Supplemental Appendix IV

This supplemental appendix provides greater detail regarding the determination that the parametric models are plausible estimations of the population distributions of similarity statistic values for both non-mated and mated friction ridge skin impressions for each quantity of features.

Non-mated distribution parameter estimation

The empirical distributions for the non-mated samples exhibit a distinct bimodal appearance for several of the lower feature quantities. As the feature quantities increase, however, the proportion of the distribution represented by the second mode decreases until the distributions for higher feature quantities appear unimodal. Although initially, the bimodal appearance may seem perplexing, it is actually quite straightforward as a mathematical consequence of the weight algorithms; however, it is outside the scope of this paper to go into innate detail on the actual weight algorithms themselves. Nevertheless, recognizing that the weighting functions are based on a mixture of functions, it seems natural that the resulting distribution is also a mixture of distributions. Taking this into consideration, the empirical distributions for all quantities of features were each modeled using k-component (where $k = 2$ or 3) mixtures of Gaussian distributions. Component weights and parameter estimates were determined using maximum likelihood estimation methods within commercially available statistical analysis software (JMP). Manual adjustments to the estimated weights and parameters were made to smooth trends among parameter estimates between feature quantities.

Although k-component Gaussian mixtures are more common, logistic distributions were applied on the basis of their heavier tails compared to Gaussian distributions. The heavier tails provide more conservative estimates of probabilities in the extreme ends of the distributions. The parameters for the logistic distribution were approximated using the estimated parameters of the Gaussian distributions. This was accomplished by setting the location parameter of the logistic distribution equal to the mean parameter of the Gaussian distribution as well as applying a coefficient to the standard deviation parameter of the Gaussian to approximate the scale parameter of the logistic distribution such that the difference between the two densities is minimized.

Prior to estimating the component weights and parameter values, the empirical distributions were partitioned into two groups. For each bin of feature quantities (ranging from 5 to 15), three-fourths ($n = 1{,}500$) of the sample was randomly selected and used to estimate the population distribution parameters. The remainder of the sample was used to evaluate the goodness of fit of the estimated parameters for the population distribution. Once the optimal parameters were estimated, a one-sample Kolmogorov-Smirnov (K-S) test was performed to evaluate the goodness of fit between the estimated theoretical logistic mixture distribution and the empirical distribution of the partition of similarity statistic values that was not used to estimate the theoretical distribution parameters. This process was repeated for each quantity of features (ranging from 5 to 15). Figures SAIV-1a through SAIV-1c illustrate the comparison between the theoretical distributions (k-component logistic mixtures) and the complete empirical distribution. Figure SAIV-1a overlays the cumulative frequency distributions, figure SAIV-1b illustrates the P-P plots between the cumulative frequency distributions, and figure SAIV-1c overlays the density distributions. Table SAIV-1 provides the K-S test statistics as well as the resulting *p*-values under the null hypothesis that the theoretical mixture distribution is representative of the distribution of which the non-modeled partition was drawn. Based on these findings, the distributions exhibit little difference and thus the parametric models are proposed as plausible estimations of the population distributions for each quantity of features.

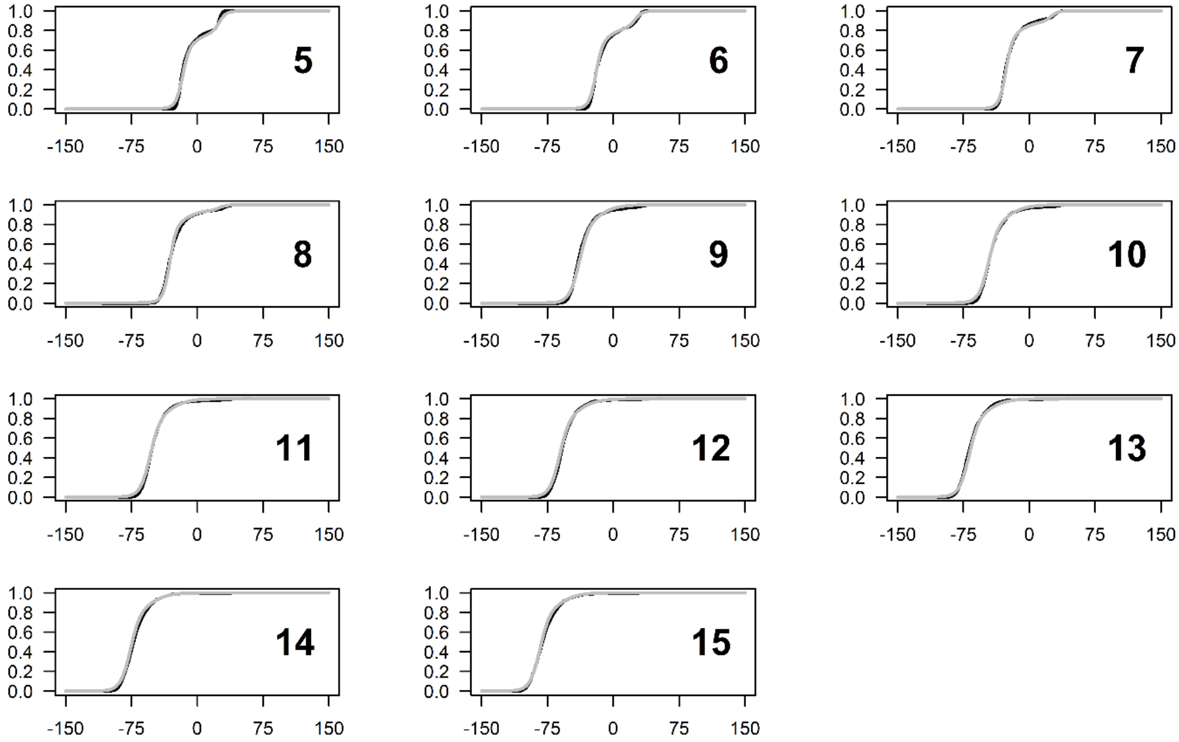

Figure SAIV-1a. Cumulative frequency distributions of the similarity statistic values for the non-mated sample (empirical) compared to the theoretical (k-component logistic mixture) distribution for each quantity of features (ranging from 5 to 15). The black line represents the empirical distribution. The grey line represents the theoretical distribution. The X-axis represents the global similarity statistic values.

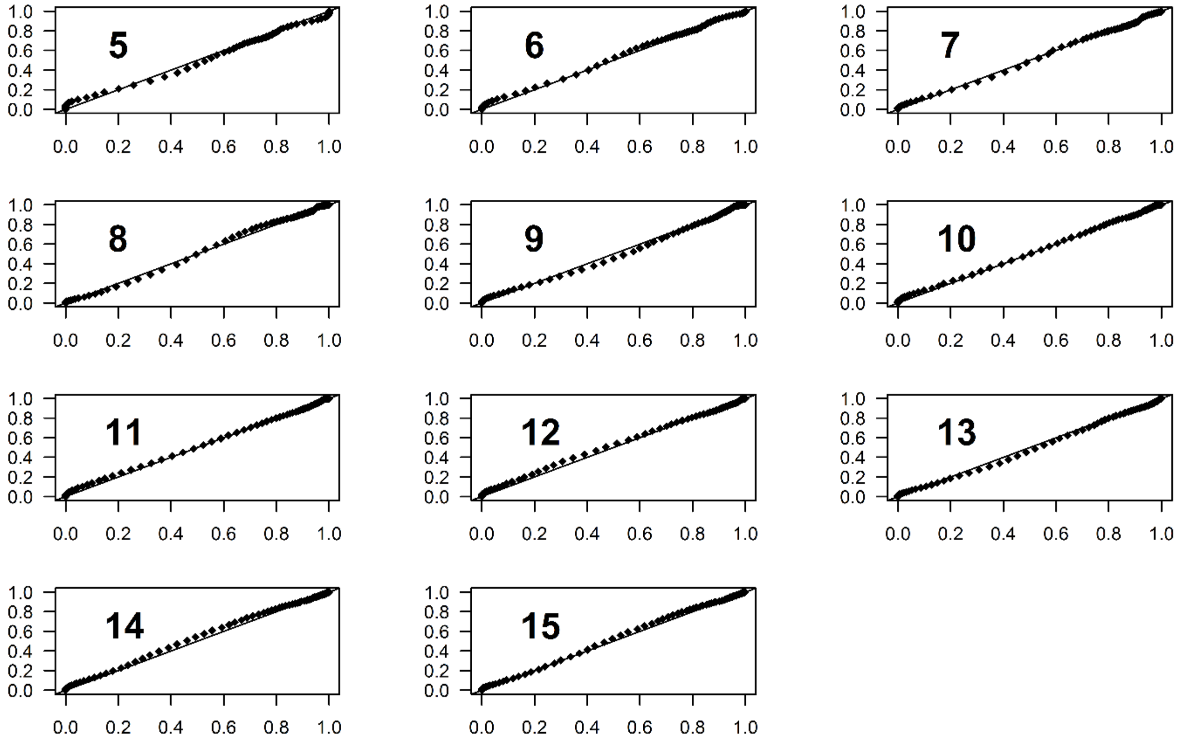

Figure SAIV-1b. P-P plots of the empirical cumulative frequency distributions of the similarity statistic values (horizontal axis) vs. theoretical (k-component logistic mixture) (vertical axis) cumulative frequency distributions for the non-mated sample for each quantity of features (ranging from 5 to 15). The black dots represent the P-P plot. The grey line represents an ideal slope of 1.

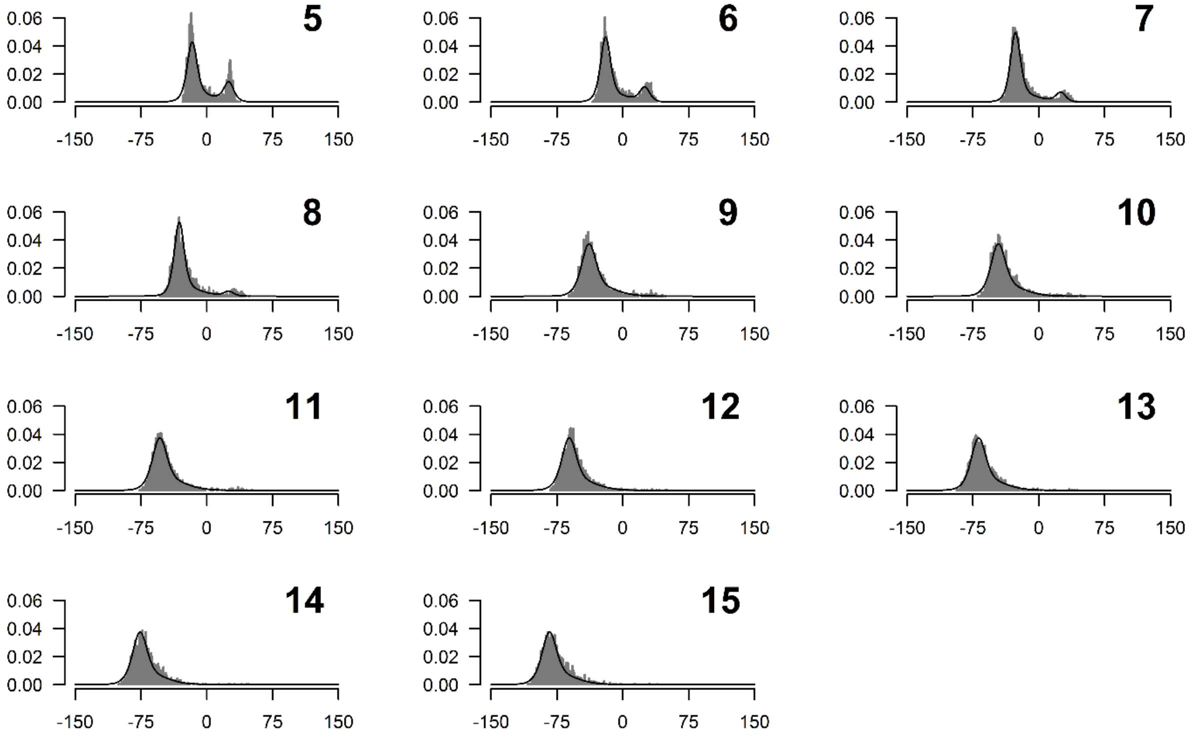

Figure SAIV-1c. Empirical density distributions of the similarity statistic values for the non-mated sample (grey) compared to the theoretical (k-component logistic mixture) distribution (black) for each quantity of features (ranging from 5 to 15). The X-axis represents the global similarity statistic values.

Feature Quantity	***n* sample** (non-mated; non-estimated partition)	**K-S test statistic**	***p (null)***
5	500	0.067	0.01 < *p* < 0.05
6	500	0.071	0.01 < *p* < 0.05
7	500	0.042	*p* > 0.05
8	500	0.077	*p* ~ 0.01
9	500	0.045	*p* > 0.05
10	500	0.041	*p* > 0.05
11	500	0.055	*p* > 0.05
12	500	0.058	*p* > 0.05
13	500	0.057	*p* > 0.05
14	500	0.058	*p* > 0.05
15	500	0.070	0.01 < *p* < 0.05

Table SAIV-1. Summary of the Kolmogorov-Smirnov test results between the distribution of similarity statistic values representing the partition not used to estimate the population parameters of the theoretical (k-component logistic mixture) distributions for each quantity of features (ranging from 5 to 15). NOTE: 1,500 sample statistic values were used to estimate the distribution parameters. The remainder of each sample was used to evaluate the goodness of fit. Statistical significance is based on a p-value decision threshold of 0.01.

Mated distribution parameter estimation

The empirical distributions for the mated samples appear unimodal; however, they exhibit a slight right-skew for several of the lesser feature quantities. As the feature quantities increase, the skew decreases, tending towards more symmetrical distributions. Recognizing that the same weighting functions were utilized for the mated source distributions, it seems natural that the resulting distributions are also a mixture of distributions. Taking this into consideration, the empirical distributions for all feature quantities were each modeled using k-component (where $k = 2$) mixtures of Gaussian distributions. Component weights and parameter estimates were determined using maximum likelihood estimation methods within commercially available statistical analysis software (JMP). Manual adjustments to the estimated weights and parameters were made to smooth trends among parameter estimates between feature quantities.

Although k-component Gaussian mixtures are more common, logistic distributions were applied on the basis for their heavier tails compared to Gaussian distributions. The heavier tails provide more conservative estimates of probabilities in the extreme ends of the distributions. The parameters for the logistic distribution were approximated using the estimated parameters of the Gaussian distributions. This was accomplished in the same manner as described above for the non-mated dataset by setting the location parameter of the logistic distribution equal to the mean parameter of the Gaussian distribution as well as applying a coefficient to the standard deviation parameter of the Gaussian to approximate the scale parameter of the logistic distribution such that the difference between the two densities is minimized.

Prior to estimating the component weights and parameter values, the empirical distributions were partitioned into two groups. For each bin of feature quantities (ranging from 5 to 14), approximately three-fourths (n = 1,500) of the sample was randomly selected and used to estimate the population distribution parameters. Due to the fewer samples in the bin for the feature quantity equal to 15, half (n = 250) were randomly selected and used to estimate the population distribution parameters. The remainder of the sample was used to evaluate the goodness of fit of the estimated parameters for the population distribution. Once the optimal parameters were estimated, a one-sample K-S test was performed to evaluate the goodness of fit between the estimated theoretical logistic mixture distribution and the empirical distribution of the partition of similarity statistic values that was not used to estimate the theoretical distribution parameters. This process was repeated for each quantity of features (ranging from 5 to 15). Figures SAIV-2a through SAIV-2c illustrate the comparison between the theoretical distributions (k-component logistic mixtures) and the complete empirical distribution. Figure SAIV-2a overlays the cumulative frequency distributions, figure SAIV-2b illustrates the P-P plots between the cumulative frequency distributions, and figure SAIV-2c overlays the density distributions. Table SAIV-2 provides the K-S test statistics as well as the resulting p-values under the null hypothesis that the theoretical mixture distribution is representative of the population distribution of which the non-modeled partition was drawn. Based on these findings, the distributions exhibit little difference and thus the parametric models are proposed as plausible estimations of the population distributions for each quantity of features.

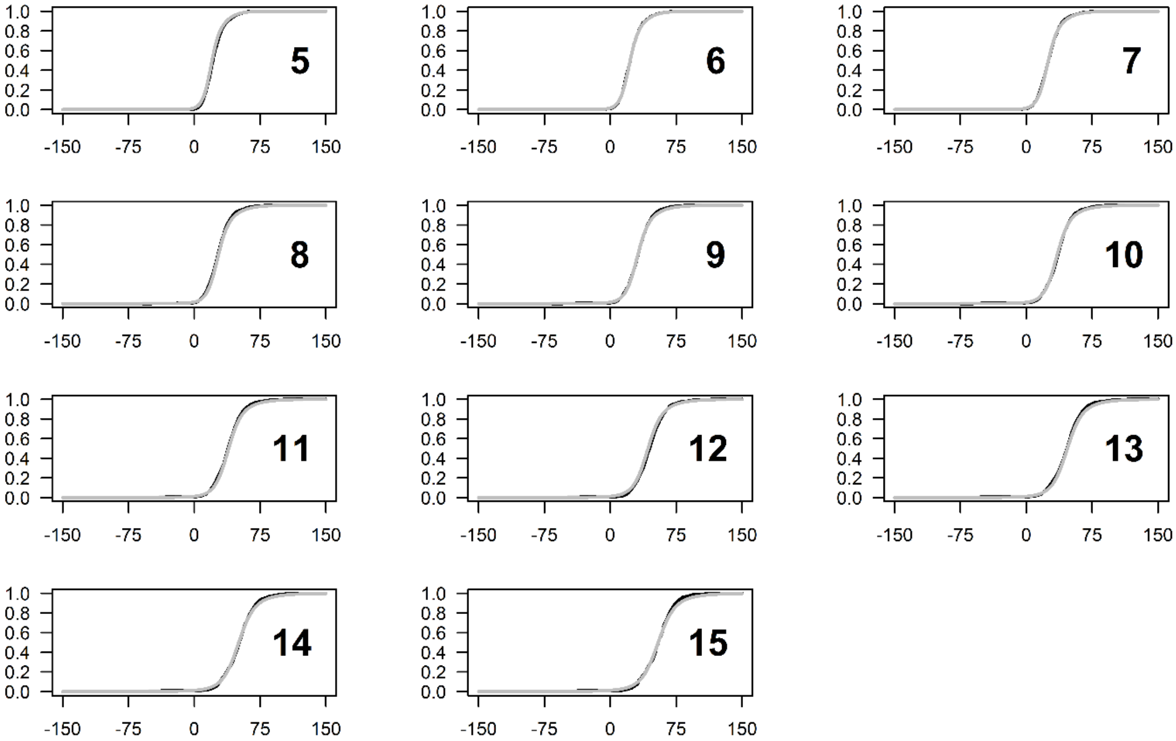

Figure SAIV-2a. Cumulative frequency distributions of the similarity statistic values for the mated sample (empirical) compared to the theoretical (k-component logistic mixture) distribution for each quantity of features (ranging from 5 to 15). The black line represents the empirical distribution. The grey line represents the theoretical distribution. The X-axis represents the global similarity statistic values.

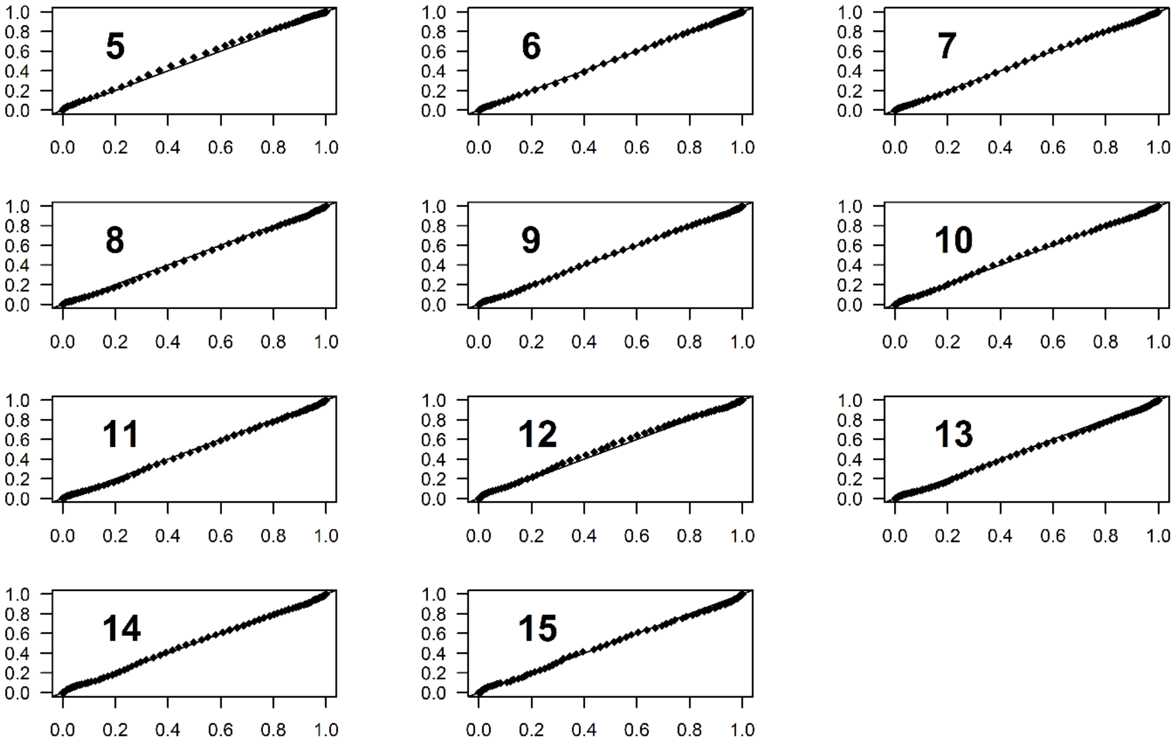

Figure SAIV-2b. P-P plots of the empirical cumulative frequency distributions of the similarity statistic values (horizontal axis) vs. theoretical (k-component logistic mixture) (vertical axis) cumulative frequency distributions for the mated sample for each quantity of features (ranging from 5 to 15). The black dots represent the P-P plot. The grey line represents an ideal slope of 1.

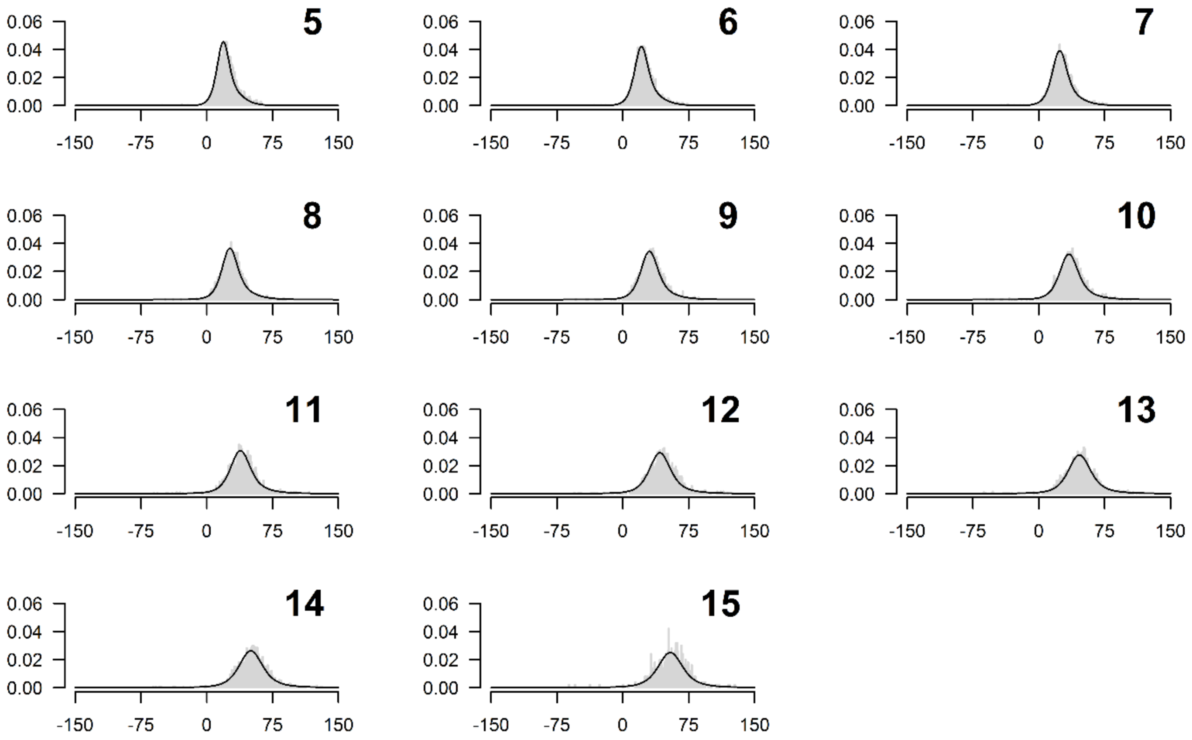

Figure SAIV-2c. Empirical density distributions of the similarity statistic values for the mated sample (grey) compared to the theoretical (k-component logistic mixture) distribution (black) for each quantity of features (ranging from 5 to 15). The X-axis represents the global similarity statistic values.

Feature Quantity	***n* sample** (mated; non-estimated partition)	K-S test statistic	***p (null)***
5	496	0.052	*p* > 0.05
6	496	0.028	*p* > 0.05
7	496	0.032	*p* > 0.05
8	496	0.053	*p* > 0.05
9	496	0.032	*p* > 0.05
10	496	0.064	0.01 < *p* < 0.05
11	496	0.049	*p* > 0.05
12	496	0.073	*p* ~ 0.01
13	496	0.048	*p* > 0.05
14	496	0.034	*p* > 0.05
15	249	0.051	*p* > 0.05

Table SAIV-2. Summary of the Kolmogorov-Smirnov test results between the distribution of similarity statistic values representing the partition not used to estimate the population parameters of the theoretical (k-component logistic mixture) distributions for each quantity of features (ranging from 5 to 15). NOTE: 1,500 sample statistic values were used to estimate the distribution parameters for feature quantities ranging from 5 to 14 and 250 were used for feature quantity = 15. The remainder of each sample was used to evaluate the goodness of fit. Statistical significance is based on a p-value decision threshold of 0.01.