The American Journal of Human Genetics, Volume *110*

Supplemental information

Challenges of accurately estimating sex-biased admixture from X chromosomal and autosomal ancestry proportions

Aaron Pfennig and Joseph Lachance

Supplemental Figures

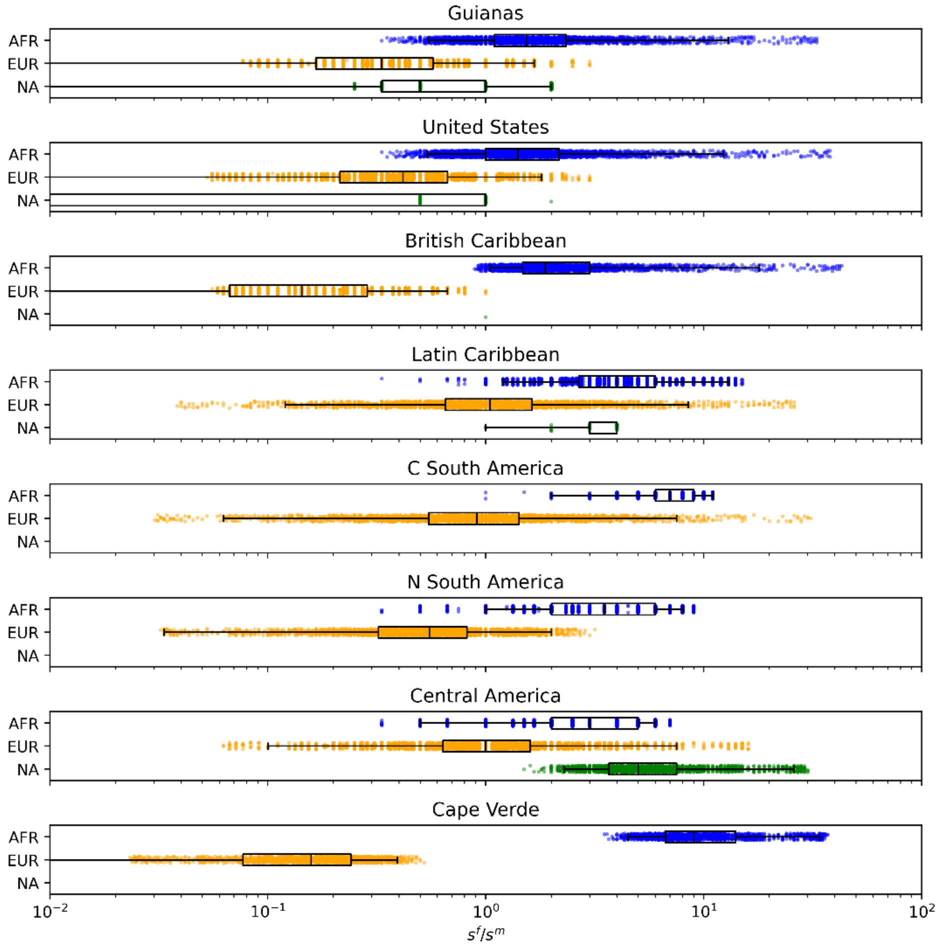

Figure S1: Possible sex ratios inferred from X chromosomal and autosomal ancestry proportions by Micheletti et al. using a model that assumes constant, nonzero admixture. Sex ratios were estimated based on ancestry proportion reported in Table S9 in Micheletti et al., for which the Euclidean distance *D* (Eq. S13) between the expected admixture proportions and reported admixture proportions was at most 0.01. Because there is not one single solution, a range of values is reported. The boxes represent the inter-quartile range (IQR), with the median sex ratio indicated by the line spanning the box. The whiskers represent the range between 2.5th and 97.5th percentile. For ancestries for which no box plot is shown, no combination of sex-specific contributions could be found that explains the observed ancestry proportions under a demographic scenario of constant, nonzero admixture**.**

Supplemental Tables

Broad Region	Region	p_f	p_m	N
Cabo Verde	Cabo Verde	0.55	0.45	121
C. South America	Rio De La Plata	0.57	0.43	29
C. South America	Brazil	0.51	0.49	1461
N. South America	Venezuela	0.48	0.52	495
Central America	Mexico	0.55	0.45	3270
Latin Caribbean	Dominican Republic	0.53	0.47	2307
Guianas	Guianas	0.57	0.43	267
British Caribbean[a]	Trinidad and Tobago	0.61	0.39	282
British Caribbean	British Leeward Islands and French Caribbean	0.45	0.55	163
British Caribbean	Haiti	0.51	0.49	596
British Caribbean	Jamaica and the Caymans	0.58	0.42	1526
British Caribbean	Bahamas	0.54	0.46	65
United States	South Atlantic	0.58	0.42	1235
United States	Gulf Coast	0.59	0.41	1411
United States	Inland Midwest	0.60	0.40	235
United States	East Inland	0.57	0.43	340
United States	Chesapeake	0.58	0.42	653
United States	Northern States	0.64	0.36	166
N. South America	Colombia	0.55	0.45	1029
Central America	Spanish Caribbean Mainland	0.53	0.47	2766
Latin Caribbean	Cuba	0.51	0.49	840
Latin Caribbean	Puerto Rico	0.55	0.45	6127
British Caribbean	British Windward Islands	0.58	0.42	293
United States	Midwest	0.58	0.42	1745
		Mean p_f	**Mean p_m**	**sum N**
Guianas		0.57	0.43	267
United States		0.58	0.42	5785
British Caribbean		0.56	0.44	2925
Latin Caribbean		0.55	0.45	9274
C. South America		0.51	0.49	1490
N. South America		0.53	0.47	1524
Central America		0.54	0.46	6036
Cabo Verde		0.55	0.45	121

Table S1. Calculation of female (p_f) and male (p_m) proportions in samples for each broad region. The proportions of females and males were inferred based on the data provided in Table S2 and S10 of Micheletti et al.

---

[a] was once listed under British Caribbean and once under Latin America in Micheletti et al. Based on the language spoken in Trinidad and Tobago, we chose British Caribbean.

♂

♀

	S_1	S_2	S_3	H	Sum female contributions
S_1	$s_1^f s_1^m$	$s_1^f s_2^m$	$s_1^f s_3^m$	$s_1^f h^m$	s_1^f
S_2	$s_2^f s_1^m \times (1-p)$	$s_2^f s_2^m \times c$	$s_2^f s_3^m \times (1-p)$	$s_2^f h^m \times (1-p)$	s_2^f
S_3	$s_3^f s_1^m$	$s_3^f s_2^m$	$s_3^f s_3^m$	$s_3^f h^m$	s_2^f
H	$h^f s_1^m$	$h^f s_2^m$	$h^f s_3^m$	$h^f h^m$	h^f
Sum male contributions	$s_1^m - p \times s_2^f s_1^m$	$s_2^m + (c-1) s_2^f s_2^m$	$s_3^m - p \times s_2^f s_3^m$	$h^m - p \times s_2^f h^m$	

Table S3. Mating matrix of sex-specific assortative mating scheme. Females from S_2 are asymmetrically more likely to mate with males from S_2. This is modeled in a way such that it does not reduce the overall likelihood of S_2 females mating. In consequence, it inflates the contributions of S_2 males and reduces the contributions of males from other populations (i.e., $S_1, S_3, \text{and } H$). p is the probability that the mating of a S_2 female with a non-S_2 male is rejected, and a male mating partner from S_2 is chosen instead. c is the corresponding amount by which the mating of an S_2 female and male is increased and is given by: $c = \frac{1 - (s_2^f s_1^m + s_2^f s_3^m + s_2^f h^m) + p(s_2^f s_1^m + s_2^f s_3^m + s_2^f h^m)}{s_2^f s_2^m}$.

Supplemental Methods

In this section, we provide the derivation of the equilibrium model for estimating sex biases from observed X chromosomal and autosomal ancestry proportions after a single admixture event and elaborate on models proposed by Goldberg and Rosenberg for estimating ancestry proportions for a more recent single admixture event as well as constant, nonzero admixture. Both models of a single admixture event were applied by Micheletti et al. to estimate magnitudes of sex-biased admixture from X chromosomal and autosomal data.[1] Because these models make a series of simplifying demographic assumptions that may have confounded the analyses of Micheletti et al.,[1] we evaluated the robustness of the models to violations of these assumptions using simulations. Here, we describe performed simulations of admixture in the Americas and subsequent interrogation of the models of sex-biased admixture regarding their robustness to sampling sizes and violations of demographic assumptions in greater detail.

Models of sex-biased admixture assuming a single admixture event

Equilibrium model

The expected X chromosomal and autosomal ancestry proportions in equilibrium after a single admixture event, i.e., an infinite number of generations of random mating within the admixed population since admixture, are easily inferred by acknowledging that the X chromosomal inheritance is sex-biased, while autosomal inheritance is not. Females contribute two-thirds of the X chromosomes and males one-third, while both sexes contribute half of the autosomes. Following the notation by Goldberg and Rosenberg, we will denote the admixed population by H and the expected ancestry proportions inherited from source population 1, S_1, on the X chromosome and the autosomes by $\mathbb{E}[H_1^X]$ and $\mathbb{E}[H_1^A]$, respectively. Given the fraction of females and males originating from S_1 (s_1^f and s_1^m, where $s_1^f + s_1^m = 1$), $\mathbb{E}[H_1^X]$ and $\mathbb{E}[H_1^A]$ are given by:[2,3]

$$\mathbb{E}[H_1^X] = \frac{2}{3}s_1^f + \frac{1}{3}s_1^m \tag{S1}$$

$$\mathbb{E}[H_1^A] = \frac{1}{2}s_1^f + \frac{1}{2}s_1^m \tag{S2}$$

Solving the Equations S1 and S2 for s_1^f and s_1^m yields the sex-specific contributions from S_1:

$$s_1^f = 3\mathbb{E}[H_1^X] - 2\mathbb{E}[H_1^A] \tag{S3}$$

$$s_1^m = 4\mathbb{E}[H_1^A] - 3\mathbb{E}[H_1^X] \tag{S4}$$

A dynamic model for a recent single admixture event

Here, we will briefly review the core ideas of the dynamic model proposed by Goldberg and Rosenberg for the X chromosomal admixture fraction of a recent single admixture event. Interested readers are referred to the original publications for more details.[2,4]

Goldberg and Rosenberg consider the female and male contributions separately, but the overall contributions of S_1 in generation g ($s_{1,g}$) are the mean of both (i.e., $(s_{1,g}^f + s_{1,g}^m)/2$). Furthermore, the female (male) contributions across all populations – all ancestral populations plus the admixed population – must sum to one.[2,4,5]

The key idea of Goldberg and Rosenberg is that $\mathbb{E}[H_1^X]$ and $\mathbb{E}[H_1^A]$ are not identically distributed because females inherit one X chromosome from each parent, while males only inherit one X chromosome from the mother, which is why the expected X chromosomal ancestry proportion in males is the expected ancestry proportion of a female X chromosome in the previous generation.[2] The authors derive a recursion for this relationship, which has a closed-form solution for a single admixture event (Equations 6-13 in Goldberg and Rosenberg).[2] Thus, the expected female and male ancestry proportions for the X chromosome in generation *g* are given by:

$$\mathbb{E}\left[H_{1,g,f}^{X}\right] = \left[2 + \left(-\frac{1}{2}\right)^{g}\right]\frac{s_{1,0}^{f}}{3} + \left[1 - \left(-\frac{1}{2}\right)^{g}\right]\frac{s_{1,0}^{m}}{3} \tag{S5}$$

$$\mathbb{E}\left[H_{1,g,m}^{X}\right] = \left[2 + \left(-\frac{1}{2}\right)^{g-1}\right]\frac{s_{1,0}^{f}}{3} + \left[1 - \left(-\frac{1}{2}\right)^{g-1}\right]\frac{s_{1,0}^{m}}{3} \tag{S6}$$

for $g \geq 1$ and $g \geq 2$ in Equation S5 and S6, respectively.[2] The mean X chromosomal ancestry proportion is then given by:

$$\mathbb{E}\left[H_{1,g}^{X}\right] = p_f\ \mathbb{E}\left[H_{1,g,f}^{X}\right] + p_m \mathbb{E}\left[H_{1,g,m}^{X}\right] \tag{S7}$$

where p_f and p_m are the fractions of females and males in the sample, respectively (see Equation 24 in Goldberg and Rosenberg).[2] Thus, estimates of initial contributions of females and males from S_1 - $s_{1,0}^{f}$ and $s_{1,0}^{m}$ - during a single admixture event g generations ago are obtained by solving Equations S2 and S7 for $s_{1,0}^{f}$ and $s_{1,0}^{m}$.

Due to different initial X chromosomal ancestry proportions in females and males, the expected ancestry proportions oscillate during the generations immediately following admixture. However, after approximately ten generations of random mating within the admixed population, the ancestry proportions converge to the expected equilibrium ancestry proportion, defined in Equations S1 and S2 (see Figure 2 in Goldberg and Rosenberg).[2]

This dynamic model and the equilibrium model defined in Equations S1 and S2 implicitly assume a constant population size of the admixed population, no subsequent gene flow, random mating, no genetic drift, and no selection. In the main text, we evaluated the robustness of the models to violations of some of these demographic assumptions using simulations. The simulations are described in detail below.

Estimating sex bias for a single admixture event

To determine if S_1's contributions to an admixed population were sex-biased, we require values for s_1^f and s_1^m, which can be obtained by applying one of the models delineated above. The contributions of S_1 are considered sex-biased if $s_1^f / s_1^m \neq 1$. For the equilibrium model, the closed-form solution for the ratio of female and male contributions is directly obtained from Equations S3 and S4 and is given by Equation 1 in the main text (for ease of notation, we refer to $\mathbb{E}[H_1^X]$ and $\mathbb{E}[H_1^A]$ as H_1^X and H_1^A in the main text). The closed-form solution for the ratio of female and male contributions using the model proposed by Goldberg and Rosenberg is more complex and is omitted here.[2]

Assuming a generation time of 25 years, the transatlantic slave trade started approximately 15 generations ago. After 15 generations of random mating within the admixed population, the model proposed by Goldberg and Rosenberg has always converged to the equilibrium model.[2] For these reasons, we report the sex ratios for generation 15 after admixture in Tables 1 & 2. We do not report sex ratios when either of the ancestry proportions is less than 0.05. Note that due to the sensitivity of the models, the results can marginally differ depending on the number of reported significant digits for ancestry proportions. Haplogroup imbalances between mitochondrial Y chromosomal haplogroups were derived from haplogroup frequencies. Code implementing these models can be found at https://github.com/LachanceLab/sex_biased_admixture.

Model failure and boundary conditions

Here, we infer conditions for model failure of the equilibrium model using the example shown in Figure 2A. Recall that we assume 0.123 of all autosomes in an admixed population are from S_1 ($H_1^A = 0.123$; this is the autosomal African ancestry proportion in central South America reported by Micheletti et al.[1]) and that there is an equal number of females and males in the admixed population. If all contributing individuals of ancestry S_1 were female, what proportion of all X chromosomes in the admixed population (H_1^X) are from ancestry S_1? In such a scenario, 12.3% of all initial individuals in the initially admixing population are females from S_1, and 37.7% are females from S_2 (i.e., 24.6% and 75.4% of the females come from S_1 and S_2, respectively). Since S_1 does not contribute any males, the remaining 50.0% of all initial individuals are males from S_2 (i.e., 100% of all males). Given that each female contributes two X chromosomes, and each male contributes a single X chromosome, the total proportion of X chromosomes from population S_1 for the equilibrium model can be calculated as follows:

$$\mathbb{E}[H_1^X] = \frac{2 \times s_1^f + s_1^m}{2 \times s_1^f + s_1^m + 2 \times s_2^f + s_2^m} = \frac{2 \times 0.246 + 0.0}{2 \times 0.246 + 0.0 + 2 \times 0.754 + 1.0} = 0.164 \qquad \text{(S8)}$$

Similarly, when all contributing individuals of ancestry S_1 were male, the expected X chromosomal ancestry proportion can be calculated as follows:

$$\mathbb{E}[H_1^X] = \frac{2 \times s_1^f + s_1^m}{2 \times s_1^f + s_1^m + 2 \times s_2^f + s_2^m} = \frac{2 \times 0.0 + 0.246}{2 \times 0.0 + 0.246 + 2 \times 1 + 0.754} = 0.082 \qquad \text{(S9)}$$

Thus, given the autosomal ancestry proportions $H_1^A = 0.123$, the model is only specified for $H_1^X \in [0.082, 0.164]$. X chromosomal ancestry proportions outside of this range cause model failure, i.e., negative sex ratios.

Alternatively, boundary conditions of ancestry proportions can also be calculated by considering conditions where either the denominator or the numerator in Equation 1 are zero. The maximum plausible X chromosome ancestry proportion (i.e., when all initial individuals from S_1 are female) is found by setting the left-hand side of Equation S4 equal to zero. Similarly, the minimum plausible X chromosome ancestry proportion (i.e., when all initial individuals from S_1 are male) is found by setting the left-hand side of Equation S3 equal to zero.

A model of sex-biased admixture assuming constant, nonzero admixture

The model

Here, we will briefly review the core ideas of the model proposed by Goldberg and Rosenberg for the X chromosomal and autosomal admixture fraction for constant, nonzero admixture. Interested readers are referred to the original publications for more details.[2,4]

The model for constant, nonzero admixture follows the basic logic of the model for a single admixture event described above. In the short term, the ancestry proportions depend on the initial admixture proportions ($s_{1,0}$ and $s_{2,0}$.) and the constant sex-specific contributions (s_1^f, s_1^m, etc.). Additionally, the sex-specific contributions from the admixed population (h^f and h^m) need to be factored in generations following initial admixture. The expected autosomal ancestry proportion is then given by Equation 37 in Goldberg, Verdu, and Rosenberg (2014):[4]

$$\mathbb{E}\left[H_{1,g}^A\right] = \begin{cases} s_{1,0}, & g = 1 \\ s_{1,0} h^{g-1} + s_1 \dfrac{1 - h^{g-1}}{1 - h}, & g \geq 2 \end{cases} \qquad \text{(S10)}$$

where s_1 and h are the mean of the constant sex-specific contributions from S_1 and H, respectively.

Initially, the X chromosomal ancestry proportion in males only depends on the initial female contributions from S_1 (i.e., $g = 1$). From the second generation on, it depends on the constant female contributions from S_1 (i.e., s_1^f) and the female contributions from the admixed population (h^f) together with the X chromosome ancestry proportion in females in the previous generation ($\mathbb{E}\left[H_{1,g-1,f}^X\right]$). Then, the recursion equation for the expected X chromosomal ancestry proportions in males is given by Equation A2 in Goldberg and Rosenberg (2015):[2]

$$\mathbb{E}\left[H^X_{1,g,m}\right] = \begin{cases} s^f_{1,0}, & g = 1 \\ s^f_1 + h^f\,\mathbb{E}\left[H^X_{1,g-1,f}\right], & g \geq 2 \end{cases} \qquad \text{(S11)}$$

The X chromosomal ancestry fraction in females depends on the overall constant contributions from source population S_1 (i.e., s_1), the female contributions from the admixed population (h^f) together with the X chromosomal ancestry proportion in females in the previous generation ($\mathbb{E}\left[H^X_{1,g-1,f}\right]$), and the male contributions from the admixed population (h^m) together with the X chromosomal ancestry proportion in males in the previous generation, which is equal to X chromosomal ancestry proportion in females two generations ago ($\mathbb{E}\left[H^X_{1,g-2,f}\right]$). The expected X chromosomal ancestry proportion in females is then given by the second-order Equation:

$$\mathbb{E}\left[H^X_{1,g,f}\right] = \begin{cases} s_{1,0} & g = 1 \\ s_1 + \frac{1}{2}\left(s_{1,0}h^f + s^f_{1,0}h^m\right) & g = 2 \\ s_1 + \frac{h^f}{2}\mathbb{E}\left[H^X_{1,g-1,f}\right] + \frac{h^m}{2}\left(s^f_1 + h^f\,\mathbb{E}\left[H^X_{1,g-2,f}\right]\right) & g \geq 3 \end{cases} \qquad \text{(S12)}$$

which is Equation A3 in Goldberg and Rosenberg (2015).[2] Goldberg and Rosenberg also derived a closed-form expressions for the expectation of X chromosomal ancestry proportion in females and males (Equations 17 and 18 in their paper), but we omit them here for simplicity.[2]

Estimating sex bias for constant, nonzero admixture

The expectations of the X chromosomal ancestry proportions in females and males depend on initial admixture proportions and constant contributions during the first few generations immediately after initial admixture. In the long-term, however, the expectations only depend on the constant contributions. Since American admixture happened approximately 15 generations ago, the effect of the initial admixture on the expected X chromosomal ancestry proportion is erased.[2] For this reason, the choices of initial sex-specific contributions have negligible effects on the inferred sex ratios. Here, we chose 0.5 for females and males.

Then, a grid search using 0.02 increments of permissible constant sex-specific contributions (i.e., $0 < h < 1$) was performed to identify combinations of s^f_1, s^m_1, s^f_2, and s^m_2 that can describe the observed ancestry proportions using Equations S10 - S12. The goodness of the parameter fit was assessed by computing the Euclidean distance between the observed ancestry proportions and the expected ancestry proportions:

$$D = \sqrt{\left[q_A - \mathbb{E}\left[H^A_{1,g}\right]\right]^2 + \left[q_X - \left(p_f\,\mathbb{E}\left[H^X_{1,g,f}\right] + p_m\,\mathbb{E}\left[H^X_{1,g,m}\right]\right)\right]^2} \qquad \text{(S13)}$$

where p_f and p_m are the fraction of females and males in the sample, respectively. Equation S13 is Equation 25 in Goldberg and Rosenberg (2015).[2] We accepted all parameter combinations for which $D \leq 0.01$. Furthermore, only parameter combinations were accepted for which the sex ratios were neither zero nor infinity (i.e., when only one sex contributed).

The model deals with a scenario of two-way admixture. We extended it to the present scenario of three-way admixture (i.e., African, European, and Native American admixture) by computing plausible sex-specific contributions for each ancestry separately while aggregating the contributions of the two other ancestries. Code implementing this model can be found at https://github.com/LachanceLab/sex_biased_admixture.

Simulations of American admixture and estimation of ancestry proportions

To evaluate how sampling sizes and violations of demographic assumptions impact the results of the above models, we simulated American admixture 15 generations ago. We used Gravel's model of African, European, and Asian demographic history to simulate ancestral continental populations.[6] Then, we simulated three-way admixture of these continental populations with admixture proportions taken from Browning and Browning (2018).[7] Because the admixture proportions were chosen arbitrarily, we replace the notion of African, European, and Asian source populations with source populations 1 (S_1), 2 (S_2), and 3 (S_3), respectively. This is to avoid any misleading associations with a specific ancestry by the reader later in the text.

American admixture

Prior to admixture, we simulated three continental ancestries using Gravel's model.[6] The ancestral S_1 population had an effective population size of 7,310 and experienced a population size expansion to 14,475 individuals 5,919 generations ago (148kya assuming 25 years per generation). The ancestral population of S_2 and S_3 split 2056 generations ago (~51kya), experiencing an initial bottleneck with a population size of 1,861. The split of S_2 and S_3 occurred 940 generations ago (~23kya). S_2 and S_3 then grew exponentially at rates of 0.38% and 0.48%, respectively. Symmetrical migration between the different populations was simulated at rates determined by Gravel et al.[6] Three-way admixture of S_1, S_2, and S_3 was simulated 15 generations ago, using the admixture proportions from Browning et al. (2018) (1/6 S_1, 1/3 S_2, and 1/2 S_3).[8] The contributions from S_1 and S_3 were simulated to be female-biased with ratios of two and 1.25 females to one male, respectively, while contributions from S_2 were simulated to be male-biased with a ratio of two males to one female. These ratios guaranteed the same number of females and males in the initially admixed population, given the admixture proportions.

Similarly to Gravel et al., we also assumed a recombination rate of 1 x 10^{-8} and a mutation rate of 2.36 x 10^{-8} per base pair per generation.[6] The mutation rate is higher than more recent estimates,[9] but we decided to stick to it to ensure realistic levels of genetic diversity.

We simulated a 100 Mb autosome, 100Mb sex chromosomes (X and Y chromosome), and 20 kb mitochondrial DNA in SLiM v3.7.1 using tree-sequence recording.[10,11] Tree-sequence recording allowed to omit neutral mutation during the forward simulations and superimpose them later, increasing computational efficiency. To ensure full coalescence, the tree sequences were first recapitated in Python3 using pyslim v0.7. Subsequently, neutral mutations were added using msprime v1.1.1 in Python3.[12,13] Finally, random individuals were sampled from each population without replacement (i.e., S_1, S_2, S_3, and the admixed population), and the corresponding data was written to a VCF file. Code implementing these simulations can be found at: https://github.com/LachanceLab/sex_biased_admixture.

Alternative demographic scenarios

As American admixture was more complex than a simple single admixture event,[8,14–19] we assessed the impact of violations of demographic assumptions on inferred sex ratios by simulating alternative scenarios, including population growth, gene flow after initial admixture, and sex-specific assortative mating.

The models assume a constant population size of the admixed population, but admixed populations in the Americas evidently experienced recent population growth.[8] Therefore, we assessed the effect of population growth on the inferred sex-specific contributions by simulating exponential population growth of the admixed population at a rate of $r = 0.05$.

Furthermore, the assumption of no gene flow after initial admixture has been violated during admixture in the Americas.[14] To assess the effect of violations of this assumption, we simulated constant, nonzero gene flow from the source population into the admixed population with a migration rate from S_1 of $m_1 = 0.05$, an S_2 migration rate of $m_2 = 0.025$, and an S_3 migration rate of $m_3 = 0.01$. Constant migration was assumed to have the same sex biases as the initial admixture event (i.e., two S_1 females for every male, two S_2 males for every female, and 1.25 S_3 females for every male). As population growth and constant gene flow both have happened to admixed populations in the Americas, we also assessed the combined effect on inferred sex ratios (i.e., $m_1 = 0.05$, $m_2 = 0.025$, $m_3 = 0.01$, and $r = 0.05$).

From historical records, it is well established that random mating did not occur in admixed populations in the Americas. For instance, European males mated with African or Native American females more frequently than European females mated with African or Native American males as a consequence of anti-miscegenation laws and sexual exploitation of enslaved women by enslavers.[15–19] In previous theoretical work, Goldberg, Rastogi, and Rosenberg showed that assortative mating does not affect sex ratios inferred from X chromosomal and autosomal data because it does not change expected mean ancestry proportions. They showed that assortative mating only increases the variances if mating preferences are symmetrical with respect to sex. However, sex-specific assortative changes expected mean ancestry proportions, and thus the appalling laws and social norms in the aftermath of the transatlantic slave trade may have affected sex ratios inferred from X chromosomal and autosomal data.[20] To evaluate the effect of sex-specific assortative mating, we performed simulations in which S_2 females are asymmetrically more likely to mate with S_2 males. This was achieved by rejecting a mating partner of an S_2 female in 40% of the cases ($p=0.4$) if the partner was a non-S_2 male and selecting a random S_2 male instead. In this mating scheme, female contributions from the different populations remain unchanged, but male contributions change, with S_2 males contributing more and males from all other populations contributing less than expected under a random mating scheme (Table S3). For example, for a single admixture event, the male contributions from S_2 are increased by approximately 0.05, while the male contributions from S_1 and S_3 are decreased by approximately 0.01 and 0.04. We also evaluated the effects of more extreme sex-specific assortative mating, rejecting non-S_2 males in 90% of the cases ($p=0.9$). Code implementing these simulations can be found at: https://github.com/LachanceLab/sex_biased_admixture.

Post-processing

The obtained VCF files were normalized using bcftools v1.14-36-g9560eb.[21] Subsequently, linkage disequilibrium (LD) pruning of SNPs with an r^2 threshold of 0.1 (*--indep-pairwise 50 kb 1 0.1*) and minor allele frequency filtering (MAF ≥ 0.01) was performed on the simulated X chromosomes and autosomes using plink v2.00a3LM.[22] X chromosomal and autosomal ancestry proportions were then inferred using ADMIXTURE v1.3.0 with *K=3*, using sampled individuals from the source populations for supervised training (*--supervised*).[23] Males were treated as haploid on the X chromosome (*--haploid="male:23"*). Admixed individuals with more than 95% of one ancestry were excluded from subsequent analyses.

Established methods for inferring mtDNA and Y chromosome haplogroups such as haplogrep2[24] and yHaplo[25] rely on human reference data, and thus cannot be used for inferring haplogroups of our simulated data. Instead, mtDNA and Y chromosome haplogroups were inferred by performing a PCA with plink v2.00a3LM and subsequent clustering of the samples of the three source populations (i.e., S_1, S_2, and S_3). For each cluster, a representative ancestral haplogroup was assigned based on the population from which most individuals in each cluster were sampled. For example, if a cluster consisted of 100 individuals, of which 90 were sampled from S_1, six from S_2, and four from S_3, the representative ancestral haplogroup of this cluster would be S_1. Admixed individuals were then assigned the ancestral haplogroup of their closest neighbor, who was sampled from one of the source populations.

We tested several clustering algorithms implemented in scikit-learn v1.0.2[26] (incl. agglomerative clustering, k-Means, DBSCAN, affinity propagation, spectral clustering, and mean shift) with various hyperparameters. We selected the algorithm and hyperparameters that produced the most homogenous clusters, i.e., clusters mainly consisted of individuals sampled from the same source population. The homogeneity of clusters was assessed using the loss function:

$$L = \sum_{i=1}^{n} \sum_{j=1}^{n} d(I, j) \times \theta \qquad \text{(S14)}$$

where *n* is the number of sampled individuals, $d(i,j)$ is the Euclidean distance between the *ith* and *jth* individual in the PCA space, and θ is a Heaviside step function, which is one if the *ith* and the *jth* individual were clustered together but were sampled from different populations, and 0 otherwise.

This clustering-based approach of assigning haplogroups is imperfect as migration can possibly bias the haplogroup assignment (this is a general caveat when using mtDNA and Y chromosome haplogroups), e.g., an individual that recently migrated from S_1 to S_2 probably has an S_1 haplogroup and not an S_2. However, in practice, we found that the low levels of simulated migrations between the source population did not bias our haplogroup assignment in admixed individuals, and the predicted frequencies of S_1, S_2, and S_3 haplogroups in the admixed population reflected the simulated admixture proportions.

Code implementing these analyses can be found at https://github.com/LachanceLab/sex_biased_admixture.

Supplemental Results

Estimating sex biases from summary statistics of Micheletti et al. using a constant admixture model

Although we did not observe sex ratios as large as those reported by Micheletti et al. or any cases of model failure when simulating constant gene flow after initial admixture (Table 2), it may have been a confounding factor in their analysis. For this reason, we estimated sex biases based on the X chromosomal and autosomal ancestry proportions reported by Micheletti et al. in Table S9[1] (also see Table 1) using a model that assumes constant, nonzero admixture.[2] This model identifies a set of constant, nonzero sex-specific contributions for each population that fit the observed X chromosomal and autosomal ancestry proportions. In general, this model yields more moderate sex ratios (Figure S1). However, often the range of possible sex ratios is wide, indicating that sex biases of various magnitudes and either female or male sex bias could explain the data. For instance, in regions with substantial amounts of unassigned ancestry (i.e., the Latin Caribbean, northern South America, central South America, and Central America), the range of possible sex ratios for European ancestry (yellow) spans a wide range, and median sex ratios are close to one, so that it is unclear whether European contributions were female- or male-biased (Figure S1). This wide range of possible sex ratios indicates that unassigned ancestry causes problems even if a model is used that assumes a more appropriate demographic scenario.

Supplemental References

1. Micheletti, S.J., Bryc, K., Ancona Esselmann, S.G., Freyman, W.A., Moreno, M.E., Poznik, G.D., Shastri, A.J., Agee, M., Aslibekyan, S., Auton, A., et al. (2020). Genetic Consequences of the Transatlantic Slave Trade in the Americas. Am. J. Hum. Genet. *107*, 265–277.

2. Goldberg, A., and Rosenberg, N.A. (2015). Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased Admixture on the X Chromosome. Genetics *201*, 263 LP – 279.

3. Lind, J.M., Hutcheson-Dilks, H.B., Williams, S.M., Moore, J.H., Essex, M., Ruiz-Pesini, E., Wallace, D.C., Tishkoff, S.A., O'Brien, S.J., and Smith, M.W. (2007). Elevated male European and female African contributions to the genomes of African American individuals. Hum. Genet. *120*, 713–722.

4. Goldberg, A., Verdu, P., and Rosenberg, N.A. (2014). Autosomal Admixture Levels Are Informative About Sex Bias in Admixed Populations. Genetics *198*, 1209 LP – 1229.

5. Verdu, P., and Rosenberg, N.A. (2011). A general mechanistic model for admixture histories of hybrid populations. Genetics *189*, 1413–1426.

6. 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. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. U. S. A. *108*, 11983–11988.

7. Browning, S.R., and Browning, B.L. (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. Am. J. Hum. Genet. *97*, 404–418.

8. Browning, S.R., Browning, B.L., Daviglus, M.L., Durazo-Arvizu, R.A., Schneiderman, N., Kaplan, R.C., and Laurie, C.C. (2018). Ancestry-specific recent effective population size in the Americas. PLOS Genet. *14*, 1–22.

9. Scally, A. (2016). The mutation rate in human evolution and demographic inference. Curr. Opin. Genet. Dev. *41*, 36–43.

10. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. Mol. Biol. Evol. *36*, 632–637.

11. Haller, B.C., Galloway, J., Kelleher, J., Messer, P.W., and Ralph, P.L. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. Mol. Ecol. Resour. *19*, 552–566.

12. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Comput. Biol. *12*, 1–22.

13. Nelson, D., Kelleher, J., Ragsdale, A.P., Moreau, C., McVean, G., and Gravel, S. (2020). Accounting for long-range correlations in genome-wide simulations of large cohorts. PLOS Genet. *16*, 1–12.

14. Jordan, I.K., Rishishwar, L., and Conley, A.B. (2019). Native American admixture recapitulates population-specific migration and settlement of the continental United States. PLOS Genet. *15*, e1008225.

15. Pascoe, P. (2009). What Comes Naturally: Miscegenation Law and the Making of Race in America. Oxford University Press.

16. Wallenstein, P. (1994). Freedom: Personal Liberty and Private Law" Race, Marriage, and the Law of Freedom: Alabama and Virginia, 1860s-1960s. Chic. Kent. Law Rev. *70*, no. 371.

17. Hartman, S. V (1997). Scenes of Subjection: Terror, Slavery, and Self-making in Nineteenth-century America. Oxford University Press.

18. Jennings, T. (1990). “Us Colored Women Had to Go Though A Plenty”: Sexual Exploitation of African-American Slave Women. J. Womens. Hist. *1*, no. 3.

19. Pokorak, J.J. (2007). Rape as a badge of slavery: The legal history of, and remedies for, prosecutorial race-of-victim changing disparities. Nevada Law Journal 7, no. 1.

20. Goldberg, A., Rastogi, A., and Rosenberg, N.A. (2020). Assortative mating by population of origin in a mechanistic model of admixture. Theor. Popul. Biol. *134*, 129–146.

21. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, Volume 10, Issue 2, February 2021, giab008.

22. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, Volume 4, Issue 1, December 2015, s13742–015–0047–8,

23. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

24. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. *44*, W58-63.

25. Poznik, G.D. (2016). Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. BioRxiv 88716. doi: https://doi.org/10.1101/088716

26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.