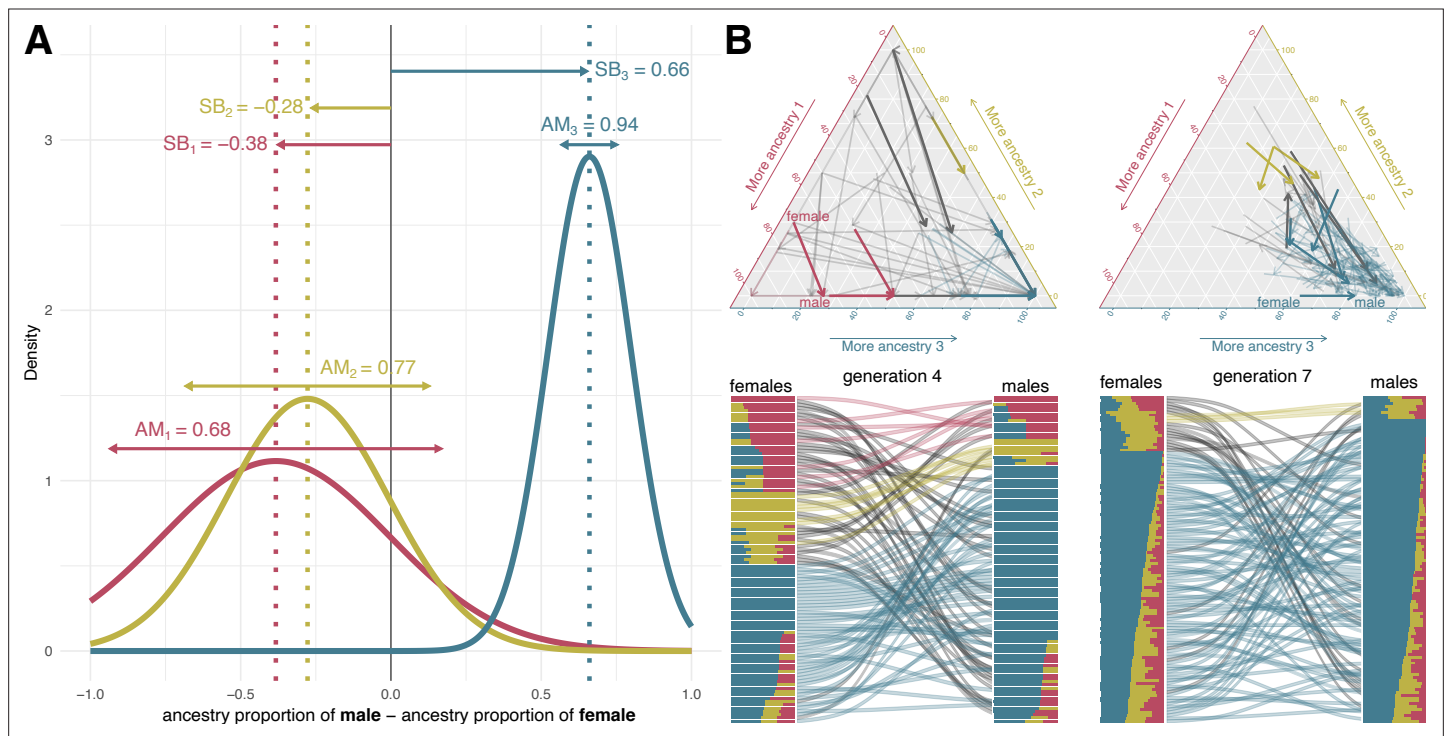


---

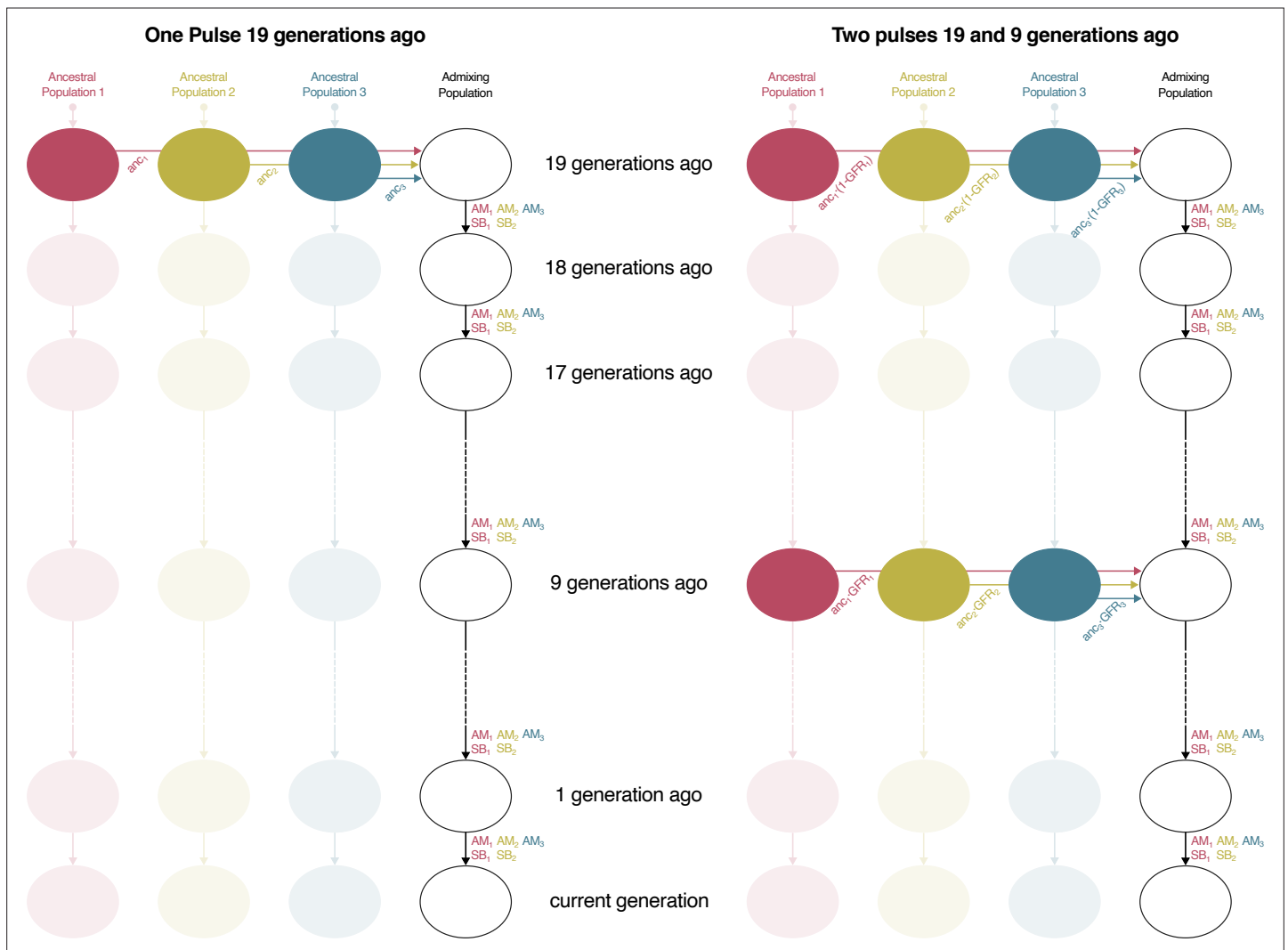
## Figures and figure supplements

The genomic footprint of social stratification in admixing American populations

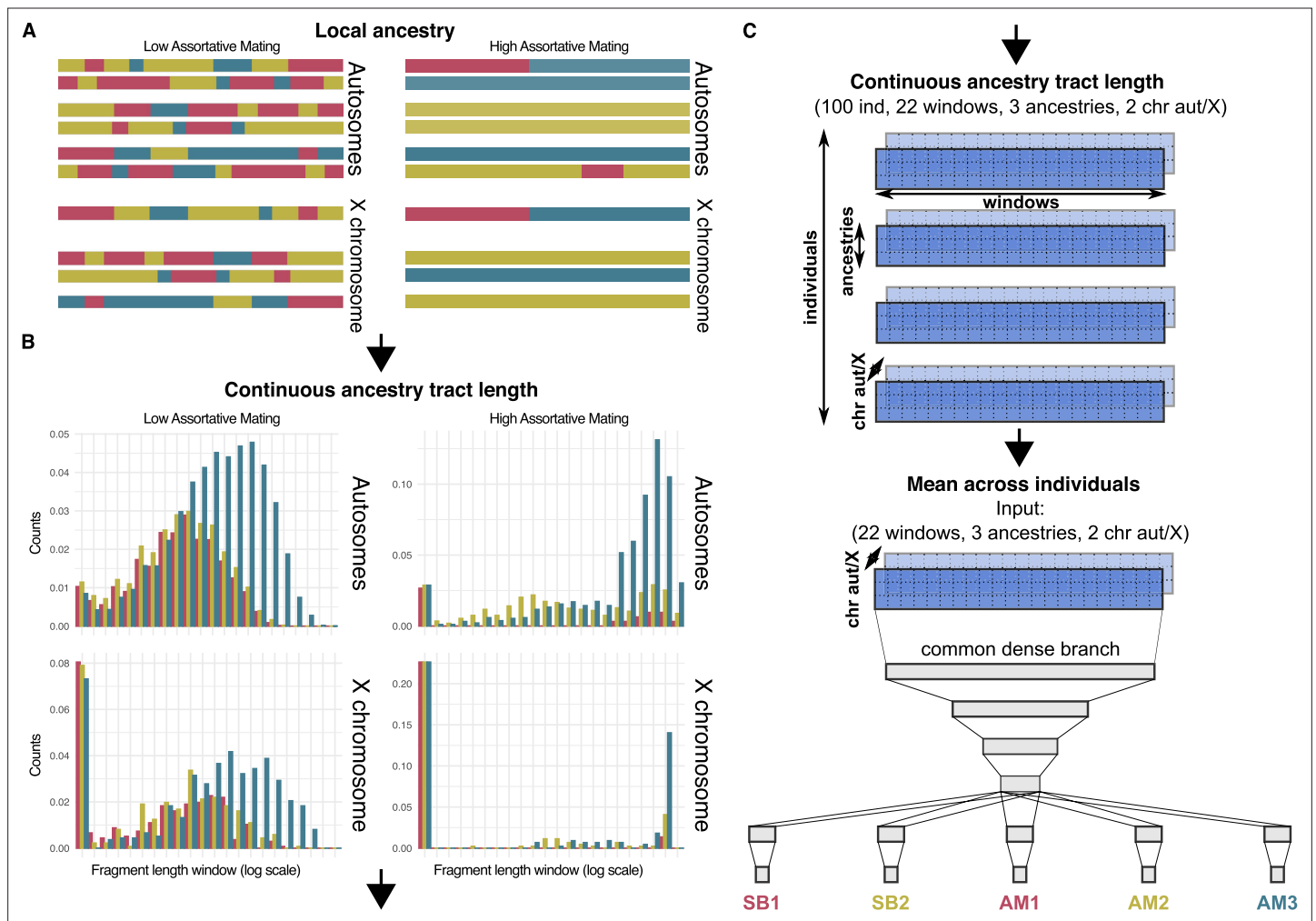
**Alex Mas-Sandoval et al.**



**Figure 1.** Mating model. **(A)** Assortative mating (AM) and sex bias (SB) values that modulate the mating probabilities in a simulation example of 19 generations from the colonization of America to nowadays. The mating probability for a given couple is set as a function of the differences in the genetic ancestry proportions for each ancestry. We assume the mating probability follow a three-dimensional normal distribution. In this normal distribution, SB sets the expected value and AM is inversely proportional to its variance. **(B)** Ancestry proportions of mating couples at generations 4 and 7 in ternary plots (top) and barplots (bottom) based on the mating probabilities defined in A. In the top plots, each arrow represents a couple. The arrow tail and head coordinates in the ternary plots show the ancestry proportions of the female and the male, respectively. In the bottom, the barplots represent male and female ancestry proportions, linked by curved lines reflecting mating. Red, yellow, and blue correspond to ancestries 1, 2, and 3. The arrows in the ternary plot and the lines between barplots representing a mating couple are colored with the color corresponding to the predominant ancestry in both male and female, and they are depicted in black if it differs between them.

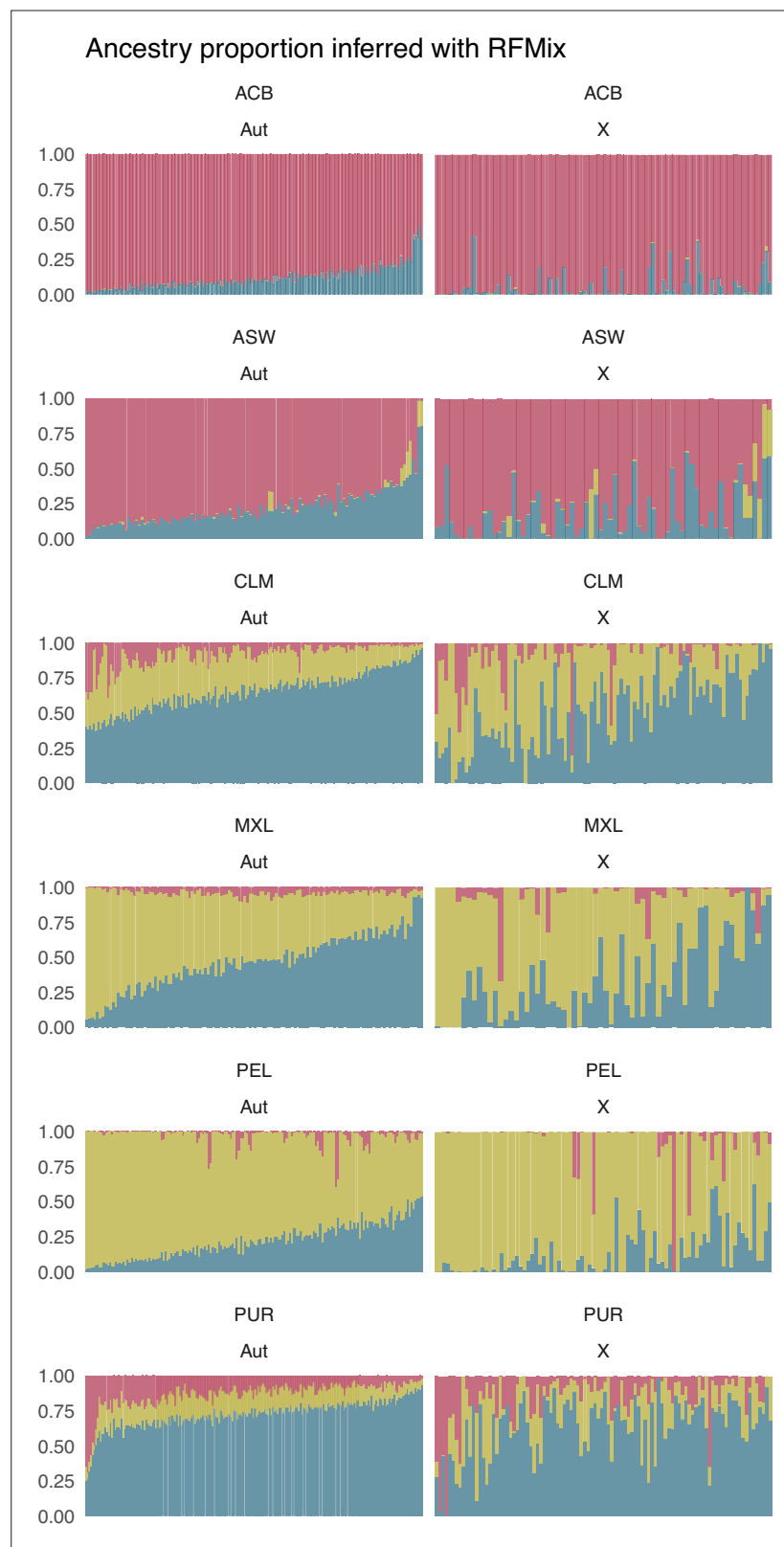


**Figure 1—figure supplement 1.** Admixture Model for One Pulse and Two Pulses. In the One Pulse model, the first generation of the admixing population is formed 19 generations ago by a migration pulse from the ancestral population that equals the ancestry proportions observed in the studied population. In the Two Pulses model, an extra pulse is modeled 9 generations ago. The proportion of migrants that arrive in the second pulse is modeled by gene flow rate (GFR). In the admixing population, there is no random mating from one generation to the other, but mating is constrained by the mating parameters ( $AM_1$ ,  $AM_2$ ,  $AM_3$ ,  $SB_1$  and  $SB_2$ ).

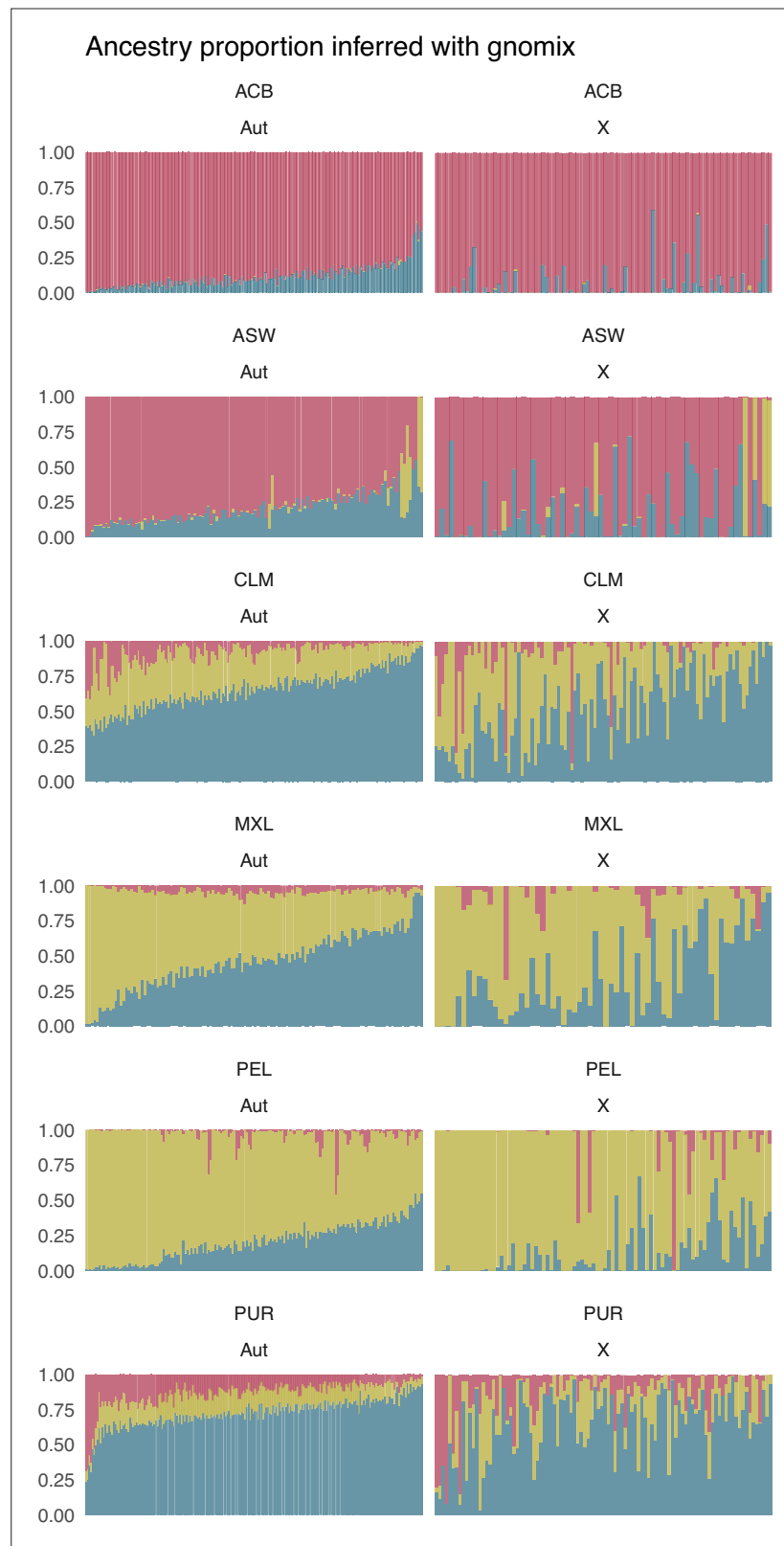


**Figure 2.** Local ancestry, Continuous ancestry tract length and Neural network. **(A)** Schematic view of the autosomal and sex chromosomes split into the continuous ancestry tracts inherited from each of the three ancestries after a local ancestry analysis with RFMix. **(B)** Continuous ancestry tract length profile displaying the number of tracts for each ancestry in each tract length bin. The break points that define the bin widths are set in a logarithmic scale. **(C)** Matrix representing the continuous ancestry tract length profile accounting for the amount of tracts in each length bin, for each ancestry in either autosomal or sex chromosome for each individual. The mean across individuals summarises the four-dimensional matrix in a population three-dimensional matrix, which is used as the input of the neural network. The neural network has four fully connected layers that split into a branch for each parameter, each one made of a last hidden layer connected to the output layer.

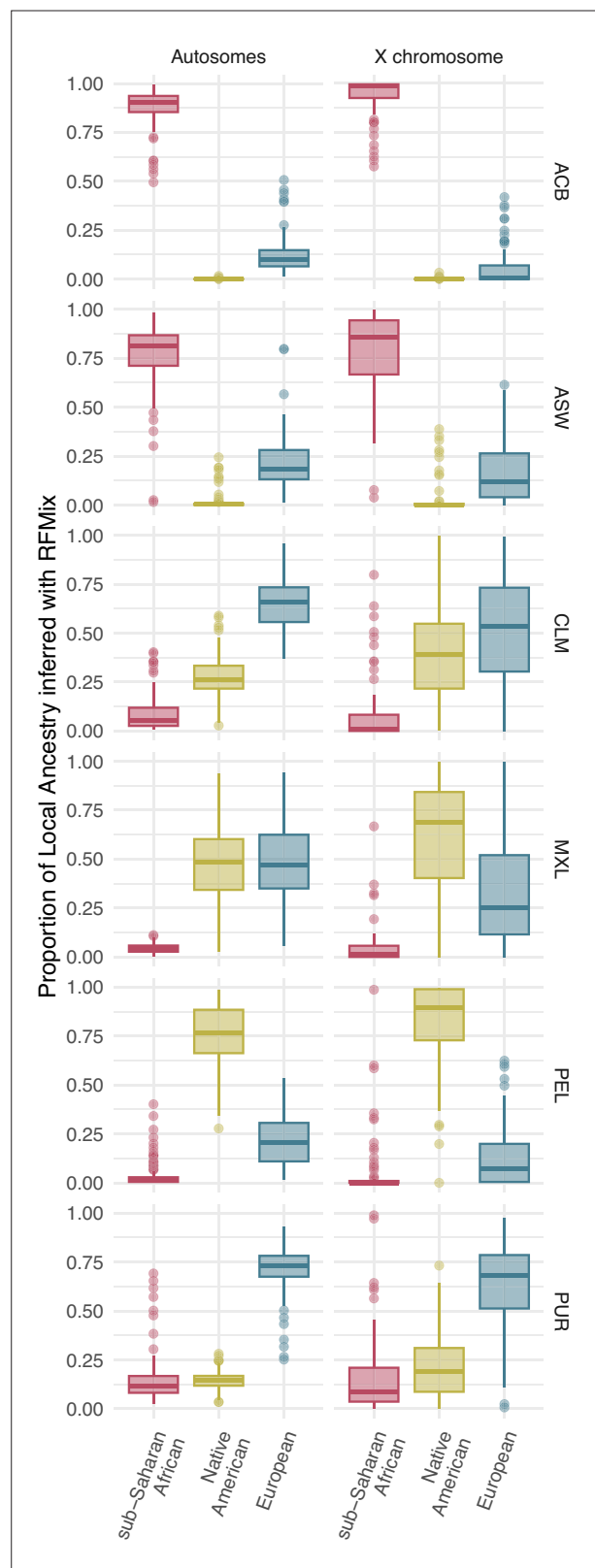




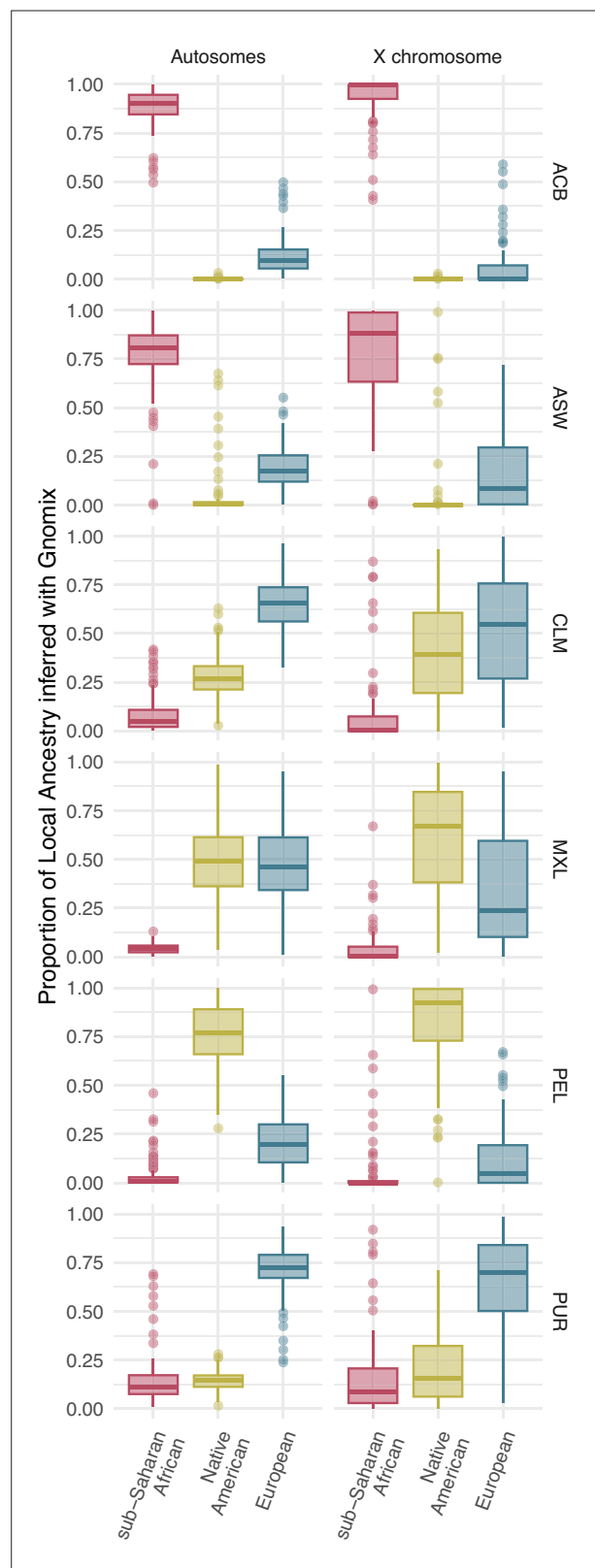
**Figure 2—figure supplement 1.** Individual proportions of sub-Saharan (red), Native American (green), and European (blue) ancestry were inferred after a Local Ancestry analysis with RFMix for autosomes (Aut), on the left, and X chromosome (X), on the right, for each population. Each vertical bar represents an individual. Individuals are sorted along the x-axis based on the autosomal proportions inferred with RFMix in both autosomes and X chromosome plots.



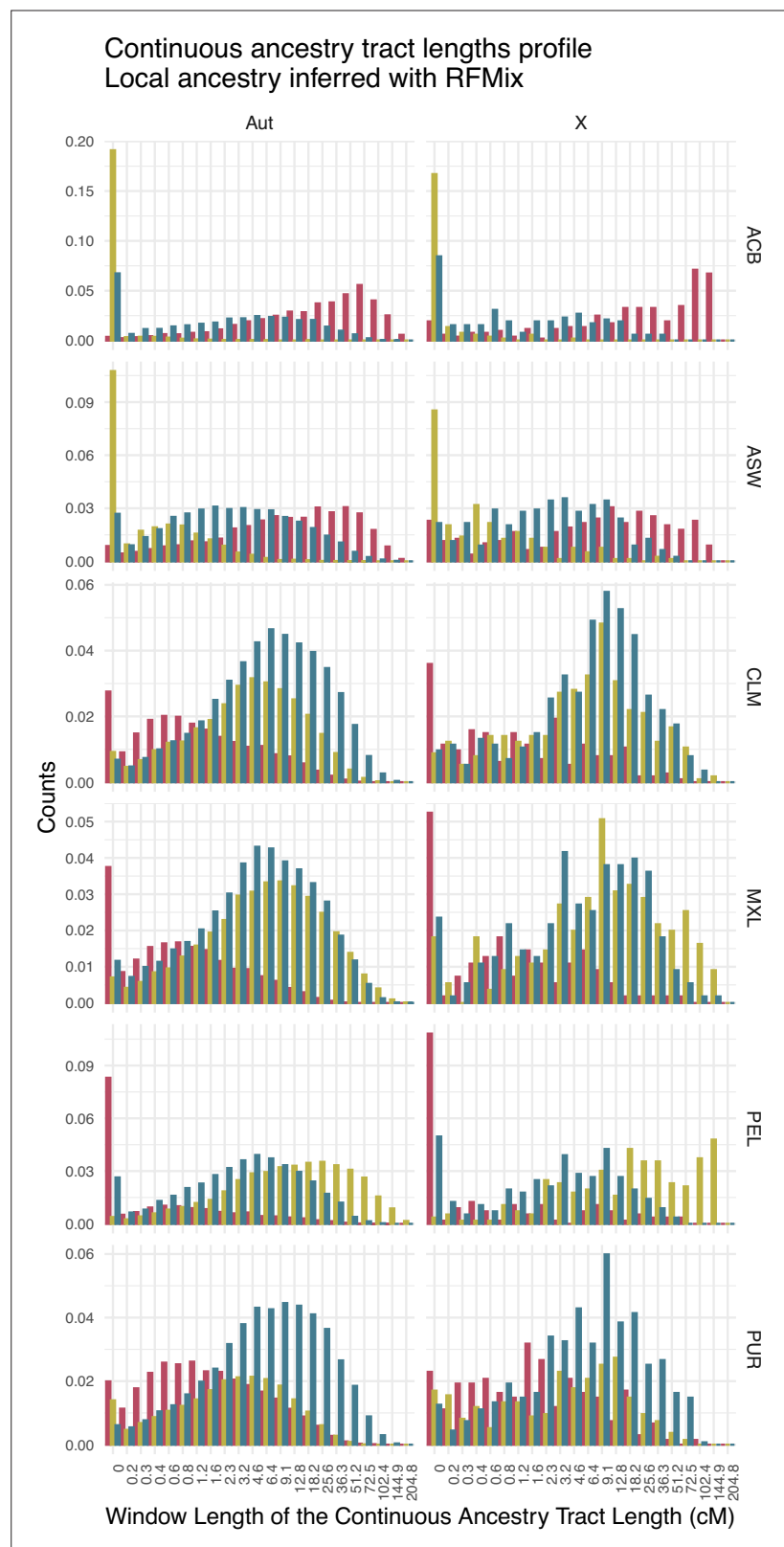
**Figure 2—figure supplement 2.** Individual proportions of sub-Saharan (red), Native American (green), and European (blue) ancestry inferred after a Local Ancestry analysis with Gnomix for autosomes (Aut), on the left, and X chromosome (X), on the right, for each population. Each vertical bar represents an individual. Individuals are sorted along the x-axis based on the autosomal proportions inferred with RFMix in both autosomes and X chromosome plots.



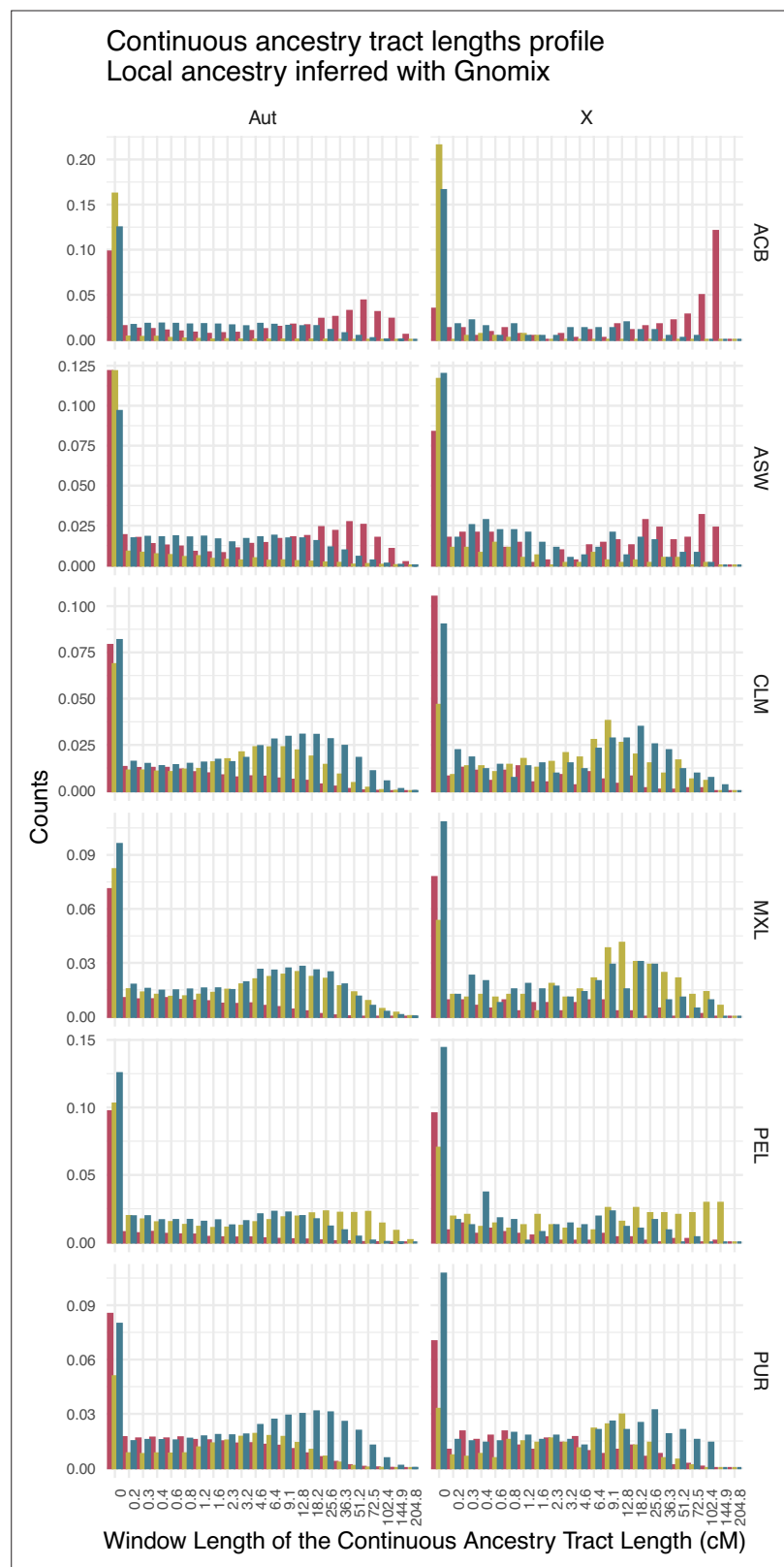
**Figure 2—figure supplement 3.** Distribution of individual proportions of sub-Saharan (red), Native American (green), and European (blue) ancestry inferred after a Local Ancestry analysis with RFMix for autosomes (Aut), on the left, and X chromosome (X), on the right, for each population. The box limits are the 25<sup>th</sup> and 75<sup>th</sup> percentiles and the points show the outliers 1.5 times the interquartile range above the 75<sup>th</sup> percentile and below the 25<sup>th</sup> percentile.



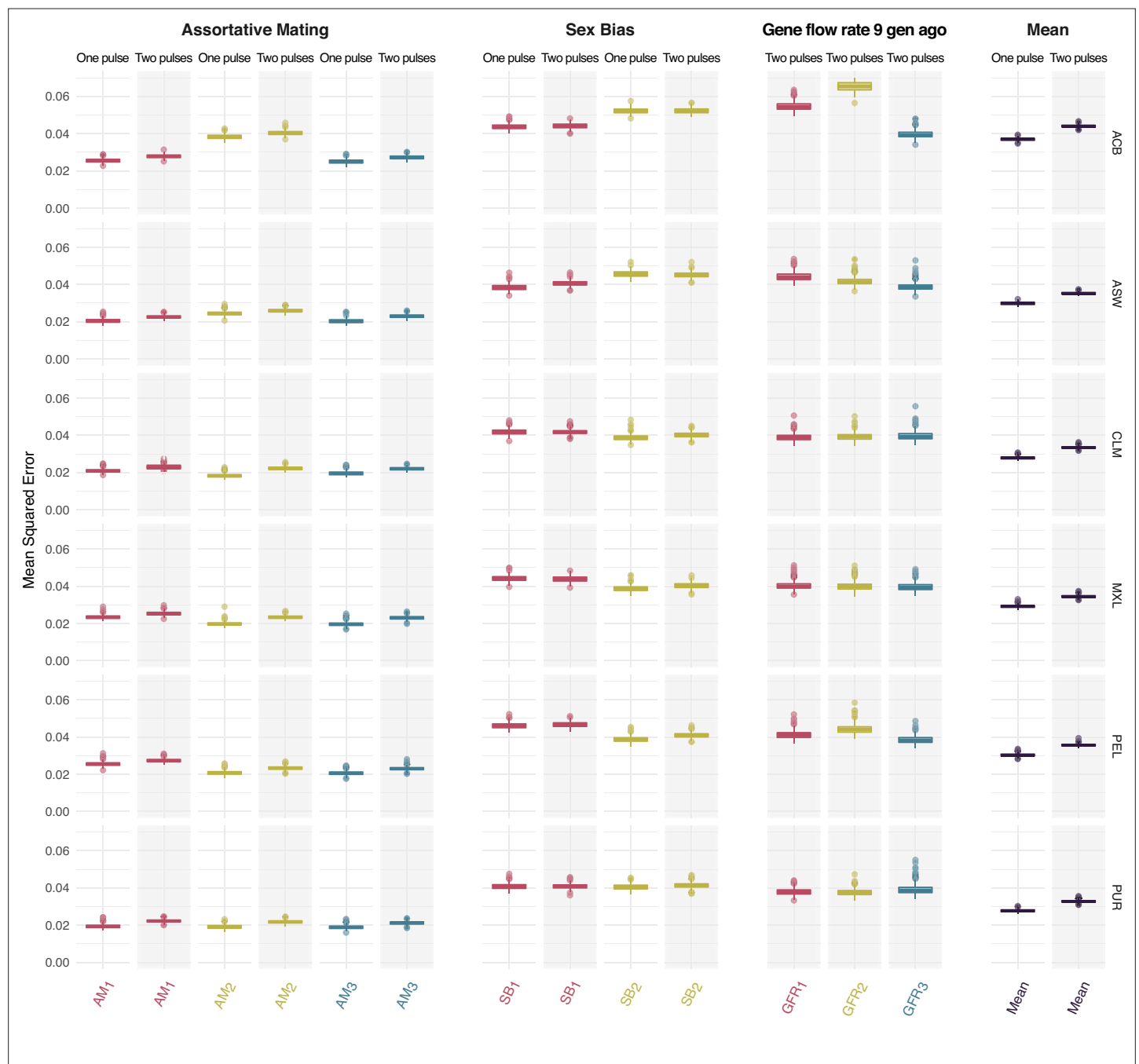
**Figure 2—figure supplement 4.** Distribution of individual proportions of sub-Saharan (red), Native American (green), and European (blue) ancestry inferred after a Local Ancestry analysis with Gnomix for autosomes (Aut), on the left, and X chromosome (X), on the right, for each population. The box limits are the 25<sup>th</sup> and 75<sup>th</sup> percentiles and the points show the outliers 1.5 times the interquartile range above the 75<sup>th</sup> percentile and below the 25<sup>th</sup> percentile.



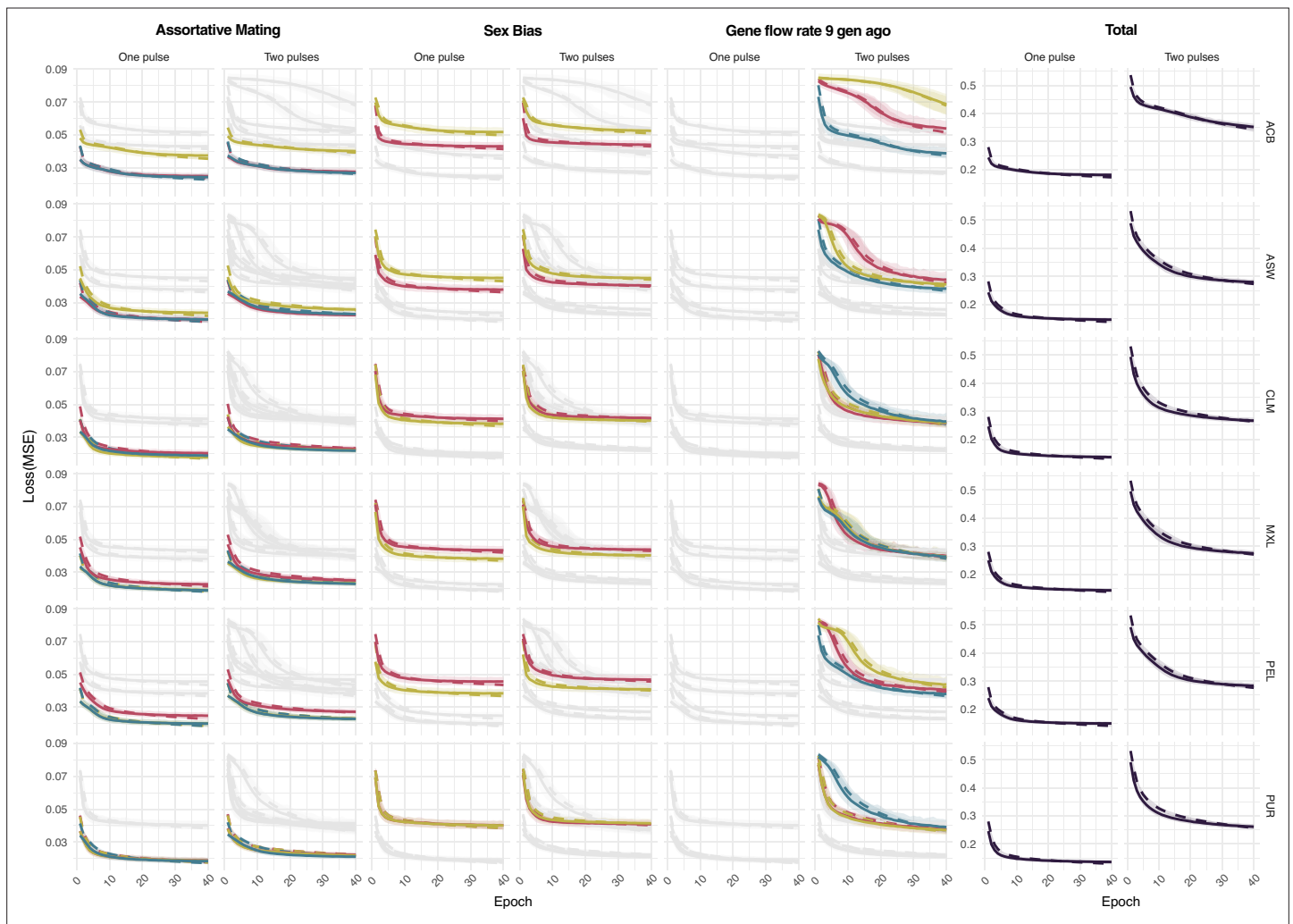
**Figure 2—figure supplement 5.** Continuous ancestry tract lengths profile showing the amount of fragments in each length window for sub-Saharan (red), Native American (green), and European (blue) ancestry for each population, after a local ancestry analysis with RFMix.



**Figure 2—figure supplement 6.** Continuous ancestry tract lengths profile showing the amount of fragments in each length window for sub-Saharan (red), Native American (green), and European (blue) ancestry for each population, after a local ancestry analysis with Gnomix.

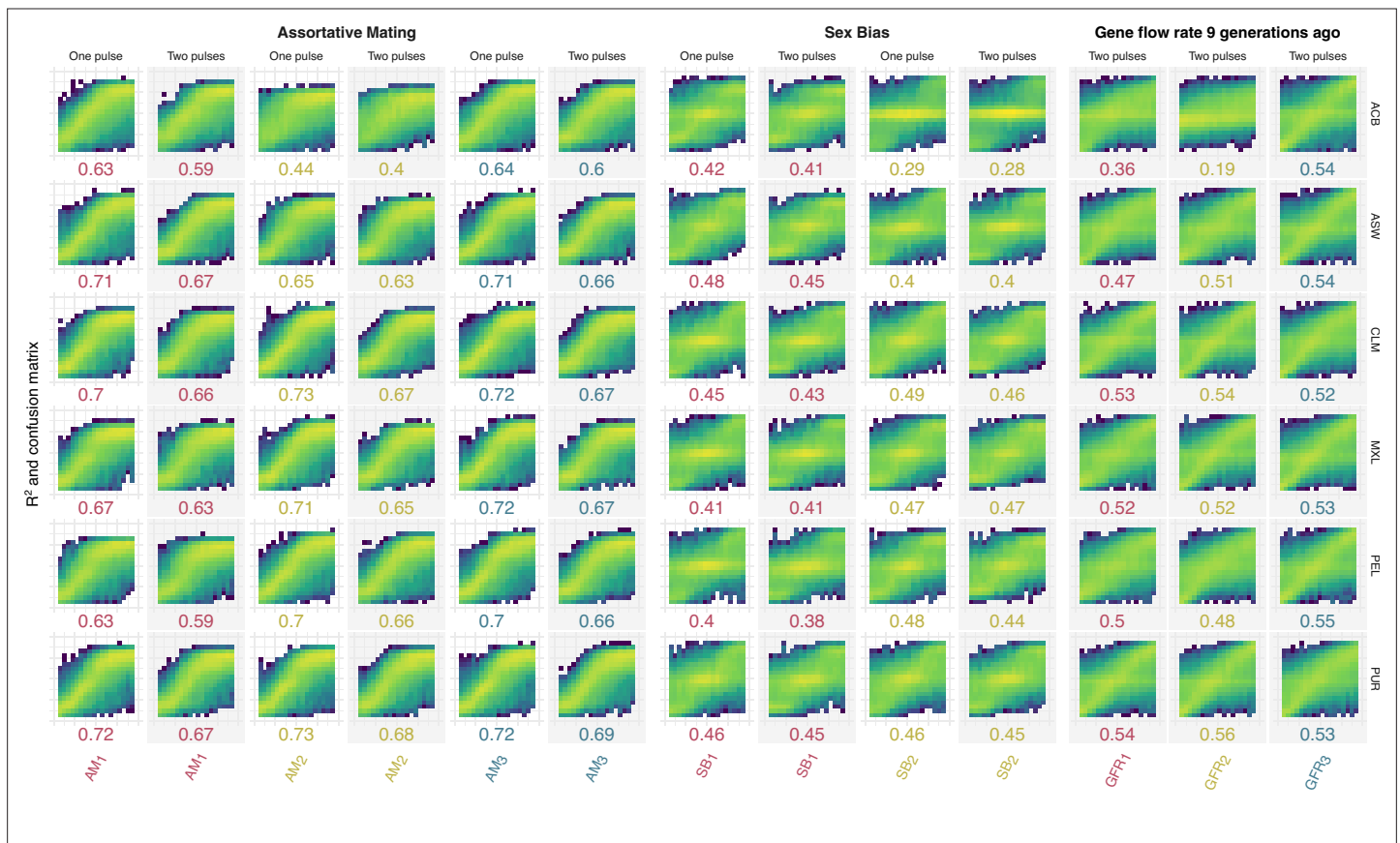


**Figure 3.** Mean squared error comparing true and predicted values at testing for the *assortative mating*, *sex bias*, *gene flow rate 9 generations ago* parameters for each ancestry for both One Pulse and Two Pulses models and mean values for each model. Each color represents the values for a different ancestry (red, yellow, and blue for ancestries 1, 2, and 3 respectively, which correspond to sub-Saharan, Native American and European ancestries). The boxplot shows the distributions of values across the 1000 trained neural networks. The box limits are the 25<sup>th</sup> and 75<sup>th</sup> percentiles and the points show the outliers 1.5 times the interquartile range above the 75<sup>th</sup> percentile and below the 25<sup>th</sup> percentile.



**Figure 3—figure supplement 1.** Loss function at training for the assortative mating, sex bias, gene flow rate 9 generations ago parameters for each ancestry for both One Pulse and Two Pulses models and mean values for each model. Each color is related to a different ancestry as in **Figure 3**.





**Figure 3—figure supplement 2.** Confusion matrices and  $R^2$  comparing true (x-axis) and predicted (y-axis) values at testing for assortative mating, sex bias, gene flow Rate 9 generations ago parameters for each ancestry for both One Pulse and Two Pulses models. Each color is related to a different ancestry as in **Figure 3**.

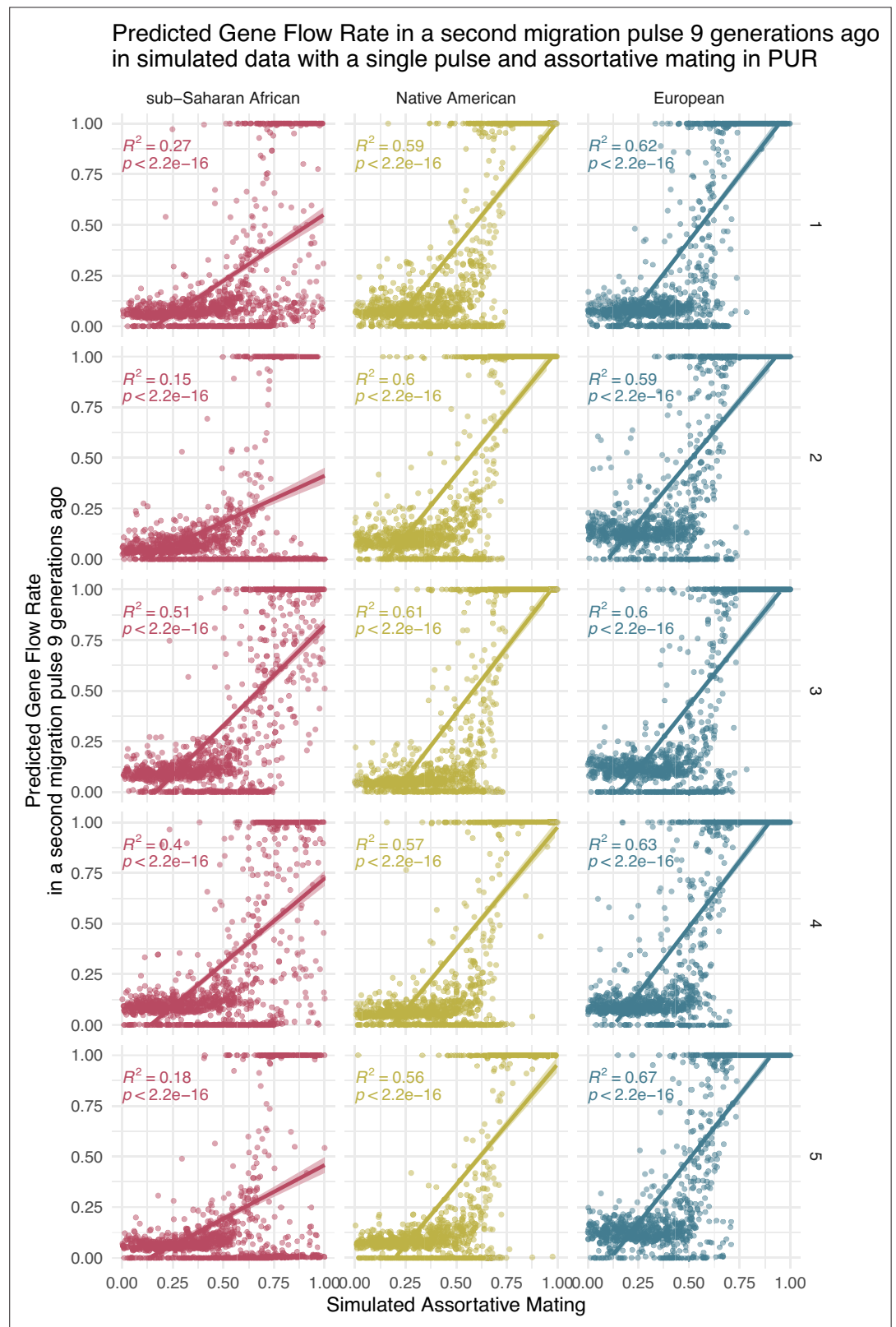


**Figure 4.** Mating probabilities, migration pulses and admixture dynamics. **(A)** Mating probabilities as a function of male and female proportions of each ancestry, for each population. **(B)** Migration pulses of each ancestry according to scenarios allowing One Pulse and Two Pulses for each population. The y-axis represents the cumulative increase in the ancestry-specific gene pool relative to the final ancestry proportions, at each generation. The ancestry proportions at generation 19 represent the observed ancestry proportions of each population in real data. The increase in the cumulative ancestry-specific gene pool is defined by gene flow rate (GFR), while the slope of the increase is represented inversely proportional to assortative mating (AM).

Figure 4 continued on next page

*Figure 4 continued*

(C) Composite likelihood ratio comparing Two Pulses model vs. One Pulse model, for each ancestry for both the X chromosome and the autosomes. In this plot, positive values show a higher likelihood of the Two Pulses model based on the fit of the real fragment lengths in the distribution of fragment lengths of simulated data under the AM and sex bias (SB) parameters predicted by the neural network. Error bars define the 95% CI obtained by bootstrapping the tract lengths profile.

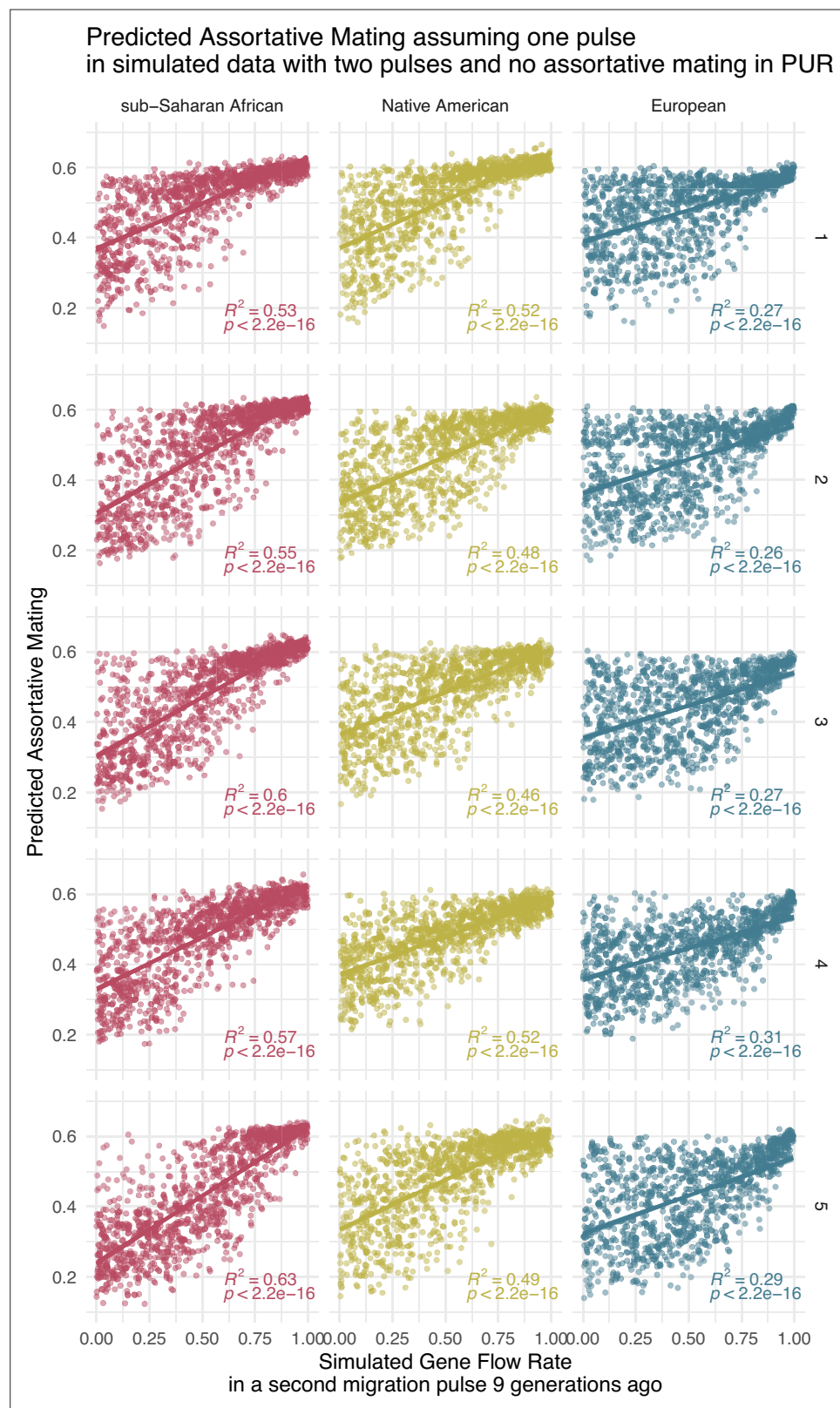


**Figure 4—figure supplement 1.** Correlation between simulated assortative mating (AM) values and predicted gene flow rate (GFR) values (Simulation only for PUR, due to large computational requirements). The Neural network for the Two Pulses model has been trained with Two Pulses simulations, but once trained it has been used to predict GFR from the output tract length of simulations of the One Pulse model that only iterates  $AM_1$ ,  $AM_2$ , and

Figure 4—figure supplement 1 continued on next page

*Figure 4—figure supplement 1 continued*

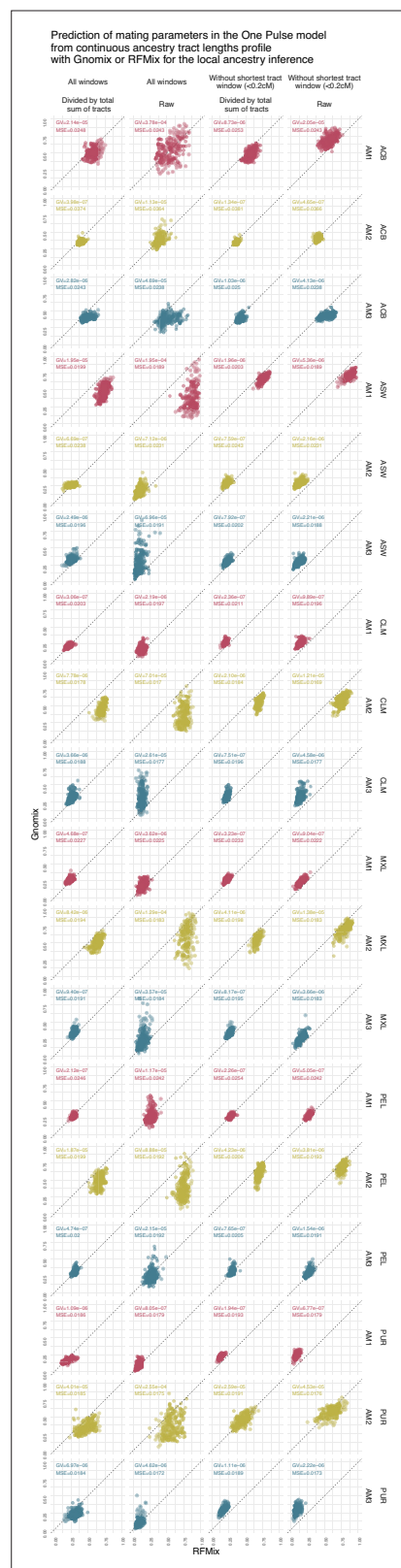
AM<sub>3</sub> but keeps SB<sub>1</sub> and SB<sub>2</sub> equal to 0. Each of the five rows of plots represents a different training of the neural network.



**Figure 4—figure supplement 2.** Correlation between simulated gene flow rate (GFR) values and predicted assortative mating (AM) values (Simulation only for PUR, due to large computational requirements). The Neural network for the One Pulse model has been trained with One Pulse simulations, but once trained it has been used to predict AM from the output tract length of simulations of the Two Pulses model that only iterates  $GFR_1$ ,  $GFR_2$ ,  
Figure 4—figure supplement 2 continued on next page

*Figure 4—figure supplement 2 continued*

and  $GFR_3$  but keeps  $SB_1$ ,  $SB_2$ ,  $AM_1$ ,  $AM_2$ , and  $AM_3$  equal to 0. Each of the five rows of plots represents a different training of the neural network.



**Figure 4—figure supplement 3.** Prediction of assortative mating in the One Pulse model using the continuous ancestry tract lengths profile from Gnomix or RFMix. Each point is the estimated

Figure 4—figure supplement 3 continued on next page

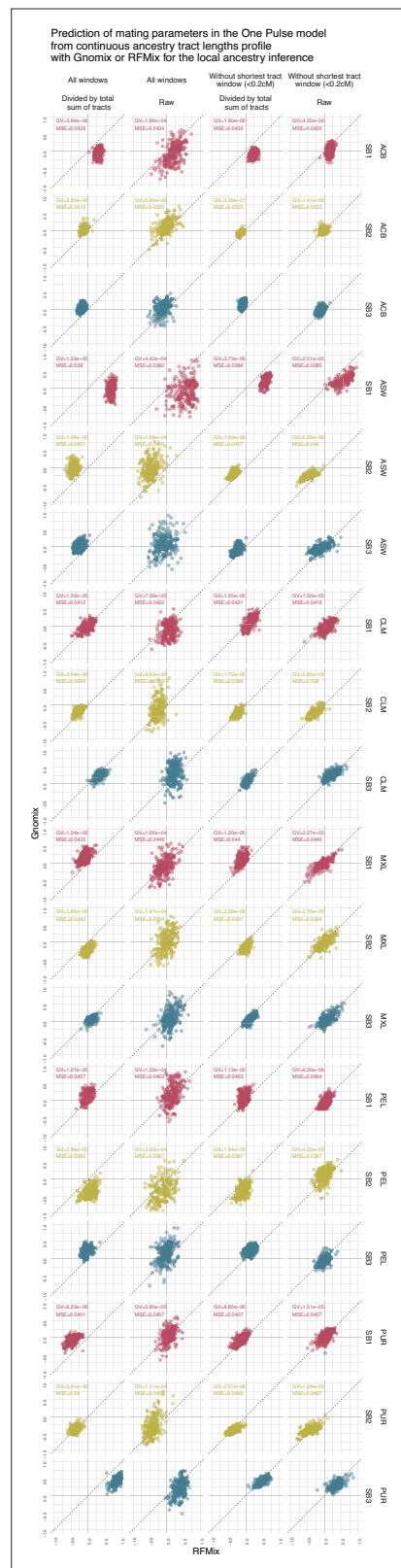


*Figure 4—figure supplement 3 continued*

parameter from each of the 1000 trainings using either RFMix or Gnomix tract length profile as input to the trained neural network. The tract length profile has been modified either by dividing each value by the total number of tracts or by removing the window corresponding to the shortest tracts, or both. We have evaluated the test MSE reported in **Figure 3** and the generalized variance (GV) in each case

$$\left( GV = \text{var}(\text{RFMix}) + \text{var}(\text{Gnomix}) - \text{cov}(\text{RFMix}, \text{Gnomix})^2 \right)$$

**Table 7** show mean GV and mean MSE for each tract length modification.

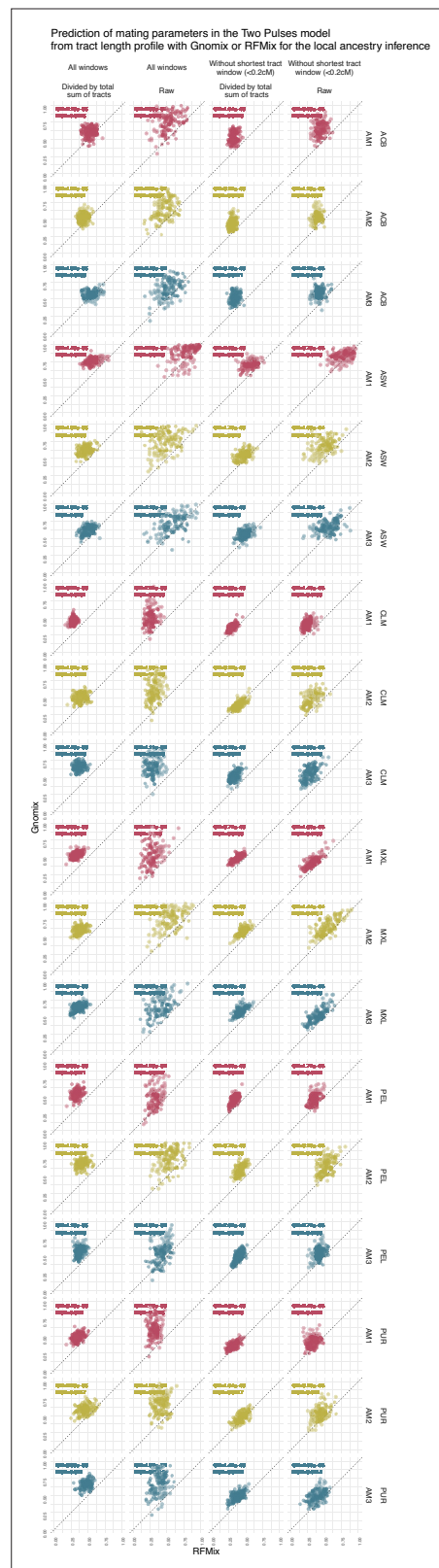


**Figure 4—figure supplement 4.** Prediction of sex bias in the One Pulse model using the continuous ancestry tract lengths profile from Gnomix or RFMix. Each point is the estimated parameter from each of the 1000

Figure 4—figure supplement 4 continued on next page

*Figure 4—figure supplement 4 continued*

trainings using either RFMix or Gnomix tract length profile as input to the trained neural network. The tract length profile has been modified either by dividing each value by the total number of tracts or by removing the window corresponding to the shortest tracts, or both. We have evaluated the test MSE reported in **Figure 3** and the generalized variance (GV) in each case ( $GV = var(RFMix) + var(Gnomix) - cov(RFMix, Gnomix)^2$ ). **Table 8** show mean GV and mean MSE for each tract length modification.



**Figure 4—figure supplement 5.** Prediction of assortative mating in the Two Pulses model using the continuous ancestry tract lengths profile from Gnomix or RFMix. Each point is the estimated

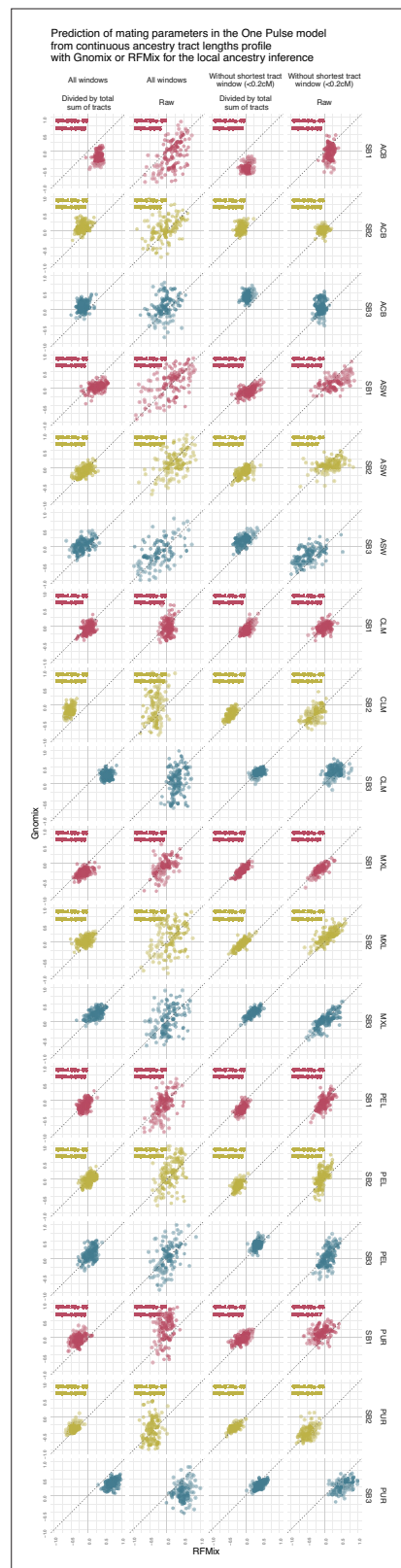
Figure 4—figure supplement 5 continued on next page

*Figure 4—figure supplement 5 continued*

parameter from each of the 1000 trainings using either RFMix or Gnomix tract length profile as input to the trained neural network. The tract length profile has been modified either by dividing each value by the total number of tracts or by removing the window corresponding to the shortest tracts, or both. We have evaluated the test MSE reported in **Figure 3** and the generalized variance (GV) in each case

$$\left( GV = \text{var}(\text{RFMix}) + \text{var}(\text{Gnomix}) - \text{cov}(\text{RFMix}, \text{Gnomix})^2 \right)$$

**Table 8** show mean GV and mean MSE for each tract length modification.

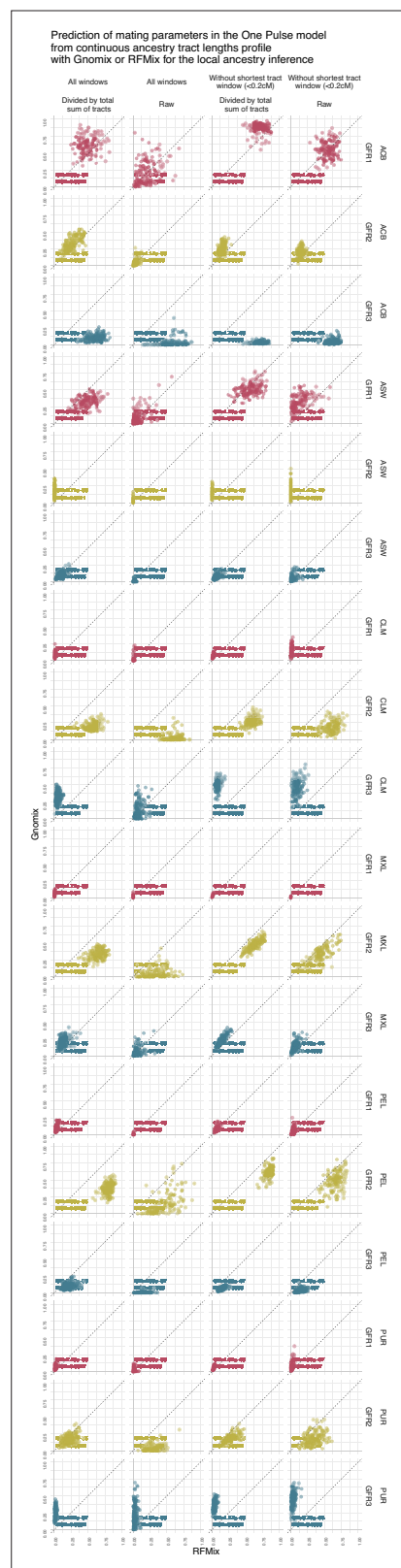


**Figure 4—figure supplement 6.** Prediction of sex bias in the Two pulses model using the continuous ancestry tract lengths profile from Gnomix or RFMix. Each point is the estimated parameter from each of the 1000

Figure 4—figure supplement 6 continued on next page

*Figure 4—figure supplement 6 continued*

trainings using either RFMix or Gnomix tract length profile as input to the trained neural network. The tract length profile has been modified either by dividing each value by the total number of tracts or by removing the window corresponding to the shortest tracts, or both. We have evaluated the test MSE reported in **Figure 3** and the generalized variance (GV) in each case ( $GV = var(RFMix) + var(Gnomix) - cov(RFMix, Gnomix)^2$ ). **Table 8** show mean GV and mean MSE for each tract length modification.



**Figure 4—figure supplement 7.** Prediction of gene flow rate in the Two Pulses model using the continuous ancestry tract lengths profile from Gnomix or RFMix. Each point is the estimated parameter

Figure 4—figure supplement 7 continued on next page



Figure 4—figure supplement 7 continued

from each of the 1000 trainings using either RFMix or Gnomix tract length profile as input to the trained neural network. The tract length profile has been modified either by dividing each value by the total number of tracts or by removing the window corresponding to the shortest tracts, or both. We have evaluated the test MSE reported in **Figure 3** and the generalized variance (GV) in each case

$$\left( GV = \text{var}(\text{RFMix}) + \text{var}(\text{Gnomix}) - \text{cov}(\text{RFMix}, \text{Gnomix})^2 \right)$$

**Table 8** show mean GV and mean MSE for each tract length modification.