Title

A hidden Markov model approach for simultaneously estimating local ancestry and admixture time using next generation sequence data in samples of arbitrary ploidy

Short Title

Estimating local ancestry and admixture time in samples of arbitrary ploidy

Authors

Russell Corbett-Detig[1,2,*] and Rasmus Nielsen[2,3]

[1]Department of Biomolecular Engineering, UC Santa Cruz, Santa Cruz, CA
[2]Department of Integrative Biology, UC Berkeley, Berkeley, CA
[3]The Natural History Museum of Denmark, University of Copenhagen, Denmark.
*Correspondence to: russcd@gmail.com

Keywords

Local Ancestry Inference, Admixture, Hidden Markov Model, *Drosophila melanogaster*, Pool-seq

Abstract

Admixture—the mixing of genomes from divergent populations—is increasingly appreciated as a central process in evolution. To characterize and quantify patterns of admixture across the genome, a number of methods have been developed for local ancestry inference. However, existing approaches have a number of shortcomings. First, all local ancestry inference methods require some prior assumption about the expected ancestry tract lengths. Second, existing methods generally require diploid genotypes, which is not feasible to obtain for many sequencing projects. Third, many methods assume samples are diploid, however a wide variety of sequencing applications will fail to meet this assumption. To address these issues, we introduce a novel hidden Markov model for estimating local ancestry that models the read pileup data, rather than genotypes, is generalized to arbitrary ploidy, and can estimate the time since admixture during local ancestry inference. We demonstrate that our method can simultaneously estimate the time since admixture and local ancestry with good accuracy, and that it performs well on samples of high ploidy—*i.e.* 100 or more chromosomes. We apply our method to pooled sequencing data derived from populations of *Drosophila melanogaster* on an ancestry cline on the east coast of North America. We find that regions of low recombination show steeper clines than regions of high recombination, suggesting that selection against foreign ancestry has had the largest effect in these regions presumably due to increased linkage between neutral and selected sites. We also identify numerous outlier loci associated with behavior suggesting selection associated with prezygotic reproductive isolation. Finally, we identify candidate genes associated with reproductive isolation between ancestral

subpopulations of *D. melanogaster*. Our results illustrate the potential of local ancestry inference for elucidating fundamental evolutionary processes.

Author Summary

When divergent populations hybridize their offspring obtain a portion of their genome from each parent population. Although the average ancestry proportion in each descendant is equal to the proportion of ancestors from each of the ancestral populations, the contribution of each ancestry type is variable across the genome. Estimating local ancestry within admixed individuals is a fundamental goal for evolutionary genetics, and here we develop a method for doing this that circumvents many of the problems associated with existing methods. Briefly, our method can use short read data, rather than genotypes and can be applied to samples with any number of chromosomes. Furthermore, our method simultaneously estimates local ancestry, and the number of generations since admixture—the time that the two ancestral populations first encountered each other. Finally, in applying our method to data from an admixture zone between ancestral populations of *Drosophila melanogaster*, we find many lines of evidence consistent with natural selection operating to against the introduction of foreign ancestry into populations of predominantly one ancestry type. Because of the generality of this method, we expect that it will be useful for a wide variety of existing and ongoing research projects.

67 # Introduction

68 Characterizing the biological consequences of admixture—the mixing of genomes from
69 divergent ancestral populations—is a fundamental and important challenge in
70 evolutionary genetics. Admixture has been reported in a variety of natural populations of
71 animals [1,2], plants [3-5] and humans [6,7], and theoretical and empirical evidence
72 suggests that admixture may affect a diverse suit of evolutionary processes. Individuals'
73 ancestry can affect disease susceptibility in admixed populations, and inferring and
74 correcting for sample population ancestries is a common practice in human genome wide
75 association studies [8-10]. More generally, admixture has the potential to influence
76 patterns of genetic variation within populations [11,12], to introduce novel adaptive
77 [13,14] and deleterious variants [7,15,16], as well as to disrupt epistatic gene networks
78 [17,18]. Therefore, developing a comprehensive understanding of the extent of admixture
79 in natural populations and resulting mosaic genome structures is essential to furthering
80 our understanding of a diverse suite of evolutionary processes.
81
82 Estimating genome-wide ancestry proportions has become a common practice in
83 population genetic inference. For example, the program STRUCTURE [19], originally
84 released in 2000, uses a Bayesian framework to model the ancestry proportions of
85 individuals derived from any number of source populations based on genotype data at a set
86 of unlinked genetic markers. More recently, this model for ancestry proportion estimation
87 has been extended to cases where individual genotypes are not known, but can be studied
88 probabilistically using low-coverage sequencing short read sequencing data [20], which is
89 an important step towards accommodating modern sequencing practices. Additionally,

90 Bergland *et. al.* [21] developed a method for estimating ancestry proportions in pooled
91 population samples of relatively high ploidy (*i.e.* 40-250 distinct chromosomes) from short
92 read sequencing data. In general, it is straightforward to estimate genome-wide ancestry
93 proportions using a number of sequencing strategies and applications.
94
95 It is substantially more challenging to accurately estimate local ancestry (LA) at markers
96 distributed along the genome of a sample. Nonetheless, analyses of LA have the potential to
97 yield more nuanced insights into our understanding of the evolutionary processes affecting
98 ancestry proportions across the genome. One of the first LAI methods was an extension of
99 the STRUCTURE [19] framework that modeled the correlation in ancestry among markers
100 due to linkage. Because the ancestry at each locus is not observed, Falush *et al.* [22]
101 suggested that a hidden Markov model is a straightforward means of inferring the ancestry
102 states at each site in the genome (which are unobserved) based on observed genotype data
103 distributed along a chromosome. Most subsequent LAI methods have also used an HMM
104 framework. The majority of LAI models that have been developed are geared towards
105 estimating LA in admixed human populations (*e.g.* [23,24]). Consequently, most existing
106 LAI methods are limited to diploid genomes with high quality genotype calls. Furthermore,
107 many methods require phased reference panels [24,25], and require the user to provide an
108 estimate of, or make implicit assumptions about, the number of generations since the initial
109 admixture event [2,23-25]. This is straightforward with human population genomic
110 samples, where abundant high quality genotyped samples are available and for which well-
111 documented demographic histories are sometimes known. However for most other species,

112 demographic histories are less well characterized, and assumptions about admixture times
113 may bias the result of LAI methods.
114
115 A number of approaches exist to estimate the time since admixture based on a well
116 characterized ancestry tract length distributions [26-29] but in general, these parameters
117 are unknown prior to LAI. We may therefore expect to improve LAI by simultaneously
118 estimating LA and demographic parameters (*e.g.* admixture time). Furthermore, in the
119 majority of sequencing applications, relatively low individual sequencing coverage is often
120 used to probabilistically estimate individual and population allele frequencies (*e.g.* [30])
121 but these data are often not sufficient to determine high confidence genotypes that are
122 required for existing LAI applications. Hence, there is a clear need for a general LAI method
123 that can accommodate genotype uncertainty and requires less advanced knowledge of
124 admixed populations' demographic histories.
125
126 Here, we introduce a framework for simultaneously estimating LA using short read pileup
127 data and the time of admixture within a population. Briefly, as with many previously
128 proposed LAI methods, we model ancestry across the genome of a sample as a hidden
129 Markov model (HMM). We estimate LA by explicitly modeling read counts as a function of
130 sample allele frequencies within an admixed population. Our method is generalized to
131 accommodate arbitrary sample ploidies, and is therefore applicable to haploid (or inbred),
132 diploid, tetraploid, as well as pooled sequencing applications. We show that this approach
133 accurately infers the time since admixture when data are simulated under the assumed
134 model. Furthermore, our method yields accurate LA estimates for simulated datasets,

135 including samples of high sample ploidy and including evolutionary scenarios that violate
136 the assumptions of the neutral demographic model. In comparisons between ours and one
137 existing LAI method, LAMPanc [23], we find that our approach offers a significant
138 improvement and is accurate over longer time scales. Furthermore, we demonstrate, using
139 a published dataset, that even state-of-the-art LAI methods can be significantly impacted by
140 assumptions about the time since admixture, and that our method provides a solution to
141 this problem.
142
143 Finally, we apply this method to a *Drosophila melanogaster* ancestry cline on the east coast
144 of North America. This species originated in sub-Saharan Africa, and approximately
145 10,000—15,000 years ago a subpopulation expanded out of the ancestral range. During
146 this expansion, the derived subpopulation experienced a population bottleneck that
147 resulted in decreased nucleotide polymorphism, extended linkage disequilibrium within
148 the derived population and substantial genetic differentiation between ancestral and
149 derived populations [2,31-35]. Hereafter, the ancestral population will be referred to as
150 "African" and the derived population as "Cosmopolitan". Following this bottleneck,
151 descendant populations of African and Cosmopolitan *D. melanogaster* have admixed in
152 numerous geographic regions [2,11,21]. Of particular relevance to this work, North
153 America was colonized recently by a population descendent from African individuals from
154 the South, and by a population descendent from cosmopolitan *D. melanogaster* in the North
155 [11,21,34]. Where these populations encountered each other in eastern North America,
156 they form an ancestry cline where southern populations have a greater contribution of
157 African ancestry than northern populations [21].

158

159 Previous work on these ancestry clines has shown that ancestry proportions vary across
160 populations with increasing proportions of cosmopolitan alleles in more temperate
161 localities. Evidence suggests spatially varying selection affects the distribution of genetic
162 variants [36-41]. Furthermore, strong epistatic reproductive isolation barriers partially
163 isolate individuals from northern and southern populations along this ancestry cline
164 [42,43]. This may be generally consistent with recent observations of ancestry-associated
165 epistatic fitness interactions within a *D. melanogaster* population in North Carolina [17],
166 and with the observation of widespread fitness epistasis between populations of this
167 species more generally [44]. There is therefore good reason to believe that natural
168 selection has acted to shape LA clines that are tightly linked to selected mutations in these
169 *D. melanogaster* populations.
170

171 Here, we show that the slopes of LA clines in North American *D. melanogaster* are
172 positively correlated with recombination rates, consistent with natural selection acting to
173 reduce African introgression into predominantly Cosmopolitan populations. We also find
174 that the X displays a higher rate of LA outlier loci, potentially consistent with a greater role
175 of the X chromosome in genetically isolating Cosmopolitan and African lineages, and we
176 identify numerous clinal outlier loci. These loci are disproportionately likely to be
177 associated with organismal behavior, and may play important roles in generating and
178 maintaining reproductive isolation between African and Cosmopolitan *D. melanogaster*
179 populations.
180

181 # Results and Discussion

182

183 **The Model**

184 Although admixed populations often are diploid, we derived a general model of ploidy in
185 which the individual has n gene copies at each locus, i.e. for diploid species $n = 2$. In
186 practice, sequences are often obtained from fully or partially inbred individuals (e.g. [35]),
187 which represent only a single uniquely derived chromosome. It is also common to pool
188 individuals prior to sequencing for allele frequency estimation, so called pool-seq (e.g.
189 [21,36,38,45-48]). If the pooling fractions are exactly equal, such a sample of b diploid
190 individuals can be treated as a sample from a single individual with ploidy $n = 2b$. Although
191 that requirement is restrictive, pool-seq has been experimentally validated as a method for
192 accurate allele frequency estimation—*i.e.* alleles are approximately binomially sampled
193 from the sample allele frequencies [49]. We therefore aimed to derive a model that can
194 accommodate arbitrary sample ploidies. In the model, we assumed that the focal
195 population was founded following a single discrete admixture event between two ancestral
196 subpopulations, labeled 0 and 1, with admixture proportions 1-m and m, respectively, at a
197 time t generations in the past. We modeled emission probabilities such that the method
198 can work directly on read pileup data, rather than high quality known genotypes. Briefly,
199 in our model, we specify an HMM $\{H_v\}$ with state space $S = \{0,1,\ldots,n\}$, where $H_v = i$, $i \in S$,
200 indicates that in the vth position i chromosomes are from population 0 and $n - i$
201 chromosomes are from population 1. In other words, this HMM enables one to estimate
202 what ancestry frequencies are present at a given site along a chromosome within a sample.
203 Importantly, we designed this method to simultaneously estimate the time of admixture,

204 which is related to the correlation between ancestry informative markers along a
205 chromosome. See Methods for a complete description of the HMM including the emissions
206 and transition probability calculations. The source code and manual are available at
207 https://github.com/russcd/Ancestry_HMM.
208

209 **Dependence on Ancestral Linkage Disequilibrium**

210 Within an admixed population, there are two sources of LD. LD that is induced due to the
211 correlation of alleles from the same ancestry type (*i.e.* admixture LD), and LD that is
212 present within each of the ancestral populations (ancestral LD). Admixture LD, is the signal
213 of LA that we seek to detect using the HMM. The second type, ancestral LD, limits the
214 independence of the ancestral information captured by each marker, and is expected to
215 confound HMM-based analyses, particularly as we aimed to estimate the time since
216 admixture within this framework. We therefore sought to quantify the effect of ancestral
217 LD by discarding one of each pair of sites in LD within either ancestral population. We
218 found that ancestral LD tends to increase admixture time estimates obtained using our
219 method, and we decreased the cutoff of the LD parameter, $|r|$, by 0.1 until the time
220 estimates obtained for single chromosomes were unbiased with respect to the true time
221 since admixture. We found that $|r| \leq 0.4$ fit this criterion, although for relatively ancient
222 admixture events with highly skewed ancestry proportions—*i.e.* $m < 0.1$ or $m > 0.9$—some
223 residual bias was apparent in the estimates of admixture time (Figure 1). This reflects the
224 fact that the SMC' ancestry tract distribution performs poorly with highly skewed ancestry
225 proportions and especially for long times since admixture [50].
226

227 Figure 1 also reveals a striking difference between otherwise equivalently skewed
228 admixture proportions. For example when m = 0.1, there was a much larger effect of
229 ancestral LD than when m = 0.9. This is due to differences in the variability and LD within
230 the ancestral populations. That is, due to the strong population bottleneck, cosmopolitan *D.*
231 *melanogaster* populations have substantially more LD and fewer polymorphic sites than
232 African *D. melanogaster* populations. Because the time estimation procedure appears to be
233 sensitive to the amount of ancestral LD present in the data, simulations of the type we
234 described here may be necessary to determine what $|r|$ cutoffs are required to produce
235 unbiased time estimates given the ancestral LD of the populations in a given analysis using
236 this method.
237
238 **Accuracy and Applications to Diploid and Pooled Samples**
239 We next sought to quantify the accuracy of our approach across varying sample ploidies
240 and times since admixture (Figure 2). Especially for moderate and short admixture times
241 (*i.e.* 0—500 generations), our method performed well for all ploidies considered and we
242 were able to accurately recover the correct admixture time with relatively little bias.
243 However, as true admixture time increases, the time estimates for pooled samples become
244 significantly less reliable and show a clear negative bias. Nonetheless, across the range of
245 times presented in Figure 2, samples of ploidy one and two showed little bias, and we
246 therefore believe our method will produce sufficiently accurate admixture time estimates
247 for a wide variety of applications.
248

249 All measures of accuracy decrease with increasing time since admixture (Figure 2).
250 However, even for relatively long times since admixture—2000 generations—and for large
251 sample ploidies, the mean posterior error remained relatively low for all ancestry
252 proportions and for long times since admixture. This indicates that this approach may be
253 sufficiently accurate for a wide variety of applications, sequencing depths, and sample
254 ploidies. Nonetheless, as the proportion of sites within the 95% credible interval decreased
255 with larger pool sizes, it is clear that for larger pools the posterior credible interval tends to
256 be too narrow, and correcting for this bias may be necessary for applications that are
257 sensitive to the accuracy of the credible interval.
258
259 **Non-Independence Among Ancestry Tracts**
260 As described above, estimates of the time of admixture demonstrate an apparent bias in
261 pools of higher ploidy (Figure 2). Specifically, time tends to be slightly overestimated for
262 relatively short admixture times and underestimated at relatively long admixture times.
263 This is particularly apparent at highly skewed ancestry proportions. Given that this bias is
264 primarily evident in pools of 10 to 20 individuals, we hypothesized that it might be due to
265 the non-independence of ancestry tracts among chromosomes, which should tend to
266 disproportionately affect samples of higher ploidy because all ancestry breakpoints are
267 assumed to be independent in our model. To test this, we simulated genotype data from
268 independent and identically distributed exponential tract lengths as is assumed by our
269 model. When we ran our HMM on this dataset, we found that no bias is evident for
270 simulations of up to 2000 generations (Figure 3), indicating that the primary cause of this
271 bias was violations in the real data of the independence of ancestry tracts that we assumed

272 when computing the transition probabilities. However, it should be possible to quantify
273 and correct for this bias in applications of this method that aim to estimate the time since
274 admixture.
275
276 **Robustness to Unknown Population Size**
277 The transition probabilities of this HMM depend on knowledge of the population size. In
278 practice, this parameter is unlikely to be known with certainty. Hence, to assess the impact
279 of misspecification of the population size, we performed simulations using a range of
280 population sizes that span three orders of magnitude (N=100, 1000, 10000, and 100000).
281 All analyses presented here were conducted by applying our HMM to haploid and diploid
282 samples, but qualitatively similar results hold for samples of larger ploidy (not shown). We
283 then analyzed these data assuming the default population size, 10000, is correct. For
284 relatively short times since admixture, there was not a clear bias for any of the true
285 population sizes considered. However, at longer true admixture times, estimated
286 admixture times for both N=100 and N=1000 asymptote at a number of generations near to
287 the population sizes. This result reflects the fact that smaller populations will tend to
288 coalesce at a portion of the loci in the genome relatively quickly, and ancestry tracts cannot
289 become smaller following coalescence. Nonetheless, the accuracy of LAI remained high
290 even when time estimates were unreliable (Figure 4) for the tested marker densities and
291 patterns of LD. Furthermore, in some cases it should be straightforward to determine if a
292 population has coalesced to either ancestry state at a large portion of the loci in the
293 genome, potentially obviating this issue.
294

295 A more subtle departure from the expectation was evident for population sizes that are
296 larger than we assumed in analyzing these data (Figure 4). This likely reflects the fact that
297 the probability of back coalescence to the previous marginal genealogy to the left after a
298 recombination event is inversely related to the population size. Hence, the rate of transition
299 between ancestry types is actually slightly higher in larger populations where back
300 coalescence is less likely than we assumed during the LAI procedure. This produced a slight
301 upward bias in the estimates of admixture time when the population was assumed to be
302 smaller than it is in reality. However, this bias appears to be relatively minor, and we
303 expect that time estimates obtained using this method will be useful so long as population
304 sizes can be approximated to within an order of magnitude. Of course, this bias is not
305 unique to our application, and it will affect methods that aim to estimate admixture time
306 after LAI as well. That is, estimating the correct effective population size is an inherent
307 problem for all admixture demographic inference methods.
308
309 **Application to Ancient Admixture**
310 Although it is clear that accurately estimating relatively ancient admixture times is
311 challenging in higher ploidy samples, we sought to determine the limits of our approach for
312 LAI and time estimation for longer admixture times for haploid sequence data. Because of
313 rapid coalescence in smaller samples (see above), we performed admixture simulations
314 with a diploid effective population size of 100,000. It is clear that there is a limit to the
315 inferences that can be made directly using our method. Like the higher ploidy samples,
316 time estimates for haploid samples departed from expectations shortly after 2,000
317 generations since admixture (Figure 5). Nonetheless, the magnitude of this bias is slight,

318 and it is likely that it could be corrected for when applying this method even for very
319 ancient admixture events. For all admixture times considered, LAI remained acceptably
320 accurate despite the slight bias in time estimates (Figure 5).
321
322 **Reference Panel Size**
323 One question is what effect varying the reference panel sizes will have on LAI inference
324 using this method. We therefore compared results from reference panels of size 10 with
325 those from panels of size 100 (Figure 6). As with results obtained for reference panels of
326 size 50, panels of size 100 were sufficient to accurately estimate admixture time and LA
327 over many generations since admixture. Whereas, when panel sizes were just 10
328 chromosomes, time estimates were clearly biased and the result was variable across
329 ancestry proportions (Figure 6). However, since there was a strong correlation between
330 true and estimated admixture times even with relatively small panel sizes, it may therefore
331 be possible to infer the correct time by quantifying this bias through simulation and
332 correcting for it. Furthermore, although LAI is clearly less reliable with smaller panels,
333 these results are not altogether discouraging and this approach, in conjunction with
334 modest reference panels may still be effective for some applications.
335
336 **High Sample Ploidy**
337 In a wide variety of pool-seq applications, samples are pooled in larger groups than we
338 have considered above (*e.g.* [36,45,47]). We are therefore interested in determining how
339 our method will perform on pools of 100 individuals. Towards this, we performed
340 simulations as before, but we designed our parameters to resemble those of the pooled

341 sequencing data that we analyze in the application of this method below. Specifically, we
342 simulated data with a mean sequencing depth of 25, a time since admixture of 1500
343 generations, and an ancestry proportion of 0.8. Consistent with results for ploidy 20, we
344 found that time tends to be dramatically underestimated (*i.e.* the mean estimate of
345 admixture time was 680 generations). However, when we provided the time since
346 admixture, our method produced reasonably accurate LAI for these samples. Although the
347 posterior credible interval was again too narrow, the mean posterior error was just 5.4 (or
348 0.054 if expressed as an ancestry frequency), indicating that this approach can produce LA
349 estimates that are close to their true values for existing sequencing datasets (*e.g.* Figure 7).
350 However, the HMM's run time increases dramatically for higher ploidy samples and higher
351 sequencing depths, a factor that may affect the utility of this program for some analyses.
352 Nonetheless, for more than 36,000 markers, a sample ploidy of 100 and a mean sequencing
353 depth of 25, the average runtime was approximately 42 hours. In contrast, for the same set
354 of parameters, but where individuals are sequenced and analyzed as diploids, the mean
355 runtime was just 8 minutes.
356
357 **Robustness to Deviations From the Neutral Demographic Model**
358 An important concern is that many biologically plausible admixture models would violate
359 the assumptions of this inference method. In particular, continuous migration and selection
360 acting on alleles from one parental population are two potential causes of deviation from
361 the expected model in the true data. To assess the extent of this potential bias, we
362 performed additional simulations. First, we considered continuous migration at a constant
363 rate that began t generations prior to sampling. In simulations with continuous migration,

364 additional non-recombinant migrants enter the population each generation. Relative to a
365 single pulse admixture model, this indicates that the ancestry tract lengths will tend to be
366 longer than those under a single pulse admixture model in which all individuals entered at
367 time *t*. Indeed, we found that admixture times tended to be underestimated with models of
368 continuous migration. However, the accuracy of LAI remained high across all situations
369 considered here (Table 1), indicating that the LAI aspect of this approach may be robust to
370 alternative demographic models.
371
372 **Table 1.** Parameter estimation and LAI when admixture occurs at a constant rate, rather
373 than in a single pulse.

Admixture Time	Number of Loci	Sample Ploidy	Estimated Time	Proportion in 95% CI	Mean 95% CI Size	Mean Posterior Error	Proportion MLE Correct
100	2	1	96	1.000	0.017	0.005	0.996
		2	98	0.999	0.082	0.022	0.986
		10	105	0.969	1.370	0.457	0.631
		20	93	0.923	2.194	0.856	0.340
	5	1	91	1.000	0.014	0.004	0.997
		2	88	0.999	0.058	0.017	0.988
		10	85	0.942	0.898	0.382	0.662
		20	65	0.949	1.450	0.834	0.302
	10	1	88	1.000	0.014	0.004	0.997
		2	86	0.999	0.060	0.016	0.989
		10	84	0.972	1.049	0.356	0.719
		20	76	0.944	1.887	0.704	0.459
	20	1	79	1.000	0.012	0.004	0.997
		2	74	0.999	0.041	0.013	0.991
		10	65	0.939	0.645	0.305	0.726
		20	53	0.779	1.082	0.704	0.396
500	2	1	521	0.998	0.096	0.027	0.980
		2	518	0.993	0.352	0.096	0.932
		10	595	0.969	2.243	0.735	0.472
		20	486	0.923	3.272	1.184	0.288
	5	1	430	0.998	0.085	0.024	0.983
		2	411	0.993	0.312	0.087	0.938

	10	287	0.955	1.842	0.639	0.523
	20	227	0.881	2.712	1.093	0.325
10	1	341	0.998	0.058	0.018	0.987
	2	303	0.994	0.192	0.059	0.956
	10	177	0.914	1.197	0.511	0.592
	20	140	0.793	1.884	0.992	0.343
20	1	272	0.999	0.040	0.011	0.992
	2	236	0.995	0.142	0.042	0.970
	10	124	0.957	0.974	0.349	0.740
	20	100	0.918	1.607	0.641	0.563

In the second set of simulations, we considered additive selection on alleles that are perfectly correlated with local ancestry in a given region (*i.e.* selected sites with frequencies 0 in population 0 and frequency 1 in population 1), and experience relatively strong selection (selective coefficients were between 0.005 and 0.05). We placed selected sites at 2, 5, 10 and 20 loci distributed randomly across the simulated chromosome, where admixture occurred through a single pulse. Ancestry tracts tend to be longer immediately surrounding selected sites, and we therefore expected admixture time to be underestimated when selection is widespread. When the number of selected loci was small, time estimates were nearly unbiased (Table 2), suggesting that our approach can yield reliable admixture time estimates despite the presence of a small number of selected loci (*i.e.* 2 selected loci on a chromosome arm). However, with more widespread selection on alleles associated with local ancestry, time estimates showed a downward bias that increased with increasing numbers of selected loci. This is likely because selected loci will tend to be associated with longer ancestry tracts due to hitchhiking. However, the accuracy of the LAI remains high for all selection scenarios that we considered here, further

391 indicating that our method can robustly delineate LA, even when the data violate

392 assumptions of the inference method (Table 1,2).

393

394 **Table 2.** Parameter estimation and LAI when a subset of loci experience natural selection

395 in the admixed population.

Admixture Time	Migration Rate	Sample Ploidy	Estimated Time	Proportion in 95% CI	Mean 95% CI Width	Mean Posterior Error	Proportion MLE Correct
100	0.0005	1	53	1.000	0.002	0.001	1.000
		2	49	1.000	0.006	0.002	0.998
		10	129	0.963	0.305	0.168	0.839
		20	98	0.545	0.328	0.661	0.353
	0.001	1	55	1.000	0.004	0.001	0.999
		2	53	1.000	0.013	0.004	0.997
		10	156	0.951	0.558	0.288	0.727
		20	90	0.551	0.719	0.858	0.179
	0.002	1	54	1.000	0.006	0.002	0.999
		2	52	0.999	0.019	0.006	0.996
		10	123	0.949	0.758	0.354	0.671
		20	74	0.679	1.115	0.889	0.176
	0.004	1	43	1.000	0.008	0.002	0.998
		2	54	0.999	0.035	0.010	0.993
		10	91	0.955	1.085	0.443	0.605
		20	75	0.860	1.788	0.889	0.248
500	0.0005	1	254	0.999	0.033	0.010	0.993
		2	250	0.997	0.121	0.036	0.974
		10	331	0.956	1.395	0.528	0.557
		20	333	0.882	2.321	1.018	0.261
	0.001	1	266	0.999	0.049	0.014	0.990
		2	268	0.996	0.198	0.055	0.962
		10	325	0.967	1.887	0.628	0.521
		20	366	0.926	3.049	1.109	0.294
	0.002	1	294	0.999	0.055	0.016	0.989
		2	297	0.996	0.238	0.064	0.956
		10	352	0.977	2.076	0.639	0.542
		20	370	0.951	3.238	1.073	0.336
	0.004	1	346	0.999	0.038	0.010	0.993
		2	350	0.997	0.164	0.042	0.973
		10	403	0.989	1.634	0.455	0.692

20	462	0.979	2.773	0.833	0.473

Comparison to LAMPanc

We next compared the results of our method to those of LAMPanc [23]. Because LAMPanc accepts only diploid genotypes, we provided this program diploid genotype data. However, for these comparisons, we still ran our method on simulated read pileups with the mean depth equal to 2. LAMPanc was originally designed for local ancestry inference in very recently admixed populations. As expected, LAMPanc performed acceptably for very short admixture times, but rapidly decreased in performance with increasing time (Figure 8). However, by default, LAMPanc removes sites in strong LD within the admixed samples, which includes ancestral LD, but also admixture LD—the exact signal LAI methods use to identify ancestry tracts.

We therefore reran LAMPanc, but instead of pruning LD within the admixed population, we removed sites in strong LD within the ancestral populations as described above in our method. With this modification, LAMPanc performs nearly as well as our method, but remains slightly less accurate especially at longer admixture times (Figure 8). This difference presumably reflects the windowed-based approach of LAMPanc. At longer times since admixture a given genomic window may overlap a breakpoint between ancestry tracts. Although the performance is nearly comparable with this modification, we emphasize that our method enables users to estimate the time since admixture, where this must be supplied for LAMPanc, and allows for LAI on read pileups, therefore incorporating

418 genotype uncertainty into the LAI procedure. Indeed our method is more accurate at longer
419 timescales even when supplied with considerably lower quality read data. However
420 LAMPanc supports LAI with multiple ancestral populations, which our method currently
421 does not (but see Conclusions). Furthermore many extensions of LAMP utilize haplotype
422 information, which may be particularly valuable in populations where LD extends across
423 large distances.
424
425 **Assessing Applications to Human Populations**
426 Given the strong interest in studying admixture and local ancestry in human populations
427 (*e.g.* [22-25]), it is useful to ask if our method can be applied to data consistent with
428 admixed populations of humans. Towards that goal, we simulated data similar to what
429 would be observed in admixture between modern European and African lineages and
430 applied our HMM to estimate admixture times and LA. We found that our method can
431 accurately estimate admixture times for relatively short times since admixture, however,
432 substantially more stringent LD pruning in the reference panels is necessary to produce
433 unbiased estimates (Figure 9). This may be expected given that linkage disequilibrium
434 extends across longer distances in human populations than it does in *D. melanogaster*. In
435 other words, the scales of ancestral LD and admixture LD become similar rapidly in
436 admixed human populations. Furthermore, this approach yields accurate time estimates
437 for shorter times since admixture than with genetic data consistent with *D. melanogaster*
438 populations. For a relatively short time since admixture, around 100 generations, it is
439 possible to obtain accurate and approximately unbiased estimates of the admixture time
440 over a wide range of ancestry proportions, indicating that this method may be applicable to

441 recently admixed human populations as well (Figure 9). Nonetheless, this result
442 underscores the need to examine biases associated with LD pruning in this approach prior
443 to application to a given dataset.
444
445 **Bias in LAI due to Uncertainty in Time of Admixture**
446 To demonstrate that assumptions about the number of generations since admixture have
447 the potential to bias LAI, we analyzed a SNP-array dataset from Greenlandic Inuits [51,52].
448 The authors had previously noted a significant impact of t on the LAI results produced
449 using RFMix [24], which we were able to reproduce here for chromosome 10 (Figure 10).
450 Indeed, even for comparisons between $t = 5$ and $t = 20$, both of which may be biologically
451 plausible for these populations, the mean difference in posterior probabilities between
452 samples estimated using RFMix was 0.0903 (Figure 10). However, when we applied our
453 method to these data, a clear optimum from t was obtained at approximately 6-7
454 generations prior to the present (Figure 10). This comparison therefore demonstrates that
455 even relatively minor changes in assumptions of t have the potential to strongly impact LAI
456 results, and underscores the importance of simultaneously performing LAI while
457 estimating t.
458
459 However, these results also indicate that our method may not be robust in situations where
460 the background LD is high and ancestry informative markers are neither common nor
461 distributed evenly across the genome. When we compared the results of our method a $t = 5$
462 and $t = 20$, we obtained similar differences in the mean posterior among individuals as with
463 RFMix. There are likely two causes. First, the datasets considered were generated with a

464 metabochip SNP-chip [53], which contains a highly non-uniform distribution of markers
465 across the genome. Second, the ancestral LD in the Inuit population is extensive [52], and
466 we could only retain a relatively small proportion of the markers after LD pruning in the
467 reference panels. These results therefore also underscore the challenges of LAI when the
468 signal to noise ratio is low.
469
470 **Patterns of LA on Inversion Bearing Chromosomes in *D. melanogaster***
471 Given their effects suppressing recombination in large genomic regions, chromosomal
472 inversions may be expected to strongly affect LAI [2,54]. Although we attempted to limit
473 the impact of chromosomal inversions by eliminating known polymorphic arrangements
474 from the reference panels (see methods), many known inversions are present within the
475 pool-seq samples we aimed to analyze [55]. We therefore focused on known inverted
476 haplotypes within the DGPR samples [54,56-58], which are comprised of inbred
477 individuals, and therefore phase is known across the entire chromosome.
478
479 In comparing LA estimates between inverted and standard arrangements, it is clear that
480 chromosomal inversions can substantially affect LA across the genomes (Figure 11). In
481 general, the chromosomal inversions considered in this work originated in African
482 populations of *D. melanogaster* [54], and consistent with this observation, most inversion
483 bearing chromosomes showed evidence for elevated African ancestry. This was
484 particularly evident in the regions surrounding breakpoints, where recombination with
485 standard arrangement chromosomes is most strongly suppressed. Importantly, this pattern
486 continued outside of inversion breakpoints as well, consistent with numerous observations

487 that recombination is repressed in heterokaryotypes in regions well outside of the
488 inversion breakpoints in *Drosophila* (*e.g.* [2,54,59]). In(3R)Mo is an exception to this
489 general pattern of elevated African ancestry within inverted arrangements (Figure 11).
490 This inversion originated within a cosmopolitan population [54], and has only rarely been
491 observed within sub-Saharan Africa [60,61]. Consistent with these observations, In(3R)Mo
492 displayed lower overall African ancestry than chromosome arm 3R than standard
493 arrangement chromosomes.
494
495 Although chromosomal inversions may affect patterns of LA in the genome on this ancestry
496 cline, we believed including chromosomal inversions in the pool-seq datasets would not
497 heavily bias our analysis of LA clines. Inversions tend to be low frequency in most
498 populations studied [55], and because they affect LA in broad swaths of the genome—
499 sometimes entire chromosome arms—including inversions is unlikely to affect LA cline
500 outlier identification which appears to affect much finer scale LA (below). Furthermore,
501 inversion breakpoint regions were not enriched for LA cline outliers in our analysis (Table
502 3), suggesting that inversions have a limited impact on overall patterns of local ancestry on
503 this cline. Nonetheless, the LAI complications associated with chromosomal inversions
504 should be considered when testing selective hypotheses for chromosomal inversions as
505 genetic differentiation may be related to LA, rather than arrangement-specific selection in
506 admixed populations such as those found in North America.
507
508 **Table 3.** LA clines in the genomic intervals immediately surrounding breakpoints of known
509 polymorphic inversions.

Inversion	Breakpoint	Rho	p-value
	Distal	0.0298	0.923
In(2L)t	Proximal	-0.297	0.325
	Proximal	-0.0615	0.849
In(2R)NS	Distal	0.254	0.426
	Proximal	-0.336	0.261
In(3R)K	Distal	-0.686	0.00958
	Proximal	-0.494	0.0858
In(3R)Mo	Distal	-0.631	0.0207
	Proximal	0.193	0.527
In(3R)P	Distal	0.0789	0.798
	Distal	-0.0711	0.836
In(3L)P	Proximal	-0.222	0.511

Application to *D. melanogaster* Ancestry Clines

Finally, we applied our method to ancestry clines between cosmopolitan and African ancestry *D. melanogaster*. Genomic variation across two ancestry clines have been studied previously [21,34,36,47]. In particular, the cline on the east coast of North America has been sampled densely by sequencing large pools of individuals to estimate allele frequencies, and previous work has shown that the overall proportion of African ancestry is strongly correlated with latitude [21]. Consistent with this observation, we found a significant negative correlation for all chromosome arms between proportion of average African ancestry and latitude (rho = -0.891, -0.561, -0.912, -0.913, and -0.755, for 2L, 2R, 3L, 3R, and X respectively).

Although global ancestry proportions have previously been investigated in populations on this ancestry cline [21,34], these analyses neglected the potentially much richer information in patterns of LA across the genome. We therefore applied our method to these

526 samples. Because of the relatively recent dual colonization history of these populations and
527 subsequent mixing of genomes, a genome-wide ancestry cline is expected [21]. However,
528 loci that depart significantly in clinality from the genome-wide background levels may
529 indicate that natural selection is operating on a site linked to that locus.
530
531 **Clinality of LA is Strongly Correlated with Recombination Rate**
532 Previous studies have shown that regions of low recombination are disproportionately
533 resistant to admixture [7,17], a pattern that may reflect the fact that loci in low
534 recombination regions are more likely to be tightly linked in an admixed population to a
535 locus that is deleterious on admixed genetic backgrounds, or where one ancestry type is
536 disfavored in the local environment of the admixed population. Consistent with these
537 observations, we observed a significant positive correlation between the mean partial
538 correlation with latitude and local recombination rates across the genome (Table 4).
539 Furthermore, for only chromosome arm 3L did the 95% bootstrap confidence interval for
540 the correlation between LA clinality and recombination rate overlap with 0 (Table 4). All
541 other chromosome arms individually showed a significant positive correlation between LA
542 clinality and recombination rates, indicating that the correlation between LA clines and
543 local recombination rates in the genome is a robust relationship.
544
545 **Table 4.** Genome-wide and arm-specific correlation between recombination and the partial
546 correlation between LA proportion and latitude, including Spearman's Correlation, the p-
547 value for the observed correlation and the 95% bootstrap confidence interval.

Chromosome Arm	rho	p	2.5% Boostrap CI	97.5% Boostrap CI

All	0.1259436	0.0001167	0.084518662	0.20262254
2L	0.2434135	0.000691	0.08819046	0.346119464
2R	0.1839665	0.01633	0.032780198	0.3305687
3L	-0.0986359	0.1213	-0.207119717	0.059341699
3R	0.1649162	0.05151	0.000177156	0.323058668
X	0.2672564	0.0002651	0.166917018	0.445182509

One interpretation of this observation that clinality of LA is strongly correlated with local recombination rates is that selection has had a substantial impact on the distribution of LA in the *D. melanogaster* ancestry cline on the east coast of North America, and lends further support to a growing consensus that low recombination regions may be especially unlikely to introgress between ancestral populations presumably because negatively selected loci have a greater effect on LA due to increased linkage between neutral and selected sites in low recombination genomic regions.

Outlier LA Clines

Selection within admixed populations may take several distinct forms. On the one hand, loci that are favorable in the admixed population—either because they are favored on an admixed genetic background, enhance reproductive success in an admixed population, or are favorable in the local environment—will tend to achieve higher frequencies, and we would expect these sites to have a more positive correlation with latitude than the genome-wide average. Conversely, loci that are disfavored within the admixed population may be expected to skew towards a more negative correlation with latitude.

Although it is not possible to distinguish between these hypotheses directly, a majority of evidence suggests that selection has primarily acted to remove African ancestry from the

568 largely Cosmopolitan genetic backgrounds found in this ancestry cline. First, abundant
569 evidence suggests pre-mating isolation barriers between some African and cosmopolitan
570 populations [62-64]. Second, there is strong post-mating isolation between populations on
571 the ends of this cline [42,43]. Third, we report here a strong negative correlation between
572 LA clines and local recombination rates (above). Finally, circumstantially, the local
573 environment on the east coast of North America is perhaps most similar to Cosmopolitan
574 than to African ancestral populations, which further suggests that Cosmopolitan alleles are
575 likely favored through locally adaptive mechanisms. We therefore examined loci that are
576 outliers for a negative partial correlation with latitude, as this is the expected pattern for
577 African alleles that are disfavored in these populations.
578
579 There is an ongoing debate about the relative merits of an outlier approach versus more
580 sophisticated models for detecting and quantifying selection in genome-wide scans. We
581 believe that the difficulties of accurately estimating demographic parameters for this
582 ancestry cline make the outlier approach most appealing for our purposes. Using our
583 outlier approach, we identified 80 loci that showed the expected negative correlation with
584 latitude (Figure 12). Although the specific statistical threshold that we employed is
585 admittedly arbitrary, given the strength of evidence indicating widespread selection on
586 ancestry in this species (above), we expected that the tail of the LA cline distribution would
587 be enriched for the genetic targets of selection.
588
589 **Differences Among Chromosome Arms**

590 Due to the differences in inheritance, evolutionary theory predicts that selection will
591 operate differently on the X chromosome relative to autosomal loci. Of specific relevance to
592 this work, the large-X effect [65,66] is the observation that loci on the X chromosome
593 contribute to reproductive isolation at a disproportionately high rate. Additionally, and
594 potentially the cause of the large-X effect, due to the hemizygosity of X-linked loci, the X
595 chromosome is expected to play a larger role in adaptive evolution, the so-called faster-X
596 effect [67]. There is therefore reason to believe that the X chromosome will play a
597 significant role in genetically isolating Cosmopolitan and African *D. melanogaster*.
598
599 Consistent with a larger role for the sex chromosomes in generating reproductive isolation
600 or selective differentiation between *D. melanogaster* ancestral populations, we found that
601 that the X chromosome has a lower mean African ancestry proportion than the autosomes
602 in all populations. Furthermore, the X displayed a significantly higher rate of outlier LA loci
603 than the autosomes (23 LA outliers on the X, 57 on the Autosomes, $p = 0.0341$, one-tailed
604 exact Poisson test). Although consistent with evolutionary theory, differences between
605 autosomal arms and the X chromosome may also be explained in part by differences in per
606 base-pair recombination rates on the X chromosome than the autosomes, differences in
607 power to identify LA clines, or by the disproportionately larger number of chromosomal
608 inversions on the autosomes than on the X chromosome in these populations [55,60].
609
610 **Biological Properties of Outlier Loci**
611 We next applied gene ontology analysis to the set of outlier genes to identify common
612 biological attributes that may suggest more specific organismal phenotypes underlying LA

613 clinal outliers. We found modest enrichments in eight GO categories without the set of
614 clinal LA genes, however none remained significant after applying a multiple testing
615 correction (Table 5). Nonetheless, given the abundance of evidence supporting a role for
616 pre-mating isolation barriers between African and Cosmopolitan flies [62-64], the GO term,
617 behavior, is of interest. Consistent with this observation, one of the strongest LA cline
618 outliers, *egh*, has been conclusively linked to strong effects on male courtship behavior
619 using a variety of genetic techniques [68]. Additionally, gene knockouts of CG43759,
620 another LA cline outlier locus, have strong effects on inter-male aggressive behavior [69],
621 and may also contribute to behavioral differences between admixed individuals. These loci
622 are therefore appealing candidate genes for functional follow-up analyses, and illustrate
623 the power of this LAI approach for identifying candidate genes associated with differential
624 selection on ancestral variation in admixed populations.
625
626 **Table 5.** Signficant gene ontology terms for LA clinal outlier loci.

GO Term	Description	p-value	enrichment score
GO:0030054	cell junction	1.33E-04	7.6
GO:0005850	eukaryotic translation initiation factor 2 complex	2.10E-04	83.58
GO:0017177	glucosidase II complex	3.49E-04	66.86
GO:0007016	cytoskeletal anchoring at plasma membrane	8.86E-05	33.43
GO:0065008	regulation of biological quality	6.89E-04	2.63
GO:0007610	behavior	7.10E-04	3.36
GO:0051049	regulation of transport	9.38E-04	5.28
GO:0032507	maintenance of protein location in cell	9.84E-04	15.2

627
628 # Conclusion

629 A growing number of next-generation sequencing projects produce low coverage data that
630 cannot be used to unambiguously assign individual genotypes, but which can be analyzed
631 probabilistically to account for uncertainty in individual genotypes [70-72]. However, most
632 existing LAI methods require genotype data derived from diploid individuals. Hence, there
633 is an apparent disconnect between existing LAI approaches and the majority of ongoing
634 sequencing efforts. In this work, we developed the first framework for applying LAI to
635 pileup read data, rather than error-free genotypes, and we have generalized this model to
636 arbitrary sample ploidies. This method therefore has immediate applications to a wide
637 variety of existing and ongoing sequencing projects, and we expect that this approach and
638 extensions thereof will be valuable to a number of researchers.
639
640 For many applications, a parameter of central biological interest is the time since
641 admixture began (t). A wide variety of approaches have been developed that aim to
642 estimate t and related parameters in admixed populations [26,28,29,73-76]. Many of these
643 methods are based on an inferred distribution of tract lengths, however, inference of the
644 ancestry tract length distribution is associated with uncertainty that is typically not
645 incorporated in currently available methods for estimating t. Furthermore, incorrect
646 assumptions regarding t has the potential to introduce biases during LAI. Hence, it is
647 preferable to estimate demographic parameters such as the admixture time during the LAI
648 procedure. Nonetheless, as noted above, although LAI using our method is robust to many
649 deviations from the assumed model, admixture time estimates are sensitive to a variety of
650 potential confounding factors and examining the resulting ancestry tract distributions after

651 LAI may be necessary to confirm that the assumed demographic model provides a
652 reasonable fit to the data.
653
654 To our knowledge, this is the first method that attempts to simultaneously link LAI and
655 population genetic parameter estimation directly, and we can envision many extensions of
656 this approach that could expand the utility of this method to a broad variety of applications.
657 For example, it is straightforward to accommodate additional reference populations (*e.g.*
658 by assuming multinomial rather than binomial read sampling). Alternatively, any
659 demographic or selective model that can be approximated as a Markov process could be
660 incorporated—in particular, it is feasible to accommodate two-pulse admixture models and
661 possibly models including ancestry tracts that are linked to positively selected sites. Such
662 methods can be used to construct likelihood ratio tests of evolutionary models and for
663 providing improved parameter estimates.
664

665 Methods

666 Constructing Emissions Probabilities

667 We model the ancestry using an HMM $\{H_v\}$ with state space $S = \{0,1,\ldots,n\}$, where $H_v = i$,
668 $i \in S$, indicates that in the vth position i chromosomes are from population 0 and $n - i$
669 chromosomes are from population 1. In the following, to simplify the notation and without
670 loss of generality, we will omit the indicator for the position in the genome as the structure
671 of the model is the same for all positions of equivalent ploidy. We assume each variant site
672 is biallelic, with two alleles A and a, and the availability of reference panels from source
673 populations 0 and 1 with total allelic counts C_{0a}, C_{1a}, C_{0A}, and C_{1A}, where the two subscripts

674 refer to population identity and allele, respectively. Also, $C_0 = C_{0A} + C_{0a}$ and $C_1 = C_{1A} + C_{1a}$.
675 Finally, we also assume we observe a pileup of r reads from the focal population, with r_A
676 and r_a reads for alleles A and a respectively ($r = r_A + r_a$). The emission probability of state
677 $i \in S$ of the process is then defined as $e_i = \Pr(r_A, C_{0A}, C_{1A} \mid r, C_0, C_1, H = i, \varepsilon)$, where
678 ε is an error rate. This probability can be calculated by summing over all possible
679 genotypes in the admixed sample and over all possible population identities of the reads, as
680 explained in the following section.
681
682 The probability of obtaining r_0 ($=r - r_1$) reads, in the admixed population, from
683 chromosomes of ancestry 0, given r and the hidden state $H = i$, and assuming no mapping or
684 sequencing biases, is binomial,
685 $$r_0 \mid H = i, n, r \sim \mathrm{Bin}(r, i/n)$$
686
687 These probabilities are pre-computed in our implementation for all possible values of $i \in S$
688 and r_0, $0 \le r_0 \le r$. Similarly, for the reference populations, for j=0,1,
689
690 $$C_{jA} \mid C_j, f_j \sim \mathrm{Bin}(C_j, f_j)$$
691
692 where f_j is the allele frequency of allele A in population j. The analogous allelic counts in the
693 admixed population, denoted C_{M0a}, C_{M1a}, C_{M0A}, and C_{M1A}, are unobserved (only reads are
694 observed for the admixed population), but are also conditionally binomially distributed,
695 *i.e.*:
696

$$C_{M0A} \mid H = i, f_0 \sim \mathrm{Bin}(i, f_0) \text{ and } C_{M1A} \mid H = i, n, f_1 \sim \mathrm{Bin}(n - i, f_1)$$

Finally, in the absence of errors, and assuming no sequencing or mapping biases, the conditional probability of obtaining r_{0A} reads of allele A in the admixed population is

$$r_{0A} \mid H = i, r_0, C_{M0A} \sim \mathrm{Bin}(r_0, C_{M0A} / i)$$

This probability can be expanded to include errors, in particular assuming a constant and symmetric error rate ε between major and minor allele, and assuming all reads with nucleotides that are not defined as major or minor are discarded, we have

$$r_{0A} | H = i, r_o, C_{M0A}, \varepsilon \sim \mathrm{Bin}(r_0, (1-\varepsilon) C_{M0A}/i + \varepsilon(1 - C_{M0A}/i)).$$

Using these expressions, and integrating over allele frequencies in the source populations, we have

$$\Pr\left(r_{0A}, C_{0A,} | r_0, C_0, n, H = i, \varepsilon\right) =$$

$$\int_0^1 \sum_{k=0}^{i} \Pr(r_{0A} | H = i, r_0, C_{M0A} = k, \varepsilon) \Pr\left(C_{M0A} = k | H = i, f_0\right) p\left(f_0\right) df_0 \quad =$$

$$\frac{C_0!\, i!}{(C_0 - C_{0A})!\, C_{0A}!\, (C_0 + i + 1)!} \sum_{k=0}^{i} \Pr(r_{0A} | H = i, r_0, C_{M0A}$$

$$= k, \varepsilon) \frac{(C_0 - C_{0A} + i - k)!\, (C_{0A} + k)!}{(i - k)!\, k!}$$

assuming a uniform [0, 1] distribution for f_0. A similar expression is obtained for $\Pr\left(r_{1A}, C_{1A,}|r_1, C_1, n, H = i, \varepsilon\right)$, assuming $f_1 \sim \mathrm{U}[0,1]$, and these expressions combine multiplicatively to give

$$\Pr\left(r_A, C_{1A,}, C_{0A,}|r_0, C_0, r_1, C_1, n, H = i, \varepsilon\right) =$$

$$\sum_{r_{0A}=\max\{0, r_A - r_1\}}^{\min\{r_0, r_A\}} \Pr\left(r_{0A}, C_{0A,}|r_0, C_0, n, H = i, \varepsilon\right) \Pr\left(r_{1A} = r_A - r_{0A}, C_{1A,}|r_1, C_1, n, H = i, \varepsilon\right),$$

and the emission probabilities become

$\Pr(r_A, C_{0A}, C_{1A} \mid r, C_0, C_1, H = i, \varepsilon) =$

$$\sum_{r_0=0}^{r} \Pr(r_0|H = i, n, r) \Pr\left(r_A, C_{1A,}, C_{0A,}|r_0, C_0, r_1 = r - r_0, C_1, n, H = i, \varepsilon\right)$$

Alternatively, if the sample genotypes are known with high confidence, i.e. $C_{MA} = C_{M0A} + C_{M1A}$ is observed, the emission probabilities are the defined as

$$\Pr\left(C_{MA}, C_{0A}, C_{1A} \mid C_0, C_1, n, H = i\right) =$$

$$\binom{C_0}{C_{0A}}\binom{C_1}{C_{1A}} \sum_{k=\max\{C_{MA}-i,0\}}^{\min\{n-i, C_{MA}\}} \int_0^1 \binom{n-i}{k} (f_0)^{C_{0A}+k} (1-f_0)^{C_0+n-i-C_{0A}-k} df_0 \int_0^1 \binom{i}{C_{MA}-k} (f_1)^{C_{MA}-k+C_{1A}} (1-f_1)^{C_1+i-C_{1A}-C_{MA}+k} df_1$$

$$= \sum_{k=\max\{C_{MA}-i,0\}}^{\min\{n-i, C_{MA}\}} \frac{C_0!C_1!i!(n-i)!(C_{MA}+C_{1A}-k)!(C_{0A}+k)!(C_1-C_{MA}-C_{1A}+i+k)!(C_0-C_{0A}-i-k+n)!}{(C_0-C_{0A})!C_{0A}!(C_1-C_{1A})!C_{1A}!(C_{MA}-k)!k!(k+i-C_{MA})!(n-k-i)!(n-i+C_0+1)!(i+C_1+1)!}$$

731 These emissions probabilities are sometimes substantially faster to compute than those for
732 short read pileups, especially when sequencing depths are high. However, the genotypes must
733 be estimated with high accuracy for this approach to be valid. For applications with low read
734 coverage, or with ploidy >2 for which many standard genotype callers are not applicable, it is
735 usually preferable to use the pileup-based approach described above.
736
737 **Constructing Transition Probabilities**
738 We assume an admixed population, of constant size, with *N* diploid individuals, in which a
739 proportion *m* of the individuals in the population where replaced with migrants *t*
740 generations before the time of sampling. Given these assumptions, and an SMC' model of
741 the ancestral recombination graph [77], the rate of transition from ancestry 0 to 1, along
742 the length of a single chromosome, is
743
744 $$\lambda_0 = 2Nm\left(1 - e^{\frac{-t}{2N}}\right)$$
745
746 per Morgan [50]. Similarly, the rate of transition from ancestry 1 to 0 on a single
747 chromosome is
748
749 $$\lambda_1 = 2N\left(1 - m\right)\left(1 - e^{\frac{-t}{2N}}\right)$$
750

751 per Morgan. Importantly, because these expressions are based on a coalescence model,
752 they account for the possibility that a recombination event occurs between two tracts of
753 the same ancestry type and the probability that the novel marginal genealogy will back-
754 coalesce with the previous genealogy [50]. Both events are expected to decrease the
755 number of ancestry switches along a chromosome and ignoring their contribution will
756 cause overestimation of the rate of change between ancestry types between adjacent
757 markers.
758
759 The transition rates are in units per Morgan, but can be converted to rates per bp, by
760 multiplying with the recombination rate in Morgans/bp, r_{bp} within a segment. The
761 transition probabilities of the HMM for a single chromosome, $\mathbf{P}(l) = \{P_{ij}(l)\}, i, j \in S,$ between
762 two markers with a distance l between each other, is then approximately
763

764 $$\mathbf{P}(l) = \begin{bmatrix} 1-\lambda_0 r_{bp} & \lambda_0 r_{bp} \\ \lambda_1 r_{bp} & 1-\lambda_1 r_{bp} \end{bmatrix}^l$$

765
766 using discrete distances, or
767
768

769 $$\mathbf{P}(l) = \begin{bmatrix} \frac{\lambda_1}{\lambda_0+\lambda_1} + \frac{\lambda_0}{\lambda_0+\lambda_1} e^{-r_{bp} l (\lambda_0+\lambda_1)} & \frac{\lambda_0}{\lambda_0+\lambda_1} - \frac{\lambda_0}{\lambda_0+\lambda_1} e^{-r_{bp} l (\lambda_0+\lambda_1)} \\ \frac{\lambda_0}{\lambda_0+\lambda_1} + \frac{\lambda_1}{\lambda_0+\lambda_1} e^{-r_{bp} l (\lambda_0+\lambda_1)} & \frac{\lambda_1}{\lambda_0+\lambda_1} - \frac{\lambda_1}{\lambda_0+\lambda_1} e^{-r_{bp} l (\lambda_0+\lambda_1)} \end{bmatrix}$$

using continuous distances along the chromosome. Here, we use the continuous representation for calculations. We emphasize that the assumption of a Markovian process is known to be incorrect [50], in fact admixture tracts tend to be more spatially correlated than predicted by a Markov model, and the degree and structure of the correlation depends on the demographic model [50]. Deviations from a Markovian process may cause biases in the estimation of parameters such as *t*.

The Markov process defined above is applicable to a single chromosome. We now want to approximate a similar process for a sample of *n* chromosomes from a single sequencing pool. The true process is quite complicated, and we choose for simplicity to approximate the process for *n* chromosomes sampled from one population, as the union of *n* independent chromosomal processes. We will later quantify biases arising due to this independence assumption using simulations. Under the independence assumption, the transition probability from *i* to *j* is simply the probability of *l* transitions from state 1 to state 0 in the marginal processes and $j - i + l$ transitions from state 0 to state 1, summed over all admissible values of *l*, i.e.,

$$\Pr\left(H_{v+k} = j \mid H_v = i\right) = \sum_{l=\max\{0,i-j\}}^{\min\{n-j,i\}} \binom{n-i}{j-i+l} \left(P_{01}(k)\right)^{j-i+l} \left(1-P_{01}(k)\right)^{n-j+i-l} \binom{i}{l} \left(P_{10}(k)\right)^{l} \left(1-P_{10}(k)\right)^{i-l}$$

790 Although this procedure can be computationally expensive when there are many markers,
791 read depths are high, and especially when n is large, in our implementation, we reduce the
792 compute time by pre-calculating and storing all binomial coefficients.
793
794 **Estimating Time Since Admixture**
795 A parameter of central biological interest, that is often unknown in practice, is the time
796 since the initial admixture event (t). We therefore use the HMM representation to provide
797 maximum likelihood estimates of t using the forward algorithm to calculate the likelihood
798 function. As this is a single parameter optimization problem for a likelihood function with a
799 single mode, optimization can be performed using a simple golden section search [78].
800 Default settings for this optimization in our software, including the search range maxima
801 defaults, t_{max} and $t_{min,}$ are documented in the C++ HMM source code provided at
802 https://github.com/russcd/Ancestry_HMM.
803
804 **Posterior Decoding**
805 After either estimating or providing a fixed value of the time since admixture to the HMM,
806 we obtained the posterior distribution for all variable sites considered in our analysis using
807 the forward-backward algorithm, and we report the full posterior distribution for each
808 marker along the chromosome.
809
810 **Simulating Ancestral Polymorphism**
811 To validate our HMM, we generated sequence data for each of two ancestral populations
812 using the coalescent simulator MACS [79]. We sought to generate data that could be

813 consistent with that observed in Cosmopolitan and African populations of *D. melanogaster*,
814 which has been studied previously in a wide variety of contexts [2,11,31-33]. We used the
815 command line "macs 400 10000000 -i 1 -h 1000 -t 0.0376 -r 0.171 -c 5 86.5 -I 2 200 200 0 -
816 en 0 2 0.183 -en 0.0037281 2 0.000377 -en 0.00381 2 1 -ej 0.00382 2 1 -eN 0.0145 0.2" to
817 generate genotype data. This will produce 200 samples of ancestry 0 and 200 samples of
818 ancestry 1 on a 10mb chromosome—*i.e.* this should resemble genotype data for about half
819 of an autosomal chromosome arm in *D. melanogaster*. Unless otherwise stated below, we
820 then sampled the first 50 chromosomes from each ancestral population as the ancestral
821 population reference panel, whose genotypes are assumed to be known with low error
822 rates. The sample size was chosen because it is close to the size of the reference panel that
823 we obtained in our application of this approach to *D. melanogaster* (below).
824
825 To evaluate the performance of our method on data consistent with human populations, we
826 simulated data that could be consistent with that observed for modern European and
827 African human populations. Specifically, we simulated the model of [80] using the
828 command line "macs 200 1e8 -I 3 100 100 0 -n 1 1.682020 -n 2 3.736830 -n 3 7.292050 -eg
829 0 2 116.010723 -eg 1e-12 3 160.246047 -ma x 0.881098 0.561966 0.881098 x 2.797460
830 0.561966 2.797460 x -ej 0.028985 3 2 -en 0.028986 2 0.287184 -ema 0.028987 3 x
831 7.293140 x 7.293140 x x x x x -ej 0.197963 2 1 -en 0.303501 1 1 -t 0.00069372 -r
832 0.00069372". Admixture between ancestral populations was then simulated as described
833 below.
834
835 **Simulating Admixed Populations**

836 Although it is commonly assumed that admixture tract lengths can be modeled as
837 independent and identically distributed exponential random variables (e.g. [26,29] and in
838 this work, above), this assumption is known to be incorrect as ancestry tracts are neither
839 exponentially distributed, independent across individuals, nor identically distributed along
840 chromosomes [50]. We therefore aim to determine what bias violations of this assumption
841 will have on inferences obtained from this model. Towards this, we used SELAM [81] to
842 simulate admixed populations under the biological model described above. Because this
843 program simulated admixture in forward time, it generates the full pedigree-based
844 ancestral recombination graph, and is therefore a conservative test of our approach
845 relative to the coalescent which is known to produce incorrect ancestry tract distributions
846 for short times [50]. Briefly, we initialized each admixed population simulation with a
847 proportion, m, of ancestry from ancestral population 1, and a proportion 1-m ancestry from
848 ancestral population 0. Unless otherwise stated, all simulations were conducted with
849 neutral admixture and a hermaphroditic diploid population of size 10,000.
850
851 We then assigned the additional, non-reference chromosomes from the coalescent
852 simulations, to each ancestry tract produced in SELAM simulations according to their local
853 ancestry along the chromosome. In this way, each chromosome is a mosaic of the two
854 ancestral subpopulations. See, *e.g.* [2], for a related approach for simulating genotype data
855 of admixed chromosomes.
856
857 **Pruning Ancestral Linkage Disequilibrium**

858 Correlations induced by LD between markers within ancestral populations violates a
859 central assumption of the Markov model framework. Although it may be feasible to
860 explicitly model linkage within ancestral populations, when ancestral populations have
861 relatively little LD, such as those of *D. melanogaster*, another effective approach is to
862 discard sites that are in strong LD in the ancestral populations. Hence, to avoid this
863 potential confounding aspect of the data, we first computed LD between all pairs of
864 markers within each reference panel that are within 0.01 centimorgans of one another. We
865 then discarded one of each pair of sites where $|r|$ in either reference panel exceeded a
866 particular threshold, and we decreased this threshold until we obtained an approximately
867 unbiased estimate of the time since admixture estimates of the HMM. This approach differs
868 from a previous method, LAMPanc, where LD is pruned from within admixed samples (see
869 also below).
870
871 **Simulating Sequence Data**
872 We first identified all sites where the allele frequencies of the ancestral populations differ
873 by at least 20% within the reference panels. We excluded weakly differentiated sites to
874 decrease runtime and because these markers carry relatively little information about the
875 LA at a given site. Then, to generate data similar to what would be produced using Illumina
876 sequencing platforms, we simulated allele counts for each sample, by first drawing the
877 depth at a given site from a Poisson distribution. In most cases and unless otherwise stated,
878 the mean of this distribution is set to be equal to the sample ploidy. We did this to ensure
879 equivalent sequencing depth per chromosome regardless of pooling strategy, and because
880 this depth is sufficiently low that high quality genotypes cannot be determined. We then

881 generated set of simulated aligned bases via binomial sampling from the sample allele
882 frequency and included a uniform error rate of 0.5% for both alleles at each site.
883
884 Unless otherwise stated, we simulated a total of 40 admixed chromosomes. The HMM can
885 perform LAI on more than one sample at a time, and we therefore included all samples
886 when running it. Hence, we used 40 haploid, 20 diploid, 4 pools of 10 chromosomes, and 2
887 pools of 20 chromosomes for most comparisons of accuracy reported below, unless
888 otherwise stated.
889
890 **Accuracy Statistics**
891 To evaluate the performance of the HMM, we computed four statistics. First, we compute
892 the proportion of sites where the true state is within the 95% posterior credible interval,
893 where ideally, this proportion would be equal to or greater than 0.95. Second, we compute
894 the mean posterior error, the average distance between the posterior distribution of
895 hidden states and the true state
896
897 $$E = \frac{\sum_{v=0}^{S}\sum_{i=0}^{n} p\left(H_v = i \mid \mathbf{r}\right)\left|i - I_v\right|}{k}$$
898
899 Here S is the total number of sites, I_v is the true state at site v, and $\mathbf{r}$ is all the combined read
900 data. Third, we also report the proportion of sites where the maximum likelihood estimate
901 of the hidden state is equal to the true ancestry state. Finally, as an indicator of the
902 specificity of our approach, we also report the average width of the 95% credible interval.

903

904 **Deviations from the Assumed Neutral Demographic Model**

905 A potential issue with this framework is that the assumptions underlying the transition
906 matrixes and related time of admixture estimation procedure is likely to be violated in a
907 number of biologically relevant circumstances. We therefore simulated populations
908 wherein individuals of ancestral population 1 began entering a population entirely
909 composed of individuals from ancestral population 0, at a time *t* generations before the
910 present, at a constant rate that is sustained across all subsequent generations until the time
911 of sampling. That is, additional unadmixed individuals of ancestry 1 migrate each
912 generation from *t* until the present.

913

914 Natural selection acting on admixed genetic regions has been inferred in a wide variety of
915 systems (*e.g.* [5,7,13,17,18]), and is expected to have pronounced effects on the distribution
916 of LA among individuals within admixed populations. Here again, this aspect of biologically
917 realistic populations will tend to violate central underlying assumptions of the model
918 assumed in this work. Towards this, we simulated admixed populations with a single pulse
919 of admixture *t* generations prior to the time of sampling. We then incorporated selection at
920 2,5,10, and 20 loci at locations uniformly distributed along the length of the chromosome
921 arm. All selected loci were assumed to be fixed within each ancestral population. Selection
922 was additive and selective coefficients were assigned based on a uniform [0.005, 0.05]
923 distribution to either ancestry 0 or 1 alleles with equal probability. As above, these
924 simulations were conducted using SELAM [81].

925

926 For both selected and continuous migration simulations, we then performed the genotype
927 and read data simulation procedure, and reran our HMM as described above. We
928 performed 10 simulations for each treatment.
929
930 **Comparisons to LAMPanc**
931 We next sought to compare our method to a commonly used local ancestry inference
932 method, LAMPanc [23]. Towards this, we again simulated data using MACS and SELAM as
933 described above. For these comparisons, the initial ancestry contribution was 0.5 and the
934 number of generations since admixture varied between 5 and 1000. For comparison, we
935 supplied LAMPanc and our program the correct time since admixture and ancestry
936 proportions, as these are required parameters for LAMPanc. We also supplied the program
937 error free genotypes, another requirement of LAMPanc, whereas we supplied our HMM
938 with read data simulated as described above. We then ran LAMPanc under default
939 parameters, and we also reran LAMPanc using LD pruning within the reference panels, as
940 we do in our method, instead of the default LD pruning implemented in LAMPanc.
941
942 **Analysis of Inuit Genotype Data**
943 To demonstrate that LAI methods can be biased by the arbitrary selection of the time since
944 admixture, we analyzed a dataset of SNP-array genotype data from Greenlandic Inuits.
945 These data are described in detail elsewhere [51,52]. This population has received some
946 admixture from a European source population, and the authors had previously used RFMix
947 [24] to perform LAI, and found some sensitivity to the assumed time since admixture (J.
948 Crawford *pers. Comm.*). We analyzed data from chromosome 10 using RFMix v1.5.4 [24] as

949 described in Moltke *et al.* [52] assuming admixture occurred either 5 or 20 generations ago.
950 We subsequently analyzed chromosome 10 using our HMM including the genotype-
951 analysis emissions probabilities and assuming a genotype error rate of 0.2%. For our
952 analysis we identified the LD cutoff that is appropriate for these data as described above.
953

954 **Generating *D. melanogaster* Reference Populations**

955 To generate reference panels, we used a subset of the high quality *D. melenaogaster*
956 assemblies that have been described previously in Pool *et al.* (2012) and Lack *et al.* (2015).
957 As in the local ancestry analysis of Pool (2015), we used the French population. For our
958 African reference panel, we selected a subset of the Eastern and Western African
959 populations (CO, RG, RC, NG, UG, GA, GU) and grouped them into a single population for the
960 purposes of our analysis. We elected to combine populations so that we would have a
961 larger reference panel of African populations for this analysis, this solution may be justified
962 because these *D. melanogaster* populations are only weakly genetically differentiated
963 [2,21,82], particularly after common inversion-bearing chromosomes are removed from
964 analyses. Specific individuals were selected for inclusion in the African reference panel if
965 previous work found they have relatively little cosmopolitan ancestry (*i.e.* below 0.2
966 genome-wide in [2]).
967

968 Because of their powerful effects on recombination, chromosomal inversions are known to
969 have substantial impacts on the distribution of genetic variants on chromosomes
970 containing chromosomal inversions in *D. melanogaster* [2,54]. For this reason, we removed
971 all common inversion-bearing chromosome arms from the reference populations [83].

972 Nonetheless, it is clear that chromosomal inversions are present in the pool-seq samples
973 [55]. Although the inversions certainly violate key assumptions of our model—particularly
974 the transmission probabilities—given that our approach is robust to a many perturbations,
975 we expect the LA within inverted haplotypes can be estimated with reasonable confidence,
976 and the overall LAI procedure will still perform adequately with low frequencies of
977 chromosomal-inversion bearing chromosomes present within these samples.
978
979 Although these reference populations are believed to have relatively little admixture, some
980 admixture is likely to remain within these samples [2]. To mitigate this potential issue, we
981 first applied our HMM to each reference population using the genotype-based emissions
982 probabilities (above). Calculated across all individuals, we found that our maximum
983 likelihood ancestry estimates were identical with those of Pool *et al.* (2012) at 96.2% of
984 markers considered in our analysis. The differences between the results of these methods
985 may reflect differences in the methodology of LAI or differences in the reference panels.
986 Nonetheless, the broad concordance suggests the two methods are yielding similar overall
987 results. We masked all sites where the posterior probability of non-native ancestry was
988 greater than 0.5 within each reference individual's genome. These masked sequences were
989 then used as the reference panel for the analyses of poolseq data below.
990
991 **Ancestry Cline Sequence Data Analysis**
992 We acquired pooled sequencing data from six populations from the east coast of the United
993 States. The generation of these samples, sequencing data, and accession numbers are
994 described in detail in [21,36]. Briefly, the samples are comprised of individuals drawn from

995 natural populations and sequenced in relatively large pools of 66-232 chromosomes. We
996 aligned all data using BWA v0.7.9a-r786 [84] using the 'MEM' function and the default
997 program parameters. For all alignments, we used version 5 of the *D. melanogaster*
998 reference genome [85] in order to make our analysis and coordinates compatible with the
999 *Drosophila* genome nexus [83]. We then realigned all reads using the indelrealigner tool
1000 within the GATK package [72], and we extracted the sequence pileup using samtools
1001 mpileup v1.1 [86] using the program's default parameters.
1002
1003 We extracted sites at ancestry informative positions within the reference panels, where we
1004 required that the reference panel have a minimum of 50% of individuals with a high quality
1005 genotype call in both Cosmopolitan and African reference populations. As above, ancestry
1006 informative sites were defined as those with a minimum of 20% difference in allele
1007 frequencies between the reference panels used, and we retained only ancestry informative
1008 sites for our analyses. We then produced global ancestry estimates for each chromosome
1009 arm separately for each sample using the method of Bergland *et al.* (2016). We ran our
1010 HMM for each chromosome arm and each population, and we provided the program this
1011 estimate of the ancestry proportion and the time since admixture, 1593 generations [17].
1012 We elected to provide the time since admixture because we have found that this parameter
1013 is difficult to estimate in relatively large pools (see Results). However, the program can
1014 accurately estimate LA in high ploidy samples even when the time since admixture cannot
1015 be estimated correctly (see Results).
1016
1017 **Correlation with Local Recombination Rates**

1018 To assess the correlation between local recombination rates and clinality of LA clines in the
1019 genome, we employed a regression approach. First, we computed the mean partial
1020 correlation between latitude and local ancestry in windows of 100 ancestry informative
1021 markers. We then annotated the recombination rate in the midpoint of that genomic
1022 window, where as above, we used the recombination rate estimates of [87]. We computed
1023 Spearman's correlation between local recombination rates and the mean partial correlation
1024 between LA and latitude for the whole genome and for each chromosome arm
1025 independently. We also estimated confidence intervals using 1000 block-bootstrap samples
1026 using window sizes of 100 SNPs.
1027
1028 **Identifying LA Cline Outliers**
1029 To detect loci that show evidence for steeper ancestry clines than the genomic average, we
1030 first computed the Spearman's rank correlation between mean ancestry proportions and
1031 latitude for each chromosome arm separately. Then, for each site for which we obtained a
1032 posterior ancestry distribution for all samples, we computed the partial Spearman's rank
1033 correlation between the posterior ancestry mean and latitude while correcting for the
1034 correlation between latitude and the overall ancestry proportion. We then computed the
1035 probability of obtaining the observed partial correlation in R, which implements the
1036 approach of [88], and we retained those sites where the probability of the partial
1037 correlation between local ancestry and latitude was less than 0.005 as significant in our
1038 analysis. Although this cutoff is arbitrary, given the strong evidence for local adaptation
1039 and reproductive isolation in these populations [42,43,89], the tail of the LA cline
1040 distribution will likely be enriched for sites experiencing selection on this ancestry

1041 gradient. Due to linkage, adjacent sites show strong autocorrelation. We therefore selected
1042 the local optima for a given clinally significant LA segment (*i.e.* a tract where all positions
1043 are significantly correlated with latitude at our threshold) and retained these for analyses
1044 of outlier loci. Finally, to further reduce the effect of autocorrelation, we retained only
1045 those local optima for which no other optimum had a stronger correlation with latitude
1046 within 100,000bp on either side on the site.
1047

1048 **Gene Ontology Analyses**

1049 We performed Gene-ontology (GO) analyses of the set of clinal outlier loci where the
1050 background set was all genes located on any euchromatic chromosome arms in the *D.*
1051 *melanogaster* genome. The foreground set was defined as the gene intersected by a LA
1052 outlier local optimum, or the nearest gene to a local optimum. All GO analyses were
1053 performed using GOrilla [90].
1054

1055 **Acknowledgements**

1056 We thank Tyler Linderoth, Shelbi Russell and members of the Nielsen and Slatkin labs for
1057 helpful comments. We also thank Jacob Crawford for providing Inuit polymorphism data
1058 and for providing the observation that LAI inference depended heavily on *t.*
1059

1060 **References**

1061 1. Kronforst MR, Young LG, Blume LM, Gilbert LE. Multilocus analyses of admixture and
1062 introgression among hybridizing *Heliconius* butterflies. Evolution. 2006;60: 1254–16.
1063 doi:10.1554/06-005.1

1064 2. Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, et al.
1065 Population genomics of sub-Saharan Drosophila melanogaster: African diversity and
1066 non-African admixture. PLoS Genet. 2012;8: e1003080–24.
1067 doi:10.1371/journal.pgen.1003080

1068 3. Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J.
1069 The genomic signature of crop-wild introgression in maize. PLoS Genet. 2013;9:
1070 e1003477–13. doi:10.1371/journal.pgen.1003477

1071 4. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and introgression
1072 between Mimulus nasutus and Mimulus guttatus. PLoS Genet. 2014;10: e1004410–
1073 15. doi:10.1371/journal.pgen.1004410

1074 5. Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, et al.
1075 Major ecological transitions in wild sunflowers facilitated by hybridization. Science.
1076 2003;301: 1211–1216. doi:10.1126/science.1086949

1077 6. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence
1078 of the neandertal genome. Science. 2010;328: 710–722.
1079 doi:10.1126/science.1188021

1080 7. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The
1081 genomic landscape of Neanderthal ancestry in present-day humans. Nature.
1082 2014;507: 354–357. doi:10.1038/nature12961

1083 8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal

1084 components analysis corrects for stratification in genome-wide association studies.
1085 Nat Genet. 2006;38: 904–909. doi:10.1038/ng1847

1086 9. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic
1087 association studies supports a contribution of common variants to susceptibility to
1088 common disease. Nat Genet. 2003;33: 177–182. doi:10.1038/ng1071

1089 10. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al.
1090 Assessing the impact of population stratification on genetic association studies. Nat
1091 Genet. 2004;36: 388–393. doi:10.1038/ng1333

1092 11. Caracristi G, Schlotterer C. Genetic differentiation between American and European
1093 Drosophila melanogaster populations could be attributed to admixture of African
1094 alleles. Mol Biol Evol. 2003;20: 792–799. doi:10.1093/molbev/msg091

1095 12. Kolbe JJ, Glor RE, Schettino LR, Lara AC, Larson A. Genetic variation increases during
1096 biological invasion by a Cuban lizard. Nature. 2004;431: 177–181.

1097 13. Consortium THG, Consortium G. Butterfly genome reveals promiscuous exchange of
1098 mimicry adaptations among species. Nature. 2012;487: 94–98.
1099 doi:10.1038/nature11041

1100 14. Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E. Evidence for archaic
1101 adaptive introgression in humans. Nat Rev Genet. 2015;16: 359–371.
1102 doi:10.1038/nrg3936

1103 15. Juric I, Aeschbacher S, Coop G. The strength of selection against Neanderthal

1104 introgression. bioRxiv. 2015 Oct pp. 1–24. doi:10.1101/030148

1105 16. Harris K, Nielsen R. The genetic cost of neanderthal introgression. Genetics. Genetics;
1106 2016;203: 881–891. doi:10.1534/genetics.116.186890

1107 17. Pool JE. The mosaic ancestry of the Drosophila Genetic Reference Panel and the D.
1108 melanogaster reference genome reveals a network of epistatic fitness interactions.
1109 Mol Biol Evol. 2015; msv194–16. doi:10.1093/molbev/msv194

1110 18. Schumer M, Cui R, Powell DL, Dresner R, Rosenthal GG, Andolfatto P. High-resolution
1111 mapping reveals hundreds of genetic incompatibilities in hybridizing fish species.
1112 eLife. 2014;3: 610–21. doi:10.7554/eLife.02535

1113 19. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using
1114 multilocus genotype data. Genetics. 2000;155: 945–959.

1115 20. Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture
1116 proportions from next generation sequencing data. Genetics. 2013;195: 693–702.
1117 doi:10.1534/genetics.113.154138/-/DC1

1118 21. Bergland AO, Tobler R, González J, Schmidt P, Petrov D. Secondary contact and local
1119 adaptation contribute to genome-wide patterns of clinal variation in Drosophila
1120 melanogaster. Mol Ecol. 2016;25: 1157–1174. doi:10.1111/mec.13455

1121 22. Falush D, Stephens M, Pritchard JK. Inference of population structure using
1122 multilocus genotype data: linked loci and correlated allele frequencies. Genetics.
1123 2003;164: 1567–1587.

1124 23. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in
1125 admixed populations. Am J Hum Genet. 2008;82: 290–303.
1126 doi:10.1016/j.ajhg.2007.09.022

1127 24. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling
1128 approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93:
1129 278–288. doi:10.1016/j.ajhg.2013.06.020

1130 25. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al. Fast and
1131 accurate inference of local ancestry in Latino populations. Bioinformatics. 2012;28:
1132 1359–1367. doi:10.1093/bioinformatics/bts144

1133 26. Pool JE, Nielsen R. Inference of historical changes in migration rate from the lengths
1134 of migrant tracts. Genetics. 2009;181: 711–719. doi:10.1534/genetics.108.098095

1135 27. Koopman W, Li Y, Coart E. Linked vs. unlinked markers: multilocus microsatellite
1136 haplotype-sharing as a tool to estimate gene flow and introgression. Mol Ecol.
1137 2007;16: 243–256. doi:10.1111/j.1365-294X.2006.03137.x

1138 28. Patterson N, Hattangadi N, Lane B. Methods for high-density admixture mapping of
1139 disease genes. Am J Hum Genet. 2004;74: 979–1000.

1140 29. Gravel S. Population Genetics Models of Local Ancestry. Genetics. 2012;191: 607–
1141 619. doi:10.1534/genetics.112.139808

1142 30. Consortium 1GP, BGI-Shenzhen, European Bioinformatics Institute, Illumina,
1143 Hospital BAW, College B, et al. An integrated map of genetic variation from 1,092

1144 human genomes. Nature. 2012. doi:10.1038/nature11632

1145 31. Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. Multilocus patterns of
1146 nucleotide variability and the demographic and selection history of Drosophila
1147 melanogaster populations. Genome Res. 2005;15: 790–799. doi:10.1101/gr.3541005

1148 32. Thornton K, Andolfatto P. Approximate Bayesian inference reveals evidence for a
1149 recent, severe bottleneck in a Netherlands population of Drosophila melanogaster.
1150 Genetics. 2006;172: 1607–1619. doi:10.1534/genetics.105.048223

1151 33. Li H, Stephan W. Inferring the demographic history and rate of adaptive substitution
1152 in Drosophila. PLoS Genet. 2006;2: e166. doi:10.1371/journal.pgen

1153 34. Duchen P, Živković D, Hutter S, Stephan W. Demographic inference reveals African
1154 and European admixture in the North American Drosophila melanogaster
1155 population. Genetics. 2013;193: 291–301. doi:10.1534/genetics.112.145912/-/DC1

1156 35. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, et al. Genomic
1157 variation in natural populations of Drosophila melanogaster. Genetics. 2012;192:
1158 533–598. doi:10.1534/genetics.112.142018

1159 36. Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. Genomic evidence of
1160 rapid and stable adaptive oscillations over seasonal time scales in Drosophila. PLoS
1161 Genet. 2014;10: e1004775–19. doi:10.1371/journal.pgen.1004775

1162 37. Schrider DR, Hahn MW, Begun DJ. Parallel evolution of copy-number variation across
1163 continents in Drosophila melanogaster. Mol Biol Evol. 2016;33: 1308–1316.

1164 doi:10.1093/molbev/msw014

1165 38. Reinhardt JA, Kolaczkowski B, Jones CD, Begun DJ, Kern AD. Parallel geographic
1166 variation in Drosophila melanogaster. Genetics. 2014;197: 361–373.
1167 doi:10.1534/genetics.114.161463/-/DC1

1168 39. Fabian DK, Kapun M, Nolte V, Kofler R, Schmidt PS, Schlotterer C, et al. Genome-wide
1169 patterns of latitudinal differentiation among populations of Drosophila
1170 melanogasterfrom North America. Mol Ecol. 2012;21: 4748–4769.
1171 doi:10.1111/j.1365-294X.2012.05731.x

1172 40. Oakeshott JG, Gibson JB, Anderson PR, Knibb WR. Alcohol dehydrogenase and
1173 glycerol-3-phosphate dehydrogenase clines in Drosophila melanogaster on different
1174 continents. Evolution. 1982;36: 86–96.

1175 41. Berry A, Kreitman M. Molecular analysis of an allozyme cline: alcohol dehydrogenase
1176 in Drosophila melanogaster on the east coast of North America. Genetics. 1993;134:
1177 869-893.

1178 42. Yukilevich R, True JR. African morphology, behavior, and phermones underlie
1179 incipient sexual isolation between US and Caribbean Drosophila melanogaster.
1180 Evolution. 2008;62: 2807–2828. doi:10.1111/j.1558-5646.2008.00488.x

1181 43. Lachance J, True JR. X-autosome incompatibilities in Drosophila melanogaster: test of
1182 Haldane's rue and geographic patterns within species. Evolution. 2010;64: 3035–
1183 3046. doi:10.1111/j.1558-5646.2010.01028.x

1184 44. Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF. Genetic incompatibilities are
1185 widespread within species. Nature. 2013;504: 135–137. doi:10.1038/nature12678

1186 45. Kofler R, Betancourt AJ, Schlotterer C. Sequencing of Pooled DNA Samples (Pool-Seq)
1187 Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila
1188 melanogaster. PLoS Genet. 2012;8: e1002487–16.
1189 doi:10.1371/journal.pgen.1002487

1190 46. terWengel PO, Kapun M, Nolte V, Kofler R, Flatt T, Schlotterer C. Adaptation of
1191 *Drosophila* to a novel laboratory environment reveals temporally heterogeneous
1192 trajectories of selected alleles. Mol Ecol. 2012;21: 4931–4941. doi:10.1111/j.1365-
1193 294X.2012.05673.x

1194 47. Kolaczkowski B, Kern AD, Holloway AK, Begun DJ. Genomic differentiation between
1195 temperate and tropical Australian populations of Drosophila melanogaster. Genetics.
1196 2011;187: 245–260. doi:10.1534/genetics.110.123059

1197 48. Kapun M, Schalkwyk H, McAllister B, Flatt T, Schlotterer C. Inference of chromosomal
1198 inversion dynamics from Pool-Seq data in natural and laboratory populations of
1199 Drosophila melanogaster. Mol Ecol. 2014;23: 1813–1827. doi:10.1111/mec.12594

1200 49. Zhu Y, Bergland AO, González J, Petrov DA. Empirical validation of pooled whole
1201 genome population re-sequencing in Drosophila melanogaster. PLoS ONE. 2012;7:
1202 e41901. doi:10.1371/journal.pone.0041901

1203 50. Liang M, Nielsen R. The lengths of admixture tracts. Genetics. 2014;197: 953–967.

1204 doi:10.1534/genetics.114.162362

1205 51. Moltke I, Grarup N, Jørgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, et al. A
1206 common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2
1207 diabetes. Nature. 2014;512: 190–193. doi:10.1038/nature13425

1208 52. Moltke I, Fumagalli M, Korneliussen TS. Uncovering the genetic history of the
1209 present-day Greenlandic population. Am J Hum Genet. 2015;96: 54–69.
1210 doi:10.1016/j.ajhg.2014.11.012

1211 53. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a
1212 custom genotyping array for genetic studies of metabolic, cardiovascular, and
1213 anthropometric traits. PLoS Genet. 2012;8: e1002793–12.
1214 doi:10.1371/journal.pgen.1002793

1215 54. Corbett-Detig RB, Hartl DL. Population genomics of inversion polymorphisms in
1216 Drosophila melanogaster. PLoS Genet. 2012;8: e1003056–15.
1217 doi:10.1371/journal.pgen.1003056

1218 55. Kapun M, Fabian DK, Goudet J, Flatt T. Genomic evidence for adaptive inversion
1219 clines in Drosophila melanogaster. Mol Biol Evol. 2016;33: 1317–1336.
1220 doi:10.1093/molbev/msw016

1221 56. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The
1222 Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482: 173–178.
1223 doi:10.1038/nature10811

1224 57. Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, et al. Natural
1225 variation in genome architecture among 205 Drosophila melanogaster Genetic
1226 Reference Panel lines. Genome Res. 2014;24: 1193–1208.
1227 doi:10.1101/gr.171546.113

1228 58. Corbett-Detig RB, Cardeno C, Langley CH. Sequence-based detection and breakpoint
1229 assembly of polymorphic inversions. Genetics. 2012;192: 131–137.
1230 doi:10.1534/genetics.112.141622

1231 59. Kulathinal RJ, Stevison LS, Noor MAF. The genomics of speciation in Drosophila:
1232 diversity, divergence, and introgression estimated using low-coverage genome
1233 sequencing. PLoS Genet. 2009;5: e1000550–7. doi:10.1371/journal.pgen.1000550

1234 60. Krimbas CB, Powell JR. Drosophila Inversion Polymorphism. CRC Press; 1992.

1235 61. Aulard S, David JR, Lemenieur F. Chromosomal inversion polymorphism in
1236 Afrotropical populations of *Drosophila melanogaster*. Genet Res. 2002;79: 49–63.
1237 doi:10.1017/S0016672301005407

1238 62. Ting CT, Takahashi A, Wu CI. Incipient speciation by sexual isolation in Drosophila:
1239 concurrent evolution at multiple loci. PNAS. 2001;98: 6709–6713.

1240 63. Hollocher H, Ting CT, Wu ML, Wu CI. Incipient speciation by sexual isolation in
1241 Drosophila melanogaster: extensive genetic divergence without reinforcement.
1242 Genetics. 1997;147: 1191–1201.

1243 64. Hollocher H, Ting CT, Pollack F, Wu CI. Incipient speciation by sexual isolation in

1244 Drosophila melanogaster: variation in mating preference and correlation between
1245 sexes. Evolution. 1997;51: 1175–1181.

1246 65. Coyne JA. Genetics and speciation. Nature. 1992;335: 511–515.

1247 66. Coyne JA, Orr HA. Patterns of speciation in Drosophila. Evolution. 1989;43: 362–381.

1248 67. Charlesworth B, Coyne JA, Barton NH. The relative rates of evolution of sex
1249 chromosomes and autosomes. American Naturalist. 1987;130: 113–146.

1250 68. Ellis LL, Carney GE. Socially-Responsive Gene Expression in Male Drosophila
1251 melanogaster Is Influenced by the Sex of the Interacting Partner. Genetics. 2011;187:
1252 157–169. doi:10.1534/genetics.110.122754

1253 69. Edwards AC, Zwarts L, Yamamoto A, Callaerts P, Mackay TF. Mutations in many genes
1254 affect aggressive behavior in Drosophila melanogaster. BMC Biol. 2009;7: 29–13.
1255 doi:10.1186/1741-7007-7-29

1256 70. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. SNP calling, genotype calling,
1257 and sample allele frequency estimation from new-generation sequencing data. PLoS
1258 ONE. 2012;7: e37558–11. doi:10.1371/journal.pone.0037558

1259 71. Fumagalli M, Vieira FG, Linderoth T, Nielsen R. ngsTools: methods for population
1260 genetics analyses from next-generation sequencing data. Bioinformatics. 2014;30:
1261 1486–1487.

1262 72. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework
1263 for variation discovery and genotyping using next-generation DNA sequencing data.

1264 Nat Genet. 2011;43: 491–498. doi:10.1038/ng.806

1265 73. Baird SJE, Barton NH, Etheridge AM. The distribution of surviving blocks of an
1266 ancestral genome. Theoretical Population Biology. 2003;64: 451–471.
1267 doi:10.1016/S0040-5809(03)00098-4

1268 74. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic ancestry
1269 of north Africans supports back-to-Africa migrations. PLoS Genet. 2012;8:
1270 e1002397–11. doi:10.1371/journal.pgen.1002397

1271 75. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The history
1272 of African gene flow into southern Europeans, Levantines, and Jews. PLoS Genet.
1273 2011;7: e1001373–13. doi:10.1371/journal.pgen.1001373

1274 76. Loh PR, Lipson M, Patterson N, Moorjani P. Inferring admixture histories of human
1275 populations using linkage disequilibrium. Genetics. 2013;193: 1233–1254.
1276 doi:10.1534/genetics.112.147330/-/DC1

1277 77. Marjoram P, Wall JD. Fast "coalecsent" simulation. BMC Genet. BioMed Central;
1278 2006;7: 16–9. doi:10.1186/1471-2156-7-16

1279 78. Kiefer J. Sequential minimax search for a maximum. Proceedings of the American
1280 Mathematical Society. 1953;4: 502–506.

1281 79. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data.
1282 Genome Res. 2008;19: 136–142. doi:10.1101/gr.083634.108

1283 80. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint

1284 demographic history of multiple populations from multidimensional SNP frequency
1285 data. PLoS Genet. 2009;5: e1000695–11. doi:10.1371/journal.pgen.1000695

1286 81. Corbett-Detig R, Jones M. SELAM: Simulation of Epistasis and Local adaptation during
1287 Admixture with Mate choice. Bioinformatics. 2016;btw365.

1288 82. Pool JE, Aquadro CF. History and Structure of Sub-Saharan Populations of Drosophila
1289 melanogaster. Genetics. 2006;174: 915–929. doi:10.1534/genetics.106.058693

1290 83. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, et al. The
1291 Drosophila Genome Nexus: a population genomic resource of 623 Drosophila
1292 melanogaster genomes, including 197 from a single ancestral range population.
1293 Genetics. 2015;199: 1229–1241. doi:10.1534/genetics.115.174664/-/DC1

1294 84. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
1295 MEM. arXiv preprint arXiv:13033997. 2013.

1296 85. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG. The genome
1297 sequence of Drosophila melanogaster. Genetics. 2000;287: 2185–2195.

1298 86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
1299 Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
1300 doi:10.1093/bioinformatics/btp352

1301 87. Comeron JM, Ratnappan R, Bailin S. The many landscapes of recombination in
1302 Drosophila melanogaster. PLoS Genet. 2012;8: e1002905.
1303 doi:10.1371/journal.pgen.1002905

1304 88. Best DJ, Roberts DE. Algorithm AS 89: the upper tail probabilities of Spearman's rho.
1305 Journal of the Royal Statistical Society Series C. 1975;24: 377–379.

1306 89. Kao JY, Lymer S, Hwang SH, Sung A, Nuzhdin SV. Postmating reproductive barriers
1307 contribute to the incipient sexual isolation of the United States and Caribbean
1308 Drosophila melanogaster. Ecol Evol. 2015;5: 3171–3182. doi:10.1002/ece3.1596

1309 90. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and
1310 visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 2013
1311 14:1. BioMed Central; 2009;10: 48–7. doi:10.1186/1471-2105-10-48

1312

1313

Figure 1. The effect of increasing stringency with ancestral LD pruning. From left to right, ancestry proportions are 0.1, 0.25, 0.5, 0.75 and 0.9. |r| cutoffs are: none (red), 1.0 (orange), 0.9 (yellow), 0.8 (green), 0.7 (dark blue), 0.6 (cyan), 0.5 (indigo), and 0.4 (violet). The solid line indicates the expectation for unbiased time estimation. All read data were simulated with ploidy = 1. True admixture time was drawn from a uniform (0, 2000) distribution.

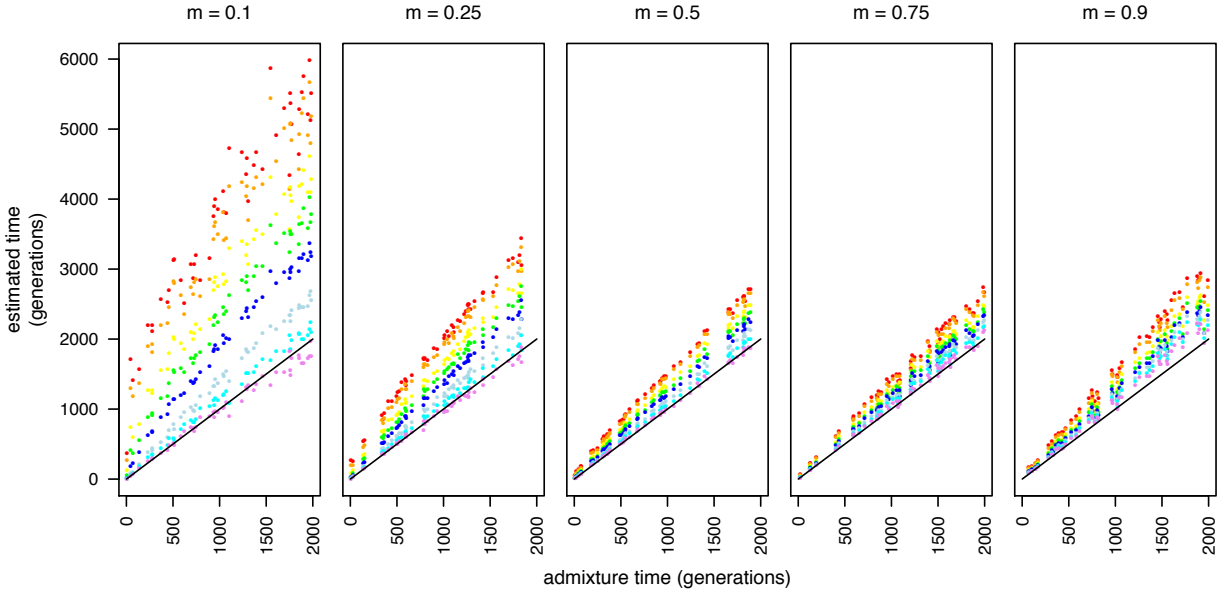
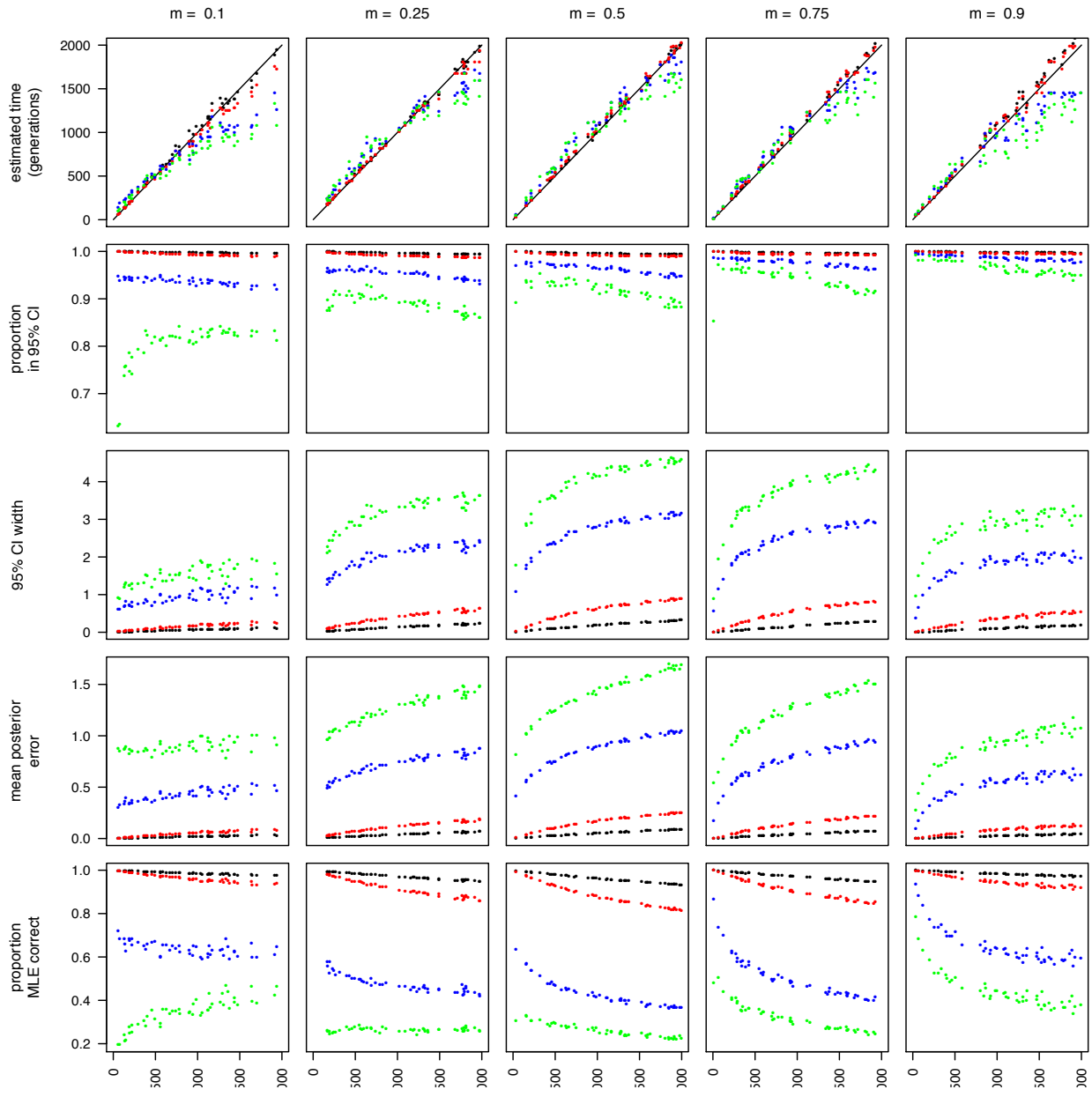
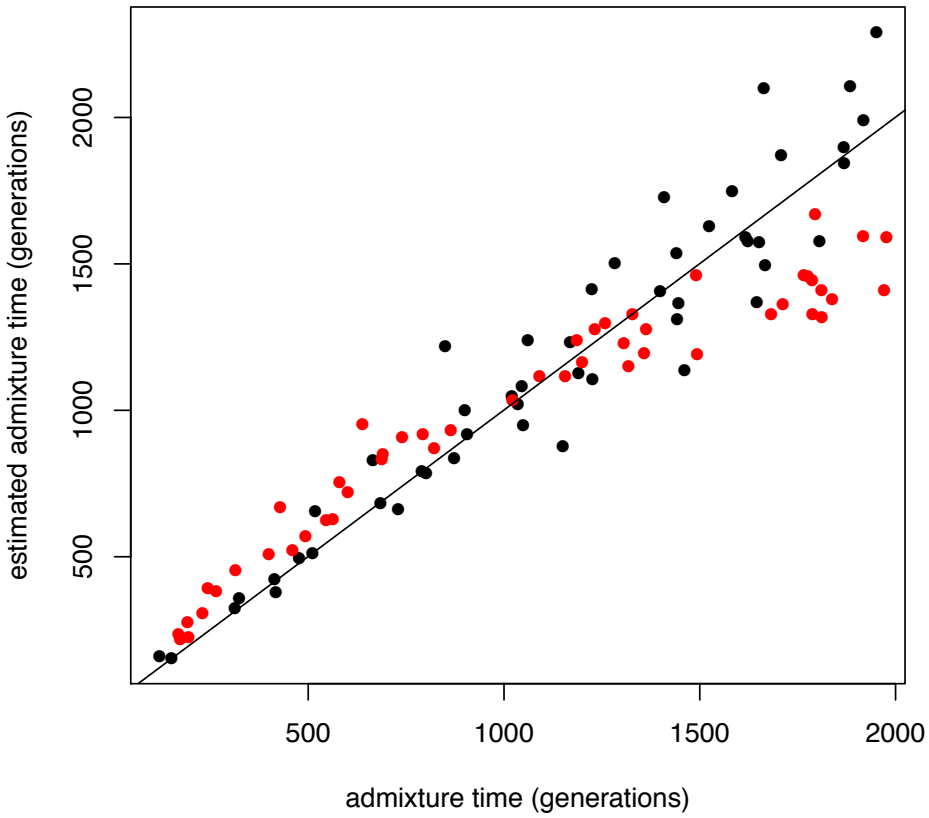


Figure 2. Time estimates and accuracy statistics for samples of varying ploidies. From left to right, ancestry proportions are 0.1, 0.25, 0.5, 0.75 and 0.9. Each sample ploidy is represented by one point color with ploidy one (black), two (red), ten (blue) and twenty (green). From top to bottom, each row is the estimated time in generations, the proportion of sites where the true state is within the 95% credible interval, the width of the 95% credible interval, the mean posterior error, and the proportion of sites where the maximum likelihood estimate is equal to the true state.



1335 **Figure 3.** Comparison between LAI using the full ancestral recombination graph via
1336 forward-time simulations (red) with those from independent and identically distributed
1337 draws from the SMC' distribution (black). Simulations were conducted using an ancestry
1338 proportion of 0.25 and population size of 10,000 hermaphroditic individuals.



1339
1340

Figure 4. Effects of unknown admixed population sizes on LAI. All LAI was conducted assuming the true population size was 10,000. Simulated population sizes were 100 (black), 1,000 (red), 10,000 (blue) and 100,000 (green). Ploidy 1 on the right, ploidy 2 on the left. From top to bottom, rows are the estimated time of admixture, the proportion of sites where the true state is within the 95% credible interval, the width of the 95% credible interval, the mean posterior error, and the proportion of times that the maximum likelihood estimate is equal to the true state. For all simulations, the ancestry proportion was equal to 0.5.

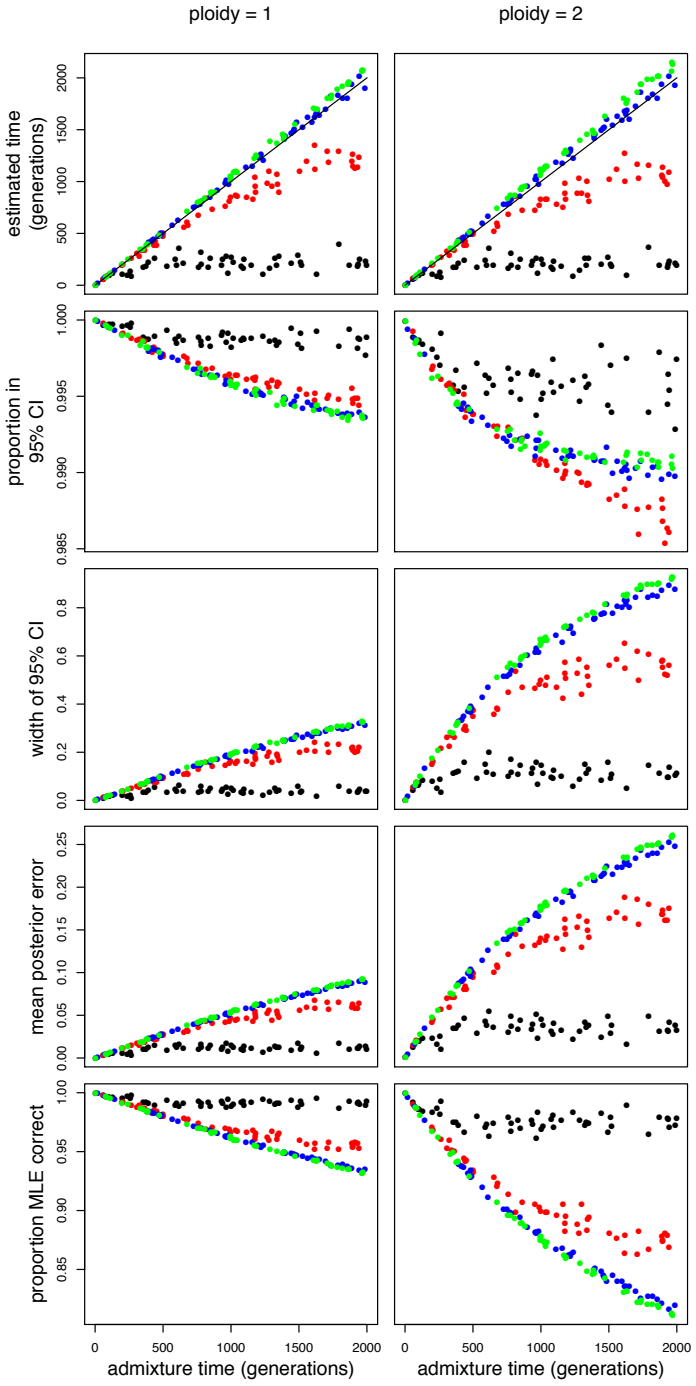


Figure 5. LAI accuracy when admixture times are increasingly ancient. Here, ancestry proportions are 0.5 (black), 0.25 (blue), 0.1 (violet), 0.75 (orange) and 0.9 (red). From top to bottom, statistics plotted are estimated time, the proportion of sites where the true ancestry frequency is within the 95% credible interval, the mean 95% credible interval width, mean posterior error, and the proportion of times that the maximum likelihood estimate is correct.

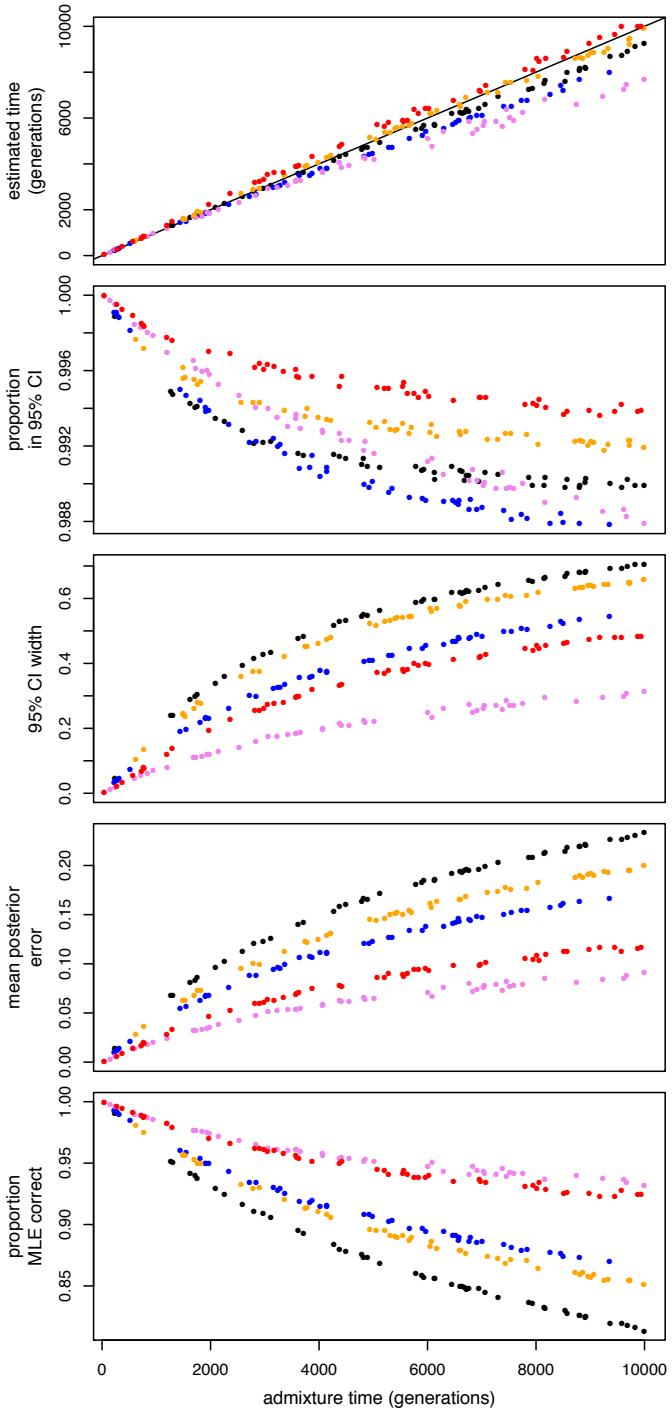


Figure 6. The effects of reference panel size on LAI and time estimation using the HMM. Here, we compare reference panels of size 100 (blue) with reference panels of size 10 (black). From left to right, ancestry proportions are 0.1, 0.25, 0.5, 0.75 and 0.9. From top to bottom the plotted statistics are estimated time, proportion in the 95% credible interval, the average width of the 95% credible interval, the mean posterior error, and the proportion of sites where the maximum likelihood ancestry estimate is correct.

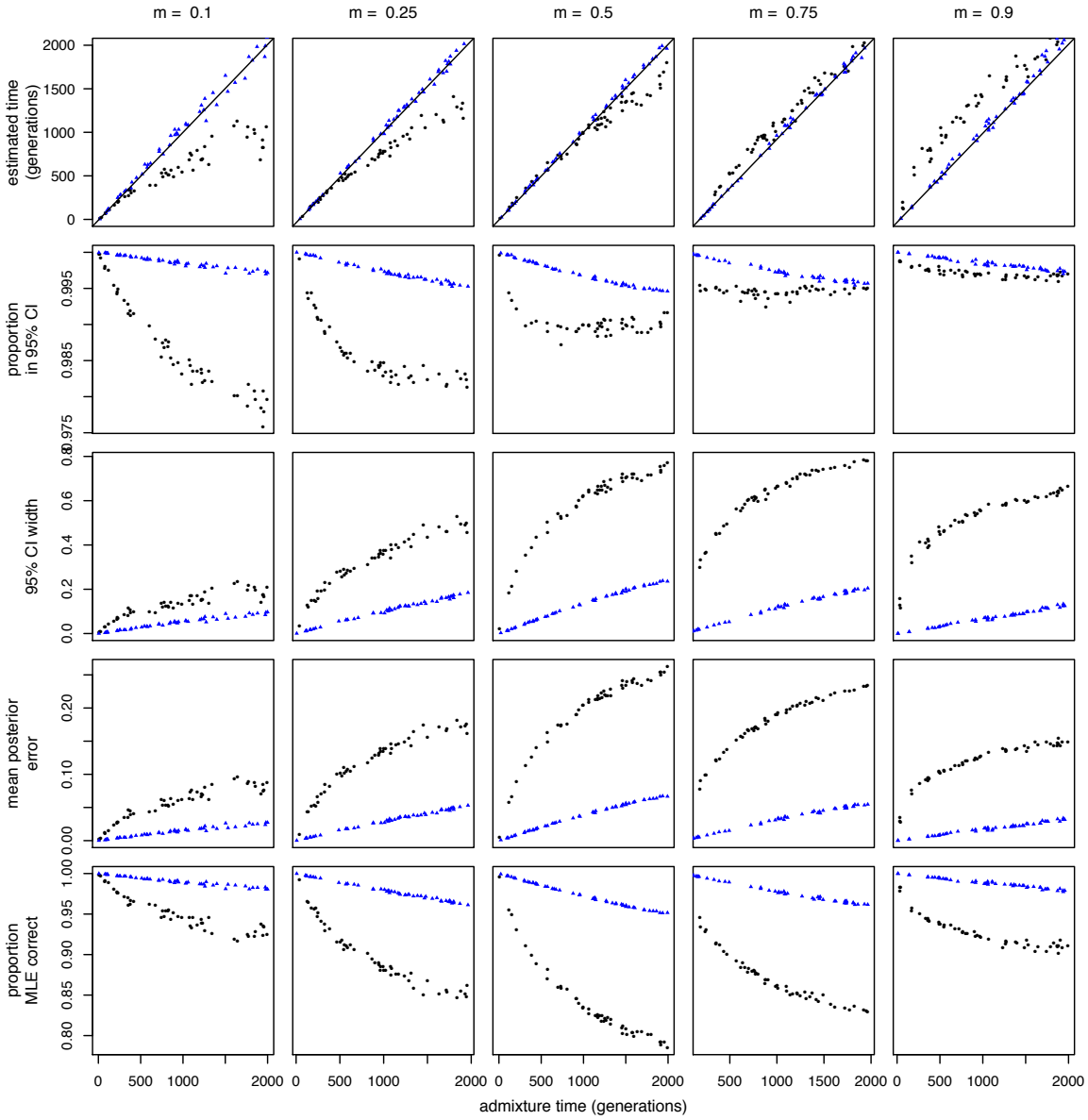
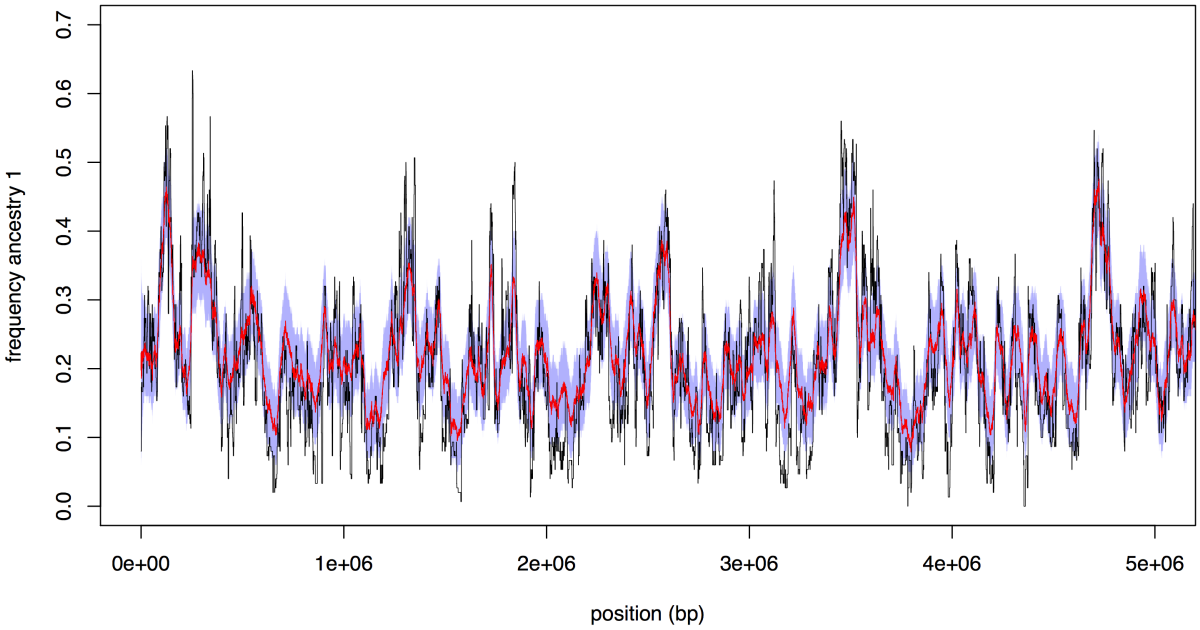
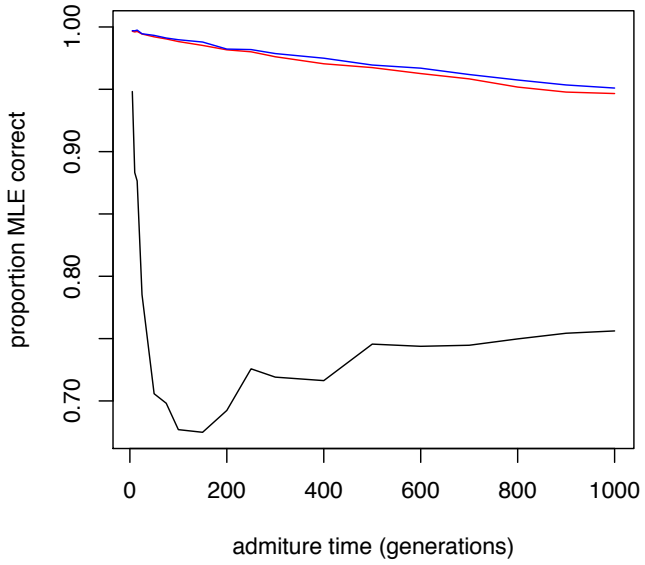


Figure 7. Accuracy of the HMM for samples of high ploidy. The 95% credible interval (shaded blue region), and the posterior mean (red) contrasted with the true ancestry frequencies (black). Simulated data were generated with an admixture time of 1500 generations, an ancestry proportion of 0.2, a sample ploidy of 100, and a mean sequencing depth of 25.

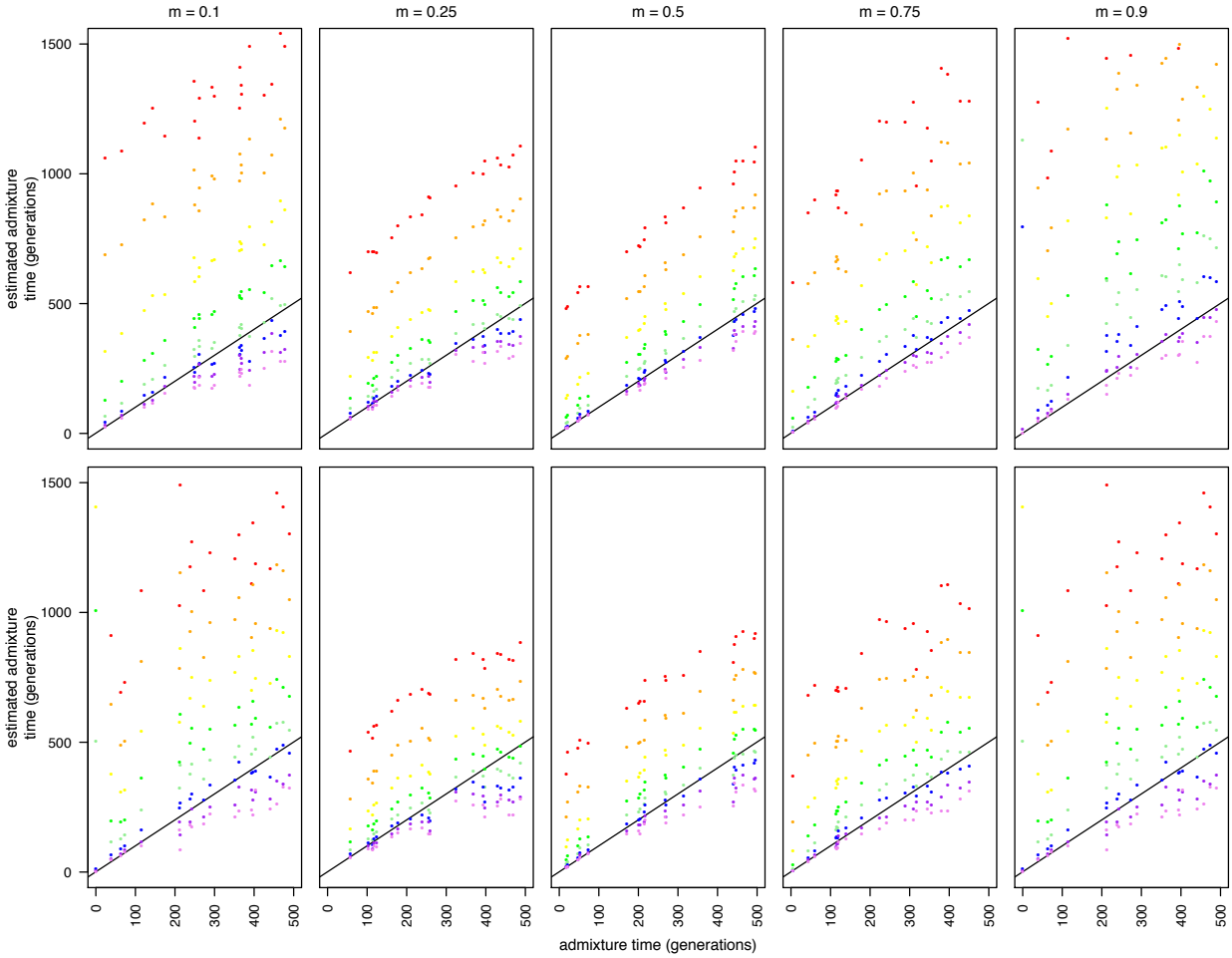


1373 **Figure 8.** Comparison of the proportion of sites where the maximum likelihood ancestry
1374 estimate of local ancestry is correct between LAMPanc and our method. LAMPanc was run
1375 with default parameters (black), and with LD pruned in the ancestral populations, but not
1376 in the admixed population (red). Our method was run with default parameters (blue), but
1377 with the time since admixture and correct ancestry proportion supplied to our program as
1378 these parameters are required by LAMPanc.
1379

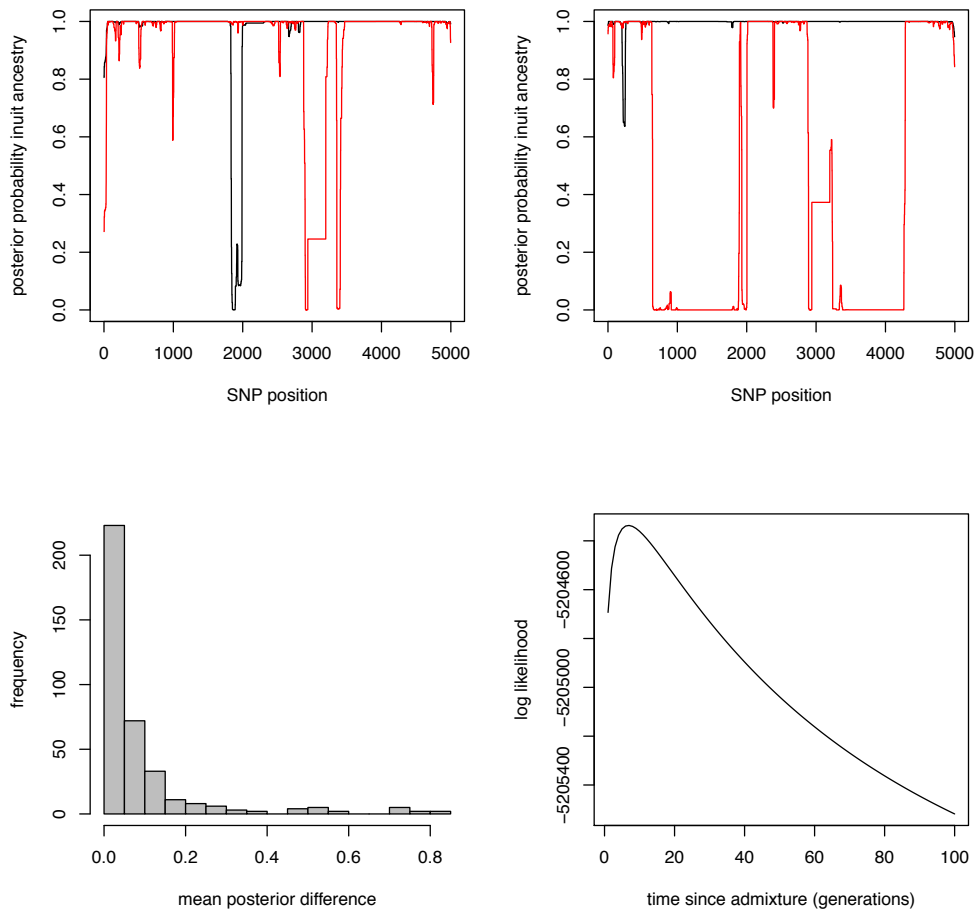


1380
1381
1382

Figure 9. Admixture time estimates for simulated data consistent with variation present in modern European and African populations. From left to right, m = 0.1, m = 0.25, m = 0.5, m = 0.75, m = 0.9. Top row is completely phased chromosomes and the bottom row is for unphased diploid data.



1389 **Figure 10.** Bias in LAI due to uncertainty in *t.* The posterior probability of European
1390 ancestry at a given site in the genome assuming t = 5 (black) and assuming t = 20 (red) for
1391 a sample representative of the average difference (top left) and a more extreme example
1392 (top right). The distribution of differences in mean Inuit ancestry for all samples (bottom
1393 left). The log likelihood of each time since admixture as computed using our method
1394 (bottom right), which shows a clear optimum at 6-7 generations since admixture. All
1395 analyses were restricted to SNPs on chromosome 10.



1396

Figure 11. Local ancestry of inversion bearing chromosomes (red) compared with those of standard arrangement chromosomes (black) for the same chromosome arm. Positions of inversion breakpoints, as reported in {CorbettDetig:2012bq} are shown as vertical dashed lines.

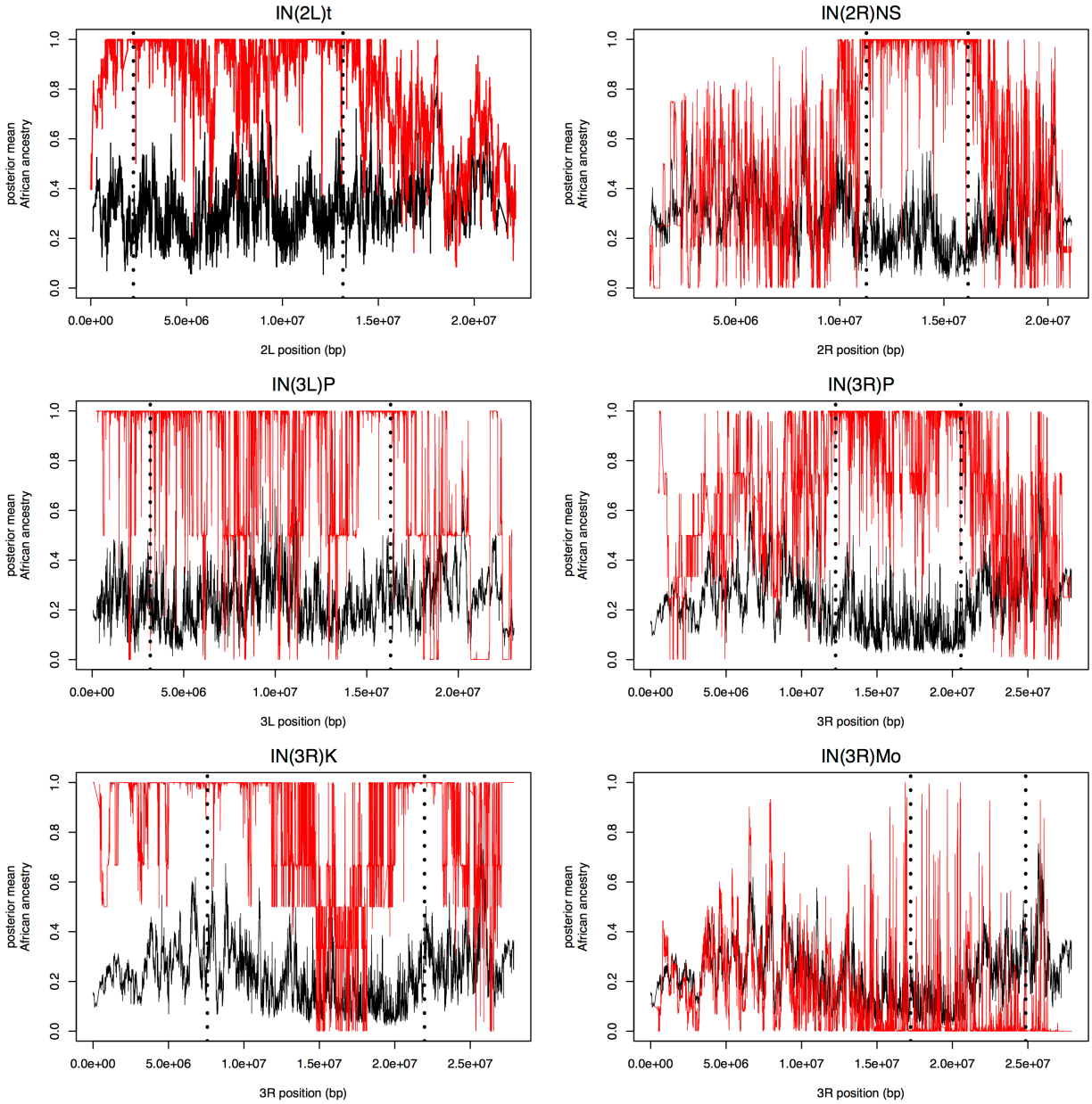
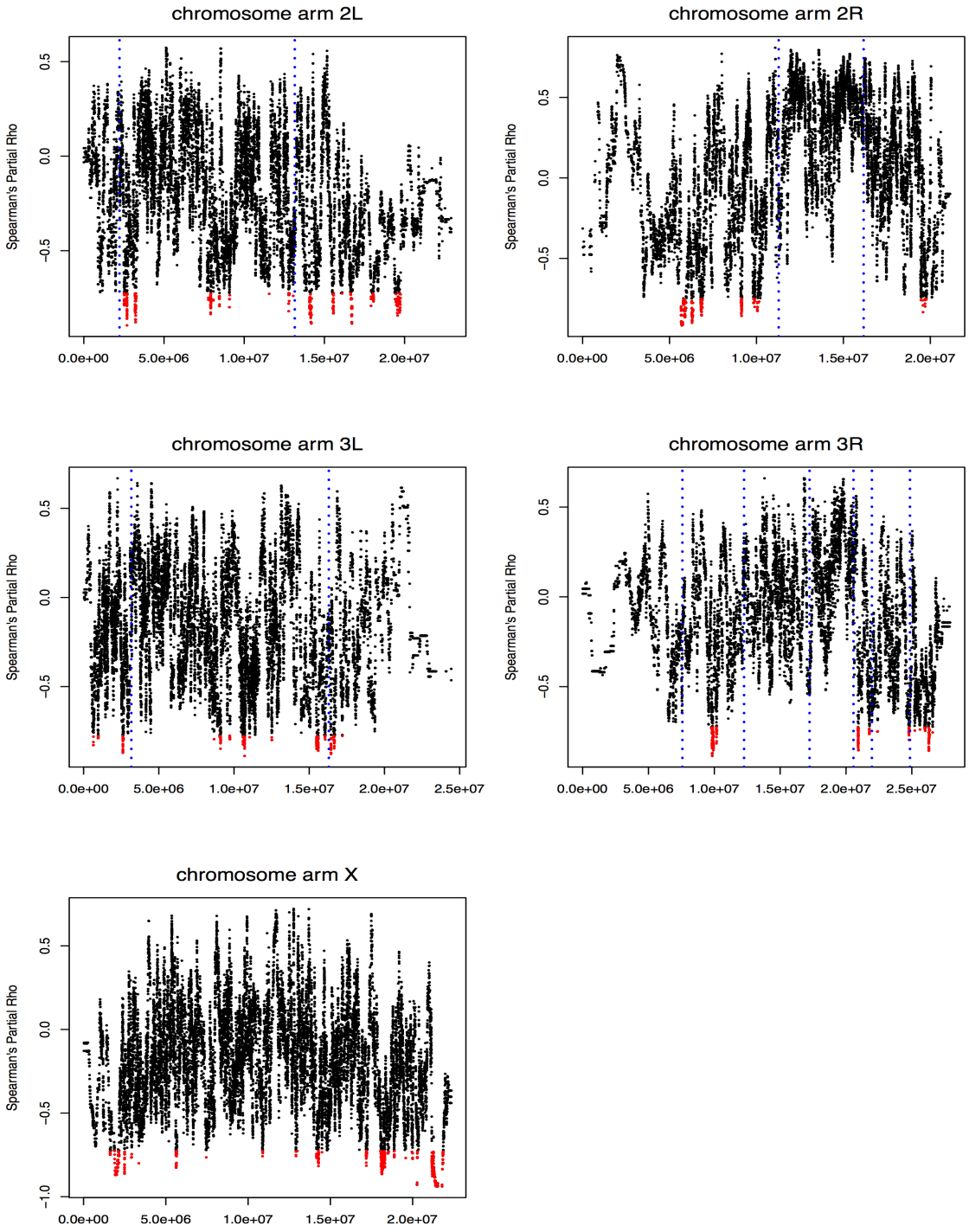


Figure 12. The partial correlation between LA and latitude with correction for chromosome-wide ancestry proportions. Sites for which the probability of the observed clinal relationship was less than 0.005 were retained as significant (red). Inversion breakpoints for inversions that are at polymorphic frequencies on this ancestry cline are shown as dotted blue lines.



1412