

1 Recovering signals of ghost archaic admixture in the genomes of 2 present-day Africans

3 Arun Durvasula¹ and Sriram Sankararaman^{1,2,3,4}

4 ¹Department of Human Genetics, David Geffen School of Medicine, University of
5 California, Los Angeles

6 ²Department of Computer Science, University of California, Los Angeles

7 ³Bioinformatics Interdepartmental Program, University of California, Los Angeles

8 ⁴Lead contact

9 March 20, 2018

10 **Abstract**

11 Analyses of Neanderthal and Denisovan genomes have characterized multiple interbreeding events
12 between archaic and modern human populations. While several studies have suggested the presence of
13 deeply diverged lineages in present-day African populations, we lack methods to precisely characterize
14 these introgression events without access to reference archaic genomes. We present a novel reference-
15 free method that combines diverse population genetic summary statistics to identify segments of archaic
16 ancestry in present-day individuals. Using this method, we find that $7.97 \pm 0.6\%$ of the genetic ancestry
17 from the West African Yoruba population traces its origin to an unidentified, archaic population (FDR
18 $\leq 20\%$). We find several loci that harbor archaic ancestry at elevated frequencies and that the archaic
19 ancestry in the Yoruba is reduced near selectively constrained regions of the genome suggesting that
20 archaic admixture has had a systematic impact on the fitness of modern human populations both within
21 and outside of Africa.

22 **Running title:** Reference-free inference of archaic introgression

23 **Correspondence:** sriram@cs.ucla.edu

24 **Introduction**

25 Admixture, the exchange of genes among previously isolated populations, is increasingly being recognized as
26 an important force in shaping genetic variation in human populations. Analyses of large collections of genome
27 sequences have shown that admixture events have been prevalent throughout human history [1]. Further,
28 these studies have shown that modern human populations outside of Africa trace a small percentage of their
29 ancestry to admixture events from populations related to archaic hominins like Neanderthals and Denisovans
30 [1, 2, 3]. Studies of the functional impacts of this introduced DNA have suggested that Neanderthal DNA
31 that segregates in modern humans contributes to phenotypic variation [4, 5].

32 Central to these studies is the problem of local archaic ancestry inference – the pinpointing of segments
33 of an individual genome that trace their ancestry to archaic hominin populations. Methods for local archaic
34 ancestry inference leverage various summary statistics computed from modern and ancient genomes. For
35 example, at a given genomic locus, individuals with archaic ancestry are expected to have high sequence
36 divergence to segments of modern human ancestry but low divergence to the archaic genome [6]. A number
37 of summary statistics have been developed to infer archaic local ancestry [7, 8, 9]. Further, statistical models
38 that can combine these summary statistics have also been proposed [2, 10, 11].

39 All of these methods are most effective in settings where reference genomes that represent genetic varia-
40 tion in the archaic population are available. For example, the analyses of Neanderthal [6, 10] and Denisovan
41 admixture events [12] relied on the genome sequences from the respective archaic populations. In a number
42 of instances, however, the archaic population is either unknown or lacks suitable reference genomes. Several
43 recent studies have found evidence for archaic introgression in present-day African populations from an un-
44 known archaic hominin [13, 14, 15] while analysis of the high-coverage Denisovan genome has suggested that
45 the sequenced individual traces a small proportion of its ancestry to a highly-diverged archaic hominin [10].

46 One of the most widely used statistics that for identifying archaic ancestry is the S^* -statistic [9], which
47 identifies highly diverged SNPs that are in high linkage disequilibrium (LD) with each other in the present-

48 day population as likely to be introgressed. The S^* -statistic is attractive as it can be applied even where no
49 reference genome is available. However, the power of the S^* -statistics tends to be low in the reference-free
50 setting [3]. Further, the value of the S^* -statistic depends on a number of parameters that need to be fixed
51 in advance.

52 Recent studies have shown that statistical predictors that combine weakly-informative summary statis-
53 tics can obtain substantially improved accuracy on a number of population genetic problems [16, 17, 18]. We
54 extend this idea to the setting of archaic local ancestry inference and present a statistical method, ARChaic
55 Introgression Explorer (ArchIE), based on a logistic regression model, that combines several population ge-
56 netic summary statistics to accurately predict archaic local ancestry. The parameters of ArchIE are estimated
57 from training data generated using coalescent simulations. We show that ArchIE obtains improved accuracy
58 in simulations over the S^* -statistic while being robust to demographic model misspecification. We apply
59 ArchIE to the 1000 Genomes Western European (CEU) population and show that the inferred segments
60 of archaic ancestry have an increased likelihood of being introgressed from Neanderthals without access to
61 Neanderthal genomes. In addition, the inferred segments of archaic ancestry in Europeans recover previ-
62 ously seen features when using reference-based methods. Specifically, using the inferences from our method,
63 we observe a decreased frequency of Neanderthal ancestry in regions of the genome with stronger selective
64 constraint [19] as well as elevated frequency of Neanderthal ancestry at the *BNC2* and *OAS* loci, both of
65 which have been previously shown to harbor Neanderthal alleles at high frequency.

66 Finally, we apply ArchIE to genomes from the West African Yoruba (YRI) population in the 1000
67 Genomes Project [20] to obtain inferences of archaic local ancestry in this population. At a precision of
68 0.80, 7.9% of the genomes of west African individuals, on average, is estimated to trace its ancestry to a
69 deeply-diverged archaic population. We enumerate 258 megabases (MB) of introgressed DNA, with 2.1 MB
70 at a high ($\geq 50\%$) frequency. We observe a decrease in the frequency of archaic ancestry in the Yoruban
71 populations in more constrained regions of the genome, suggesting that these archaic alleles have been subject
72 to the effects of purifying selection similar to the deleterious consequences of Neanderthal and Denisovan
73 alleles in the modern human genetic background. On the other hand, we find several loci that harbor archaic
74 haplotypes at elevated frequencies ($> 60\%$). These results highlight the landscape of archaic introgression

75 into African populations and provide insight into how modern humans evolved as a species.

76 Results

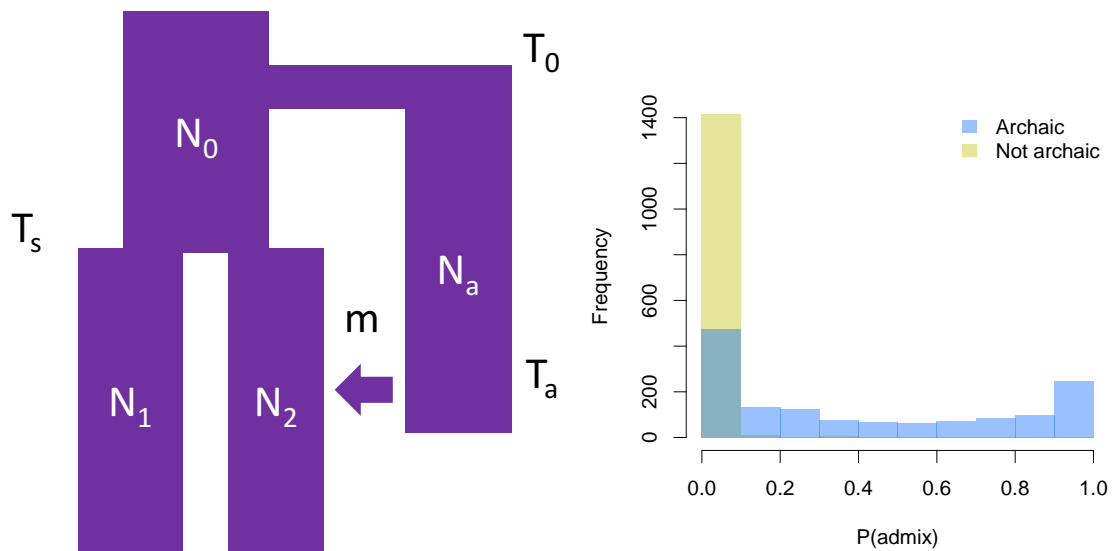
77 Model overview

78 Our method, ArchIE, aims to predict the archaic local ancestry state in a given window along an individual
79 haploid genome. This prediction is performed using a binary logistic regression model given a set of features
80 computed within this window. Estimating the parameters of this model requires labeled training data *i.e.*, a
81 dataset containing pairs of features and the archaic local ancestry state for a given window along an individual
82 genome. To this end, we simulate data under a demographic model that includes archaic introgression, label
83 windows as archaic or not, calculate a set of features that are potentially informative of introgression, and
84 estimate the parameters of our predictor on the resulting data (Figure 1A, Methods).

85 We simulate training data using a modified version of the coalescent simulator, ms [21], which allows
86 us to track each individual's ancestry. We use the demographic model used in Sankararaman *et al.* 2014
87 [2]. In this model, an ancestral population splits T_0 generations ago forming archaic and modern human
88 populations. The modern human population splits into two populations at T_s , one of which then mixes
89 with the archaic population (referred to as the target population) while the other does not (the reference
90 population). We simulate one haploid genome (haplotype) in the archaic population, 100 haplotypes in the
91 target population and 100 haplotypes in the reference population. In the results below, we simulate 10,000
92 replicates of 50,000 base pairs each (bp), resulting in 1,000,000 training examples.

93 We summarize the resulting data using features that are likely to be informative of archaic admixture.
94 Since we are interested in the probability of archaic ancestry for a given haplotype, we use features that are
95 specific for each haplotype, *i.e.*, the focal haplotype. First, for a focal haplotype, we calculate an individual
96 frequency spectrum (IFS), which is a vector of length n , the sample size of the target population. Each
97 entry in the vector counts the number of mutations on the focal haplotype that are segregating in the target
98 population with a specific count of derived alleles. Due to the accumulation of private mutations in the
99 archaic population, we expect an excess of alleles segregating at frequencies close to the admixture fraction

Figure 1: (A) **Outline of the demographic model used for training ArchIE.** We simulate a population starting at size N_0 and splitting into archaic and modern human (MH) populations at time T_0 . The MH population splits into an MH reference and target population of size N_1 and N_2 , respectively, at time T_s . Then, at time T_a , the archaic population admixes with the target population with an association admixture proportion m . We use data simulated from this model to train a logistic regression classifier. (B) **Distribution of predictions for ArchIE on test data.** Haplotypes predicted to have a low probability of being archaic in origin are enriched in truly non-archaic haplotypes, while truly archaic haplotypes are enriched for higher probabilities.



100 in the introgressed population. The IFS is expected to capture this signal.

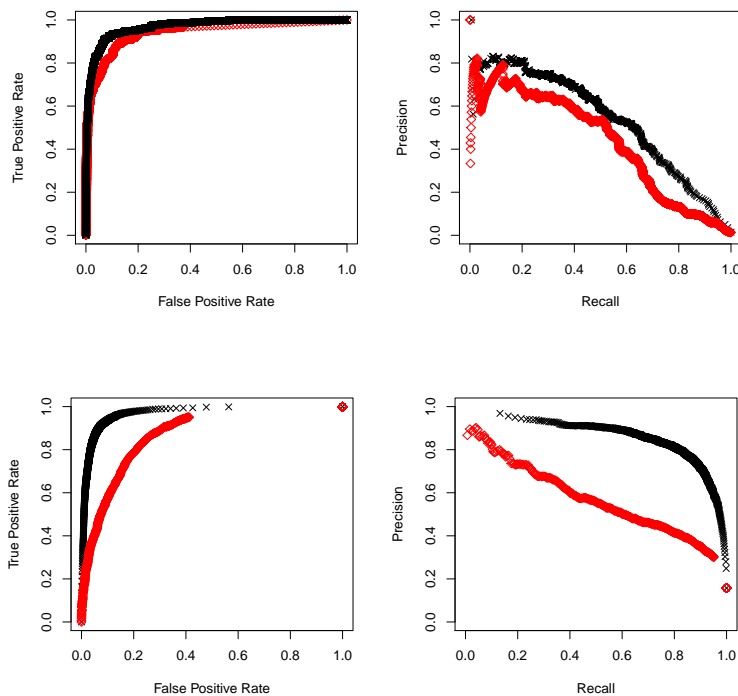
101 Next, we calculate the Euclidean distance between the focal haplotype and all other haplotypes, resulting
102 in a vector of length n . Under a scenario of archaic admixture, the distribution of pairwise differences is
103 expected to differ when we compare two haplotypes that are both modern human or archaic versus when we
104 compare an archaic haplotype to a modern human haplotype.

105 The next set of features rely on a present-day reference human population that has a different demographic
106 history compared to the target population. We term this population the *MH reference* to make it clear that
107 our method does not rely on an archaic reference. The choice of the MH reference will alter the specific
108 admixture events that our method is sensitive to: we expect the method to be sensitive to admixture events
109 in the history of the target population since its divergence from the MH reference. While our method can
110 also be applied in the setting where no such reference population exists, in the context of human populations
111 where genomes from a diverse set of populations is available [1], the use of the MH reference can improve the
112 accuracy and the interpretability of our predictions. Given a reference population, we compute the minimum
113 distance of the focal haplotype to all haplotypes in the reference population. A larger distance is suggestive
114 of admixture from a population that diverged from the ancestor of the target and reference population before
115 the reference and target populations split.

116 We also calculate the number of SNPs private to the focal haplotype, removing SNPs shared with the MH
117 reference, as these SNPs are suggestive of an introgressed haplotype. Finally, we calculate S^* [9], a feature
118 designed for detecting archaic admixture by looking for enrichments of long stretches of derived alleles in
119 LD.

120 Using these features, we train a logistic regression classifier to distinguish between archaic and non
121 archaic segments. In our training data, we define archaic haplotypes as those that contain $\geq 70\%$ of bases
122 with archaic ancestry and non-archaic as those that contain $\leq 30\%$ archaic ancestry. We discard haplotypes
123 that fall in-between those values in both the training the test datasets. We simulated 1,000,000 haplotypes
124 for the training data and 100,000 haplotypes for the test data which resulted in 988,372 training examples
125 and 98,922 test examples after filtering. Figure 1B shows the distribution of predicted probabilities of archaic
126 ancestry on test data. Archaic haplotypes tend to be associated with a high probability of being archaic

Figure 2: **ArchIE is more accurate than the S*-statistic** (A) Receiver Operator Characteristic (ROC) and (B) precision-recall (PR) curves for ArchIE (black crosses) and S* (red diamonds) in a 2% admixture scenario with a Human-Neanderthal demography. (C) ROC and (D) PR curves for a 20% admixture scenario.



127 while non archaic haplotypes are enriched for low probabilities of being archaic.

128 Simulation results

129 We tested the accuracy of ArchIE by simulating data under a demography reflective of the history of Ne-
130 anderthals and present-day humans [2]. We began by simulating an admixture event with 2% Neanderthal
131 ancestry that occurred 2,000 generations ago and simulated 1,000,000 haplotypes under 10,000 different repli-
132 cates (100 haplotypes per replicate). We compute Receiver Operator Characteristic (ROC) and Precision
133 Recall (PR) curves by varying the threshold at which we call a haplotype archaic and calculating the true
134 positive rate (TPR), false positive rate (FPR), precision, and recall (Figure 2).

135 We compared these results to an implementation of the S* algorithm similar to Vernot and Akey [3].
136 First, we calculate S* in a cohort of 100 haplotypes from both the reference and target populations. Then,

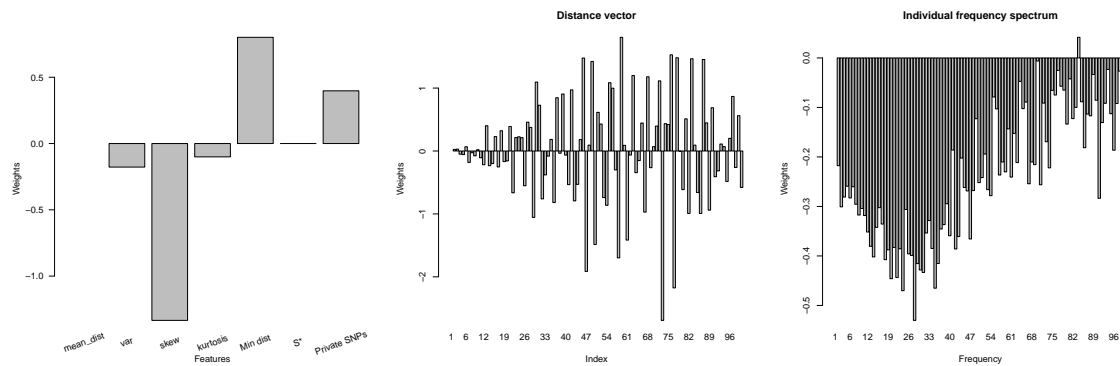
137 we convert the S^* scores into a rank between [0-1] using the empirical cumulative distribution. We obtain
138 precision and recall values for varying thresholds between [0-1] and report these values. At a 2% admixture
139 fraction, ArchIE outperforms the S^* statistic across all thresholds (Figure 2AB). At a fixed precision of 0.8,
140 *i.e.*, false discovery rate of 0.20, ArchIE obtains a recall of 0.21, while S^* obtains a recall of 0.024. The area
141 under the ROC curve is 0.943 for S^* and 0.969 for ArchIE and the area under the PR curve is 0.431 for S^*
142 and 0.535 for ArchIE. We also note that while the ROC curves are quite similar, the PR curve show a large
143 difference, indicative of the utility of PR curves in class imbalance problems.

144 Our simulation setup is challenging partly due to the low admixture fraction resulting in a large class
145 imbalance across the archaic and non-archaic classes. In this setting, we find fewer than 20,000 positive
146 examples. We compared the two methods using an admixture fraction of 20%, thereby increasing the
147 number of positive examples in training and test data. While the accuracy of S^* only slightly improved, the
148 logistic regression classifier shows a much larger improvement (Figure 2CD). While this admixture fraction is
149 higher than the data suggests for Human-Neanderthal introgression, it is not implausible in other species or
150 admixture events [22, 23, 24, 25, 26, 18]. This example indicates the utility of the parameterized statistical
151 model underlying ArchIE that can be tuned to accurately infer archaic ancestry under plausible demographic
152 settings.

153 **ArchIE learns informative features**

154 We examined the weights learned by ArchIE to understand the features that contribute substantially to
155 its predictions. Examining single features, we find that the minimum distance between the focal haplotype
156 and each of the reference MH haplotypes, as well as the number of private SNPs are the most positively
157 correlated with a high probability of being archaic (Figure 3A). Intuitively, as a larger distance to a reference
158 population and a larger number of private SNPs should both indicate archaic ancestry. The next largest
159 single statistic was the skew of the distance vector, which was negatively correlated with archaic ancestry.
160 Under a simple scenario of admixture, we expect a bimodal distribution of pairwise distances. However, when
161 there is little archaic ancestry, the distribution will be skewed towards 0, resulting in a negative relationship
162 between skew and archaic ancestry. Examining the distance vector itself, the weights are often flipped

Figure 3: **Relative importance of the features used as input to ArchIE.** We examined the weights associated with each of the features included in the logistic regression model underlying ArchIE. (A) The first four entries indicate the moments of the distance vector. The skew has the largest weight associated with it, indicating that this is the most important feature in the distance vector. The S*-statistic has a very small weight, while the minimum distance and the number of private SNPs have larger weights. (B) The distance vector has a mix of positive and negative weights, suggesting uninformative statistics. (C) The individual frequency spectrum mostly has negative weights and lower frequency entries generally have larger weights associated with them.



163 positive to negative from one entry to another (Figure 3B) suggesting there is little signal in the distance
164 vector. On the other hand, the IFS mostly contains negative weights, suggesting that values in these entries
165 are negatively correlated with archaic ancestry (Figure 3C). Notably, S^* makes little contribution to the
166 model likely because it is correlated with the other features included in the model.

167 **Classifier robustness**

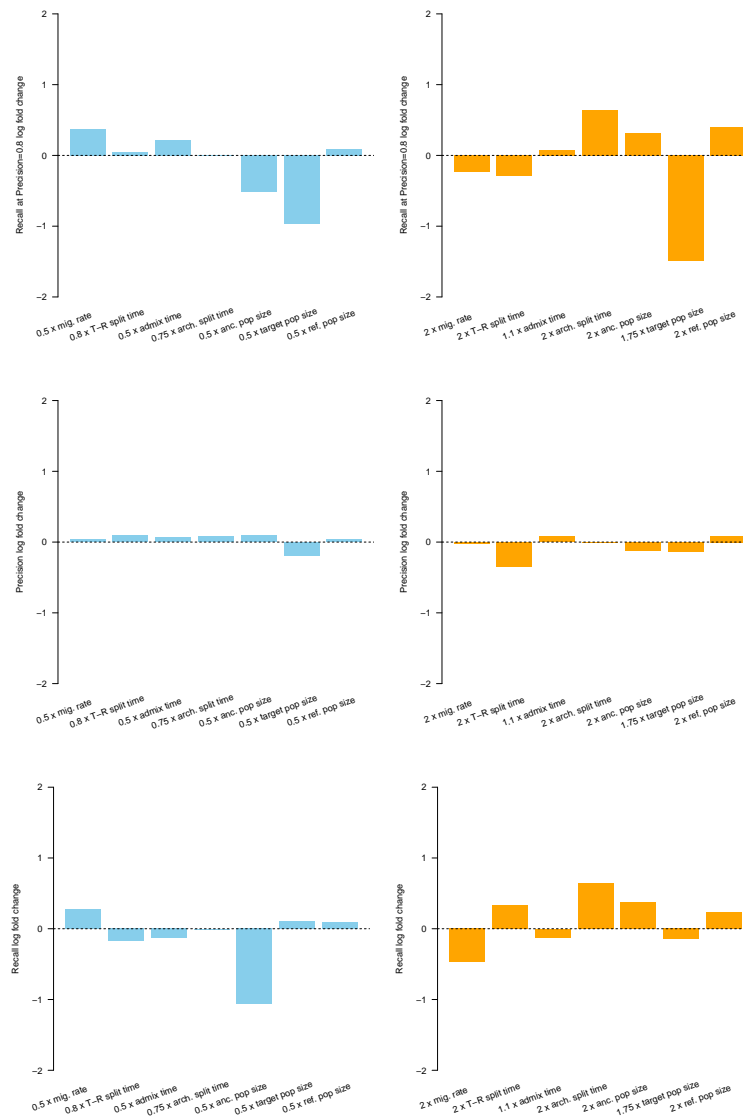
168 Our method relies on simulating data from a demography where the parameters are known. In practice, these
169 parameters are inferred from data with some uncertainty. Thus, we wanted to determine the sensitivity of our
170 method to demographic uncertainty. An exhaustive exploration of demographic uncertainty is challenging
171 given the number of demographic parameters associated with even the simplest demographic models. As an
172 alternative to an exhaustive exploration, we systematically perturbed each parameter at a time, simulated
173 data using the perturbed model, and evaluated the performance of our classifier in terms of changes in
174 precision and recall (trained on the unperturbed parameters corresponding to the Neanderthal demographic
175 history).

176 When the parameters are not perturbed, the recall at a precision of 0.8 is 0.21 (Figure 2B). The accuracy
177 of the classifier is unchanged or increased under many parameter changes (Figure 4). For example, increasing
178 the split time of the modern and archaic populations (T_0) greatly increases the accuracy because there is
179 more time for private mutations to accumulate on the archaic branch. Reducing this time by 25% does not
180 result in a drop in accuracy. For time of admixture (T_a), decreasing the value results in improved accuracy,
181 likely as a result of having longer haplotype blocks and less time for recombination to spread private variants
182 across the haplotypes in the population. Decreasing the split time of the reference and target populations
183 (T_s) largely leaves precision and recall the same.

184 Changing the population sizes of the reference (N_1) populations does not result in large differences in
185 accuracy, while changing the target population size does result in decreased accuracy.

186 Finally, increasing the admixture fraction reduces accuracy, while decreasing it has the opposite effect.
187 While this may seem to contradict the increased accuracy in Figure 2B when simulating a 20% admixture
188 scenario, this is likely due to the fact that the IFS is shifted more by a 2X increase in admixture than a 0.5X

Figure 4: **ArchIE is robust to misspecification in the demographic model.** We tested ArchIE on data simulated after perturbing single demographic parameters lower (left, blue) and higher (right, orange) relative to their values in the training data. Values are reported as log fold changes compared to the baseline model performance (dashed line). We report (A, B) recall at a precision of 0.8 under different parameter misspecification scenarios. (C, D) precision at the threshold that gives a precision of 0.8 ($P(\text{archaic}) = 0.862$). (E, F) recall at the threshold that gives a precision of 0.8 ($P(\text{archaic}) = 0.862$).



189 decrease. Thus there are more simulations in the unperturbed training data that contain samples that are
190 similar to the 0.5X than to the 2X.

191 Recently, Hsieh *et al* [15] used a demographic model estimated from data to infer archaic admixture in
192 African Pygmy populations. They modeled gene flow from an archaic human population that split off 24,137
193 generations ago into a Pygmy population 5,344 generations ago at a frequency of 2%.

194 In addition to the perturbations we performed here, we wanted to see how ArchIE would perform under an
195 alternative model unrelated to what it had been trained on. Importantly, this model contains many different
196 values across all parameter changes, which provides a different test than systematically perturbing single
197 parameters. We simulated data under the Hsieh *et al* model and found that at the 80% precision threshold,
198 ArchIE attained a precision of 0.998 and a recall of 0.700. This high performance is partly due to the fact
199 that under this demographic model, the archaic split time is nearly double that of the Human-Neanderthal
200 split time, allowing additional time for the archaic and modern human lineages to differentiate.

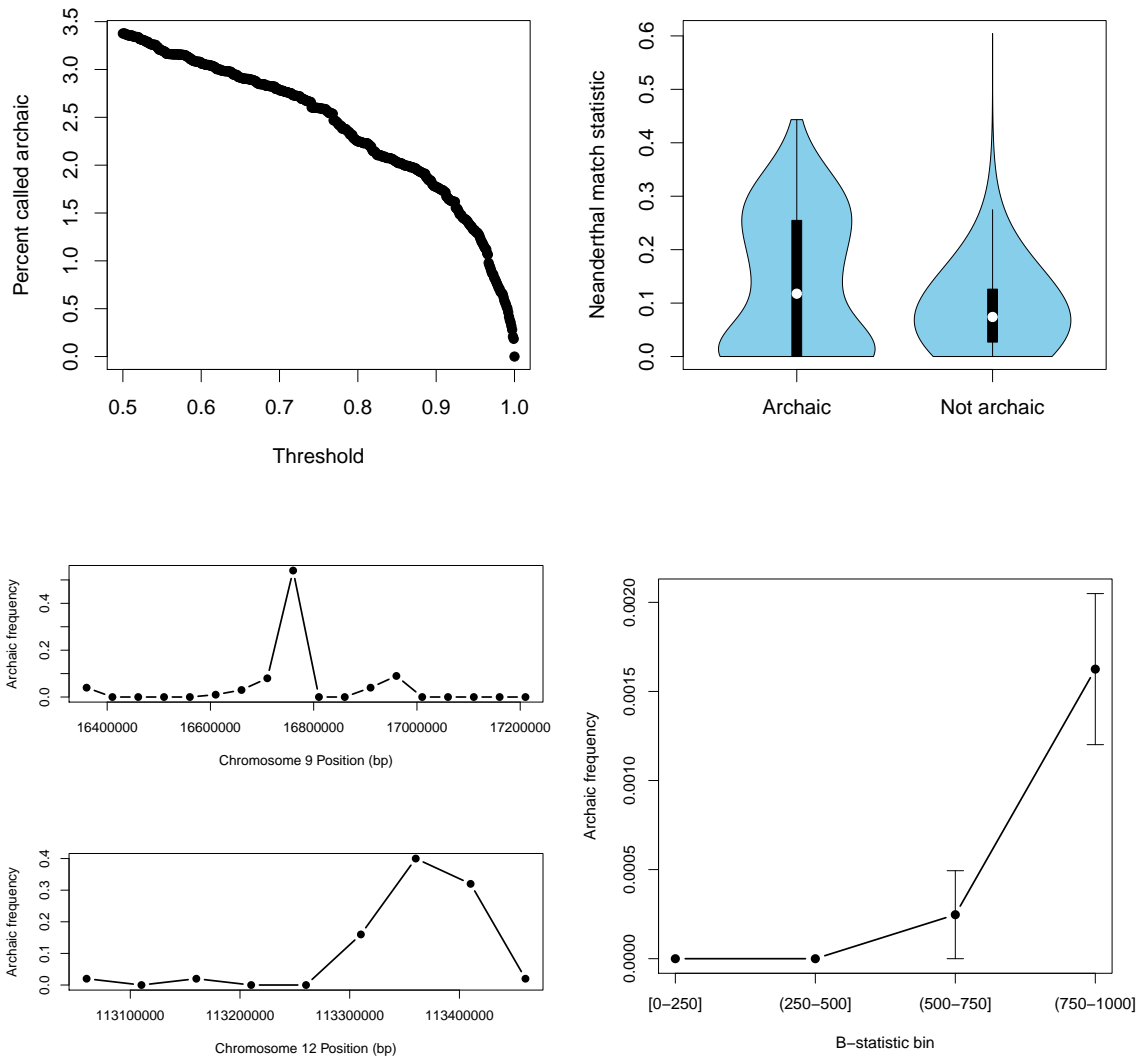
201 **ArchIE detects Neanderthal introgression in European genomes**

202 To validate our method on a realistic setting, we applied our method to data from Phase 3 of the 1000
203 Genomes Project[20] to detect regions of Neanderthal introgression in the European populations without us-
204 ing any of the Neanderthal genomes [6, 10, 27]. We compared our inferences to results from a previous method
205 that inferred Neanderthal ancestry using the high-coverage Altai Neanderthal genome as a reference[10] and
206 trained on data simulated under a demographic model with parameters described in [2].

207 We randomly selected 50 individuals from a European (CEU) population as our target individuals and 50
208 individuals from an African (YRI) population as a reference and calculated the summary statistics described
209 above. We evaluated the average percent of windows inferred as archaic as a function of the calling threshold
210 (Figure 5A). On average, we inferred 1.99% (jackknife SE= 0.3%) of the genome as archaic at a precision of
211 0.8 ($P(\text{archaic})= 0.862$) of the genome as archaic, which is in line with proportion of Neanderthal ancestry
212 from previous analyses [2, 6, 10].

213 Next, we sought to determine whether the archaic haplotypes inferred by our model are enriched for
214 introgressed Neanderthal variants. For each 50 kb window, we computed a Neanderthal match statistic

Figure 5: **Application of ArchIE to 1000 Genomes European (CEU).** (A) Percentage of genome called archaic as a function of probability threshold. (B) Neanderthal match statistic for haplotypes inferred as archaic vs non-archaic. (C) Frequency of haplotypes labeled as archaic near *BNC2* gene and (D) the *OAS* gene cluster. (E) Mean frequency of archaic ancestry increases with B-statistic. A B-statistic near 0 denotes more selectively constrained regions. Lines indicate standard error of the mean.



215 (NMS) as the number of shared variants between an individual haplotype and the Altai Neanderthal reference
216 genome sequence [10] divided by the total number of segregating sites in that window. We see that the archaic
217 regions confidently inferred by our method ($P(\text{archaic}) \geq 99.99\%$) have a higher NMS suggesting that, even
218 in the absence of a reference genome, the archaic ancestry segments identified by the classifier are likely to
219 represent introgressed Neanderthal sequence (Figure 5B; P value = 1.87×10^{-11} via block jackknife).

220 We then focused on two genomic regions at which the frequency of Neanderthal ancestry in European
221 individuals has been found to be relatively high: the *BNC2* gene (Chromosome 9:16,409,501-16,870,786)
222 [2] and the *OAS* gene cluster (Chromosome 12:113,344,739-113,357,712) [7]. The frequency of confidently
223 inferred archaic ancestry is substantially increased in both these genes (Figure 5C, D).

224 Finally, we analyzed the correlation between a measure of selective constraint of a given genomic region
225 (B-value [19]) and frequency of confidently inferred archaic segments in the CEU population in the same
226 region. Sankararaman *et al.* 2014 [2] describe a relationship where more constrained regions (lower B-value)
227 have a lower frequency of archaic ancestry. We observe the same trend where more neutral regions (B-value
228 ≥ 750) contain more archaic ancestry than constrained regions (B-value ≤ 250), consistent with selection
229 against the archaic ancestry (P value = 8.49×10^{-4} via block jackknife; Figure 5E).

230 These analyses suggest that ArchIE obtains results concordant with those from a previous reference-aware
231 method indicating that the inferences from ArchIE are reasonable. We caution, however, that the observed
232 concordance can be inflated due to any biases shared by the two methods.

233 **Ghost admixture into the Yoruba**

234 We next turned our attention to inferring archaic ancestry across the genomes of 50 Yoruban individuals
235 (YRI) from the 1000 Genomes Project [20]. We set the CEU population as the MH reference population
236 and inferred archaic ancestry in 50KB windows. We applied the logistic regression predictor trained using
237 the parameters from the modern Human-Neanderthal demography. While this predictor is likely to be most
238 accurate if the introgressing archaic population had a similar relationship to the Yoruba as the Neanderthals
239 to the CEU, we expect the predictor to be sensitive to introgression events from populations that diverged
240 from the ancestors of the Yoruba before the YRI-CEU split and then introgressed after. We found that at a

241 precision of 0.8 ($P(\text{archaic})= 0.862$), 7.97% (jackknife SE= 0.6%) of the genome is inferred to harbor archaic
242 ancestry on average.

243 To understand the source of archaic ancestry in the Yoruba, we first computed a Neanderthal Match
244 Statistic as before and found a significant enrichment of Neanderthal matching windows (Figure 6B, P
245 value = 5.87×10^{-37} via block jackknife). It is plausible that this archaic ancestry is, at least partly, the
246 result of Neanderthal introgression into the Yoruba mediated by more recent west Eurasian gene flow into
247 Yoruba [10]. However, the proportion of Neanderthal ancestry in Yoruba is very small (about 2×10^{-4}) [10]
248 so that we would not expect this small proportion to explain our signal.

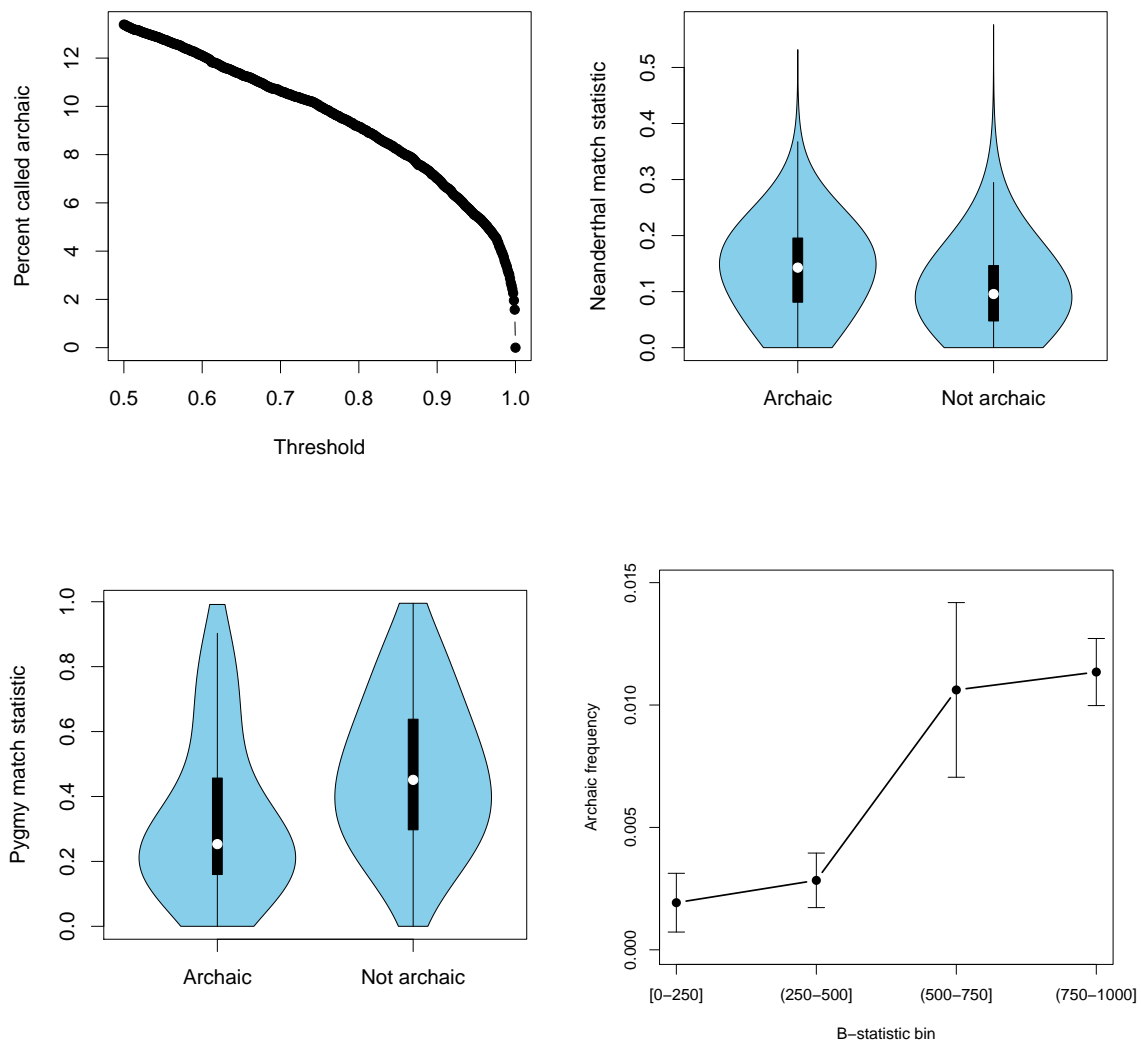
249 Another potential explanation for this archaic signal is the result of admixture with an extant but highly-
250 diverged African population. Recent studies have provided evidence for recent gene flow between Yoruba
251 and western Central African Pygmy populations[15]. To test the hypothesis that the archaic ancestry is
252 due to admixture with an ancestral population related to present-day Pygmy populations, we ran a similar
253 matching statistic using a genome from the Biaka Pygmy as the reference (Figure 6C). While there is much
254 more matching in non-archaic haplotypes, consistent with a more recent divergence, there is a depletion of
255 matching with archaic haplotypes suggesting that the Biaka are not the source of admixture (P value =
256 3.23×10^{-16} via block jackknife). Thus, the source of archaic ancestry in the Yoruba does not appear to be
257 well-represented by extant populations.

258 **The genomic distribution of archaic ancestry in the Yoruba**

259 We examined the relationship between B-value and archaic frequency to test where selectively constrained
260 regions are less likely to contain archaic ancestry. More constrained regions (B-value ≤ 250) harbor less
261 archaic ancestry than more neutrally evolving regions (B-value ≥ 750) (P value = 3.01×10^{-9} via block
262 jackknife; Figure 6D) indicating that archaic alleles that introgressed into the Yoruban population were
263 deleterious on average.

264 On the other hand, we also find evidence for loci at which the introgressed archaic alleles are segregating
265 at substantially elevated frequencies with 2.1 MB of introgressed sequence at high-frequency ($\geq 50\%$) and
266 258 MB of introgressed sequence total (Figure 8). Previous studies [9, 28] found evidence for introgres-

Figure 6: **Application of ArchIE to 1000 Genomes Yoruba (YRI)**. (A) Average percentage of genome called archaic as a function of calling threshold for 50 Yoruban individuals. (B) Neanderthal match statistics for archaic regions and non archaic called in Yoruba. (C) Pygmy match statistic for archaic and non archaic regions in Yoruba. (D) B-value versus archaic ancestry frequency. Lines indicate standard error of the mean.



267 sion into Yoruban individuals using the S^* -statistic on 135 loci in 12 individuals. Based on this limited
268 data, they found evidence for introgression in several genes. We used our map of archaic introgression
269 in the Yoruba to confirm their top three hits (ranked by P -value), *XRCC4* (Chromosome 5:82,373,317-
270 82,649,579), *TJP1* (Chromosome 15:29,992,338-30,261,038), *DUT* (Chromosome 15:48,623,215-48,635,570)
271 (Figure 7), validating their results. Further, we found several genes at high frequency including *NF1*, a
272 tumor suppressor gene, *HSD17B2*, a gene involved with hormone regulation, and *KCNIP4*, which is a gene
273 involved with potassium channels (Table 1). We also find genes coding for a transcription factor and an
274 miRNA, suggesting a role for transcription regulation in these introgressed genes. Of the genes where the
275 archaic haplotype is at high frequency, several have been found in previous scans for positive selection in the
276 Yoruban population, including *NF1* [29, 30], *KCNIP4* [31], and *TRPS1* [32].

277 Taken together, these results suggest that introgression from one or more deeply diverged populations
278 has shaped the genomes of a modern human population in Africa. Further, we find that natural selection
279 has altered the frequency of these introgressed haplotypes, suggesting there are possible functional impacts
280 of this introgression.

Chromosome	Gene name	Frequency	Gene type
chr17	<i>KRT18P61</i>	0.84	pseudogene
chr1	<i>RP11-286M16</i>	0.84	lincRNA
chr17	<i>NF1</i>	0.83	protein coding
chr21	<i>MIR125B2</i>	0.76	miRNA
chr16	<i>HSD17B2</i>	0.74	protein coding
chr1	<i>RN7SKP160</i>	0.74	pseudogene
chr4	<i>KCNIP4</i>	0.73	protein coding
chr8	<i>TRPS1</i>	0.71	protein coding
chr17	<i>RP1115E18</i>	0.67	pseudogene
chr6	<i>MTFR2</i>	0.67	protein coding

Table 1: Genomics regions with a high frequency of archaic ancestry in the Yoruba.

Figure 7: Validation of previously found genes suggestive of archaic introgression [9]. Top to bottom: *XRCC4*, *TJP1*, *DUT*.

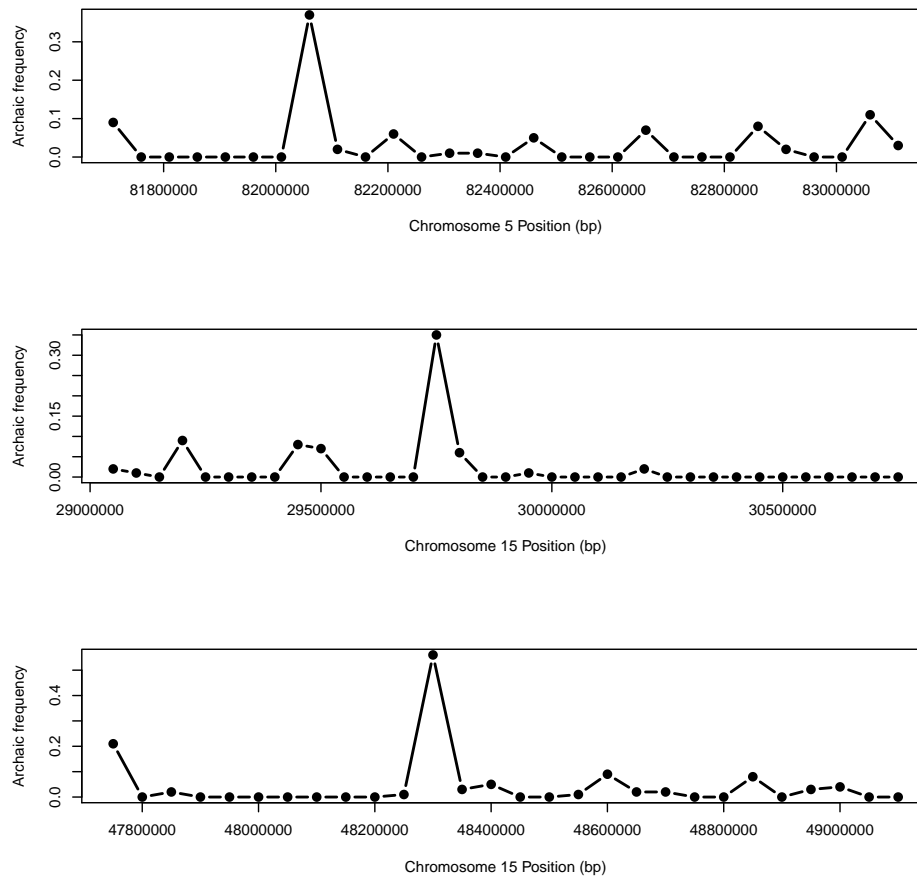
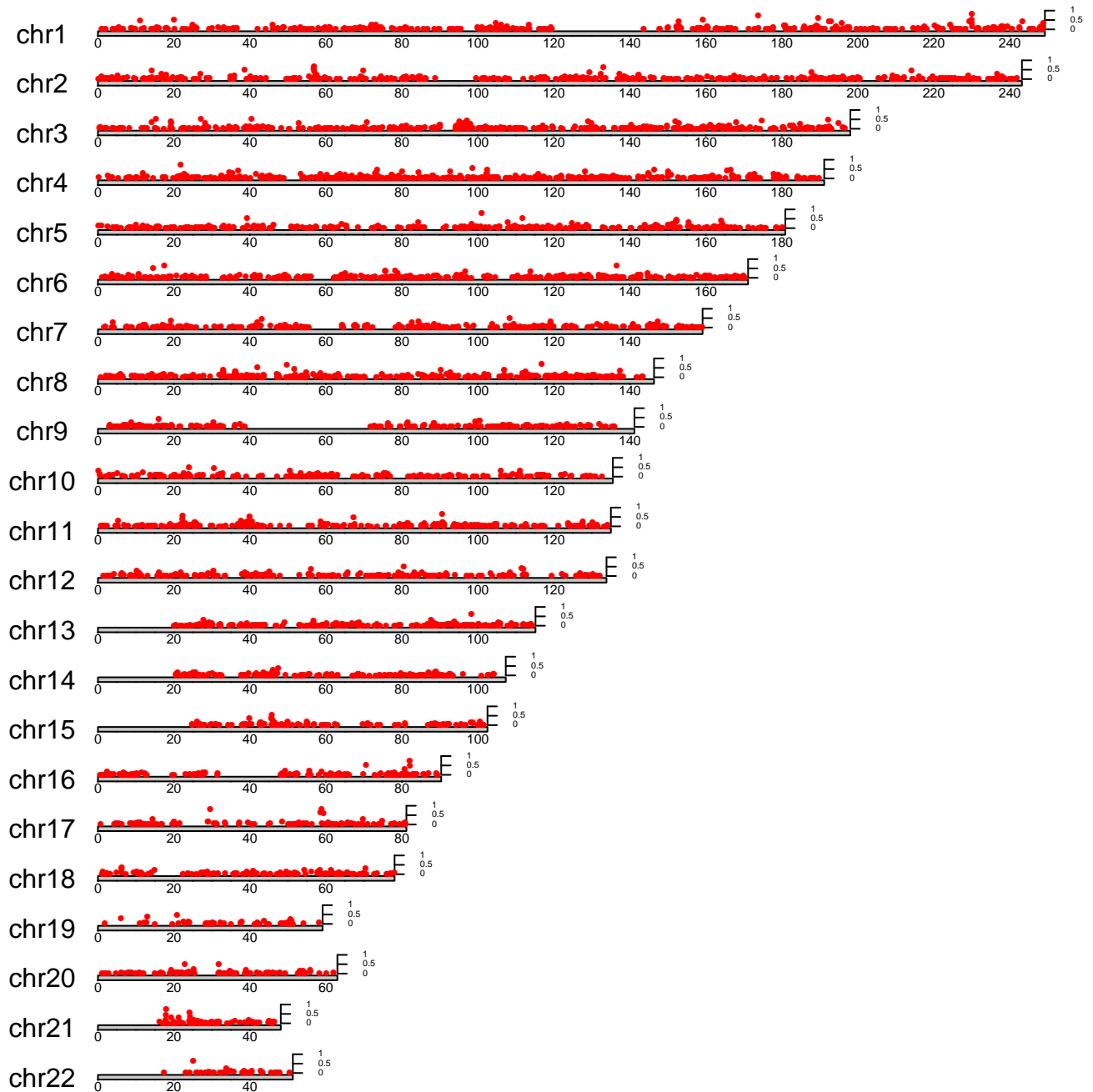


Figure 8: **Genome wide map of archaic ancestry in Yoruba.** Y-axis denotes the frequency of archaic haplotypes.



281 Discussion

282 Detecting archaic admixture and characterizing its impact on genetic variation is an important problem
283 in human population genetics. Here, we present a classification approach to detecting regions of archaic
284 local ancestry without the need for an archaic reference sequence. Our method combines weakly informative
285 signals across a wide range of statistics to create a more powerful predictor.

286 Our results suggest that Yoruban individuals trace about 7.9% of their genomes to an as yet unidentified
287 archaic population. This is in agreement with some results from previous papers in other African populations
288 such as the Biaka and the Baka [15], suggesting that there was a rich diversity of hominin species within Africa
289 and that introgression was commonplace. Using our inferred segments of archaic ancestry in the Yoruba,
290 we find that there are regions of the genome that are under higher selective constraint have reduced archaic
291 ancestry on average indicating that the archaic alleles were deleterious in the hybrid population. More data
292 is needed for a complete picture of these ghost populations. For example, it is unclear whether the archaic
293 signatures found here are from the same as those found in other African populations[13, 14, 15, 33].

294 One advantage of our approach is that the learning algorithm is general allowing it to be applied broadly to
295 diverse prediction problems as well as input features while its simplicity allows for a transparent interpretation
296 of the features and the model. It is possible for us to examine the weights and determine the relative
297 contribution the algorithm learns to place on each feature. In doing so, we find that there is moderate
298 weight on each value of the individual frequency spectrum and more weight is placed on the skew of the
299 distance vector, the number of private SNPs, and the minimum distance to the MH reference.

300 There are several future directions we propose based off these results. First, it is possible to train more
301 complex models like deep neural networks, which can learn and capture non-linear relationships between
302 features and tend not to suffer from the curse of dimensionality [16]. These methods have been used to great
303 success in tasks such as image classification [34] and we anticipate their use in population genetics could
304 improve predictive power. Preliminary results applying deep learning to this problem with the features used
305 here are promising, motivating future work. A related direction could be to automatically learn features
306 from raw sequencing data. Our method relies on hand crafted features that are informed by population
307 genetics theory, similar to other methods that have been proposed in population genetics [16, 17, 35, 36].

308 In conclusion, our method improves on previous methods for reference-free inference of archaic ancestry by
309 combining informative summary statistics in a statistical learning framework. We anticipate that this method
310 will be informative not only in human populations where questions about admixture with other hominins
311 abound, but also in other species and systems where pervasive admixture has shaped the distribution of
312 genetic variation.

313 **Methods**

314 **Simulating training data**

315 We simulated training and test data sets using a modified version of ms [21] that tracks ancestry of each
316 site in each individual genome. Using a previously proposed demographic model relating modern humans
317 and Neanderthals [2], we sampled 100 haplotypes from from population 2 (the target), and 100 haplotypes
318 from population 1 (MH reference) over a region of length 50 kb. We use a mutation rate $\mu = 1.25 \times 10^{-8}$
319 and a recombination rate $r = 1 \times 10^{-8}$. This was used as training data for both the CEU-Neanderthal
320 introgression inference and the YRI-Ghost introgression inference.

321 The general demography is as follows: a population of size N_a splits from a population of size N_0 T_0
322 generations ago. Then, at T_S , two populations split off from the ancestral population that are size N_1 and
323 N_2 . Then, at time T_A , the archaic population migrates into P_2 with a rate of m . See Figure 1A for a
324 graphical outline.

325 **Feature calculation**

326 Each simulation at a given locus generates 100 haplotypes in the target population. For each haplotype, we
327 calculate the following classes of summary statistics: haplotype level frequency spectrum, distance vector to
328 all haplotypes within the test population, minimum distance to haplotypes in population 1, the number of
329 private SNPs, and the S*-statistic.

330 The individual frequency spectrum is created as follows: given a sample of n haplotypes, for each hap-
331 lotype j , we construct a vector X of length n where entry X_i counts the number of derived alleles in the

332 focal haplotype j at frequency i . For example, the first entry counts the number of singletons present in the
333 haplotype, the second entry counts the number of doubletons and so on until n .

334 The distance vector is a vector of length n where each entry is the Euclidean distance from haplotype x
335 to haplotype y_i over all sites, where x is the focal haplotype and y_i is the haplotype being compared.

336 This results in 208 features per example (a 50kb window for a single haploid genome), with 100 exam-
337 ples per locus and 10,000 loci resulting in 1,000,000 examples for training before filtering haplotypes with
338 intermediate levels of admixture.

339 Learning algorithm

340 We tested the ability of a logistic regression framework to classify archaic ancestry from non-archaic ancestry.
341 We used the “glm” function in R to construct a logistic model using the family=binomial(“logit”) option.
342 We used the predict function to obtain a prediction and converted it to a probability using the “plogis”
343 function.

344 We evaluated the performance using Precision-Recall (PR) curves. We calculated precision and recall as:

$$Recall(t) = \frac{TP(t)}{TP(t) + FN(t)}$$
$$Precision(t) = \frac{TP(t)}{TP(t) + FP(t)}$$

345 Where $TP(t)$ is the number of true positives at threshold t , $FN(t)$ is the number of false negatives at
346 threshold t , and $FP(t)$ is the number of false positives at threshold t . In this case, a true positive is a
347 haplotype that traces ancestry back to the archaic population that we call as archaic. A false negative is
348 a haplotype that also traces ancestry back to the archaic population, but does not pass our threshold for
349 calling archaic. A false positive is a haplotype that passes our threshold for being called archaic, but traces
350 its ancestry back to the target population rather than the archaic population.

351 Robustness

352 We examined the robustness of ArchIE to a specified demographic model by systematically perturbing one
353 parameter at a time, simulating a dataset, and evaluating ArchIE’s performance. We doubled and halved

354 the parameters, except when doing so would produce a demographic model that is not sensible.

355 Neanderthal introgression

356 We validated our method using the Neanderthal introgression scenario as a test case. We downloaded phased
357 CEU genomes from the 1000 Genomes Phase 3 dataset [20] and calculated the features mentioned above in
358 50KB windows. For each individual haplotype, we inferred the probability that the window is archaic. We
359 then intersected our calls with the 1000 Genomes strict mask using BEDtools v2.26.0 [37], removing regions
360 that are difficult to map to.

361 We calculated a Neanderthal match statistic (NMS) for focal haplotype i as the number of segregating
362 sites in a window shared with the Altai Neanderthal [10] genome:

$$NMS_i = \frac{S_i}{N_i + H_i}$$

363 where S_i denotes the number of segregating sites between the focal haplotype and the Neanderthal
364 genome, N_i denotes the number of Neanderthal segregating sites, and H_i denotes the number of human
365 segregating sites. Since the Neanderthal genome is not phased, we counted a site as matching if it contained
366 at least one single matching allele or more. We dropped sites with missing data in the Altai reference genome.

367 We tested for significant differences between windows we call archaic and non archaic using a 1 megabase
368 (MB) block jackknife. For each window, we compute the mean NMS for archaic and non archaic haplotypes,
369 take the difference, and then divide by the ungrouped mean NMS to control for mutation rate heterogene-
370 ity. To compute significance, we drop 1 MB windows (non-overlapping) and recalculate the genome wide
371 difference in means.

372 Background selection

373 In order to assess the relationship between background selection and inferred archaic ancestry, we use the
374 B-values from McVicker *et al.* 2009 [19] and intersected them with our calls. For visualization, we binned
375 the B-values into 4 bins, [0-250], (250-500], (500-750], and (750-1000].

376 We tested for significant differences in allele frequency between the lowest and highest bins using a block
377 jackknife, dropping each window and recalculating the difference in allele frequency.

378 **Ghost admixture**

379 We evaluated the presence of ghost admixture into the Yoruba population by sampling 50 individuals from
380 the 1000 Genomes project phase 3 data set [20]. As before, we calculated features in 50 KB windows,
381 intersected the calls with the 1000 Genomes strict mask, this time using the CEU population as the MH
382 reference. We calculated NMS on confidently archaic and non archaic haplotypes as above and calculated a
383 Pygmy match statistic (PMS) using a single Biaka genome as the reference [1].

384 **Acknowledgements**

385 We would like to thank Emilia Huerta Sánchez and Benjamin Vernot for help with S*, members of the
386 Lohmueller and Sankararaman labs, the UCLA Medical and Population Genetics group for helpful dis-
387 cussions, and Alec Chiu for comments on a draft of the paper. SS is supported in part by NIH grants
388 R00GM111744, R35GM125055, an Alfred P. Sloan Research Fellowship, and a gift from the Okawa Founda-
389 tion.

390 **References**

- 391 [1] Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.
392 *Nature* **538**, 201 (2016). URL <https://www.nature.com/articles/nature18964>.
- 393 [2] Sankararaman, S. *et al.* The genomic landscape of Neanderthal an-
394 cestry in present-day humans. *Nature* **507**, 354–357 (2014). URL
395 <https://www.nature.com/nature/journal/v507/n7492/full/nature12961.html>.
- 396 [3] Vernot, B. & Akey, J. M. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes.
397 *Science* **343**, 1017–1021 (2014). URL <http://science.sciencemag.org/content/343/6174/1017>.

- 398 [4] Simonti, C. N. *et al.* The phenotypic legacy of admixture between modern humans and Neandertals.
399 *Science* **351**, 737–741 (2016). URL <http://science.sciencemag.org/content/351/6274/737>.
- 400 [5] McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-Introgressed Se-
401 quences on the Landscape of Human Gene Expression. *Cell* **168**, 916–927.e12 (2017). URL
402 <http://www.sciencedirect.com/science/article/pii/S0092867417301289>.
- 403 [6] Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710–722 (2010). URL
404 <http://science.sciencemag.org/content/328/5979/710>.
- 405 [7] Mendez, F. L., Watkins, J. C. & Hammer, M. F. Neandertal origin of genetic variation at the cluster
406 of OAS immunity genes. *Molecular Biology and Evolution* **30**, 798–801 (2013).
- 407 [8] Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
- 408 [9] Plagnol, V. & Wall, J. D. Possible Ancestral Structure in Hu-
409 man Populations. *PLOS Genetics* **2**, e105 (2006). URL
410 <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0020105>.
- 411 [10] Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**,
412 43–49 (2014). URL <http://www.nature.com/nature/journal/v505/n7481/abs/nature12886.html>.
- 413 [11] Seguin-Orlando, A. *et al.* Genomic structure in Europeans dating back at least 36,200 years. *Science*
414 **346**, 1113–1118 (2014). URL <http://science.sciencemag.org/content/346/6213/1113>.
- 415 [12] Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*
416 **468**, 1053–1060 (2010).
- 417 [13] Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C. & Wall, J. D. Genetic evidence for
418 archaic admixture in Africa. *Proceedings of the National Academy of Sciences* **108**, 15123–15128 (2011).
419 URL <http://www.pnas.org/content/108/37/15123>.
- 420 [14] Lachance, J. *et al.* Evolutionary History and Adaptation from High-Coverage Whole-
421 Genome Sequences of Diverse African Hunter-Gatherers. *Cell* **150**, 457–469 (2012). URL
422 [http://www.cell.com/cell/abstract/S0092-8674\(12\)00831-8](http://www.cell.com/cell/abstract/S0092-8674(12)00831-8).

- 423 [15] Hsieh, P. *et al.* Model-based analyses of whole-genome data reveal a complex evolutionary his-
424 tory involving archaic introgression in Central African Pygmies. *Genome Research* (2016). URL
425 <http://genome.cshlp.org/content/early/2016/02/08/gr.196634.115>.
- 426 [16] Sheehan, S. & Song, Y. S. Deep Learning for Population Genetic In-
427 ference. *PLOS Computational Biology* **12**, e1004845 (2016). URL
428 <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004845>.
- 429 [17] Schrider, D. R. & Kern, A. D. S/HIC: Robust Identification of Soft and Hard
430 Sweeps Using Machine Learning. *PLOS Genetics* **12**, e1005928 (2016). URL
431 <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005928>.
- 432 [18] Schrider, D., Ayroles, J., Matute, D. R. & Kern, A. D. Supervised machine learning reveals intro-
433 gressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *bioRxiv* 170670 (2017). URL
434 <https://www.biorxiv.org/content/early/2017/09/25/170670>.
- 435 [19] McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread Genomic Signatures
436 of Natural Selection in Hominid Evolution. *PLOS Genetics* **5**, e1000471 (2009). URL
437 <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000471>.
- 438 [20] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–
439 74 (2015). URL <https://www.nature.com/nature/journal/v526/n7571/full/nature15393.html>.
- 440 [21] Hudson, R. R. Generating samples under a Wright–Fisher neutral
441 model of genetic variation. *Bioinformatics* **18**, 337–338 (2002). URL
442 <https://academic.oup.com/bioinformatics/article/18/2/337/225783>.
- 443 [22] Heerwaarden, J. v. *et al.* Genetic signals of origin, spread, and introgression in a large sample of
444 maize landraces. *Proceedings of the National Academy of Sciences* **108**, 1088–1092 (2011). URL
445 <http://www.pnas.org/content/108/3/1088>.

- 446 [23] Hufford, M. B. *et al.* The Genomic Signature of Crop-Wild In-
447 trogression in Maize. *PLOS Genetics* **9**, e1003477 (2013). URL
448 <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003477>.
- 449 [24] Brandvain, Y., Kenney, A. M., Flagel, L., Coop, G. & Sweigart, A. L. Speciation and Introgres-
450 sion between *Mimulus nasutus* and *Mimulus guttatus*. *PLOS Genetics* **10**, e1004410 (2014). URL
451 <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004410>.
- 452 [25] Novikova, P. Y. *et al.* Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating
453 speciation and abundant trans-specific polymorphism. *Nature Genetics* **48**, 1077–1082 (2016). URL
454 <https://www.nature.com/ng/journal/v48/n9/full/ng.3617.html>.
- 455 [26] Kryvokhyzha, D. *et al.* Parental legacy, demography, and introgression influenced the evolution of the
456 two subgenomes of the tetraploid *Capsella bursa-pastoris* (Brassicaceae). *bioRxiv* 234096 (2017). URL
457 <https://www.biorxiv.org/content/early/2017/12/13/234096>.
- 458 [27] Meyer, M. *et al.* A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*
459 **338**, 222–226 (2012). URL <http://science.sciencemag.org/content/338/6104/222>.
- 460 [28] Wall, J. D., Lohmueller, K. E. & Plagnol, V. Detecting Ancient Admixture and Estimating Demographic
461 Parameters in Multiple Human Populations. *Molecular Biology and Evolution* **26**, 1823–1827 (2009).
462 URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2734152/>.
- 463 [29] Kudaravalli, S., Veyrieras, J.-B., Stranger, B. E., Dermitzakis, E. T. & Pritchard, J. K. Gene expression
464 levels are a target of recent natural selection in the human genome. *Molecular Biology and Evolution*
465 **26**, 649–658 (2009).
- 466 [30] Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713
467 (2013).
- 468 [31] International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million
469 SNPs. *Nature* **449**, 851–861 (2007).

- 470 [32] Barreiro, L. B., Laval, G., Quach, H., Patin, E. & Quintana-Murci, L. Natural selection has
471 driven population differentiation in modern humans. *Nature Genetics* **40**, 340–345 (2008). URL
472 <https://www.nature.com/articles/ng.78>.
- 473 [33] Xu, D. *et al.* Archaic Hominin Introgression in Africa Contributes to Functional Salivary
474 MUC7 Genetic Variation. *Molecular Biology and Evolution* **34**, 2704–2715 (2017). URL
475 <https://academic.oup.com/mbe/article/34/10/2704/3988100>.
- 476 [34] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015). URL
477 <http://www.nature.com/nature/journal/v521/n7553/full/nature14539.html>.
- 478 [35] Schrider, D. R. & Kern, A. D. Machine Learning for Population Genetics: A New Paradigm. *bioRxiv*
479 206482 (2017). URL <https://www.biorxiv.org/content/early/2017/10/20/206482>.
- 480 [36] Chan, J. *et al.* A Likelihood-Free Inference Framework for Population Ge-
481 netic Data using Exchangeable Neural Networks. *bioRxiv* 267211 (2018). URL
482 <https://www.biorxiv.org/content/early/2018/02/18/267211>.
- 483 [37] Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities
484 for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). URL
485 <https://academic.oup.com/bioinformatics/article/26/6/841/244688>.