequency estimates. Ten replicate data sets were simulated with an admixed population size of $N = 500$ and $t = 20$, 50, or 200 generations since admixture ($\frac{t}{N} = 0.04$, 0.10, or 0.40) with $m = 0$ (no ongoing gene flow). Another series of simulations was used to quantify the effect of selection on ancestry inference. Here, each individual's fitness was determined by its ancestry at L loci, with individuals having mixed ancestry at these loci suffering reduced fitness (i.e., underdominance was assumed). Fitness was multiplicative such that $w_j = (1-s)^{l_j}$, where w_j is the relative fitness of individual j and s is the selection coefficient, and l_j is the number of loci (out of L) where individual j has one allele copy from each source population. Ten replicate data sets were generated with diffuse ($s = 0.03$ and $L = 20$ with 10 underdominant loci per chromosome) or strong ($s = 0.3$ and $L = 2$ with both underdominant loci on the same chromosome) selection. A third series of data sets was simulated with low ($m = 0.005$) or high ($m = 0.05$) rates of ongoing gene flow following the initial admixture event. Here m denotes the proportion of the admixed population composed of immigrants each generation. Ten replicate data sets were simulated for each migration rate and $t = 200$ generations with no selection.

The proposed method was then used to estimate population ancestry frequencies for each simulated data set. Data sets were analyzed using local ancestry windows that included 4, 10, or 20 SNPs on either side of the focal SNP. Two MCMC runs were conducted for each data set and window size, each with a 10,000 iteration burn-in, 30,000 total MCMC steps and a thinning interval of 5. Likely convergence to the stationary and adequate MCMC mixing were evaluated by calculating the Gelman and Rubin's potential scale reduction factor and the effective sample size for each parameter. As a comparison, each data set was also analyzed using the correlated local ancestry HMM (i.e., linkage model) implemented in `structure` (version 2.3.4 [32]). Unlike many local ancestry inference approaches, the linkage model in `structure` can analyze un-phased genotypic data, and thus represents a valid comparison to the proposed

method. Posterior estimates of local ancestry were based on two MCMC runs each with 10,000 iterations for sampling and 10,000 iteration burn-ins. Reference samples were specified, and were used as the sole source of information to infer source population allele frequencies. Population ancestry frequencies were then inferred *post hoc* from local ancestry estimates by equating the sample mean with the population ancestry frequency.

Approximate posterior distributions for population ancestry frequencies from the proposed CCBPM were summarized by calculating the posterior mean (point estimate) and 95% equal-tail probability intervals (ETPIs). Discrepancies between true ancestry frequencies and estimates from both the proposed method (posterior mean) and the linkage model in `structure` were quantified by calculating the root-mean-square deviation (RMSD). The coefficient of variation (CV) of the RMSD was then obtained by dividing the RMSD by the true parameter value. In addition, the CVRMSD was calculated for a null model where genome-average ancestry (calculated as the average of the local ancestry frequencies) was used as the estimate of the ancestry frequency for each locus. Model adequacy was also assessed by determining the frequency with which the true population ancestry frequency was included in the 95% ETPIs and by calculating the correlation between the true and estimated parameter values.

Application to human genetic data

The proposed method was also applied to the Uyghur, which are a human population in Xinjiang, China known to be historically admixed with western Eurasian and Asian ancestry [7, 60]. Published results suggest that admixture in this group occurred approximately 790 (±60) years ago, or about 27 generations ago (assuming 29 years per generation, [7]; an alternative approach suggests admixture occurred 126 generations ago [60]). This places the time since admixture within the realm where the proposed methods should be applicable. Previous results indicate that the population admixture proportion for Uyghur (i.e., genome-average ancestry at the population-level) is 45.2% to 52.5% west Eurasian [7, 60, 61].

The data analyzed here come from the curated version of the Harvard HGDP-CEPH Genotypes for Population Genetics Analyses Supplement 10 that was released with `admixtools` [7]. Han (32 individuals) and French (26 individuals) were chosen as source populations for Uyghur (as suggested by [7]). The full data set of 621,038 SNPs was filtered to retain only those SNPs that were variable in the source populations with a minor allele frequency greater than 0.1, and to discard tightly linked, redundant SNPs (every third SNP was retained). This left a data set of 116,871 SNPs across the 22 human autosomes, which I analyzed with the proposed CCBPM. Each chromosome was analyzed separately for computational efficiency. Parameter estimates were obtained from two MCMC runs, each consisting of 30,000 iterations, a 10,000 iteration burn-in and a thinning interval of 5. I used a 15 SNP window, which previous results suggest should rarely span ancestry breakpoints (average ancestry block lengths are 2.43 to 4.07 cM [61]).

Application to a mouse hybrid zone

The house mouse species *Mus domesticus* and *M. musculus* diverged about 500,000 years ago [62, 63], but now hybridize along a narrow hybrid zone in central Europe that formed a few thousand years ago [64]. Weak assortative mating and reduced hybrid fertility, particularly in males, limit gene flow across the hybrid zone, although the severity of each varies among populations and individuals (e.g., [65–67]). Because this hybrid zone is wide relative to dispersal distance, the hybrid zone consists of numerous admixed populations that differ in their genome composition but exhibit relatively little variation in genome-average ancestry within populations [67, 68]. Given these low dispersal rates (i.e., limited ongoing gene flow), ancestry

frequencies as inferred from the proposed CCBPM should be a useful summary of genetic ancestry in these admixed populations. I focused on three admixed populations from the Bavarian transect through this hybrid zone, which differ in the genomic contribution of *M. domesticus* (as measured by a hybrid index, h): Tütenhausen (TU: $\bar{h} = 0.27$, range = [0.18 – 0.36]), Haindlfing (HA: $\bar{h} = 0.25$, range = [0.17 – 0.38]), and Neufahrn bei Freising (FS: $\bar{h} = 0.58$, range = [0.45 – 0.68]) [16, 67, 69].

The data analyzed here are from captive-bred first generation offspring of wild-caught mice (only mice with both parents from a single locality were included; [69]); this included 21 mice from TU, 32 from HA, and 31 from FS. Smaller sample sizes were available for source (i.e., reference) populations: five *M. domesticus* from SO and ST (Pelka and Pallhausen) and five *M. musculus* from GL, RE and RF (Emling, Neufahrn bei Erding, and Finsingermoos) [67, 69]. The genetic data comprised 93,699 SNPs (from the Mouse Diversity Genotyping Array) from the 23 autosomal chromosomes (SNPs with very high LD were not included in the data set [69]). I analyzed each admixed populations separately using the proposed CCBPM. Parameter estimates were obtained from two MCMC runs, each consisting of 30,000 iterations, a 10,000 iteration burn-in and a thinning interval of 5. A window size of 4 SNPs was used because of the age of the hybrid zone and to account for the low parental sample sizes (i.e., to ensure that the sample size exceeded the number of variables for the DA). A key strength of this data set was that it allowed me to contrast patterns of variation in ancestry frequencies across multiple admixed populations, and thus ask about the consistency of these patterns.

Results

Performance on simulated data

Variation in local ancestry increased with time and selection, as expected (because dynamics depend on $\frac{t}{N}$, variation in ancestry would also scale with population size, but this was kept constant; [21, 44]). Specifically, ancestry frequencies varied most across the genome in the 200 generation simulations, followed by the 50 generation simulations with diffuse and strong selection; gene flow reduced variation in ancestry frequencies (Fig 2). However, the simulated populations were generally segregating for local ancestry across the genome. In particular the mean proportion of genetic loci where ancestry from one source population reached fixation varied from 0.0% (20 and 50 generation simulations without selection or with diffuse selection) to 0.3% (s.d. = 0.2%; 200 generation simulations) or 2.3% (s.d. = 1.5%; 50 generation simulations with strong selection).

Accuracy was affected by window size and time since admixture, such that ancestry frequencies were estimated more accurately with larger windows. And there was a greater correlation between true and estimated ancestry frequencies with more time since admixture, but the CVRMSD was lower when admixture occurred more recently (e.g., t = 20, ±4 SNPs, r = 0.62, CVRMSD = 0.16 vs. t = 20, ±20 SNPs, r = 0.77, CVRMSD = 0.10 vs. t = 200, ±20 SNPs, r = 0.89, CVRMSD = 0.20; Figs 3, 4, & 5; Table 1). The HMM in `structure` performed similarly to or slightly better than the proposed method with a window size of ±4 SNPs, but the proposed CCBPM out-performed this method when windows of ±10 or ±20 SNPs were used, particularly when admixture occurred 200 generations ago (Fig 4; Table 1). Both the proposed CCBPM and the `structure` HMM were considerably more accurate than the null model of no ancestry frequency variation (Fig 4). Time since admixture and window size also affected coverage of the true parameter value by the 95% ETPIs, with better coverage for more recent admixture and smaller windows (e.g., t = 20, ±4 SNPs, coverage = 64% vs. t = 200, ±4 SNPs, coverage = 52% vs. t = 200, ±20 SNPs, coverage = 41%).

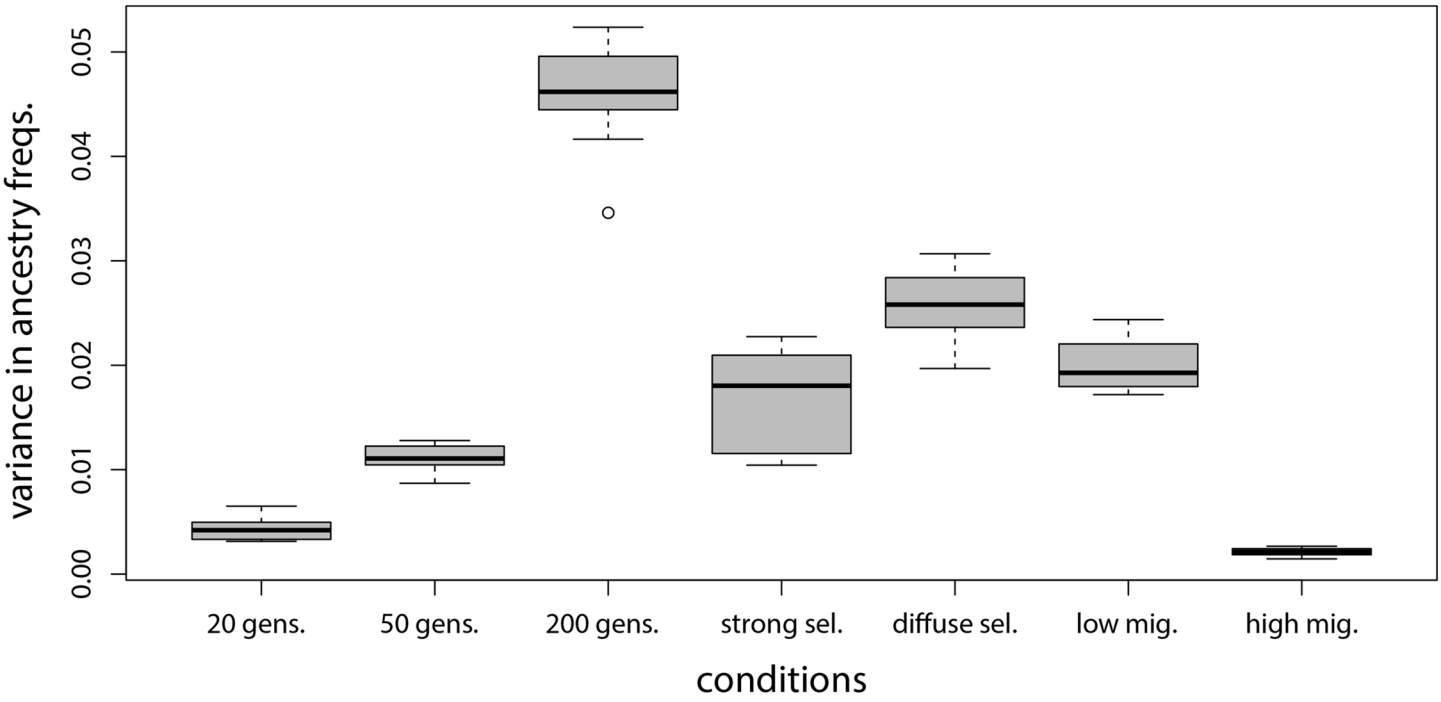

Fig 2. Variation in ancestry frequencies. Boxplots summarize the variance in local ancestry frequencies calculated across 20,002 genetic loci and 10 replicate simulations. This includes 20, 50 and 200 generation simulations without selection, 50 generation simulations with strong or diffuse selection, and 200 generations with low (m = 0.005) or high (m = 0.05) migration.

doi:10.1371/journal.pone.0151047.g002

Results with selection were generally similar, but the improved performance of the proposed CCBPM relative to the linkage HMM in `structure` was more evident (Figs 6 & 7; Table 1). Once again, more accurate estimates were obtained with larger SNP windows. Parameter estimates were more strongly correlated with their true values when selection was strong (e.g., r: diffuse selection, ±4 SNPs = 0.77, ±20 SNPs = 0.91; strong selection, ±4 SNPs = 0.83, ±20 SNPs = 0.94; Fig 3; Table 1), but the CVRMSD was lower when selection was diffuse (e.g., CVRMSD: diffuse selection, ±4 SNPs = 0.17, ±20 SNPs = 0.097; strong selection, ±4 SNPs = 0.21, ±20 SNPs = 0.12; Fig 6; Table 1). Nearly identical results were obtained when the SNPs nearest to each selected locus were removed from the analysis (performance metrics were indistinguishable from those in Table 1), thus the improved performance of the CCBPM was not driven by these loci but rather by the overall effect of selection on variation in ancestry frequencies across the genome.

A low rate of ongoing gene flow after admixture (i.e., $m = 0.005$) did not noticeably degrade the performance of the CCBPM in general or relative to the linkage HMM in `structure` (Figs 3, 8, & 9; Table 1). However, parameter estimates were less strongly correlated with their true values when ongoing gene flow occurred at a higher rate (r: ±4 SNPs = 0.51, ±10 SNPs = 0.62, ±20 SNPs = 0.66). But, even under these conditions, ancestry estimates from the CCBPM were more strongly correlated with their true values than the estimates from the HMM, particularly when a window size of ±10 or ±20 was used (Fig 3; Table 1).

Admixture in humans

The genome-average ancestry frequencies in the Uyghur population were 52.9% Han and 47.1% French, which is consistent with [7]. Average ancestry frequencies varied modestly

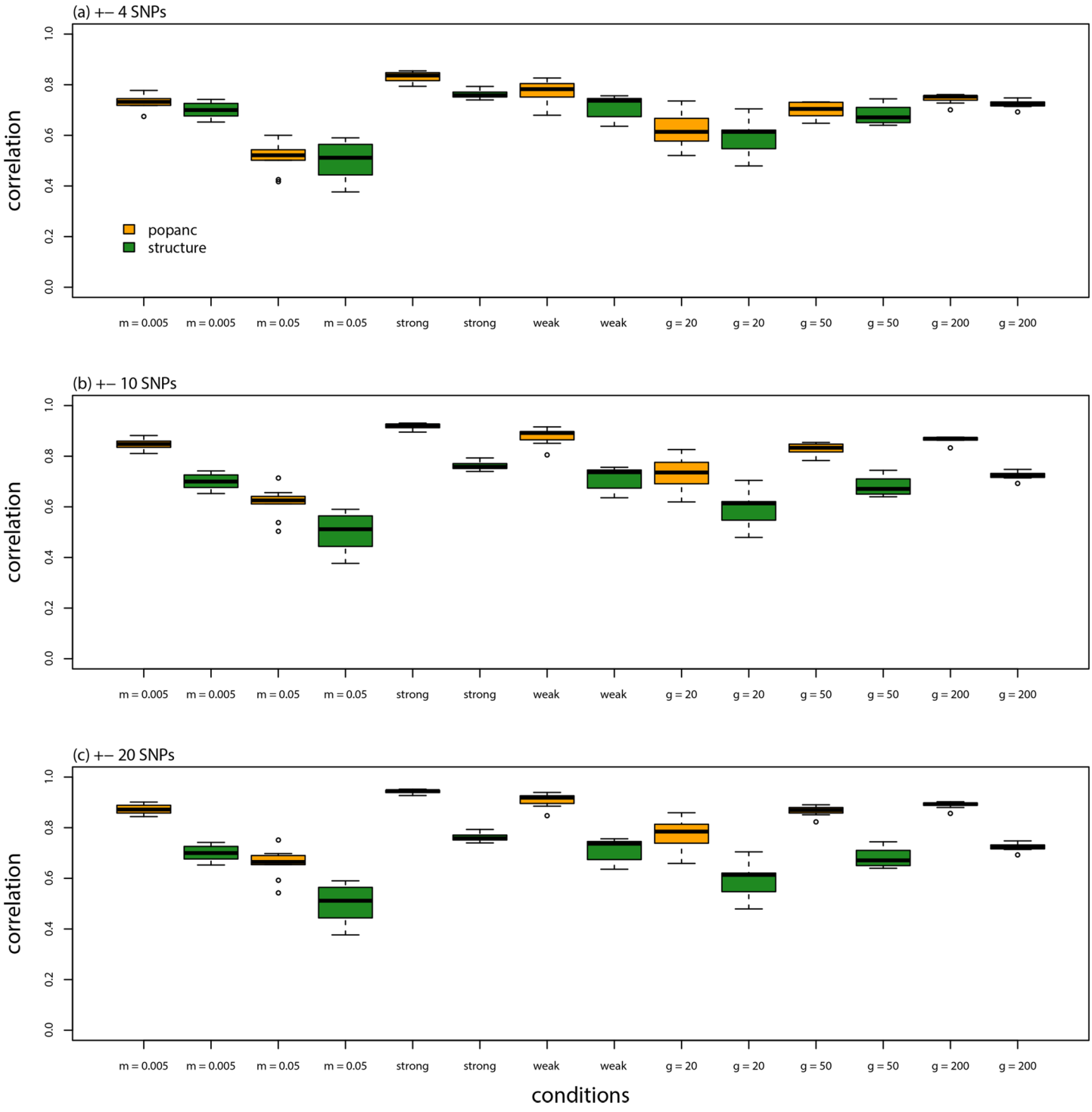

Fig 3. Model evaluation. Boxplots show the distribution of correlations (Pearson correlation coefficient) between true and inferred ancestry frequencies from each of 10 replicate simulations analyzed with different methods, window sizes (for `popanc`), generations since admixture, selection regimes, and migration rates.

doi:10.1371/journal.pone.0151047.g003

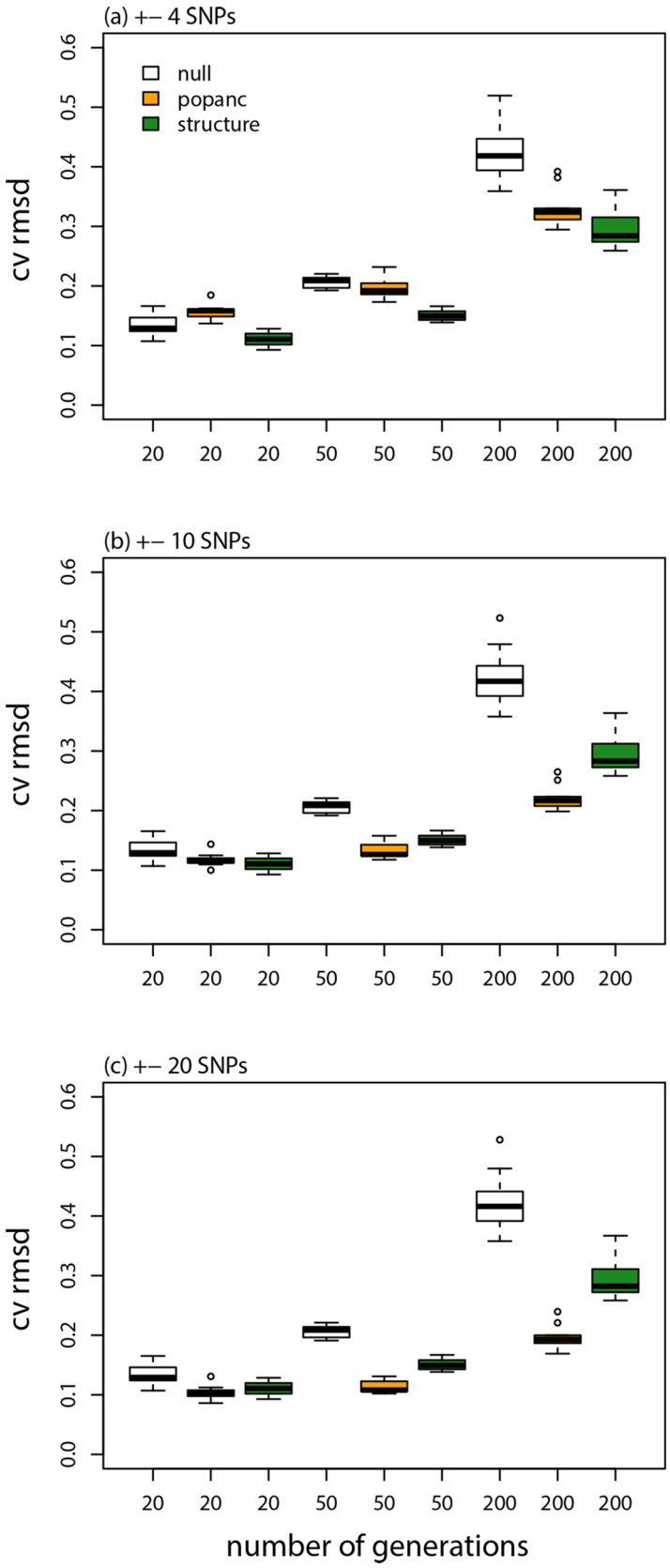
(a) +– 4 SNPs
null
popanc
structure
cv rmsd
0.0
0.1
0.2
0.3
0.4
0.5
0.6
20
20
20
50
50
50
200
200
200
(b) +– 10 SNPs
cv rmsd
0.0
0.1
0.2
0.3
0.4
0.5
0.6
20
20
20
50
50
50
200
200
200
(c) +– 20 SNPs
cv rmsd
0.0
0.1
0.2
0.3
0.4
0.5
0.6
20
20
20
50
50
50
200
200
200
number of generations

Fig 4. Model evaluation. Boxplots show the average CVRMSD for local ancestry frequencies from each of 10 replicate simulations analyzed with different methods, generations since admixture, and window sizes (for `popanc`). The CVRMSD for a null model that equates the ancestry frequencies at each locus with the genome-average is also shown.

doi:10.1371/journal.pone.0151047.g004

among chromosomes, ranging from 47.0% Han on chromosome 18 to 58.6% Han on chromosome 20. Even more variation in ancestry frequencies was observed within chromosomes (average ancestry frequency variance within chromosomes = 0.016, s.d. = 0.003; Fig 10). Indeed, for 55.1% of loci, the chromosome-average ancestry frequency was outside of the local ancestry frequency. Although no regions of fixed ancestry were detected, ancestry frequencies as high as 96.1% Han or 95.6% French were observed.

Genetic loci with the greatest excess or deficit of Han or French ancestry, that is, the 118 SNPs below the 0.05^{th} (9.0% Han) or above the 99.95^{th} (92.7% Han) empirical quantile for ancestry frequencies, were analyzed in more detail. These 118 SNPs formed three contiguous genetic regions with excess Han ancestry (two on chromosome 2 and one on chromosome 10) and four contiguous regions with excess French ancestry (one on chromosome 2 and three on chromosome 11; mean size of region = 408.9 kbp; Table 2). Most of these genetic regions contain one or more known genes. For example, a region of excess French ancestry on chromosome 11 includes *bridging integrator 1* (*BIN1*), and variation in this gene has repeatedly been associated with Alzheimer's disease [70–72]. We also found evidence consistent with the hypothesis of selection on this gene in the French: F_{ST} between French and Han was elevated for the two SNPs in BIN1 relative to the rest of chromosome 2 (F_{ST} for BIN1 = 0.23, mean for chromosome 2 = 0.073, randomization test, $p = 0.0617$), and genetic diversity was significantly reduced in the French population (heterozygosity for BIN1 = 0.00, mean for chromosome 2 = 0.37, randomization test, $p < 0.0001$).

Mouse hybrid zone

Genome-average ancestry frequencies varied among admixed mouse populations from 19.0% (s.d. = 9.5%) and 21.9% (s.d. = 10.4%) *M. domesticus* in TU and HA to 35.6% (s.d. = 13.7%) *M. domesticus* in FS (Fig 11). This differed somewhat from, but were correlated with previous estimates of hybrid index which were based on different reference samples and different SNP markers [67]. As with the Uyghur, average ancestry frequencies differed substantially among chromosomes in each admixed population (ranges: TU = 13.8% *M. domesticus* to 22.5%, HA 18.6% to 26.9%, FS = 25.1% to 45.4%). In each population, some genetic loci were fixed or nearly fixed for *M. musculus* ancestry (number of genetic loci with *M. domesticus* ancestry frequencies < 1%: TU = 202, HA = 96, FS = 2), while maximum frequencies of *M. domesticus* ancestry were high, but always less than one (TU = 84.5%, HA = 76.8%, FS = 92.1%).

Ancestry frequencies were correlated between pairs of admixed populations (TU × HA: $r = 0.47$, $p < 0.001$; TU × FS: $r = 0.33$, $p < 0.001$; HA × FS: $r = 0.33$, $p < 0.001$; Fig 11). Moreover, several of the same genetic loci had high *M. domesticus* ancestry in more than one admixed population. For example, considering the 0.5% of SNPs with the highest *M. domesticus* frequency in each admixed population (469 SNPs), 173 high *M. domesticus* ancestry SNPs were shared between two populations and six were shared among all three admixed populations. This is significantly more overlap than would be expected under a null hypothesis of independence among populations (null for two pops.: expected = 7.4, ×-fold enrichment [observed/expected] = 23.3×, $p < 0.001$; null for all three pops." expected < 0.01, ×-fold

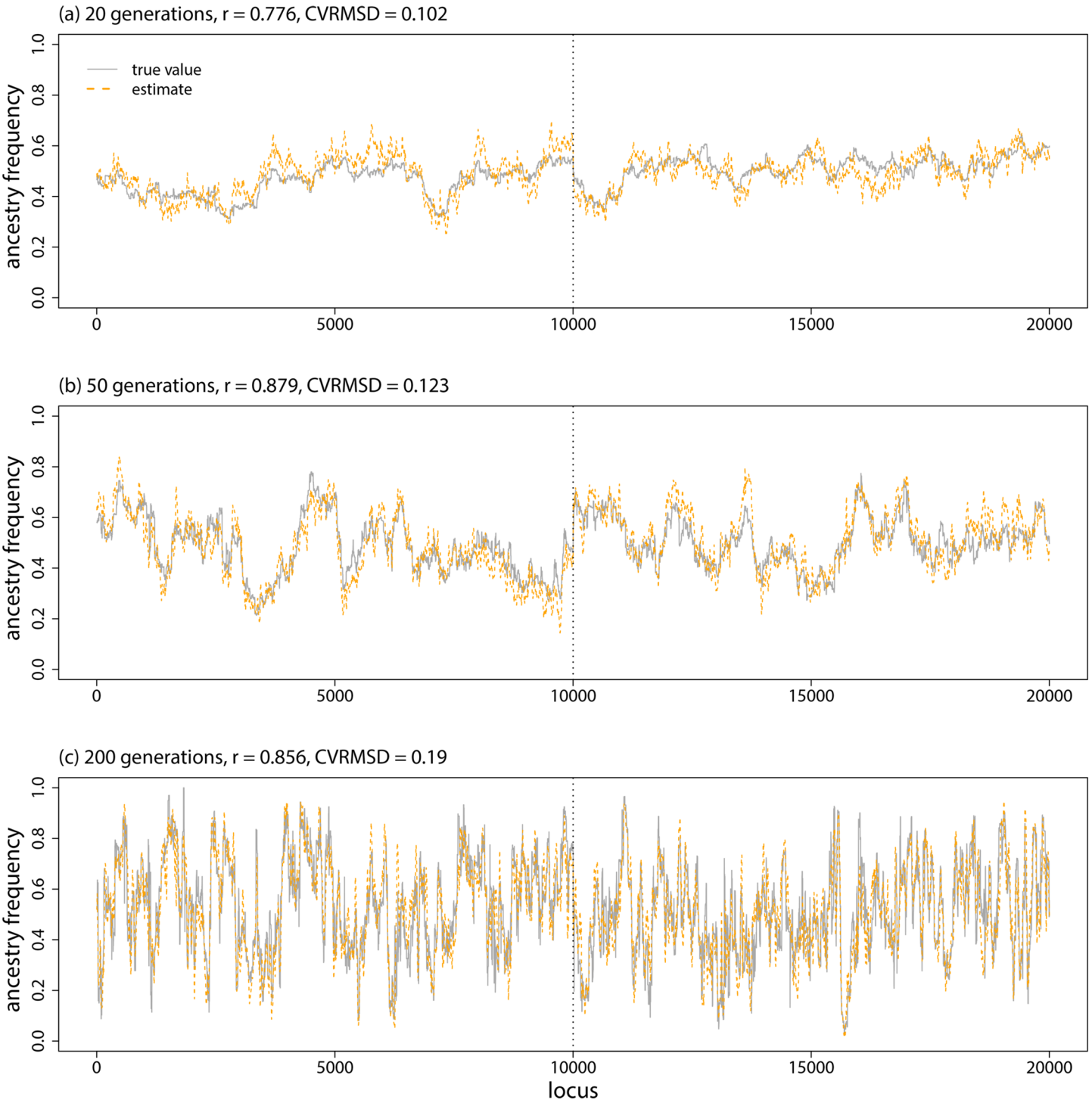

Fig 5. Genetic ancestry estimates from simulated data. Plots show the true and inferred (posterior mean from `popanc`) ancestry frequencies from one representative data set from the 20, 50, and 200 generation simulations. Results with a window size of ±20 SNPs are shown. The dashed vertical lines delineate the two distinct chromosomes.

doi:10.1371/journal.pone.0151047.g005

Table 1. Performance of methods on simulated data. Results shown for local ancestry frequencies estimates in 20, 50, or 200 generation (gens.) simulations with or without selection or ongoing gene flow and analyzed with the proposed method (`popanc`) or the HMM in `structure`: Correlation between true and estimated parameter (correlation), normalized RMSD (CVRMSD), and proportion of cases where the true value was included in the 95% ETPIs (95$ ETPI cov.).

gens.	selection	migration	method	correlation (*r*)	CVRSMD	95% ETPI cov.
20	none	none	`popanc` ±4 SNPs	0.62	0.16	0.64
20	none	none	`popanc` ±10 SNPs	0.73	0.12	0.56
20	none	none	`popanc` ±20 SNPs	0.77	0.10	0.60
20	none	none	`structure` HMM	0.60	0.11	NA
50	none	none	`popanc` ±4 SNPs	0.70	0.20	0.61
50	none	none	`popanc` ±10 SNPs	0.83	0.13	0.51
50	none	none	`popanc` ±20 SNPs	0.87	0.11	0.56
50	none	none	`structure` HMM	0.68	0.15	NA
200	none	none	`popanc` ±4 SNPs	0.74	0.33	0.52
200	none	none	`popanc` ±10 SNPs	0.87	0.22	0.46
200	none	none	`popanc` ±20 SNPs	0.89	0.20	0.41
200	none	none	`structure` HMM	0.72	0.29	NA
50	diffuse	none	`popanc` ±4 SNPs	0.77	0.17	0.61
50	diffuse	none	`popanc` ±10 SNPs	0.88	0.11	0.51
50	diffuse	none	`popanc` ±20 SNPs	0.91	0.10	0.56
50	diffuse	none	`structure` HMM	0.72	0.16	NA
50	strong	none	`popanc` ±4 SNPs	0.83	0.21	0.60
50	strong	none	`popanc` ±10 SNPs	0.92	0.14	0.48
50	strong	none	`popanc` ±20 SNPs	0.94	0.12	0.53
50	strong	none	`structure` HMM	0.76	0.23	NA
200	none	0.005	`popanc` ±4 SNPs	0.73	0.24	0.58
200	none	0.005	`popanc` ±10 SNPs	0.85	0.16	0.47
200	none	0.005	`popanc` ±20 SNPs	0.87	0.15	0.48
200	none	0.005	`structure` HMM	0.70	0.21	NA
200	none	0.05	`popanc` ±4 SNPs	0.51	0.14	0.67
200	none	0.05	`popanc` ±10 SNPs	0.62	0.11	0.61
200	none	0.05	`popanc` ±20 SNPs	0.66	0.10	0.64
200	none	0.05	`structure` HMM	0.50	0.09	NA

doi:10.1371/journal.pone.0151047.t001

enrichment = 1500×, $p < 0.001$). However, because of gene flow, strict independence would not be expected anyway. A similar analysis was not conducted for SNPs with excess *M. musculus* ancestry, because TU and HA had high frequencies of *M. musculus* ancestry overall and were fixed or nearly fixed for *M. musculus* ancestry at dozens of SNPs (as described in the previous paragraph). Consistent results were found when considering the 0.05% SNPs with the greatest excess of *M. domesticus* ancestry (47 SNPs above the 99.95th empirical quantile) in each population (Table 3). While none of these were shared among populations, neighboring regions of excess *M. domesticus* were detected in the three populations (chromosome 1, 21, 910–22, 026 kb in TU, 20, 0825–21, 726 kb in FS, and 24,848–25, 420 kb in HA). Moreover, regions with the greatest excess *M. domesticus* ancestry in any one population had elevated *M. domesticus* ancestry frequencies in the others (randomization test, TU SNPs: mean in other pops. = 0.455, ratio of observed to null expectation = 1.58×, $p = 0.002$; HA SNPs: mean in other pops. = 0.417, ratio of observed to null expectation = 1.54×, $p = 0.012$; FS SNPs: mean in other pops. = 0.392, ratio of observed to null expectation = 1.91×, $p < 0.001$).

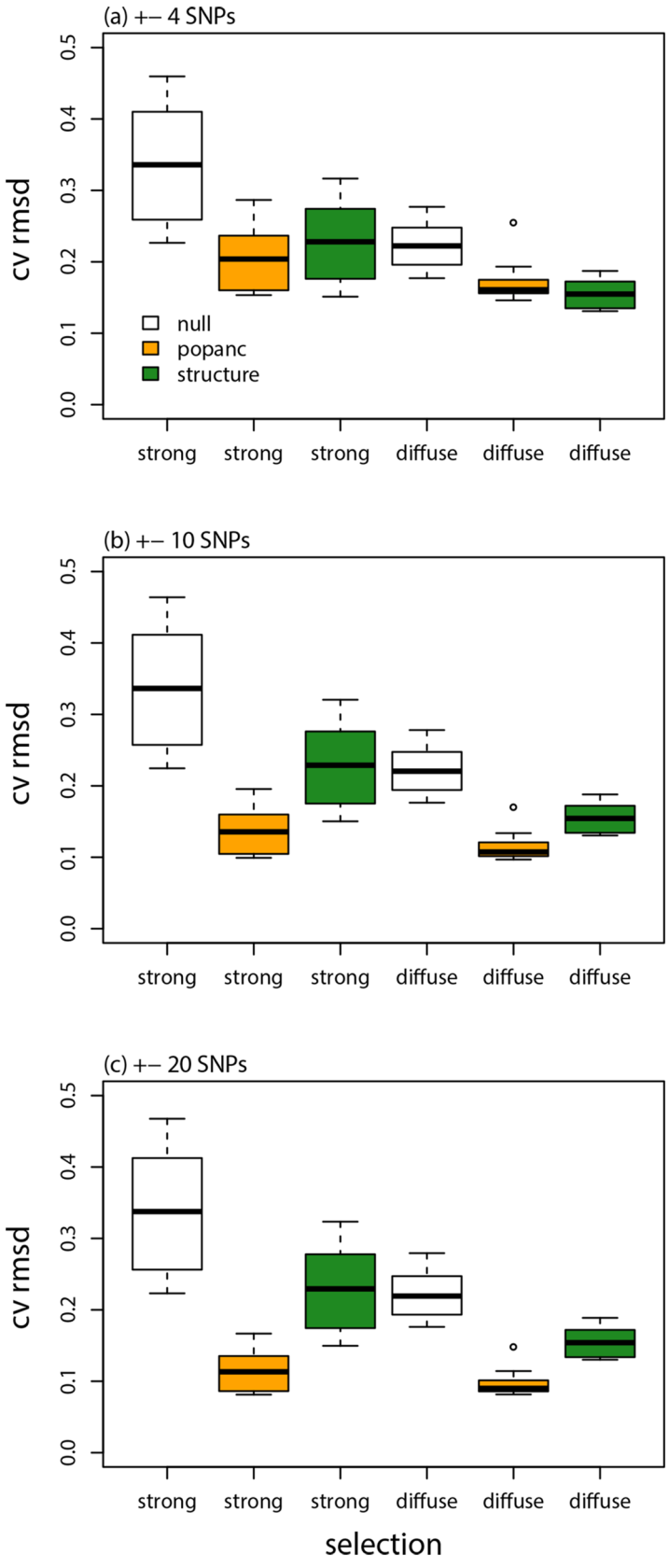
(a) +– 4 SNPs
(b) +– 10 SNPs
(c) +– 20 SNPs
cv rmsd
0.0
0.1
0.2
0.3
0.4
0.5
null
popanc
structure
strong
strong
strong
diffuse
diffuse
diffuse
selection

Fig 6. Model evaluation with selection. Boxplots show the average CVRMSD for local ancestry frequencies for each of 10 replicate simulations analyzed with different methods, selection regimes, and window sizes (for `popanc`). The CVRMSD for a null model that equates the ancestry frequencies at each locus with the genome-average is also shown.

doi:10.1371/journal.pone.0151047.g006

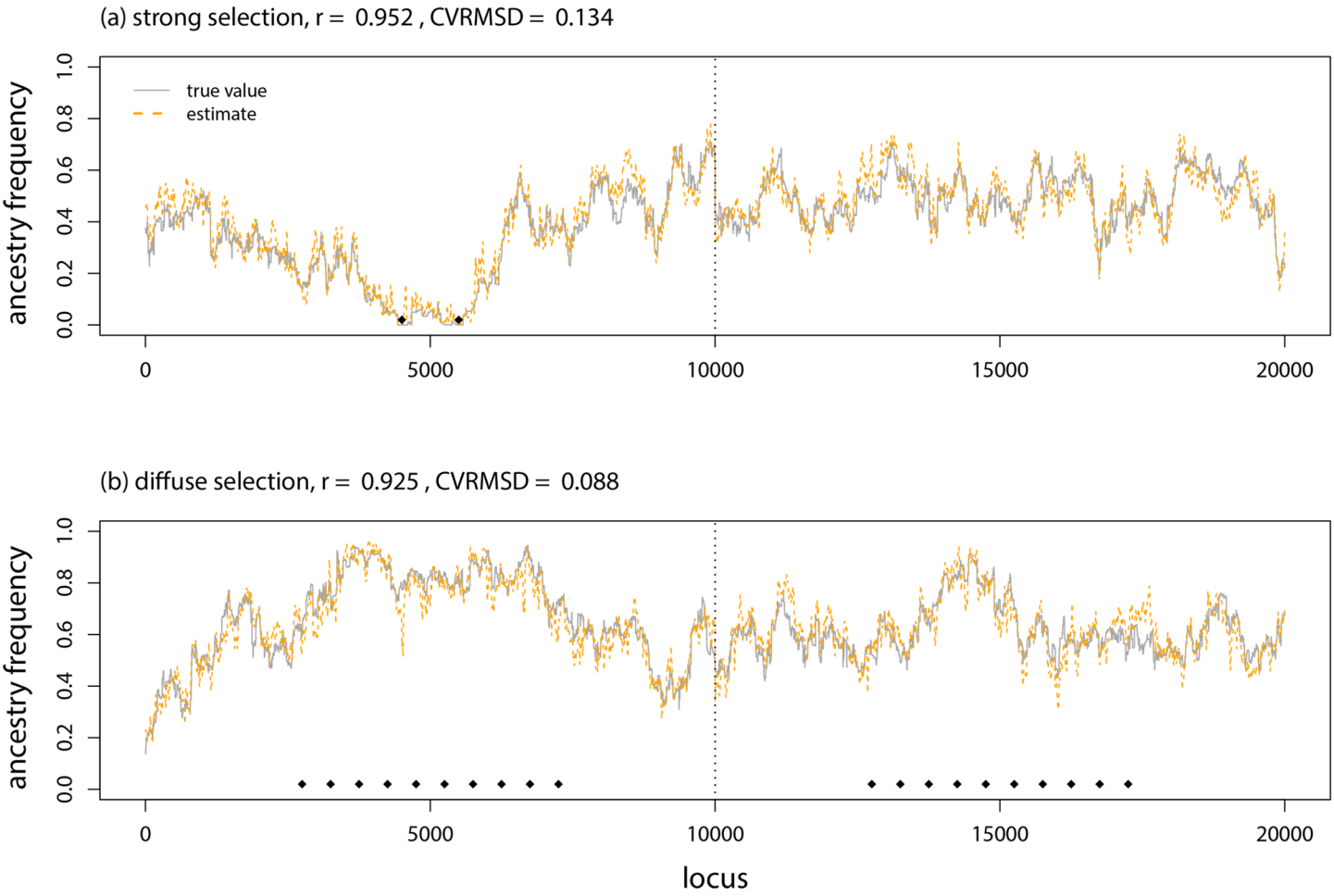

Fig 7. Genetic ancestry estimates from simulated data with selection. Plots show the true and inferred (posterior mean from `popanc`) ancestry frequencies from one representative data set from the strong and diffuse selection simulations. Results with a window size of ±20 SNPs are shown. The dashed vertical lines delineate the two distinct chromosomes. Black diamonds indicate the genome positions that affected fitness (i.e., the direct targets of selection against inter-population ancestry).

doi:10.1371/journal.pone.0151047.g007

Discussion

Performance and utility

I have described a new method to infer local ancestry frequencies from un-phased genotypic data using a continuous correlated beta process model (CCBPM). The proposed method produced accurate estimates of ancestry frequencies, outperforming a traditional HMM (and a null model of no variation in ancestry frequencies) under most conditions examined (e.g., Figs 3, 4, 6, & 8, Table 1). As expected, the improved performance of this new approach relative to

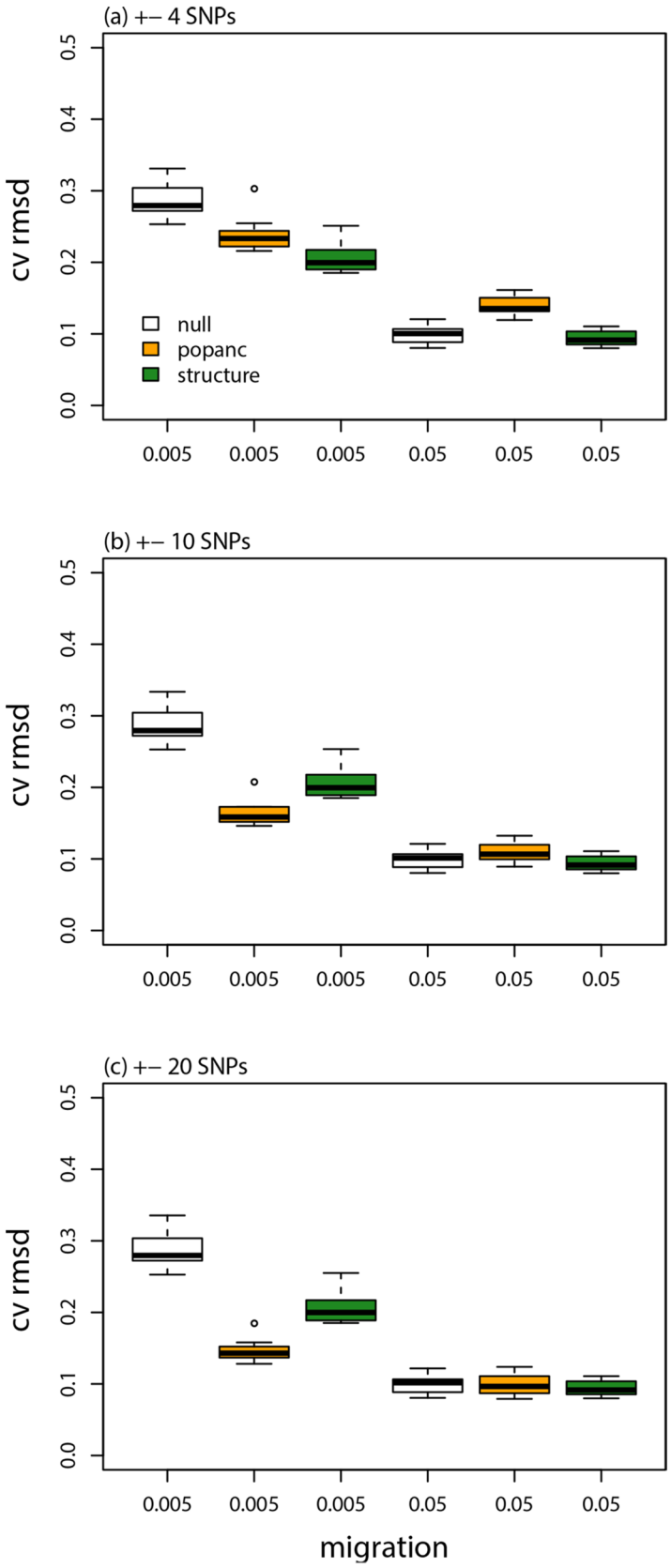
(a) +– 4 SNPs
(b) +– 10 SNPs
(c) +– 20 SNPs
cv rmsd
0.0
0.1
0.2
0.3
0.4
0.5
null
popanc
structure
0.005
0.005
0.005
0.05
0.05
0.05
migration

Fig 8. Model evaluation with on-going gene flow. Boxplots show the average CVRMSD for local ancestry frequencies for each of 10 replicate simulations analyzed with different methods, migration rates, and window sizes (for `popanc`). The CVRMSD for a null model that equates the ancestry frequencies at each locus with the genome-average is also shown.

doi:10.1371/journal.pone.0151047.g008

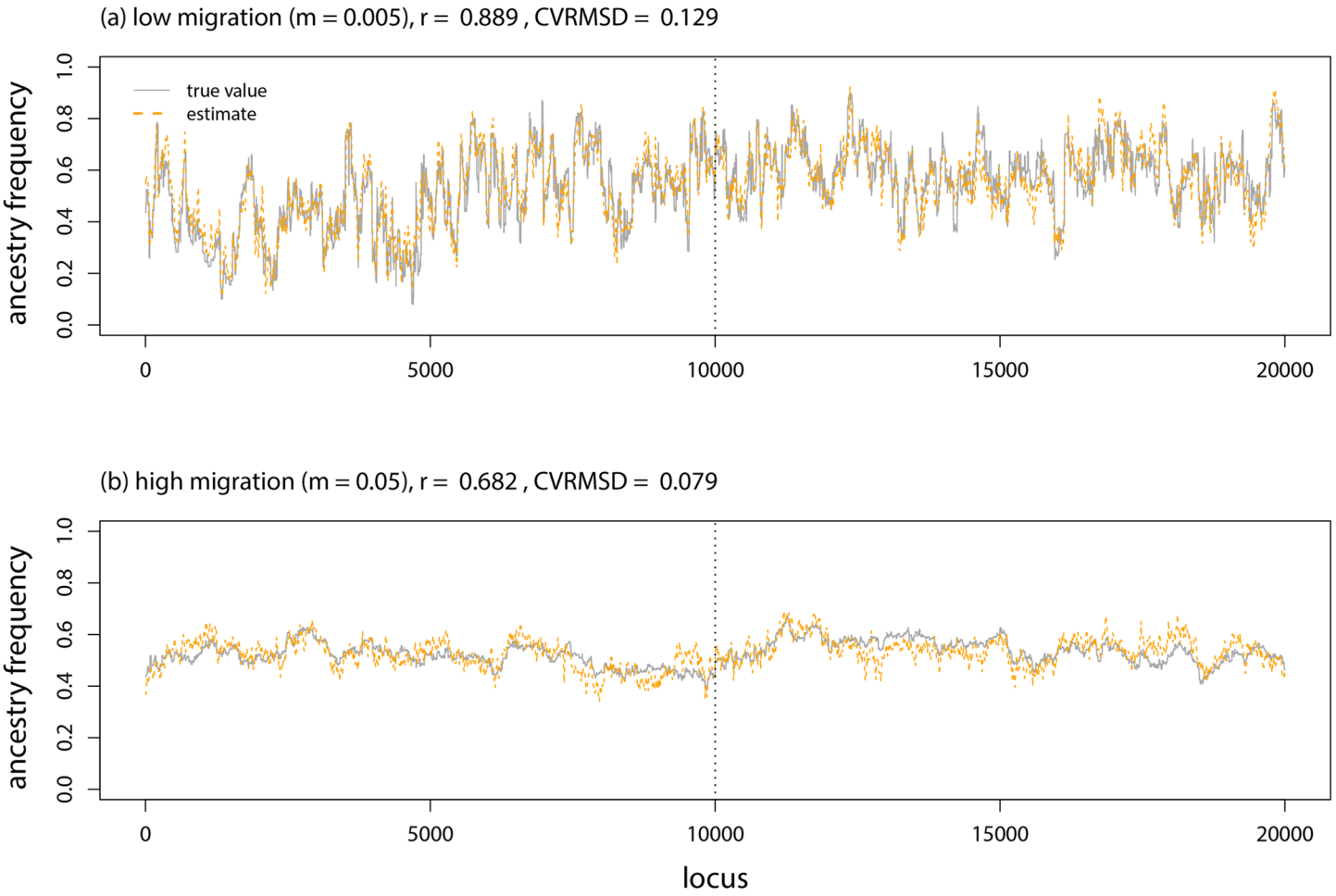

Fig 9. Genetic ancestry estimates from simulated data with on-going gene flow. Plots show the true and inferred (posterior mean from `popanc`) ancestry frequencies from one representative data set from simulations with low or high gene flow. Results with a window size of ±20 SNPs are shown. The dashed vertical lines delineate the two distinct chromosomes.

doi:10.1371/journal.pone.0151047.g009

the HMM in `structure` was more pronounced when ancestry frequencies varied more across the genome, either because of selection, more ancient admixture, or little to no gene flow post admixture. The poorer performance of the homogeneous HMM under these conditions likely reflects the fact that the HMM makes the *a priori* assumption that ancestry frequencies do not vary. This should bias local ancestry inference towards the genome-average unless the genetic data are perfectly informative of ancestry. Moreover, the difference in performance between these methods was not trivial; for example, the CVRMSD for ancestry frequencies in

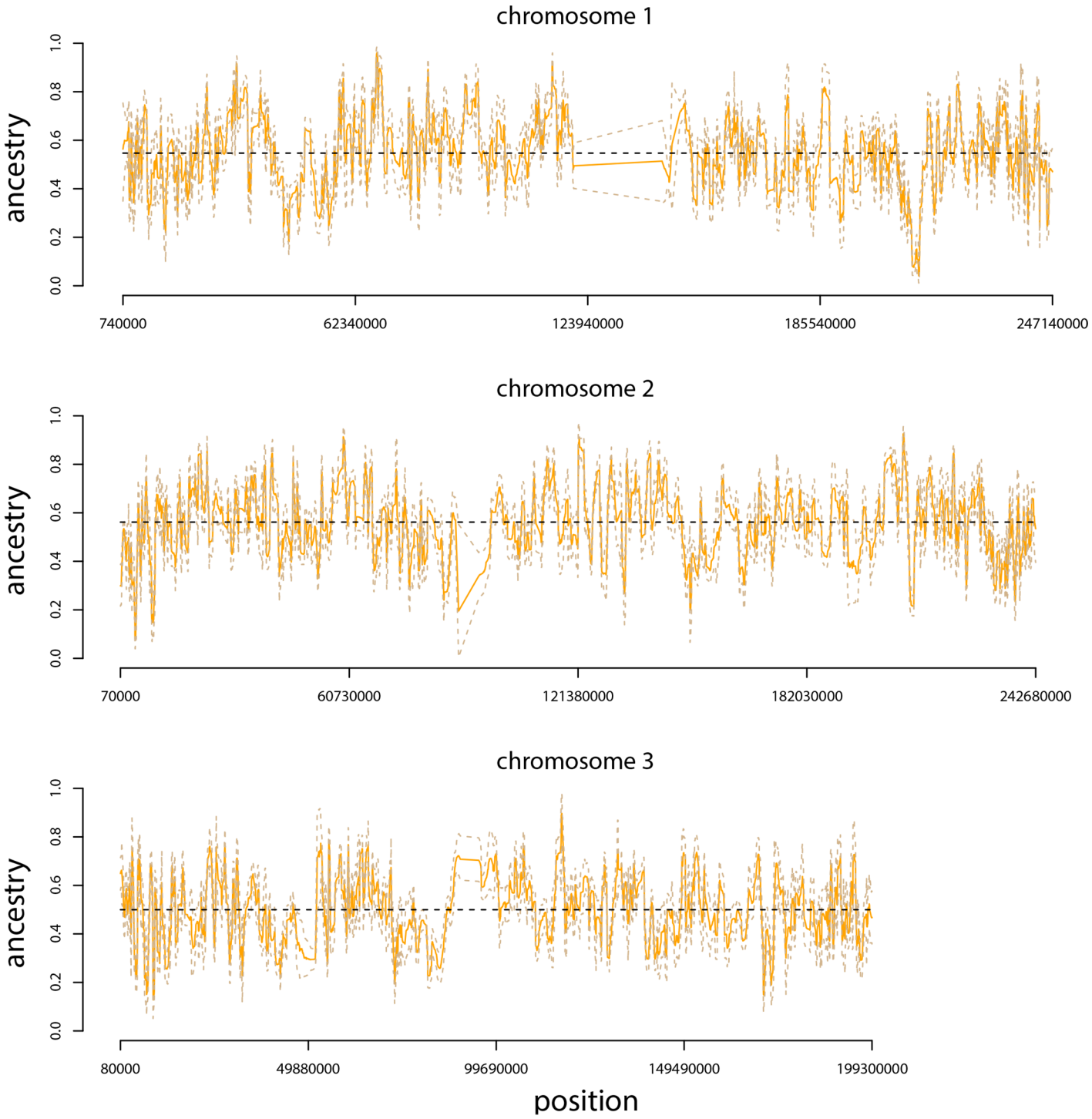

Fig 10. Genetic ancestry in Uyghur. Plots show ancestry frequency estimates from three human chromosomes. The posterior median (solid, orange line) and 95% ETPIs (dashed, tan line) for Han ancestry are given. The average ancestry frequency for each chromosome is shown by the dashed, black line.

doi:10.1371/journal.pone.0151047.g010

Table 2. Top excess ancestry regions in the Uyghur. Chromosome and position (start and end) for seven regions of excess Han or French ancestry in the Uyghur. Genes in these regions were identified using the UCSC Genome Browser on the Human Feb. 2009 (GRCh37/hg19) Assembly.

chrom.	start (kbp)	end (kbp)	no. SNPs	excess	genes
2	155508	157032	43	Han	KCNJ3
2	182130	182175	3	Han	AK125001
10	72176	72303	13	Han	EIF4EBP2, NODAL, PALD1
2	127590	127875	17	French	BIN1
11	26727	27268	24	French	BBOX1, FIBIN, SLC5A12
11	28373	28657	16	French	none
11	89859	89915	2	French	NAALAD2

doi:10.1371/journal.pone.0151047.t002

the 200 generation simulations with ±20 SNP windows for the CCBPM was only $\sim \frac{2}{3}$ that of the `structure` HMM. Another key distinction between the approaches is that CCBPM generates measures of uncertainty in the ancestry frequencies (as captured in the approximate posterior distribution), which account for uncertainty in local ancestry within individuals, whereas *post hoc* estimates of ancestry frequencies from deconvolutions of local ancestry within individuals do not. However, this benefit is lessened by the fact that the 95% ETPIs inferred from the CCBPM appear to routinely underestimate uncertainty in local ancestry (the cause and a potential solution for this are discussed more below; Table 1). Finally, computational methods used for the proposed method are relatively efficient, and thus run times should not be prohibitive for the analysis of large GBS, SNP, or even whole genome sequence data (for the results presented here runs on standard Linux compute nodes took 2-5 hours, and different chromosomes can be analyzed separately).

By applying the proposed method to a case of historical admixture in a human population, I further documented the reliability and utility of the approach. The method generated estimates of genome-average ancestry consistent with previous studies [7], but also showed that ancestry frequencies varied substantially both within and among chromosomes (Fig 10; [60] also reported chromosome-average ancestries which were similar to those obtained with `popanc`, $r = 0.895$, $p < 0.0001$). Ancestry frequency variation in the Uyghur likely reflects the effects of drift and selection, but parsing these effects remains difficult, as drift can have a substantial effect on ancestry frequencies given sufficient time (e.g., Fig 1). Here, comparisons with other admixed human populations could be useful, as selection would be more likely to generate consistent excess ancestry in the same regions of the genome (e.g., [73–75]). Analyzing such variation in ancestry can also be important for uncovering the basis for variation in traits and the prevalence of diseases among different human populations (e.g., [26, 27, 76–79]). Thus, ancestry frequency inference can be viewed as complementary to admixture mapping approaches used in recently admixed populations (e.g., African Americans or Hispanic or Latino populations), which utilize variation in local ancestry blocks and disease risk among individuals within a population (e.g., [26, 80, 81]). Along these lines, here I found a 285 kb region on chromosome 11 in the Uyghur population where greater than 90% of gene copies harbored French ancestry, whereas the genome-average frequency of French ancestry was only 47.1%. This region also contained a gene that has been repeatedly associated with Alzheimer's risk [70, 71], and showed patterns of genetic differentiation and variation in the source populations consistent with a history of selection in the French. Thus, it is possible that risk for this or a related disease was affected by this gene and varied between the ancestral source populations of the Uyghur, though further work would be required to test this hypothesis.

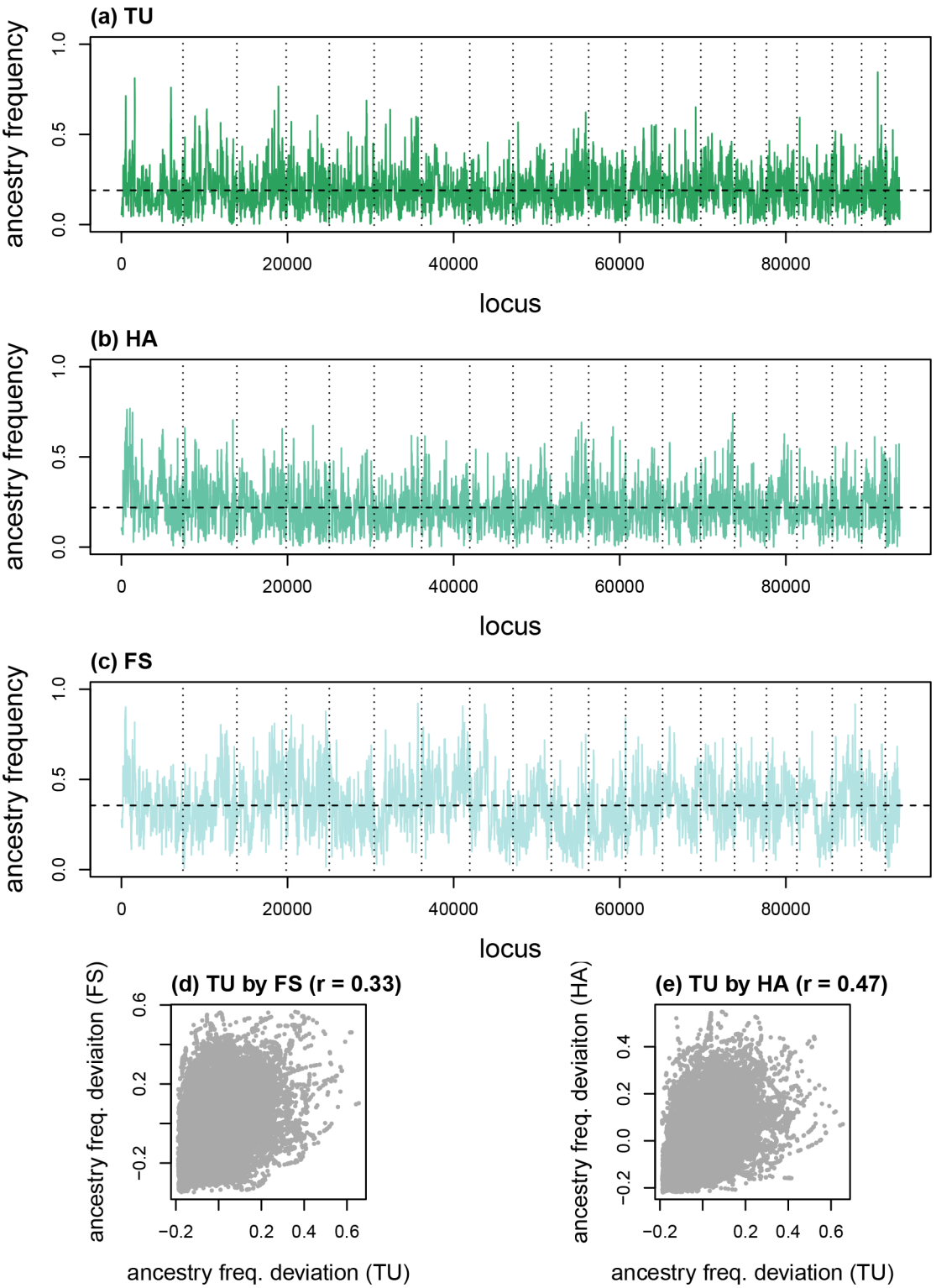

Fig 11. Genetic ancestry in a mouse hybrid zone. Plots show ancestry frequency estimates from three admixed populations. Solid lines in (a-c) give ancestry frequencies along chromosomes for TU, HA, and FS populations. Vertical dotted lines delineate chromosomes and the average ancestry frequency for each chromosome is shown by the dashed, black line. Scatterplots in (e) and (f) depict correlations in ancestry frequencies (relative to the mean for each population) for pairs of populations.

doi:10.1371/journal.pone.0151047.g011

Table 3. Top excess ancestry regions in house mice. Chromosome and position (start and end) for regions with the greatest excess *M. domesticus* ancestry in each of the three admixed house mouse populations. Only regions comprised of more than one SNP were reported.

population	chrom.	start (kbp)	end (kbp)	no. SNPs
TU	1	21910	22026	8
TU	1	47056	47096	3
TU	1	168139	168693	18
TU	3	139005	139401	15
TU	19	44236	44237	2
HA	1	24848	25420	26
HA	1	35230	35298	6
HA	1	40639	40774	9
HA	2	170032	170038	3
HA	14	120403	120433	4
FU*	1	20825	21726	23
FU	6	141892	142305	10
FU	7	138382	138450	3
FU	8	46928	46988	9
FU	18	73871	73891	2

* this range represents a composite of a few nearly contiguous regions.

doi:10.1371/journal.pone.0151047.t003

The proposed CCBPM was also applied to genetic data from three admixed populations that were part of the Bavarian transect through the central European *M. domesticus* × *M. musculus* hybrid zone. As with the human data, notable variation in ancestry frequencies was detected within and among chromosomes. Genetic regions with high or very high *M. domesticus* ancestry frequencies in one population tended to have higher *M. domesticus* ancestry frequencies in the other populations as well. Often such patterns of consistency or parallelism across replicate populations are interpreted as evidence of selection [14, 73]. While this could be correct here, gene flow among admixed populations could also explain this pattern. Indeed, comparative analyses of distant transects through this hybrid zone have shown that patterns of introgression vary considerably in different parts of this hybrid zone [68]. Of course, gene flow and selection are not mutually exclusive hypotheses. And additional data support the hypothesis that at least one of the excess *M. domesticus* ancestry regions (chromosome 1, 168, 139–168, 693 kb in TU) was likely affected by selection as it has been associated with a trait (aberrant RNA expression patterns in the testis) that likely contributes to reduced fertility in hybrids [69], and coincides with a marker with a putative epistatic effect on fitness in this hybrid zone [82].

Results from these empirical studies raise an important question: to what extent can genomic variation in ancestry frequencies be interpreted as evidence for past selection? Clearly, selection can drive extreme ancestry frequencies (Fig 7; [39, 44]). In particular, if selection in an admixed population favors a generally beneficial allele that had a higher frequency in one source population, the local ancestry frequency of the chromosomal segment containing that allele should increase. Underdominance and epistasis will have a similar effect, but the favored allele and ancestry type should exhibit positive frequency dependence, and thus, the outcome will depend on the initial conditions (e.g., with underdominance the marginal fitness of the more common ancestry type will be higher because it will occur proportionally less often in heterozygotes; [44]). However, genetic drift, particularly in small or old admixed populations,

can also cause substantial variation in ancestry frequencies (Fig 5). The effects of drift and selection can be hard to disentangle, particularly when selection is weak. This could be further confounded by variation in recombination rates, which would cause some genetic regions to be more or less influenced by the indirect effects of selection [83–85]. This will be particularly pronounced in admixed populations, because of admixture linkage disequilibrium. Thus, while it would be possible to develop an explicit test for selection-based patterns of ancestry frequency variation (e.g., [39, 40]), I avoid that here. Rather, genetic regions with extreme ancestry frequencies (relative to the overall variance across the genome) should be viewed as potential regions of interest that are likely enriched for targets of selection, but one cannot assume that all or most of these have in fact experienced substantial selection. Contrasts among independent admixed populations (e.g., [68, 86, 87]) or between admixed and allopatric source populations (e.g., [18, 88]) could be used to further test the hypothesis of selection, but this often assumes selection acts similarly across populations and environments.

Methodological considerations

As demonstrated here, window size affects the accuracy of ancestry inference. In particular, because of drift in admixed populations and shared variation between source populations, information must be extracted from a series of SNPs to obtain accurate estimates of local ancestry and ancestry frequencies (hence the popularity of window-based methods for local ancestry inference, e.g., [34, 55, 89]). Thus, for the simulated data sets analyzed here, ancestry frequencies were better estimated with the larger ±10 and ±20 SNP windows than the ±4 SNP windows (Figs 4 & 6, Table 1). However, very large windows will also lead to errors, because windows will frequently span ancestry breakpoints, and thus will include a mixture of SNPs with different ancestry. Note that window size and the scale parameter from the kernel function are related but distinct. Window size reflects ancestry blocks within individuals, whereas the scale parameter captures autocorrelation in ancestry frequencies at the population-level and is inferred from the data.

A few different approaches could be used to select an appropriate window size. First, if the time since admixture is known, the expected density of ancestry breakpoints and thus size of ancestry blocks can be calculated. For example, if one assumes that recombination operates as a random Poisson process and that half of all recombination events cause ancestry breakpoints (ancestry breakpoints are only generated by recombination between chromosomes with different local ancestry [11]), then the density of ancestry breakpoints for a pair of homologous chromosomes (i.e., in a diploid individual) should be $\frac{L^{chrom}}{t}$, where L^{chrom} is the map size of the chromosome in Morgans (M) and t is the time since admixture. Assuming one has genotypes from L^{SNP} SNPs on each chromosome, and that each chromosome is 1 M in length, ancestry blocks should thus contain an average of $\frac{L^{SNP}}{t}$ SNPs. Therefore, window sizes smaller than $\frac{L^{SNP}}{t}$ SNPs should be used. When admixture is very old or the contributions of the source populations differ substantially, this equation will underestimate ancestry block size and thus the approach is conservative (this occurs because recombination is less likely to generate breakpoints when ancestry frequencies are farther from 0.5). Applying this equation to the 20, 50 and 200 generation simulations analyzed here, one would expect average ancestry block sizes of 500 (t = 20), 200 (t = 50), and 50 (t = 200) SNPs. As expected, these numbers slightly underestimate the true block sizes, which were 540.3 (t = 20), 213.0 (t = 50), and 59.4 (t = 200) SNPs. Alternatively, if the time since admixture is not known, one can obtain a reasonable estimate of the size of ancestry blocks by first inferring local ancestry using a HMM, such as the linkage model in `structure` [32], or one of the other available HMMs (e.g., [33, 36]). Moreover, data sets can be analyzed using a series of window sizes to evaluate the robustness of the results

to this parameter. One should then be able to identify a range of reasonable window sizes that give consistent estimates of local ancestry frequencies.

The proposed method uses DA to calculate the likelihood of local ancestry given the genetic data, that is, $\Pr(s_{xj}|z_{xj})$ (Eq 6). However, the proposed CCBPM could be combined with ancestry likelihoods obtained from other discriminant methods, such as the random forest algorithm implemented in `RFMix` [37]. With that said, DA is rapid and can be used for integer or non-integer valued genetic data. The latter is a particularly nice feature, as this means that posterior mean genotypes, such as those obtained from imputation or low to moderate coverage GBS data [6, 90, 91], can readily be handled by the proposed method with DA. Thus, this is an important feature to ensure that the proposed method can be applied to non-model systems.

A final issue of note is that the approximate 95% ETPIs for local ancestry were consistently too narrow, and thus often failed to contain true parameter value; this was true despite low overall errors and high correlations between parameter values and their estimates (Table 1). This is a predictable outcome of the generalized Bayesian update for the CCBPM specified in Eq (1). By combining information across genetic loci, the CCBPM generates more accurate estimates of ancestry frequencies that account for autocorrelations along chromosomes, but this also unduly narrows the posterior distribution for these parameters. This problem could be circumvented by obtaining point estimates from the CCBPM posterior distribution (i.e., by using Eq (1)), but also sampling from beta($\alpha = z_x + \alpha_0, \beta = 2n_x - z_x + \beta_0$) to obtain 95% ETPIs for that parameter. This option is available in the `popanc` software, and greatly increased the proportion of the time the 95% ETPIs contained the true parameter value for a strong selection data set that I re-analyzed (95% ETIP coverage: CCBPM = 0.48, simple beta-binomial model = 0.78).

Conclusions

Existing ancestry deconvolution methods are best suited for relatively recent or very ancient admixture. In the former case, ancestry frequencies should not vary much across the genome and instead variation in genome-average ancestry or local ancestry-blocks are of interest, whereas in the latter case one can assume that admixed population or species are fixed for ancestry blocks. However, admixed populations between these two extremes exist [6], and should exhibit substantial variation in ancestry frequencies (Fig 1). The method proposed and evaluated in this paper can be used to estimate ancestry frequencies under such conditions. Ancestry frequencies can then be examined to describe the genomic composition of admixed populations and to evaluate progress towards genome stabilization (i.e., the loss of segregating variation for local ancestry). Moreover, comparisons of admixture frequencies across multiple admixed populations could provide critical information on whether hybridization has repeatable, predictable outcomes, and thus on the relative roles of deterministic and stochastic processes in shaping genome composition [92]. Additionally, admixture mapping methods could condition on local ancestry frequencies and thereby gain power to map important trait variation in admixed populations. And finally, ancestry frequencies could help identify regions of the genome that have been targets of selection in hybrids, which is particularly relevant for understanding the genetic basis of reproductive isolation between incipient species [15, 39, 93]. Thus, by filling what has been an analytical gap, the proposed CCBPM should be a useful tool for a variety of biologists.

Acknowledgments

This manuscript was improved by comments from A. Buerkle, C. Nice, and an anonymous reviewer. A. Buerkle provided the program used to simulated the data sets (`simadmix`,

available from http://www.uwyo.edu/buerkle/software/dfuse/). B. Harr kindly provided the data from the mouse hybrid zone. Computing, storage, and other resources from the Division of Research Computing in the Office of Research and Graduate Studies at Utah State University are gratefully acknowledged.

Author Contributions

Conceived and designed the experiments: ZG. Performed the experiments: ZG. Analyzed the data: ZG. Contributed reagents/materials/analysis tools: ZG. Wrote the paper: ZG. Designed the software used in analysis: ZG.

References

1. Stebbins GL. Variation and Evolution in Plants. New York: Columbia University Press; 1950.
2. Whittemore AT, Schaal BA. Interspecific gene flow in sympatric oaks. Proceedings of the National Academy of Sciences. 1991; 88(6):2540–2544. doi: 10.1073/pnas.88.6.2540
3. Arnold ML. Natural Hybridization and Evolution. Oxford University Press, New York; 1997.
4. Mallet J. Hybridization as an invasion of the genome. Trends in Ecology and Evolution. 2005 MAY; 20 (5):229–237. doi: 10.1016/j.tree.2005.02.010 PMID: 16701374
5. Hermansen JS, Sæther SA, Elgvin TO, Borge T, Hjelle E, Sætre GP. Hybrid speciation in sparrows I: phenotypic intermediacy, genetic admixture and barriers to gene flow. Molecular Ecology. 2011; 20:3812–3822. doi: 10.1111/j.1365-294X.2011.05183.x PMID: 21771138
6. Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC. Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. Molecular Ecology. 2014; 23(18):4555–4573. doi: 10.1111/mec.12811 PMID: 24866941
7. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient Admixture in Human History. Genetics. 2012; 192(3):1065–1093. doi: 10.1534/genetics.112.145037 PMID: 22960212
8. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, et al. Genetic evidence for recent population mixture in India. The American Journal of Human Genetics. 2013; 93(3):422–438. doi: 10.1016/j.ajhg.2013.07.006 PMID: 23932107
9. Vernot B, Akey JM. Human Evolution: Genomic Gifts from Archaic Hominins. Current Biology. 2014; 24 (18):R845–R848. doi: 10.1016/j.cub.2014.07.079 PMID: 25247359
10. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nature Reviews Genetics. 2015; 16(6):359–371. doi: 10.1038/nrg3936 PMID: 25963373
11. Buerkle CA, Rieseberg LH. The rate of genome stabilization in homoploid hybrid species. Evolution. 2008; 62(2):266–275. doi: 10.1111/j.1558-5646.2007.00267.x PMID: 18039323
12. Gompert Z, Buerkle CA. Analyses of genetic ancestry enable key insights for molecular ecology. Molecular Ecology. 2013; 22:5278–5294. Available from: http://dx.doi.org/10.1111/mec.12488 doi: 10.1111/mec.12488 PMID: 24103088
13. Barton NH, Hewitt GM. Analysis of hybrid zones. Annual Review of Ecology and Systematics. 1985; 16:113–148. doi: 10.1146/annurev.es.16.110185.000553
14. Buerkle CA, Rieseberg LH. Low intraspecific variation for genomic isolation between hybridizing sunflower species. Evolution. 2001; 55:684–691. doi: 10.1111/j.0014-3820.2001.tb00804.x PMID: 11392386
15. Payseur BA, Krenz JG, Nachman MW. Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. Evolution. 2004; 58:2064–2078. doi: 10.1111/j.0014-3820.2004.tb00490.x PMID: 15521462
16. Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE, et al. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome Research. 2008; 18(1):67–76. doi: 10.1101/gr.6757907 PMID: 18025268
17. Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012; 487:94–98.
18. Gompert Z, Lucas LK, Nice CC, Fordyce JA, Forister ML, Buerkle CA. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. Evolution. 2012; 66 (7):2167–2181. doi: 10.1111/j.1558-5646.2012.01587.x PMID: 22759293

19. Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J. The Genomic Signature of Crop-Wild Introgression in Maize. PLoS Genetics. 2013 05; 9(5):e1003477. doi: 10.1371/journal.pgen.1003477 PMID: 23671421

20. Chapman NH, Thompson EA. The effect of population history on the lengths of ancestral chromosome segments. Genetics. 2002; 162:449–458. PMID: 12242253

21. Gravel S. Population Genetics Models of Local Ancestry. Genetics. 2012; 191:607–619. doi: 10.1534/genetics.112.139808 PMID: 22491189

22. Liang M, Nielsen R. The lengths of admixture tracts. Genetics. 2014; 197(3):953–967. doi: 10.1534/genetics.114.162362 PMID: 24770332

23. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. PMID: 10835412

24. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, et al. Control of confounding of genetic associations in stratified populations [Article]. American Journal of Human Genetics. 2003 JUN; 72(6):1492–1504. doi: 10.1086/375613 PMID: 12817591

25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38(8):904–909. doi: 10.1038/ng1847 PMID: 16862161

26. Shriner D, Adeyemo A, Rotimi CN. Joint Ancestry and Association Testing in Admixed Individuals [Article]. PLoS Computational Biology. 2011 DEC; 7(12):e1002325. doi: 10.1371/journal.pcbi.1002325 PMID: 22216000

27. Zaitlen N, Pasaniuc B, Sankararaman S, Bhatia G, Zhang J, Gusev A, et al. Leveraging population admixture to characterize the heritability of complex traits. Nature Genetics. 2014; 46(12):1356–1362. doi: 10.1038/ng.3139 PMID: 25383972

28. Seehausen O, Van Alphen JJ, Witte F. Cichlid fish diversity threatened by eutrophication that curbs sexual selection. Science. 1997; 277(5333):1808–1811. doi: 10.1126/science.277.5333.1808

29. Taylor EB, Boughman JW, Groenenboom M, Sniatynski M, Schluter D, Gow JL. Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (*Gasterosteus aculeatus*) species pair. Molecular Ecology. 2006; 15(2):343–355. doi: 10.1111/j.1365-294X.2005.02794.x PMID: 16448405

30. Fitzpatrick BM, Johnson JR, Kump DK, Smith JJ, Voss SR, Shaffer HB. Rapid spread of invasive genes into a threatened native species [Article]. Proceedings of National Academy of Sciences. 2010 FEB 23; 107(8):3606–3610. doi: 10.1073/pnas.0911802107

31. Gese EM, Knowlton FF, Adams JR, Beck K, Fuller TK, Murray DL, et al. Managing hybridization of a recovering endangered species: The red wolf *Canis rufus* as a case study. Curr Zool. 2015; 61:191–203. doi: 10.1093/czoolo/61.1.191

32. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164:1567–1587. PMID: 12930761

33. Tang H, Coram M, Wang P, Zhu XF, Risch N. Reconstructing genetic ancestry blocks in admixed individuals [Article]. American Journal of Human Genetics. 2006 JUL; 79(1):1–12. doi: 10.1086/504302 PMID: 16773560

34. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating Local Ancestry in Admixed Populations. American Journal of Human Genetics. 2008; 82:290–303. doi: 10.1016/j.ajhg.2007.09.022 PMID: 18252211

35. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genetics. 2009 06; 5(6): e1000519. doi: 10.1371/journal.pgen.1000519 PMID: 19543370

36. Sohn KA, Ghahramani Z, Xing EP. Robust Estimation of Local Genetic Ancestry in Admixed Populations using a Non-parametric Bayesian Approach. Genetics. 2012;.

37. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. The American Journal of Human Genetics. 2013; 93 (2):278–288. doi: 10.1016/j.ajhg.2013.06.020 PMID: 23910464

38. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. PLoS Genetics. 2012 01; 8(1):e1002453. doi: 10.1371/journal.pgen.1002453 PMID: 22291602

39. Gompert Z, Buerkle CA. A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. Molecular Ecology. 2009; 18:1207–1224. doi: 10.1111/j.1365-294X.2009.04098.x PMID: 19243513

40. Gompert Z, Buerkle CA. Bayesian estimation of genomic clines [Article]. Molecular Ecology. 2011 MAY; 20(10):2111–2127. doi: 10.1111/j.1365-294X.2011.05074.x PMID: 21453352

41. Vaha J, Primmer C. Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci [Article]. Molecular Ecology. 2006 JAN; 15(1):63–72. doi: 10.1111/j.1365-294X.2005.02773.x PMID: 16367830

42. Barton NH, Hewitt GM. Adaptation, speciation and hybrid zones. Nature. 1989; 341:497–503. doi: 10.1038/341497a0 PMID: 2677747

43. Rieseberg LH, Whitton J, Gardner K. Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. Genetics. 1999; 152:713–727. PMID: 10353912

44. Gompert Z, Parchman TL, Buerkle CA. Genomics of isolation in hybrids. Philosophical Transactions of the Royal Society B: Biological Sciences. 2012; 367(1587):439–450. doi: 10.1098/rstb.2011.0196

45. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014; 507(7492):354–357. doi: 10.1038/nature12961 PMID: 24476815

46. da Barbiano LA, Gompert Z, Aspbury AS, Gabor CR, Nice CC. Population genomics reveals a possible history of backcrossing and recombination in the gynogenetic fish *Poecilia formosa*. Proceedings of the National Academy of Sciences. 2013; 110(34):13797–13802. doi: 10.1073/pnas.1303730110

47. Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, et al. Major ecological transitions in wild sunflowers facilitated by hybridization. Science. 2003; 301:1211–1216. doi: 10.1126/science.1086949 PMID: 12907807

48. Bertorelle G, Excoffier L. Inferring admixture proportions from molecular data. Molecular Biology and Evolution. 1998; 15(10):1298–1311. doi: 10.1093/oxfordjournals.molbev.a025858 PMID: 9787436

49. Nice CC, Gompert Z, Fordyce JA, Forister ML, Lucas LK, Buerkle CA. Hybrid speciation and independent evolution in lineages of alpine butterflies. Evolution. 2013; 67(4):1055–1068. doi: 10.1111/evo.12019 PMID: 23550755

50. Ungerer MC, Baird SJE, Pan J, Rieseberg LH. Rapid hybrid speciation in wild sunflowers. Proceedings of the National Academy of Sciences. 1998; 95:11757–11762. doi: 10.1073/pnas.95.20.11757

51. Goetschalckx R, Poupart P, Hoey J. Continuous correlated beta processes. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence. vol. 22. Citeseer; 2011. p. 1269.

52. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE. 2011 05; 6(5):e19379. doi: 10.1371/journal.pone.0019379 PMID: 21573248

53. Parchman TL, Benkman CW, Jenkins B, Buerkle CA. Low levels of population genetic structure in *Pinus contorta* (Pinaceae) across a geographic mosaic of co-evolution. American Journal of Botany. 2011; 98(4):669–679. doi: 10.3732/ajb.1000378 PMID: 21613166

54. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics. 2010; 11(1):94. Available from: http://www.biomedcentral.com/1471-2156/11/94 doi: 10.1186/1471-2156-11-94 PMID: 20950446

55. Paşaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific ancestry in closely related populations. Bioinformatics. 2009; 25(12):i213–i221. doi: 10.1093/bioinformatics/btp197 PMID: 19477991

56. Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, et al. GNU Scientific Library: Reference Manual. Network Theory Ltd.; 2009.

57. The HDF5 Group. Hierarchical data format version 5, 2000-2010; 2010. http://www.hdfgroup.org/HDF5

58. Nicholson G, Smith AV, Jonsson F, Gustafsson O, Stefansson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data [Proceedings Paper]. Journal of the Royal Statistical Society Series B-Methodological. 2002; 64(Part 4):695–715. doi: 10.1111/1467-9868.00357

59. Ewens WJ. Mathematical Population Genetics: I. Theoretical Introduction. vol. 27. Springer Science & Business Media; 2004.

60. Xu S, Huang W, Qian J, Jin L. Analysis of genomic admixture in Uyghur and its implication in mapping strategy. The American Journal of Human Genetics. 2008; 82(4):883–894. doi: 10.1016/j.ajhg.2008.01.017 PMID: 18355773

61. Xu S, Jin L. A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. The American Journal of Human Genetics. 2008; 83(3):322–336. doi: 10.1016/j.ajhg.2008.08.001 PMID: 18760393

62. Salcedo T, Geraldes A, Nachman MW. Nucleotide variation in wild and inbred mice. Genetics. 2007; 177(4):2277–2291. doi: 10.1534/genetics.107.079988 PMID: 18073432

63. Geraldes A, Basset P, Gibson B, Smith KL, Harr B, Yu HT, et al. Inferring the history of speciation in house mice from autosomal, X-linked, Y-linked and mitochondrial genes. Molecular Ecology. 2008; 17 (24):5349–5363. doi: 10.1111/j.1365-294X.2008.04005.x PMID: 19121002

64. Cucchi T, Vigne JD, Auffray JC. First occurrence of the house mouse (*Mus musculus domesticus* Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. Biological Journal of the Linnean Society. 2005; 84(3):429–445. doi: 10.1111/j.1095-8312.2005.00445.x

65. Smadja C, Catalan J, Ganem G. Strong premating divergence in a unimodal hybrid zone between two subspecies of the house mouse. Journal of evolutionary biology. 2004; 17(1):165–176. doi: 10.1046/j.1420-9101.2003.00647.x PMID: 15000659

66. Good JM, Dean MD, Nachman MW. A complex genetic basis to X-linked hybrid male sterility between two species of house mice. Genetics. 2008; 179(4):2213–2228. doi: 10.1534/genetics.107.085340 PMID: 18689897

67. Turner LM, Schwahn DJ, Harr B. Reduced male fertility is common but highly variable in form and severity in a natural house mouse hybrid zone. Evolution. 2012; 66(2):443–458. doi: 10.1111/j.1558-5646.2011.01445.x PMID: 22276540

68. Teeter KC, Thibodeau LM, Gompert Z, Buerkle CA, Nachman MW, Tucker PK. The variable genomic architecture of isolation between hybridizing species of house mouse. Evolution. 2010; 64(2):472–485. doi: 10.1111/j.1558-5646.2009.00846.x PMID: 19796152

69. Turner LM, Harr B. Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. eLife. 2014; 3:e02504. doi: 10.7554/eLife.02504

70. Hu X, Pickering E, Liu YC, Hall S, Fournier H, Katz E, et al. Meta-analysis for genome-wide association study identifies multiple variants at the BIN1 locus associated with late-onset Alzheimer's disease. PloS ONE. 2011; 6(2):e16616. doi: 10.1371/journal.pone.0016616 PMID: 21390209

71. Wijsman EM, Pankratz ND, Choi Y, Rothstein JH, Faber KM, Cheng R, et al. Genome-wide association of familial late-onset Alzheimer's disease replicates BIN1 and CLU and nominates CUGBP2 in interaction with APOE. PLoS Genetics. 2011; 7(2):e1001308. doi: 10.1371/journal.pgen.1001308 PMID: 21379329

72. Piñero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015;2015. Available from: http://database.oxfordjournals.org/content/2015/bav028.abstract doi: 10.1093/database/bav028 PMID: 25877637

73. Schluter D, Nagel LM. Parallel speciation by natural selection. American Naturalist. 1995; 146:292–301. doi: 10.1086/285799

74. Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, et al. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles [Article]. Science. 2005 MAR 25; 307(5717):1928–1933. doi: 10.1126/science.1107239 PMID: 15790847

75. Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, et al. Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. Science. 2014; 344(6185):738–742. doi: 10.1126/science.1252136 PMID: 24833390

76. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. PNAS. 2006; 103:14068–14073. doi: 10.1073/pnas.0605832103 PMID: 16945910

77. Deo RC, Patterson N, Tandon A, McDonald GJ, Haiman CA, Ardlie K, et al. A High-Density Admixture Scan in 1,670 African Americans with Hypertension. PLoS Genetics. 2007; 3:e196. doi: 10.1371/journal.pgen.0030196 PMID: 18020707

78. Nalls MA, Wilson JG, Patterson NJ, Tandon A, Zmuda JM, Huntsman S, et al. Admixture Mapping of White Cell Count: Genetic Locus Responsible for Lower White Blood Cell Count in the Health ABC and Jackson Heart Studies. American Journal of Human Genetics. 2008; 82(1):81–87. doi: 10.1016/j.ajhg.2007.09.003 PMID: 18179887

79. Molineros JE, Maiti AK, Sun C, Looger LL, Han S, Kim-Howard X, et al. Admixture Mapping in Lupus Identifies Multiple Functional Variants within IFIH1 Associated with Apoptosis, Inflammation, and Autoantibody Production. PLoS Genetics. 2013 02; 9(2):e1003222. doi: 10.1371/journal.pgen.1003222 PMID: 23441136

80. McKeigue PM. Prospects for Admixture Mapping of Complex Traits. The American Journal of Human Genetics. 2005; 76(1):1–7. doi: 10.1086/426949 PMID: 15540159

81. Buerkle CA, Lexer C. Admixture as the basis for genetic mapping. Trends in Ecology & Evolution. 2008; 23(12):686–694. doi: 10.1016/j.tree.2008.07.008

82. Janoušek V, Wang L, Luzynski K, Dufková P, Vyskočilová MM, Nachman MW, et al. Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between *Mus musculus musculus* and *M. m. domesticus*. Molecular Ecology. 2012; 21(12):3032–3047. doi: 10.1111/j.1365-294X.2012.05583.x PMID: 22582810

83. Charlesworth D, Charlesworth B, Morgan MT. The pattern of neutral molecular variation under the background selection model. Genetics. 1995; 141:1619–1632. PMID: 8601499

84. Charlesworth B. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. Genetical Research. 1996; 68(2):131–150. doi: 10.1017/S0016672300034029 PMID: 8940902

85. Hahn MW. Toward a selection theory of molecular evolution. Evolution. 2008; 62(2):255–265. doi: 10.1111/j.1558-5646.2007.00308.x PMID: 18302709

86. Nolte AW, Gompert Z, Buerkle CA. Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. Molecular Ecology. 2009; 18:2615–2627. doi: 10.1111/j.1365-294X.2009.04208.x PMID: 19457191

87. Dufková P, Macholan M, Pialek J. Inference of selection and stochastic effects in the house mouse hybrid zone. Evolution. 2011; 65(4):993–1010. doi: 10.1111/j.1558-5646.2011.01222.x PMID: 21463294

88. Parchman TL, Gompert Z, Braun MJ, Brumfield RT, McDonald DB, Uy JAC, et al. The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. Molecular Ecology. 2013; 22:3304–3317. doi: 10.1111/mec.12201 PMID: 23441849

89. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics. 2012; 28(10):1359–1367. doi: 10.1093/bioinformatics/bts144 PMID: 22495753

90. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012; 44:821–824. doi: 10.1038/ng.2310 PMID: 22706312

91. Gompert Z, Jahner JP, Scholl CF, Wilson JS, Lucas LK, Soria-Carrasco V, et al. The evolution of novel host use is unlikely to be constrained by trade-offs or a lack of genetic variation. Molecular Ecology. 2015; 24(11):2777–2793. doi: 10.1111/mec.13199 PMID: 25877787

92. Rieseberg LH, Widmer A, Arntz AM, Burke JM. The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations. Philosophical Transactions of the Royal Society London B. 2003; 358:1141–1147. doi: 10.1098/rstb.2003.1283

93. Barton N, Bengtsson B. The barrier to genetic exchange between hybridizing populations. Heredity. 1986 DEC; 57(3):357–376. doi: 10.1038/hdy.1986.135 PMID: 3804765