Supplementary Material

Fernando A. Villanea and Joshua G. Schraiber

Compiled on June 8, 2018.

1 Analytical theory

1.1 One pulse model

An introgression of intensity f can be modeled as an injection of alleles at frequency f into a population. Each allele represents an introgressed haplotype, which will then undergo genetic drift until the present, at which time it is sampled at some (random) frequency. The Wright-Fisher diffusion model of genetic drift enables us to calculate the sampling probabilities after drift by computing

$$p_{n,k}(t;f) = \int_0^1 \binom{n}{k} y^k (1-y)^{n-k} \phi(f,y;t)dy$$

where $\phi(f,y;t)$ is the probability that a haplotype has gone from frequency f to frequency y in $2N_e t$ generations. Using well known results [Ewens, 2012], we obtain a differential equation for the frequency dependent part, $\mu_{n,k}(t) \equiv \int_0^1 y^k(1-y)^{n-k}\phi(f,y;t)dy$,

$$\frac{d}{dt}\mu_{n,k} = \frac{k(k-1)}{2}\mu_{n,k-1} - k(n-k)\mu_{n,k} + \frac{(n-k)(n-k-1)}{2}\mu_{n,k+1}.$$

This is a linear system of differential equations and can be solved by matrix exponentiation. Thus,

$$p_{n,k}(t;f) = \binom{n}{k} e^{Qt}\mathbf{f}$$

where Q is the matrix of coefficients of the system of differential equations and $\mathbf{f} = ((1-f)^n, f(1-f)^{n-1}, \ldots, f^n)^T$. This approach is similar to that used in Kamm et al. [2018] and Jouganous et al. [2017].

1.2 Two pulse model

We can apply a similar logic to the two pulse model, and again obtain an approximate formula. Working in a similar setting to before, we suppose that an admixture of intensity

f_1 occurred, then t_1 generations more recently was followed by a second admixture of intensity f_2, which was t_2 generations more ancient than the present. Then, we want to evaluate the integral

$$\begin{aligned} p_{n,k}(t_1, t_2; f_1, f_2) &= \int_0^1 \int_0^1 \binom{n}{k} y^k (1-y)^{n-k} \phi(f_1, z; t_1) \phi(f_2 + (1-f_2)z, y; t_2) dz dy \\ &= \int_0^1 \left(\int_0^1 \binom{n}{k} y^k (1-y)^{n-k} \phi(f_2 + (1-f_2)z, y; t_2) dy \right) \phi(f_1, z; t_1) dz \end{aligned}$$

because we need to integrate over all possible allele frequencies at the time of the second pulse of admixture.

The internal integral can be solved much the same way the one pulse model, however, the initial allele frequency needs to be adjusted to $f_2 + (1-f_2)z$. Thus, we need to derive a differential equation for

$$\eta_{n,k}(t) \equiv \int_0^1 (f_2 + (1-f_2)z)^k (1 - f_2 - (1-f_2)z)^{n-k} \phi(f, z; t) dz.$$

Defining $d_{n,k} = (f_2 + (1-f_2)z)^k (1 - f_2 - (1-f_2)z)^{n-k}$ and applying the Wright-Fisher generator to $d_{n,k}$, we get

$$\begin{aligned} \mathcal{L} d_{n,k} &= \frac{1}{2} x(1-x) \frac{d^2}{dx^2} d_{n,k} \\ &= \frac{1}{2} (1-f_2)^2 x(1-x) \left(k(k-1) d_{n-2,k-2} - 2k(n-k) d_{n-2,k-1} \right. \\ &\quad \left. + (n-k)(n-k-1) d_{n-2,k} \right), \end{aligned}$$

which is actually identical to the dilution model, but with $d_{n,k}$ in place of $c_{n,k}$.

Now, put $D = k(k-1)d_{n-2,k-2} - 2k(n-k)d_{n-2,k-1} + (n-k)(n-k-1)d_{n-2,k}$, and write

$$\begin{aligned} (1-f_2)^2 x(1-x) D &= (f_2 + (1-f_2)x - f)(1 - f_2 - (1-f_2)x) D \\ &= (f_2 + (1-f_2)x)(1 - f_2 - (1-f_2)x) D - f_2 (1 - f_2 - (1-f_2)x) D. \end{aligned}$$

Now, the first term looks like

$$k(k-1) d_{n,k-1} - 2k(n-k) d_{n,k} + (n-k)(n-k-1) d_{n,k+1},$$

which is the same as the one pulse model. However, the second term will be

$$k(k-1) d_{n-1,k-2} - 2k(n-k) d_{n-1,k-1} + (n-k)(n-k-1) d_{n-1,k}.$$

Note that the second term is *not* the same as the first term with $n \mapsto n-1$. However, we will apply the approximation, that $d_{n,k} \approx d_{n-1,k}$ for large n, Thus, we have

$$\begin{aligned}\frac{d}{dt}\eta_{n,k} = & \frac{k(k-1)}{2}\eta_{n,k-1} - k(n-k)\eta_{n,k} + \frac{(n-k)(n-k-1)}{2}\eta_{n,k+1} \\ & - f\left(\frac{k(k-1)}{2}\eta_{n,k-2} - k(n-k)\eta_{n,k-1} + \frac{(n-k)(n-k-1)}{2}\eta_{n,k}\right).\end{aligned}$$

The first line is simply the same differential equation as the one pulse case, while the second line is shifted down one term. Defining the matrix corresponding to that differential equation as Q_m, we see that

$$p_{n,k}(t_1, t_2; f_1, f_2) \approx \binom{n}{k} e^{Qt_2} e^{(Q - fQ_m)t_1} \mathbf{f}_t$$

where $\mathbf{f}_t = ((1 - f_2 - (1 - f_2)f_1)^n, (f_2 + (1 - f_2)f_1)(1 - f_2 - (1 - f_2)f_1)^{n-1}, \ldots, (f_2 + (1 - f_2)f_1)^n)^T$.

1.3 Dilution model

Under this model, an admixture of intensity f_1 occurs, then t_1 more generations more recently, an unadmixed group contributes to the population at hand with intensity f_2 t_2 generations in the past. Again, we can write down an integral to solve,

$$\begin{aligned}p_{n,k}(t_1, t_2; f_1, f_2) &= \int_0^1 \int_0^1 \binom{n}{k} y^k (1-y)^{n-k} \phi(f_1, z; t_1)\phi((1 - f_2)z, y; t_2) dz dy \\ &= \int_0^1 \left(\int_0^1 \binom{n}{k} y^k (1-y)^{n-k} \phi((1 - f_2)z, y; t_2) dy\right) \phi(f_1, z; t_1) dz\end{aligned}$$

The internal integral is simple: it's the same as the one admixture pulse model, except that the initial allele frequency is $(1 - f_2)z$. Evidently, the result that integral will be a function of terms $c_{n,k} = ((1 - f_2)z)^k (1 - (1 - f_2)z)^{n-k}$, so we need to solve integrals of the form

$$\nu_{n,k} = \int_0^1 ((1 - f_2)z)^k (1 - (1 - f_2)z)^{n-k} \phi(f_1, z; t_1) dz.$$

Applying the generator of the Wright-Fisher diffusion to the function $c_{n,k}$ we see that

$$\begin{aligned}\mathcal{L}c_{n,k} &= \frac{1}{2}x(1-x)\frac{d^2}{dx^2}c_{n,k} \\ &= \frac{1}{2}(1-f)^2 x(1-x)\left((k(k-1)c_{n-2,k-2} - 2k(n-k)c_{n-2,k-1}\right. \\ &\quad \left. + (n-k)(n-k-1)c_{n-2,k}\right).\end{aligned}$$

Let $C = k(k-1)c_{n-2,k-2} - 2k(n-k)c_{n-2,k-1} + (n-k)(n-k-1)c_{n-2,k}$, then note that

$$\begin{aligned}(1-f)^2x(1-x)C &= ((1-f)x)(1-(1-f)x-f)C \\ &= ((1-f)x)(1-(1-f)x)C - f((1-f)x)C.\end{aligned}$$

Multiplying through, we see that the first term looks like

$$k(k-1)c_{n,k-1} - 2k(n-k)c_{n,k} + (n-k)(n-k-1)c_{n,k+1},$$

while the second term will be

$$k(k-1)c_{n-1,k-1} - 2k(n-k)c_{n-1,k} + (n-k)(n-k-1)c_{n-1,k+1},$$

i.e. it is the same except with $n \mapsto n-1$. Making an approximation that $c_{n,k} \approx c_{n-1,k}$ for large n, we can pull out a factor of $(1-f_2)$ and obtain a system of differential equations,

$$\frac{d}{dt}\nu_{n,k} \approx (1-f_2)\left(\frac{k(k-1)}{2}\nu_{n,k-1} - k(n-k)\nu_{n,k} + \frac{(n-k)(n-k-1)}{2}\nu_{n,k+1}\right).$$

Noting that this is essentially the same differential equation as the one pulse model, we have that

$$\begin{aligned}p_{n,k}(t_1,t_2;f_1,f_2) &\approx \binom{n}{k}e^{Qt_2}e^{(1-f_2)Qt_1}\mathbf{f}_d \\ &= \binom{n}{k}e^{Q((1-f_2)t_1+t_2)}\mathbf{f}_d\end{aligned}$$

where now $\mathbf{f}_d = ((1-(1-f_2)f_1)^n, ((1-f_2)f_1)(1-(1-f_2)f_1)^{n-1}, \ldots, ((1-f_2)f_1)^n)^T$.

Note that this surprisingly simple form suggests that a dilution can be understood as an admixture of intensity $(1-f_2)f_1$ occurring $(1-f_2)t_1+t_2$ generations ago.

2 Error model

To incorporate false negative and false positive calls into our model, assume that there are independent false negative and false positive calls with rates ϵ_+ and ϵ_-, respectively. Specifically, every individual that has a fragment is called negative independently with probability ϵ_- and every individual that doesn't is called positive with probability ϵ_+. Define $b(k;N,p)$ to the probability mass function of a binomial random variable with size N and probability p. Also define $f(k;N_1,N_2,p_1,p_2)$ to be the distribution of the difference of two binomial random variables. Then we have that

$$f(k;N_1,N_2,p_1,p_2) = \sum_{i=0}^{N_2} b(k+i;N_1,p_1)b(i;N_2,p_2)$$

by a simple argument. Thus, we have that the probability of observing a fragment at frequency k is

$$\tilde{p}_{n,k} = \sum_{i=0}^{n} f(k-i; n-i, i, \epsilon_+, \epsilon_-) p_{n,i}$$

because if we have i introgressed fragments, we independently take false positives out of the $n-i$ non-introgressed fragments and false negatives out of the i introgressed fragments. If the number of false positives minus false negatives is d, then we have $i+d$ total fragments after errors, and thus need $d = k-i$ to end up with exactly k fragments. We then sum over all i.

To quantify the impact of errors in calling fragments, we generated an expected FFS under a 1 pulse model with $f = 0.02$ and $t = 0.1$ diffusion time units. We then computed the Kullback-Leibler divergence between the true FFS and the an FFS with a given false positive and false negative rate. Supplementary Figure 5 shows that, for low amounts of admixture as we simulated here, the impact of false positives is much larger than that of false negatives, due largely to the fact that most of the genome is a true negative. Nonetheless, even with relatively high false negative and false positive rates, such as $\epsilon_- = 0.1$ and $\epsilon_+ = 0.01$, the Kullback-Leibler divergence is only ~ 0.005, indicating that false positives and false negatives do not have a substantial effect on the FFS.

3 Results from maximum likelihood fitting

Supplementary Table 1 and 2 shows the parameter estimates from maximum likelihood fitting of the Asian and European data, respectively, across a variety of cutoffs from the Steinrücken data.

4 Admixture constraints

Given European and East Asian mixture proportions, f_{EUR} and f_{ASN}, respectively, we constrain mixture proportions by constraining

$$a = \frac{f_{\text{ASN}} + f_{\text{EUR}}}{2}$$

and

$$d = f_{\text{ASN}} - f_{\text{EUR}}$$

we then express f_{ASN} and f_{EUR} in terms of the model parameters, and solve for model parameters that will adhere to the constraints.

In the one pulse model, f_1 is a single pulse of Neandertal introgression into the ancestral Eurasian population. So,

$$f_{\text{ASN}} = f_1$$

and

$$f_{\mathrm{EUR}} = f_1.$$

Thus, we see that

$$f_1 = a$$

In the two pulse model, f_1 is a first pulse of Neandertal introgression into the ancestral Eurasian population, and f_2 is a second pulse of introgression into the East Asian population. This results in

$$f_{\mathrm{ASN}} = f_2 + (1 - f_2)f_1$$

and

$$f_{\mathrm{EUR}} = f_1,$$

which yields

$$f_1 = a - d/2$$

and

$$f_2 = \frac{d}{1 + d/2 - a}.$$

For the three pulse model, f_1 is a first pulse of Neandertal introgression into the ancestral Eurasian population, f_2 is a second pulse of introgression into the East Asian population, and f_3 is a second pulse of introgression into the European population. Thus,

$$f_{\mathrm{ASN}} = f_2 + (1 - f_2)f_1$$

and

$$f_{\mathrm{EUR}} = f_3 + (1 - f_3)f_1,$$

Note that in this model, we have more free parameters than constraints, so we sample f_1 from a uniform distribution between 0 and $a - d/2$, and then solve to obtain

$$f_2 = \frac{a + d/2 - f_1}{1 - f_1}$$

and

$$f_3 = \frac{-a + d/2 + f_1}{f_1 - 1}.$$

For the dilution model, f_1 is a single pulse of Neandertal introgression into the ancestral Eurasian population, and f_4 represents the dilution from the Basal Eurasian population into the European population. This yields admixture proportions

$$f_{\mathrm{ASN}} = f_1$$

and

$$f_{\mathrm{EUR}} = (1 - f_4)f_1,$$

resulting in

$$f_1 = a + d/2$$

and

$$f_4 = \frac{d}{a + d/2}$$

In the model with 3 pulses of Neandertal admixture and dilution, f_1 is a first pulse of Neandertal introgression into the ancestral Eurasian population, f_2 is a second pulse of introgression into the East Asian population, and f_3 is a second pulse of introgression into the European population, while f_4 represents the dilution from the Basal Eurasian population into the European population. Further, we always assume that dilution is more recent than the second pulse in Europe. In this case, admixture proportions are

$$f_{\text{ASN}} = f_2 + (1 - f_2)f_1$$

and

$$f_{\text{EUR}} = (f_3 + (1 - f_3)f_1)(1 - f_4)$$

Again, we have more free parameters than constraints so we first draw f_1 from a uniform distribution between 0 and $a - d/2$ and f_4 from a uniform distribution between 0 and 0.5. Then, we set

$$f_2 = \frac{a + d/2 - f_1}{1 - f_1}$$

and

$$f3 = \frac{a - d/2 + f_1(1 - f_4)}{(1 - f_1)(1 - f_4)}$$

5 Neural Network Weights

In order to quantify the impact of different frequency spectrum categories to the inferences of the FCNN, we computed the matrix product of the weights across the fully connected layers. Note that we flatten our input FFS matrix to a single vector of length $m_1 = 64 \times 64 = 4096$ initially. Then, we compute the weighted sum of the weights leading to the nodes in the subsequent layers of the FCNN. Specifically, if layer i has m_i nodes, and w_i is the $m_i \times m_{i+1}$ matrix where $w_{i,j,k}$ provides the weights from node j in layer i to node k in layer $i + 1$, then we compute the matrix product

$$M = w_i w_{i+1} \cdots w_n$$

when we have n layers.

The resulting matrix M will be 4096×5, with each of the 5 columns corresponding to one of the different models. Thus, we map each of the columns back into the original 64×64 matrix, resulting in the panels shown in Supplementary Figure 6.

6 Robustness of FCNN results

The Neandertal fragment calls from Steinrücken et al. [2018] are based on the posterior probability of introgression at each position estimated using diCal-admix. Because each position has a different probability of introgression, defining a global cut-off is necessary to obtain a consensus across the genome, such that a higher cut-off results in a higher certainty of the calls (i.e. higher precision), but more false negatives (i.e. lower recall). For the analysis presented in the main text, we used a cut-off of 0.45, recommended in Steinrücken et al. [2018] as it provides the best balance across performance metrics based on their precision-recall curves. However, to test the robustness of the selected cut-off, we generated FFS based on a range of cut-offs and analyzed them using the fully-connected neural network. Supplementary Figure 4 shows that our results are consistent across the entire range of cutoffs.

In addition, we had access to the introgressed fragments calls from Sankararaman et al. [2014], which were ascertained independently using a conditional random field method. We converted these fragment calls into introgressed site calls by looking at the same positions along 100kb windows used previously, and counting how many individuals presented an introgressed fragment which overlapped with that site. We used these SNP calls as independent confirmation of all results (Figure 7).

References

Warren J Ewens. *Mathematical population genetics 1: theoretical introduction*, volume 27. Springer Science & Business Media, 2012.

Julien Jouganous, Will Long, Aaron P Ragsdale, and Simon Gravel. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*, pages genetics–117, 2017.

John A. Kamm, Jonathan Terhorst, Richard Durbin, and Yun S. Song. Efficiently inferring the demographic history of many populations with allele count data. *bioRxiv*, 2018. doi: 10.1101/287268. URL `https://www.biorxiv.org/content/early/2018/03/23/287268`.

Sriram Sankararaman, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich. The genomic landscape of neanderthal ancestry in present-day humans. *Nature*, 507(7492):354, 2014.

Matthias Steinrücken, Jeffrey P. Spence, John A. Kamm, Emilia Wieczorek, and Yun S. Song. Modelbased detection and analysis of introgressed neanderthal ancestry in modern humans. *Molecular Ecology*, 2018. doi: 10.1111/mec.14565.

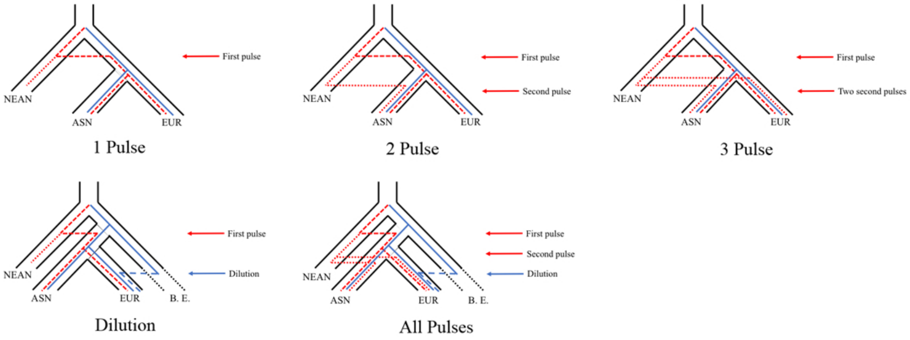

Figure 1: Representation of the five different demographic models simulated in MSprime.

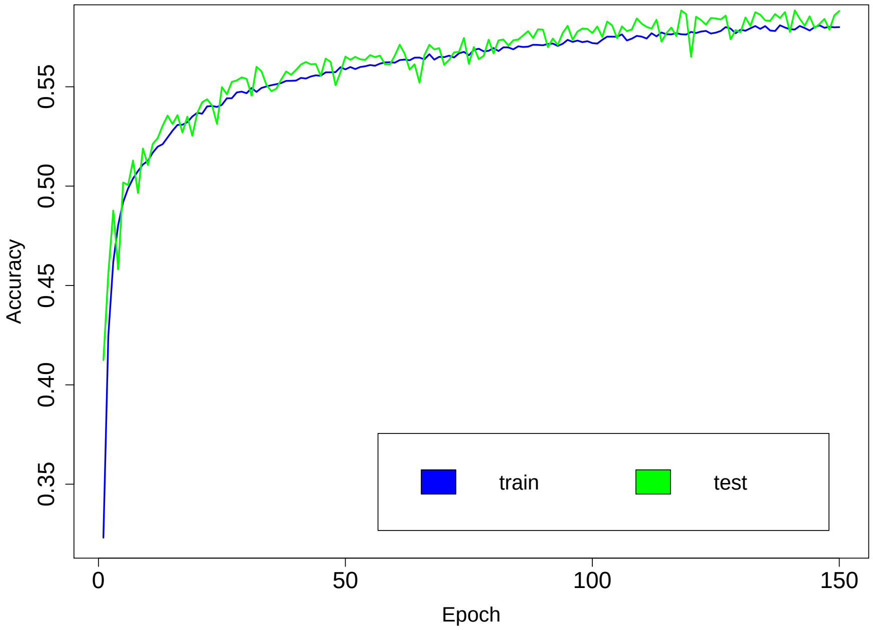

Figure 2: Training and Validation accuracy of FCNN as it trained over 150 epochs. The x axis indicates the training epoch (i.e. one pass through the whole dataset) while the y-axis shows the accuracy on either training data (blue) or validation data (green).

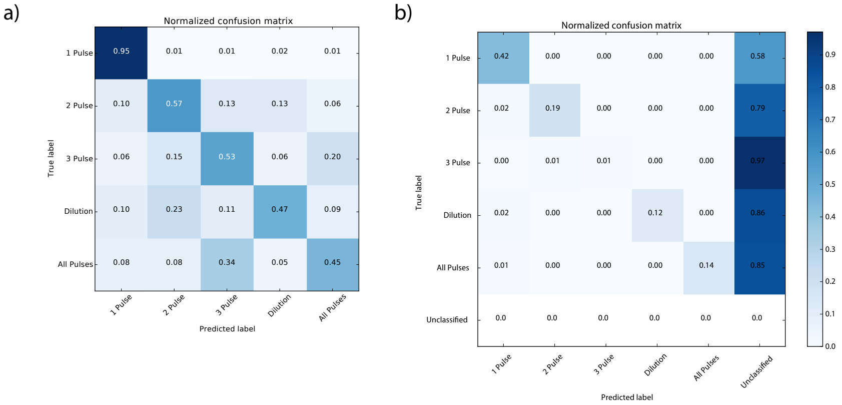

Figure 3: a) Confusion matrix of simulated data categorized by the FCNN. b) Confusion matrix of simulated data categorized by the FCNN when accepting results with a probability cut-off of 0.8.

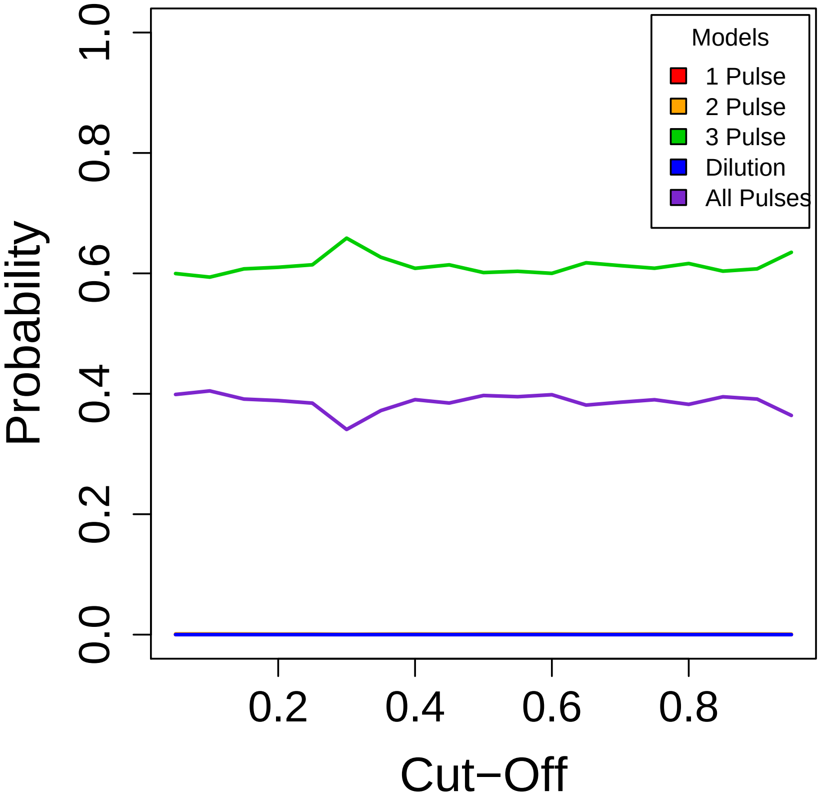

Figure 4: Posterior probability of the empirical introgression data from the FCNN classifier under different cut-offs of the posterior probability of introgression in the Steinrücken et al. [2018] data. The x axis indicates the posterior probability cutoff, and the y axis the model probability according to the FCNN. Each line corresponds to a different model.

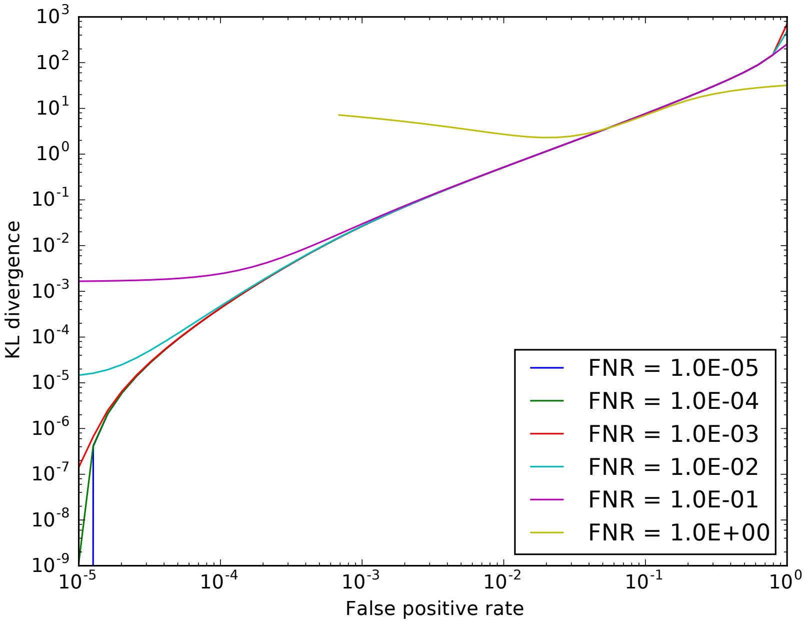

Figure 5: The impact of false positive and false negative fragment calls on the FFS. The x axis shows the false positive rate, and the y axis the Kullback-Liebler divergence of the observed FFS to the true FFS (larger values indicate more difference). Each line corresponds to a different false negative rate.

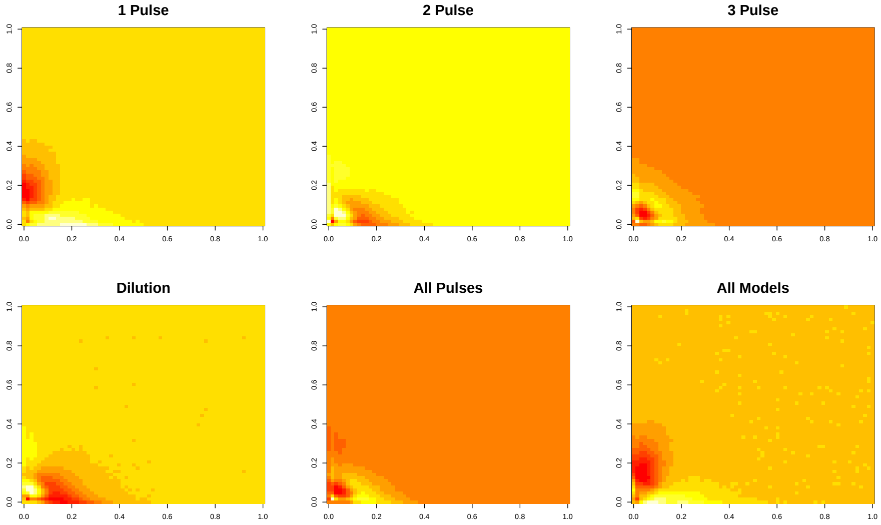

Figure 6: Weights projected across layers into the final dense layer, representing the relative importance of each position along the FFS matrix when classifying a FFS into one of the five final categories.

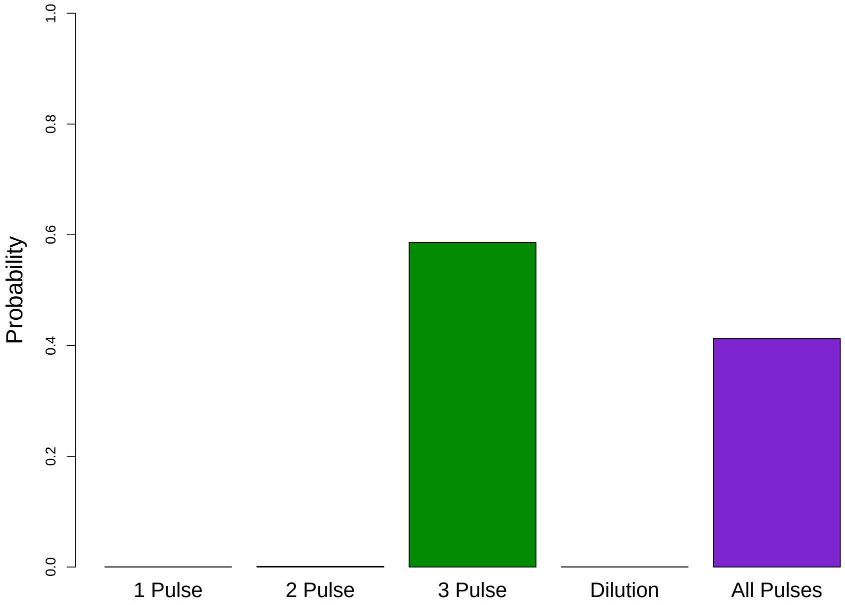

Figure 7: Posterior probability of the empirical introgression data from Sankararaman et al., 2014 matching each of the five demographic models, determined by the FCNN classifier.