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 genotypes, which is not feasible to obtain for many next-generation 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. As this method is very general, we expect it will be useful for local ancestry inference in a wider variety of populations than what previously has been possible. We then applied 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 local recombination rates are negatively correlated with the proportion of African ancestry, suggesting that selection against foreign ancestry is the least efficient in low recombination regions. Finally we show that clinal outlier loci are enriched for genes associated with gene regulatory functions, consistent with a role of regulatory evolution in ecological adaptation of admixed *D. melanogaster*

populations. Our results illustrate the potential of local ancestry inference for elucidating fundamental evolutionary processes.

Author Summary

When divergent populations hybridize, their offspring obtain portions of their genomes 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 one predominant 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.

68 # Introduction

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

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

113 demographic histories are less well characterized, and assumptions about admixture times
114 may bias the result of LAI methods.
115
116 A number of approaches exist to estimate the time since admixture based on well
117 characterized ancestry tract length distributions [26-29] but in general, these parameters
118 are unknown prior to LAI. Conversely, another class of methods can be used to estimate the
119 time of admixture based on the decay of linkage disequilibrium without performing LAI
120 [30-32]; however as with LAI procedure above, these approaches are also limited to diploid
121 genotype data. We may therefore expect to improve LAI by simultaneously estimating LA
122 and demographic parameters (*e.g.* admixture time). Furthermore, in the majority of
123 sequencing applications, relatively low individual sequencing coverage is often used to
124 probabilistically estimate individual and population allele frequencies (*e.g.* [33]) but these
125 data are often not sufficient to determine high confidence genotypes that are required for
126 existing LAI applications. Hence, there is a clear need for a general LAI method that can
127 accommodate genotype uncertainty and requires less advanced knowledge of admixed
128 populations' demographic histories.
129
130 Here, we introduce a framework for simultaneously estimating LA using short read pileup
131 data and the time of admixture within a population. Briefly, as with many previously
132 proposed LAI methods, we model ancestry across the genome of a sample as a HMM. We
133 estimate LA by explicitly modeling read counts as a function of sample allele frequencies
134 within an admixed population. Our method is generalized to accommodate arbitrary
135 sample ploidies, and is therefore applicable to haploid (or inbred), diploid, tetraploid, as

136 well as pooled sequencing applications. We show that this approach accurately infers the
137 time since admixture when data are simulated under the assumed model. Furthermore, our
138 method yields accurate LA estimates for simulated datasets, including samples of high
139 sample ploidy and including evolutionary scenarios that violate the assumptions of the
140 neutral demographic model. In comparisons between ours and an existing LAI method,
141 WINPOP [34], we find that our approach offers a significant improvement and is accurate
142 over longer time scales. Furthermore, we demonstrate, using a published dataset, that even
143 state-of-the-art LAI methods can be significantly impacted by assumptions about the time
144 since admixture, and that our method provides a solution to this problem.
145
146 Finally, we apply this method to a *Drosophila melanogaster* ancestry cline on the east coast
147 of North America. This species originated in sub-Saharan Africa, and approximately
148 10,000—15,000 years ago a subpopulation expanded out of the ancestral range. During
149 this expansion, the derived subpopulation experienced a population bottleneck that
150 resulted in decreased nucleotide polymorphism, extended linkage disequilibrium within
151 the derived population and substantial genetic differentiation between ancestral and
152 derived populations [2,35-39]. Hereafter, the ancestral population will be referred to as
153 "African" and the derived population as "Cosmopolitan". Following this bottleneck,
154 descendant populations of African and Cosmopolitan *D. melanogaster* have admixed in
155 numerous geographic regions [2,11,21]. Of particular relevance to this work, North
156 America was colonized recently by a population descendent from African individuals from
157 the South, and by a population descendent from cosmopolitan *D. melanogaster* in the North
158 [11,21,38]. Where these populations encountered each other in eastern North America,

159 they form an ancestry cline where southern populations have a greater contribution of
160 African ancestry than northern populations [21].
161
162 Previous work on these ancestry clines has shown that ancestry proportions vary across
163 populations with increasing proportions of cosmopolitan alleles in more temperate
164 localities. Evidence suggests spatially varying selection affects the distribution of genetic
165 variants [40-45]. Furthermore, strong epistatic reproductive isolation barriers partially
166 isolate individuals from northern and southern populations along this ancestry cline
167 [46,47]. This may be generally consistent with recent observations of ancestry-associated
168 epistatic fitness interactions within a *D. melanogaster* population in North Carolina [17],
169 and with the observation of widespread fitness epistasis between populations of this
170 species more generally [48]. There is therefore good reason to believe that natural
171 selection has acted to shape LA clines that are tightly linked to selected mutations in these
172 *D. melanogaster* populations.
173
174 Here, we show that African ancestry in North American *D. melanogaster* populations is
175 negatively correlated with recombination rates, consistent with more efficient selection
176 against foreign ancestry in high recombination rate regions of the genome. We also find
177 that the X chromosome displays a higher rate of LA outlier loci, potentially consistent with
178 a greater role of the X chromosome in clinal adaptation. Clinal loci are disproportionately
179 likely to be associated with high level gene regulatory protein complexes, and may play
180 important roles in ecological divergence between African and Cosmopolitan *D.*
181 *melanogaster* populations. Furthermore, we identify numerous loci with decreased African

182 ancestry across all populations, which suggests that these alleles that are disfavored on
183 predominantly cosmopolitan genetic backgrounds. This subset of loci is enriched for genes
184 related to oogenesis, potentially consistent with epistatic interactions that affect female
185 reproductive success in these populations.
186

187 # Results and Discussion

188

189 **The Model**

190 Although admixed populations often are diploid, we derived a general model of ploidy in
191 which the individual has n gene copies at each locus, i.e. for diploid species $n = 2$. In
192 practice, sequences are often obtained from fully or partially inbred individuals (e.g.
193 [39,49]), which represent only a single uniquely derived chromosome. It is also common to
194 pool individuals prior to sequencing for allele frequency estimation, so called pool-seq (e.g.
195 [21,40,42,50-53]). If the pooling fractions are exactly equal, such a sample of b diploid
196 individuals can be treated as a sample from a single individual with ploidy $n = 2b$. Although
197 that requirement is restrictive, pool-seq has been experimentally validated as a method for
198 accurate allele frequency estimation—*i.e.* alleles are approximately binomially sampled
199 from the sample allele frequencies [54]. We therefore aimed to derive a model that can
200 accommodate arbitrary sample ploidies. In the model, we assumed that the focal
201 population was founded following a single discrete admixture event between two ancestral
202 subpopulations, labeled 0 and 1, with admixture proportions $1-m$ and m, respectively, at a
203 time t generations in the past. We modeled emission probabilities such that the method
204 can work directly on read pileup data, rather than high quality known genotypes. Briefly,

205 in our model, we specify an HMM $\{H_v\}$ with state space $S = \{0,1,\ldots,n\}$, where $H_v = i$, $i \in S$,
206 indicates that in the vth position i chromosomes are from population 0 and $n - i$
207 chromosomes are from population 1. In other words, this HMM enables one to estimate
208 what ancestry frequencies are present at a given site along a chromosome within a sample.
209 Importantly, we designed this method to simultaneously estimate the time of admixture,
210 which is related to the correlation between ancestry informative markers along a
211 chromosome. See Methods for a complete description of the HMM including the emissions
212 and transition probability calculations. The source code and manual are available at
213 https://github.com/russcd/Ancestry_HMM. For this model, it is assumed that the number
214 of chromosomes present in a sample, n, is known and that the global ancestry proportion,
215 m, is known. As there are many methods for accurately estimating m in a wide variety of
216 contexts implemented in standard population genetic analysis pipelines [19,20], we believe
217 this assumption is not too restrictive.
218

219 **Admixture Simulation Framework**

220 In order to test our method with data of known provenance, we also developed an
221 approach for simulating chromosomes sampled from admixed populations. Briefly, we first
222 simulated genetic diversity consistent with ancestral populations using a coalescent
223 simulation method [55]. We then generated ancestry tracts consistent with admixture
224 models developed to test our inference method using the forward-time admixture
225 simulation program, SELAM [56]. We retained a portion of each coalescent-generated
226 population to serve as a reference panel for allele frequency and LD estimation. We then
227 took the remaining chromosomes and placed them on the appropriate ancestry tracts in

228 admixed chromosomes. Finally, we generated read counts for these chromosomes, or pools
229 of chromosomes for samples with ploidy greater than one, via binomial sampling from the
230 genotype frequencies of the sample. Implicitly, this procedure assumes that the allele
231 frequencies in the reference panel and the admixed individuals whose ancestry is from a
232 given reference panels are equivalent. For large, well-mixed populations such as those of *D.*
233 *melanogaster*, this is likely to be a reasonable assumption. Nonetheless, below we assess
234 the impact of differences in the ancestral allele frequencies for plausible demographic
235 models in this species.
236
237 **Dependence on Ancestral Linkage Disequilibrium**
238 Within an admixed population, there are two sources of LD. LD that is induced due to the
239 correlation of alleles from the same ancestry type (*i.e.* admixture LD), and LD that is
240 present within each of the ancestral populations (ancestral LD). Admixture LD, is the signal
241 of LA that we seek to detect using the HMM. The second type, ancestral LD, limits the
242 independence of the ancestral information captured by each marker, and is expected to
243 confound HMM-based analyses, particularly as we aimed to estimate the time since
244 admixture within this framework. We therefore sought to quantify the effect of ancestral
245 LD by discarding one of each pair of sites in LD within either ancestral population. We
246 found that ancestral LD tends to increase admixture time estimates obtained using our
247 method, and we decreased the cutoff of the LD parameter, $|r|$, by 0.1 until the time
248 estimates obtained for single chromosomes were unbiased with respect to the true time
249 since admixture. We found that $|r| \leq 0.4$ fit this criterion, although for relatively ancient
250 admixture events with highly skewed ancestry proportions—*i.e.* $m < 0.1$ or $m > 0.9$—some

251 residual bias was apparent in the estimates of admixture time (Figure 1). This reflects the
252 fact that the SMC' ancestry tract distribution performs poorly with highly skewed ancestry
253 proportions and especially for long times since admixture [57].
254
255 Figure 1 also reveals a striking difference between otherwise equivalently skewed
256 admixture proportions. For example when $m = 0.1$, there was a much larger effect of
257 ancestral LD than when $m = 0.9$. This is due to differences in the variability of LD within the
258 ancestral populations. That is, due to the strong population bottleneck, cosmopolitan *D.*
259 *melanogaster* populations have substantially more LD and fewer polymorphic sites than
260 African *D. melanogaster* populations. Because the time estimation procedure appears to be
261 sensitive to the amount of ancestral LD present in the data, simulations of the type we
262 described here may be necessary to determine what $|r|$ cutoffs are required to produce
263 unbiased time estimates given the ancestral LD of the populations in a given analysis using
264 this method.
265
266 **Accuracy and Applications to Diploid and Pooled Samples**
267 We next sought to quantify the accuracy of our approach across varying sample ploidies
268 and times since admixture (Figure 2). Especially for moderate and short admixture times
269 (*i.e.* 0—500 generations), our method performed well for all ploidies considered and we
270 were able to accurately recover the correct admixture time with relatively little bias.
271 However, as true admixture time increases, the time estimates for pooled samples become
272 significantly less reliable and show a clear negative bias. Nonetheless, across the range of
273 times presented in Figure 2, samples of ploidy one and two showed little bias, and we

274 therefore believe our method will produce sufficiently accurate admixture time estimates
275 for a wide variety of applications.
276
277 All measures of accuracy decrease with increasing time since admixture (Figure 2).
278 However, even for relatively long times since admixture—2000 generations—and for large
279 sample ploidies, the mean posterior error remained relatively low for all ancestry
280 proportions and for long times since admixture. This indicates that this approach may be
281 sufficiently accurate for a wide variety of applications, sequencing depths, and sample
282 ploidies. Nonetheless, the proportion of sites within the 95% credible interval decreased
283 with larger pool sizes and it is clear that for larger pools the posterior credible interval
284 tends to be too narrow. Therefore, correcting for this bias may be necessary for
285 applications that are sensitive to the accuracy of the credible interval.
286
287 An important consideration is that estimates of *t* will be reliable only if the local
288 recombination rates are known with reasonably high accuracy [58]. In many species, an
289 accurate broad-scale map is available. However, fine-scale variation in recombination rates
290 has only been documented for a few model species. Therefore, for relatively short to
291 moderate times since admixture, error in the genetic map is expected to have a limited
292 impact on date estimates. However, for longer times since admixture, this factor has the
293 potential to bias estimates of *t* [58], particularly in species with large variance in local
294 recombination rates (*e.g.* due to hotspots). Since *D. melanogaster* has one of the best
295 recombination maps currently available in any species [59] and because we do not aim to
296 estimate time in our applications, we do not believe this will heavily impact the analyses

297 we present below. However, for most applications, it will be necessary to consider the
298 impact of error in the assumed genetic map to accurately interpret estimates of *t* obtained
299 using this method. We emphasize that this challenge is not unique to this application, but
300 will impact virtually all ancestry estimation methods that rely on a genetic map for
301 estimating the time since admixture.
302
303 **Non-Independence Among Ancestry Tracts**
304 As described above, estimates of the time of admixture demonstrate an apparent bias in
305 pools of higher ploidy (Figure 2). Specifically, time tends to be slightly overestimated for
306 relatively short admixture times and underestimated at relatively long admixture times.
307 This is particularly apparent at highly skewed ancestry proportions. Given that this bias is
308 primarily evident in pools of 10 to 20 individuals, we hypothesized that it might be due to
309 the non-independence of ancestry tracts among chromosomes, which should tend to
310 disproportionately affect samples of higher ploidy because all ancestry breakpoints are
311 assumed to be independent in our model. To test this, we simulated genotype data from
312 independent and identically distributed exponential tract lengths as is assumed by our
313 model. When we ran our HMM on this dataset, we found that no bias is evident for
314 simulations of up to 2000 generations (Supplemental Figure S1), indicating that the
315 primary cause of this bias was violations in the real data of the independence of ancestry
316 tracts that we assumed when computing the transition probabilities. However, it should be
317 possible to quantify and correct for this bias in applications of this method that aim to
318 estimate the time since admixture.
319

320 **Robustness to Unknown Population Size**

321 The transition probabilities of this HMM depend on knowledge of the population size. In
322 practice, this parameter is unlikely to be known with certainty. Hence, to assess the impact
323 of misspecification of the population size, we performed simulations using a range of
324 population sizes that span three orders of magnitude (N=100, 1000, 10000, and 100000).
325 All analyses presented here were conducted by applying our HMM to haploid and diploid
326 samples, but qualitatively similar results hold for samples of larger ploidy. We then
327 analyzed these data assuming the default population size, 10000, is correct. For relatively
328 short times since admixture, there was not a clear bias for any of the true population sizes
329 considered. However, at longer true admixture times, estimated admixture times for both
330 N=100 and N=1000 asymptote at a number of generations near to the population sizes.
331 This result reflects the fact that smaller populations will tend to coalesce at a portion of the
332 loci in the genome relatively quickly, and ancestry tracts cannot become smaller following
333 coalescence. Nonetheless, the accuracy of LAI remained high even when time estimates
334 were unreliable (Supplemental Figure S2) for the tested marker densities and patterns of
335 LD. Furthermore, in some cases it should be straightforward to determine if a population
336 has coalesced to either ancestry state at a large portion of the loci in the genome,
337 potentially obviating this issue.

338

339 A more subtle departure from the expectation was evident for population sizes that are
340 larger than we assumed in analyzing these data (Supplemental Figure S2). This likely
341 reflects the fact that the probability of back coalescence to the previous marginal genealogy
342 to the left after a recombination event is inversely related to the population size. Hence, the

343 rate of transition between ancestry types is actually slightly higher in larger populations
344 where back coalescence is less likely than we assumed during the LAI procedure. This
345 produced a slight upward bias in the estimates of admixture time when the population was
346 assumed to be smaller than it is in reality. However, this bias appears to be relatively
347 minor, and we expect that time estimates obtained using this method will be useful so long
348 as population sizes can be approximated to within an order of magnitude. Of course, this
349 bias is not unique to our application, and it will affect methods that aim to estimate
350 admixture time after LAI as well. That is, estimating the correct effective population size is
351 an inherent problem for all admixture demographic inference methods.
352
353 **Application to Ancient Admixture**
354 Although it is clear that accurately estimating relatively ancient admixture times is
355 challenging in higher ploidy samples, we sought to determine the limits of our approach for
356 LAI and time estimation for longer admixture times for haploid sequence data. Because of
357 rapid coalescence in smaller samples (see above), we performed admixture simulations
358 with a diploid effective population size of 100,000. It is clear that there is a limit to the
359 inferences that can be made directly using our method. Like the higher ploidy samples,
360 time estimates for haploid samples departed from expectations shortly after 2,000
361 generations since admixture (Supplemental Figure S3). Nonetheless, the magnitude of this
362 bias is slight, and it is likely that it could be corrected for when applying this method even
363 for very ancient admixture events. For all admixture times considered, LAI remained
364 acceptably accurate despite the slight bias in time estimates (Supplemental Figure S3).
365

366 **Reference Panel Size**

367 One question is what effect varying the reference panel sizes will have on LAI inference
368 using this method. We therefore compared results from reference panels of size 10
369 chromosomes with those from panels of size 100 chromosomes (Supplemental Figure S4).
370 As with results obtained for reference panels of size 50, panels of size 100 were sufficient
371 to accurately estimate admixture time and LA over many generations since admixture.
372 Whereas, when panel sizes were just 10 chromosomes, time estimates were clearly biased
373 and the result was variable across ancestry proportions (Supplemental Figure S4).
374 However, since there was a strong correlation between true and estimated admixture
375 times even with relatively small panel sizes, it may therefore be possible to infer the
376 correct time by quantifying this bias through simulation and correcting for it. Furthermore,
377 although LAI is clearly less reliable with smaller panels, these results are not altogether
378 discouraging and this approach, in conjunction with modest reference panels may still be
379 effective for some applications.

380

381 **Allele Frequency Differences Between Ancestral and Admixed Populations**

382 Ultimately, there are three reasons why allele frequencies in the reference panels and in
383 the admixed population panel would be expected to differ beyond that expected from
384 binomial samples with the same mean. First, some amount of genetic drift may have
385 occurred in the ancestral population and in the admixed population in the time since the
386 admixed population was founded. Second, in some cases, it is infeasible to sample the
387 ancestral population of an admixed group, and a genetically divergent population must
388 suffice as the reference panel if this method is to be used. Third, divergent selection may

389 quickly modify allele frequencies between admixed and ancestral populations. Hence,
390 genetic divergence between reference and admixed populations may be an important
391 challenge for this method.
392
393 To address this, we simulated the second scenario, where increasingly divergent
394 populations are used as the reference panels to study admixed populations. In order to
395 make this relevant to the application to *D. melanogaster* populations, below, we selected
396 times for divergence that might be consistent with differences across continental
397 populations in Sub-Saharan Africa and in Cosmopolitan populations. Although time
398 estimates obtained using this approach are weakly positively biased with increasing
399 divergence between the ancestral population and reference panels, the accuracy of this LAI
400 method is largely unaffected (Supplemental Figure S5). Hence, for biological scenarios
401 potentially consistent with those of *D. melanogaster* ancestral populations, we do not
402 expect this challenge to strongly bias our method. Nonetheless, in applications to other
403 populations, with potentially differently structured ancestral populations, it would be
404 necessary to examine the effects of this bias in detail.
405
406 **High Sample Ploidy**
407 In a wide variety of pool-seq applications, samples are pooled in larger groups than we
408 have considered above (*e.g.* [40,50,52]). We are therefore interested in determining how
409 our method will perform on pools of 100 individuals. Towards this, we performed
410 simulations as before, but we designed our parameters to resemble those of the pooled
411 sequencing data that we analyze in the application of this method below. Specifically, we

412 simulated data with a mean sequencing depth of 25, a time since admixture of 1500
413 generations, and an ancestry proportion of 0.8. Consistent with results for ploidy 20, we
414 found that time tends to be dramatically underestimated (*i.e.* the mean estimate of
415 admixture time was 680 generations). However, when we provided the time since
416 admixture, our method produced reasonably accurate LAIs for these samples. Although the
417 posterior credible interval was again too narrow, the mean posterior error was just 0.053
418 when expressed as an ancestry frequency, indicating that this approach can produce LA
419 estimates that are close to their true values for existing sequencing datasets (*e.g.* Figure 3).
420 However, the HMM's run time increases dramatically for higher ploidy samples and higher
421 sequencing depths, a factor that may affect the utility of this program for some analyses.
422 Nonetheless, for more than 36,000 markers, a sample ploidy of 100 and a mean sequencing
423 depth of 25, the average runtime was approximately 42 hours. In contrast, for the same set
424 of parameters, but where individuals are sequenced and analyzed as diploids, the mean
425 runtime was just 8 minutes (See Supplemental Table S1 for a comparison of run times
426 across many parameter sets).
427
428 **Robustness to Deviations From the Neutral Demographic Model**
429 An important concern is that many biologically plausible admixture models would violate
430 the assumptions of this inference method. In particular, continuous migration and selection
431 acting on alleles from one parental population are two potential causes of deviation from
432 the expected model in the true data. To assess the extent of this potential bias, we
433 performed additional simulations. First, we considered continuous migration at a constant
434 rate that began t generations prior to sampling. In simulations with continuous migration,

435 additional non-recombinant migrants enter the population each generation. Relative to a
436 single pulse admixture model, this indicates that the ancestry tract lengths will tend to be
437 longer than those under a single pulse admixture model in which all individuals entered at
438 time *t*. Indeed, we found that admixture times tended to be underestimated with models of
439 continuous migration. However, the accuracy of LAI remained high across all situations
440 considered here (Table 1), indicating that the LAI aspect of this approach may be robust to
441 alternative demographic models.
442
443 **Table 1.** Parameter estimation and LAI when admixture occurs at a constant rate, rather
444 than in a single pulse.

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	0.001	0.998
		10	129	0.963	0.305	0.017	0.839
		20	98	0.545	0.328	0.033	0.353
	0.001	1	55	1.000	0.004	0.001	0.999
		2	53	1.000	0.013	0.002	0.997
		10	156	0.951	0.558	0.028	0.727
		20	90	0.551	0.719	0.043	0.179
	0.002	1	54	1.000	0.006	0.002	0.999
		2	52	0.999	0.019	0.003	0.996
		10	123	0.949	0.758	0.035	0.671
		20	74	0.679	1.115	0.045	0.176
	0.004	1	43	1.000	0.008	0.002	0.998
		2	54	0.999	0.035	0.005	0.993
		10	91	0.955	1.085	0.044	0.605
		20	75	0.860	1.788	0.045	0.248
500	0.0005	1	254	0.999	0.033	0.010	0.993
		2	250	0.997	0.121	0.018	0.974
		10	331	0.956	1.395	0.027	0.557
		20	333	0.882	2.321	0.051	0.261
	0.001	1	266	0.999	0.049	0.014	0.990
		2	268	0.996	0.198	0.027	0.962

	10	325	0.967	1.887	0.063	0.521
	20	366	0.926	3.049	0.055	0.294
0.002	1	294	0.999	0.055	0.016	0.989
	2	297	0.996	0.238	0.032	0.956
	10	352	0.977	2.076	0.064	0.542
	20	370	0.951	3.238	0.054	0.336
0.004	1	346	0.999	0.038	0.010	0.993
	2	350	0.997	0.164	0.021	0.973
	10	403	0.989	1.634	0.045	0.692
	20	462	0.979	2.773	0.041	0.473

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 indicating that our method can robustly delineate LA, even when the data violate assumptions of the inference method (Table 1,2).

463

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

465 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	0.001	0.998
		10	129	0.963	0.305	0.017	0.839
		20	98	0.545	0.328	0.033	0.353
	0.001	1	55	1.000	0.004	0.001	0.999
		2	53	1.000	0.013	0.002	0.997
		10	156	0.951	0.558	0.028	0.727
		20	90	0.551	0.719	0.043	0.179
	0.002	1	54	1.000	0.006	0.002	0.999
		2	52	0.999	0.019	0.003	0.996
		10	123	0.949	0.758	0.035	0.671
		20	74	0.679	1.115	0.045	0.176
	0.004	1	43	1.000	0.008	0.002	0.998
		2	54	0.999	0.035	0.005	0.993
		10	91	0.955	1.085	0.044	0.605
		20	75	0.860	1.788	0.045	0.248
500	0.0005	1	254	0.999	0.033	0.010	0.993
		2	250	0.997	0.121	0.018	0.974
		10	331	0.956	1.395	0.027	0.557
		20	333	0.882	2.321	0.051	0.261
	0.001	1	266	0.999	0.049	0.014	0.990
		2	268	0.996	0.198	0.027	0.962
		10	325	0.967	1.887	0.063	0.521
		20	366	0.926	3.049	0.055	0.294
	0.002	1	294	0.999	0.055	0.016	0.989
		2	297	0.996	0.238	0.032	0.956
		10	352	0.977	2.076	0.064	0.542
		20	370	0.951	3.238	0.054	0.336
	0.004	1	346	0.999	0.038	0.010	0.993
		2	350	0.997	0.164	0.021	0.973
		10	403	0.989	1.634	0.045	0.692
		20	462	0.979	2.773	0.041	0.473

466

467

468 **Comparison to WinPop**

469 We next compared the results of our method to those of WinPop [34]. Because WinPop
470 accepts only diploid genotypes, we provided this program diploid genotype data. However,
471 for these comparisons, we still ran our method on simulated read pileups with the mean
472 depth equal to 2. WinPop was originally designed for local ancestry inference in very
473 recently admixed populations. As expected, WinPop performed acceptably for very short
474 admixture times, but rapidly decreased in performance with increasing time (Supplemental
475 Figure S6). However, by default, WinPop removes sites in strong LD within the admixed
476 samples, which includes ancestral LD, but also admixture LD—the exact signal LAI methods
477 use to identify ancestry tracts.

478

479 We therefore reran WinPop, but instead of pruning LD within the admixed population, we
480 removed sites in strong LD within the ancestral populations as described above in our
481 method. With this modification, WinPop performs nearly as well as our method, but
482 remains slightly less accurate especially at longer admixture times (Supplemental Figure
483 S6). This difference presumably reflects the windowed-based approach of WinPop. At
484 longer times since admixture a given genomic window may overlap a breakpoint between
485 ancestry tracts. Although the performance is nearly comparable with this modification, we
486 emphasize that our method enables users to estimate the time since admixture, where this
487 must be supplied for WinPop, and allows for LAI on read pileups, therefore incorporating
488 genotype uncertainty into the LAI procedure. Indeed our method is more accurate at longer
489 timescales even when supplied with considerably lower quality read data. However
490 WinPop supports LAI with multiple ancestral populations, which our method currently

491 does not (but see Conclusions). Furthermore many LAI algorithms utilize haplotype
492 information, which may be particularly valuable in populations where LD extends across
493 large distances as in *e.g.* human populations.
494
495 **Assessing Applications to Human Populations**
496 Given the strong interest in studying admixture and local ancestry in human populations
497 (*e.g.* [22-25]), it is useful to ask if our method can be applied to data consistent with
498 admixed populations of humans. Towards that goal, we simulated data similar to what
499 would be observed in admixture between modern European and African lineages and
500 applied our HMM to estimate admixture times and LA. We found that our method can
501 accurately estimate admixture times for relatively short times since admixture, however,
502 substantially more stringent LD pruning in the reference panels is necessary to produce
503 unbiased estimates (Figure 4). This may be expected given that linkage disequilibrium
504 extends across longer distances in human populations than it does in *D. melanogaster*. In
505 other words, the scales of ancestral LD and admixture LD become similar rapidly in
506 admixed human populations. Furthermore, this approach yields accurate time estimates
507 for shorter times since admixture than with genetic data consistent with *D. melanogaster*
508 populations. For a relatively short time since admixture, around 100 generations, it is
509 possible to obtain accurate and approximately unbiased estimates of the admixture time
510 over a wide range of ancestry proportions, indicating that this method may be applicable to
511 recently admixed human populations as well (Figure 4). Nonetheless, this result
512 underscores the need to examine biases associated with LD pruning in this approach prior
513 to application to a given dataset.

514

515 **Bias in LAI due to Uncertainty in Time of Admixture**

516 To demonstrate that assumptions about the number of generations since admixture have
517 the potential to bias LAI, we analyzed a SNP-array dataset from Greenlandic Inuits [60,61].
518 The authors had previously noted a significant impact of *t* on the LAI results produced
519 using RFMix [24], which we were able to reproduce here for chromosome 10
520 (Supplemental Figure S7). Indeed, even for comparisons between $t = 5$ and $t = 20$, both of
521 which may be biologically plausible for these populations, the mean difference in posterior
522 probabilities between samples estimated using RFMix was 0.0903 (Supplemental Figure
523 S7). However, when we applied our method to these data, a clear optimum from *t* was
524 obtained at approximately 6-7 generations prior to the present, which is close to the
525 plausible times of admixture for these populations (Supplemental Figure S7). This
526 comparison therefore demonstrates that even relatively minor changes in assumptions of *t*
527 have the potential to strongly impact LAI results, and underscores the importance of
528 simultaneously performing LAI while estimating *t*.

529

530 However, these results also indicate that our method may not be robust in situations where
531 the background LD is high and ancestry informative markers are neither common nor
532 distributed evenly across the genome. When we compared the results of our method at $t =$
533 5 and at $t = 20$, we also obtained differences in the mean posterior among individuals as
534 with RFMix. However, one notable difference is that the mean posterior difference using
535 RFMix has a particularly high variance and therefore higher mean error (Supplemental
536 Figure S7), but actually a lower median difference than we found using our method. There

537 are likely two causes for differences observed in the mean ancestry posterior among
538 individuals. First, the datasets considered were generated with a metabochip SNP-chip
539 [62], which contains a highly non-uniform distribution of markers across the genome.
540 Second, the ancestral LD in the Inuit population is extensive [61], and we could only retain
541 a relatively small proportion of the markers after LD pruning in the reference panels. These
542 results therefore also underscore the challenges of LAI when the signal to noise ratio is low
543 as may be the case in some human populations, for which LD is extensive, and for some
544 sequencing strategies.
545
546 **Bias due to Incorrect Estimates of *t* and *m***
547 Although in general it is straightforward to estimate *m* from genome-wide data, in some
548 cases this parameter may be misestimated prior to LAI. We therefore sought to quantify
549 this potential effect by performing LAI after supplying incorrect values of m. In general, we
550 found that values close to the true range, *i.e.* within 0.05 of the true *m*, tend to yield
551 reasonably accurate time estimates. However, increasingly incorrect values produce
552 sharply downwardly biased time estimates and this effect is especially pronounced for
553 highly skewed true *m* (Supplemental Figure S8). As could be expected given the robustness
554 of LAI to many perturbations (above), when the incorrect *t* is supplied to the program, the
555 LA results remain reasonable. However it is worth noting that the penalty appears to be
556 greatest when *t* is too small rather than too large (Supplemental Figure S9).
557
558 **Estimating Confidence Intervals for *t***

559 Although this is not a primary focus for this work, for some users it may be of interest to
560 construct confidence intervals for estimates of *t*. We recommend the block bootstrap as the
561 preferred method for estimating confidence interval for *t*, and we have written a script that
562 will produce these (available on the github page for this project:
563 https://github.com/russcd/Ancestry_HMM). Simulations confirm that this can produce
564 confidence intervals overlapping the true *t* (Supplemental Figure S10), but bias in *t*
565 estimates for higher ploidy samples may still be apparent in some cases.
566

567 **Patterns of LA on Inversion Bearing Chromosomes in *D. melanogaster***

568 Given their effects suppressing recombination in large genomic regions, chromosomal
569 inversions may be expected to strongly affect LAI [2,63]. Although we attempted to limit
570 the impact of chromosomal inversions by eliminating known polymorphic arrangements
571 from the reference panels (see methods), many known inversions are present within the
572 pool-seq samples we aimed to analyze [64]. We therefore focused on known inverted
573 haplotypes within the DGPR samples [63,65-67], which are comprised of inbred
574 individuals, and therefore phase is known across the entire chromosome.
575

576 In comparing LA estimates between inverted and standard arrangements, it is clear that
577 chromosomal inversions can substantially affect LA across the genomes (Figure 5). In
578 general, the chromosomal inversions considered in this work originated in African
579 populations of *D. melanogaster* [63], and consistent with this observation, most inversion
580 bearing chromosomes showed evidence for elevated African ancestry. This was
581 particularly evident in the regions surrounding breakpoints, where recombination with

582 standard arrangement chromosomes is most strongly suppressed. Importantly, this pattern
583 continued outside of inversion breakpoints as well, consistent with numerous observations
584 that recombination is repressed in heterokaryotypes in regions well outside of the
585 inversion breakpoints in *Drosophila* (*e.g.* [2,63,68]). In(3R)Mo is an exception to this
586 general pattern of elevated African ancestry within inverted arrangements (Figure 5). This
587 inversion originated within a cosmopolitan population [63], and has only rarely been
588 observed within sub-Saharan Africa [69,70]. Consistent with these observations, In(3R)Mo
589 displayed lower overall African ancestry than chromosome arm 3R than standard
590 arrangement chromosomes.
591
592 Although chromosomal inversions may affect patterns of LA in the genome on this ancestry
593 cline, we believed including chromosomal inversions in the pool-seq datasets would not
594 heavily bias our analysis of LA clines. Inversions tend to be low frequency in most
595 populations studied [64], and because they affect LA in broad swaths of the genome—
596 sometimes entire chromosome arms—including inversions is unlikely to affect LA cline
597 outlier identification which appears to affect much finer scale LA (below). Furthermore,
598 inversion breakpoint regions were not enriched for LA cline outliers in our analysis
599 (Supplemental Table S2), suggesting that inversions have a limited impact on overall
600 patterns of local ancestry on this cline. Nonetheless, the LAI complications associated with
601 chromosomal inversions should be considered when testing selective hypotheses for
602 chromosomal inversions as genetic differentiation may be related to LA, rather than
603 arrangement-specific selection in admixed populations such as those found in North
604 America.

605

606 **Application to *D. melanogaster* Ancestry Clines**

607 Finally, we applied our method to ancestry clines between cosmopolitan and African
608 ancestry *D. melanogaster*. Genomic variation across two ancestry clines have been studied
609 previously [21,38,40,52]. In particular, the cline on the east coast of North America has
610 been sampled densely by sequencing large pools of individuals to estimate allele
611 frequencies, and previous work has shown that the overall proportion of African ancestry
612 is strongly correlated with latitude [21]. Consistent with this observation, we found a
613 significant negative correlation for all chromosome arms between proportion of average
614 African ancestry and latitude (rho = -0.891, -0.561, -0.912, -0.913, and -0.755, for 2L, 2R,
615 3L, 3R, and X respectively).

616

617 Although global ancestry proportions have previously been investigated in populations on
618 this ancestry cline [21,38], these analyses neglected the potentially much richer
619 information in patterns of LA across the genome. We therefore applied our method to these
620 samples. Because of the relatively recent dual colonization history of these populations and
621 subsequent mixing of genomes, a genome-wide ancestry cline is expected [21]. However,
622 loci that depart significantly in clinality from the genome-wide background levels may
623 indicate that natural selection is operating on a site linked to that locus.

624

625 **LA is Correlated with Recombination Rate**

626 Previously Pool (2015) found that regions of low recombination are disproportionately
627 enriched for African ancestry in the Raleigh, NC population [17]. Here, we find a similar

628 pattern and we further find that is replicated across all populations that were assayed on
629 this ancestry cline. Specifically, in all populations studied the proportion of African ancestry
630 is significantly negatively correlated with local recombination rates (Figure 6). Ultimately,
631 this correlation may have two causes. First, if selection is more efficient at purging African
632 alleles in high recombination regions, these loci will tend to be removed preferentially in
633 those genomic regions. An alternative explanation is that introgressing African alleles that
634 are favored by selection would tend to bring larger linkage blocks along with them in the
635 predominantly low recombination regions. Regardless of the specific source of natural
636 selection, a neutral admixture model would not predict this robust correlation between LA
637 and recombination rates within all populations, indicating that natural selection has played
638 an important role in shaping LA on this ancestry cline.
639
640 **Robustness of LAI to Genomic Heterogeneity**
641 Previous studies have found that heterogeneity in the genome with respect to ancestry
642 informative markers may impact the accuracy of LAI [71]. To assess this possibility, we
643 computed the mean difference between posterior mean estimates for the two samples from
644 Florida and between the two samples from Maine. Importantly, because these pooled
645 samples were created using different isofemale lines [40], this is a conservative test of our
646 method since there will be true biological differences as well as stochastic sequencing
647 differences between replicates from each population. We found no correlation between the
648 mean difference of the posterior means and local recombination rates (P = 0.2353 and P =
649 0.7529, Spearman's rank correlation for Florida and Maine respectively), indicating that
650 the correlation observed between local recombination rates and LA is unlikely to be an

651 artifact of differential accuracy of LAI in different genomic regions. However, it should be
652 acknowledged that in some genomic regions it maybe challenging to unambiguously infer
653 LA [17,71].
654
655 **Outlier LA Clines**
656 Selection within admixed populations may take several distinct forms. On the one hand,
657 loci that are favorable in the admixed population—either because they are favored on an
658 admixed genetic background, enhance reproductive success in an admixed population, or
659 are favorable in the local environment—will tend to achieve higher frequencies, and we
660 would expect these sites to have a more positive correlation with latitude than the genome-
661 wide average. Conversely, loci that are disfavored within the admixed population may be
662 expected to skew towards a more negative correlation with latitude.
663
664 Although it is not possible to distinguish between these hypotheses directly, a majority of
665 evidence suggests that selection has primarily acted to remove African ancestry from the
666 largely Cosmopolitan genetic backgrounds found in the Northern portion of this ancestry
667 cline. First, abundant evidence suggests pre-mating isolation barriers between some
668 African and cosmopolitan populations [72-74]. Second, there is strong post-mating
669 isolation between populations on the ends of this cline [46,47]. Third, we report here a
670 strong negative correlation between LA frequency and local recombination rates (above).
671 Finally, circumstantially, the local environment on the east coast of North America is
672 perhaps most similar to the environment of Cosmopolitan compared to African ancestral
673 populations, which further suggests that Cosmopolitan alleles are likely favored through

674 locally adaptive mechanisms. For these reasons, we therefore examined loci that are
675 outliers for a negative partial correlation with latitude, as this is the expected pattern for
676 African alleles that are disfavored in more temperate populations. In other words, the
677 outlier regions show a significantly stronger negative correlation between local African
678 ancestry and latitude than the chromosome arm does on average.
679
680 There is an ongoing debate about the relative merits of an outlier approach versus more
681 sophisticated models for detecting and quantifying selection in genome-wide scans. We
682 believe that the difficulties of accurately estimating demographic parameters for this
683 ancestry cline make the outlier approach most feasible for our purposes. Using our outlier
684 approach, we identified 80 loci that showed the expected negative partial correlation with
685 latitude (Figure 7). Although the specific statistical threshold that we employed is
686 admittedly arbitrary, given the strength of evidence indicating widespread selection on
687 local ancestry in this species (above), we expected that the tail of the LA cline distribution
688 would be enriched for the genetic targets of selection.
689
690 **Differences Among Chromosome Arms**
691 Due to the differences in inheritance, evolutionary theory predicts that selection will
692 operate differently on the X chromosome relative to autosomal loci. Of specific relevance to
693 this work, the large-X effect [75,76] is the observation that loci on the X chromosome
694 contribute to reproductive isolation at a disproportionately high rate. Additionally, and
695 potentially the cause of the large-X effect, due to the hemizygosity of X-linked loci, the X
696 chromosome is expected to play a larger role in adaptive evolution, the so-called faster-X

697 effect [77]. There is therefore reason to believe that the X chromosome will play a
698 significant role in genetically isolating Cosmopolitan and African *D. melanogaster*.
699
700 Consistent with a larger role for the sex chromosomes in generating reproductive isolation
701 or selective differentiation between *D. melanogaster* ancestral populations, we found that
702 that the X chromosome has a lower mean African ancestry proportion than the autosomes
703 in all populations. Furthermore, the X displays a stronger correlation between local
704 recombination rates and the frequency of African ancestry than the autosomes in all 14
705 populations samples, potentially indicating that selection has had a disproportionately
706 strong effect shaping patterns of local ancestry on this chromosome than on the autosomes.
707 In addition, the X has a significantly higher rate of outlier LA clinal loci than the autosomes
708 (23 LA outliers on the X, 57 on the Autosomes, $p = 0.0341$, one-tailed exact Poisson test).
709 Although consistent with evolutionary theory, differences between autosomal arms and the
710 X chromosome may also be explained in part by differences in effective recombination
711 rates on the X chromosome than the autosomes, differences in power to identify LA clines
712 associated with chromosome arm specific patterns, or by the disproportionately larger
713 number of chromosomal inversions on the autosomes than on the X chromosome in these
714 populations [64,69]. Distinguishing between this hypothesis and confounding factors will
715 be central to determining whether key results from speciation research are replicated in
716 much more recently diverged populations.
717
718 **Biological Properties of Outlier LA Clinal Loci**

719 We next applied gene ontology analysis to the set of outlier genes to identify common
720 biological attributes that may suggest more specific organismal phenotypes underlying LA
721 clinal outliers. In total, we identified seven GO terms that remained significant after
722 applying a 5% FDR correction (Table S3). These GO terms reflect the presence of two
723 primary clusters of genes. The first, which corresponds broadly to histone acetylation, may
724 be related to chromatin remodeling and therefore is expected to effect gene expression
725 levels across a large number of loci. Previous work focused on this ancestry cline has
726 identified chromatin remodeling genes as a potentially important component locally
727 adaptive variation on this ancestry cline [78]. This may indicate that this previous efforts to
728 identify spatially varying selection in these populations may have been detecting selection
729 on local ancestry components associated with ecological adaptation in ancestral
730 populations. The second GO cluster, eukaryotic translation initiation factor 2 complex, also
731 appears to implicate a central role of clinal LA outliers on the regulation of gene expression.
732 One plausible explanation of these observations is that gene expression, particularly high
733 level regulation of gene expression, may be especially likely to contribute to epistatic
734 interactions as these proteins will inherently interact with a diverse set of loci throughout
735 the genome. Given that two distinct gene clusters related to gene expression are identified
736 by this analysis, gene expression would appear to be a plausible candidate phenotype to
737 investigate in future work on ecological divergence and isolating factors in admixed *D.*
738 *melanogaster* populations. Testing this prediction empirically through expression profiling
739 may therefore offer fruitful grounds of understanding the earliest stages of reproductive
740 isolation.
741

742 **Regions of Decreased African Ancestry**

743 Another subset of loci that we may wish to identify using these data are those that
744 contribute to reproductive isolation between African and Cosmopolitan *D. melanogaster*
745 populations and would therefore be removed by selection from most populations on this
746 ancestry cline. Although it is possible that Cosmopolitan alleles would be disfavored in an
747 admixed background as well, because these populations are predominantly Cosmopolitan,
748 we expect that the majority of selection on negatively epistatically interacting loci would
749 remove African alleles from populations. To identify these loci, we first computed the mean
750 African ancestry across all populations, and we then identified the subset of loci that were
751 in the lowest 5% tail. From those loci, we selected the loci minima from adjacent genomic
752 windows (see Methods, Figure 8), and we obtained a total of 84 local ancestry outliers.
753

754 As with the clinal outlier analysis above, to identify commonalities in the types of loci
755 identified by this analysis, we performed GO analysis on the set of loci that are outliers for
756 the mean proportion of African ancestry. After a 5% FDR correction, there are again several
757 gene clusters that are significantly enriched in this set of outlier loci (Supplemental Table
758 S4). Of particular interest is the GO term oogenesis, which may indicate that female
759 reproduction is affected during admixture between cosmopolitan and African populations
760 of *D. melanogaster*. This finding is particularly interesting in light of the fact that female
761 fertility is strongly affected when autosomal chromosomes from one end of this ancestry
762 cline are made homozygous on a genetic background carrying the X chromosome from the
763 other end of this ancestry cline [47]. Hence, the effects of combining divergence ancestry
764 types on female fertility, and specifically the genetic basis of oogenesis, may be an

765 appealing phenotype to characterize in detail in attempting to clarify the genetic effects
766 that isolate African and Cosmopolitan *D. melanogaster* populations.
767
768 **Candidate Behavioral Reproductive Isolation Genes**
769 Given the abundance of evidence supporting a role for pre-mating isolation barriers
770 between African and Cosmopolitan flies [72-74], we are interested in highlighting genes
771 potentially related to behavioral isolation between ancestral populations of *D.*
772 *melanogaster*. Consistent with this observation, one of the strongest LA cline outliers, *egh*,
773 has been conclusively linked to strong effects on male courtship behavior using a variety of
774 genetic techniques [79]. Additionally, gene knockouts of CG43759, another LA cline outlier
775 locus, have strong effects on inter-male aggressive behavior [80], and may also contribute
776 to behavioral differences between admixed individuals. These loci are therefore appealing
777 candidate genes for functional follow-up analyses, and illustrate the power of this LAI
778 approach for identifying candidate genes that are potentially associated with well
779 characterized phenotypic differences between ancestral populations.
780
781 **Little Evidence for Seasonal LA Outliers**
782 The Pennsylvania population included in this study has been sampled extensively,
783 including several paired fall and spring samples across three consecutive years. Previously,
784 Bergland et al. [40] identified numerous SNPs that showed recurrent and rapid seasonal
785 frequency changes in the Pennsylvania populations included in this study. They concluded
786 that these sites are experiencing recurrent selection associated with recurrent
787 environmental seasonal changes. To determine if LA across the *D. melanogaster* genome

788 might also experience selection associated with seasonal frequency shifts, we searched for
789 loci that showed a strong recurrent seasonal shift in LA. However, we identified fewer
790 significantly seasonal sites than we would expect to by chance (the proportion of
791 significant site at the alpha = 0.05 level of significance is 0.041). Furthermore, after
792 applying a false discovery rate correction [81], there are no sites that are significantly
793 seasonal at the q = 0.1 level. Collectively, these results indicate that LA within the
794 Pennsylvania populations of *D. melanogaster* remains remarkably stable during seasonal
795 environmental cycles.
796
797 Although this observation may, to a first approximation, appear to be at odds with the
798 results reported in Bergland et al. [40], we believe that it is consistent with the model
799 proposed in that work. Specifically, the authors suggested that long term balancing
800 selection may maintain these seasonally favorable polymorphisms in diverse *D.*
801 *melanogaster* populations and even in the ancestors of *D. melanogaster* and *D. simulants*
802 [40]. We therefore may expect that these polymorphisms will be maintained at similar
803 frequencies in African and Cosmopolitan populations. Hence, although the seasonal SNPs
804 change rapidly in frequency between spring and fall [40], the LA at these sites can remain
805 stable during seasonal fluctuations.
806
807 **Conclusion**
808 A growing number of next-generation sequencing projects produce low coverage data that
809 cannot be used to unambiguously assign individual genotypes, but which can be analyzed
810 probabilistically to account for uncertainty in individual genotypes [82-84]. However, most

811 existing LAI methods require genotype data derived from diploid individuals. Hence, there
812 is an apparent disconnect between existing LAI approaches and the majority of ongoing
813 sequencing efforts. In this work, we developed the first framework for applying LAI to
814 pileup read data, rather than genotypes, and we have generalized this model to arbitrary
815 sample ploidies. This method therefore has immediate applications to a wide variety of
816 existing and ongoing sequencing projects, and we expect that this approach and extensions
817 thereof will be valuable to a number of researchers. Although evaluating this application is
818 beyond the scope of this work, one particularly enticing potential use of this method is LAI
819 in ancient DNA samples for which sequencing depths often preclude accurate genotype
820 calling. Importantly, it would be straightforward to model site-specific errors in this
821 framework, which could be particularly important for ancient DNA applications [6].
822
823 For many applications, a parameter of central biological interest is the time since
824 admixture began (t). A wide variety of approaches have been developed that aim to
825 estimate t and related parameters in admixed populations [26,28-31,85,86]. Many of these
826 methods are based on an inferred distribution of tract lengths, however, inference of the
827 ancestry tract length distribution is associated with uncertainty that is typically not
828 incorporated in currently available methods for estimating t. Furthermore, incorrect
829 assumptions regarding t have the potential to introduce biases during LAI. Hence, it is
830 preferable to estimate demographic parameters such as the admixture time during the LAI
831 procedure. Nonetheless, as noted above, although LAI using our method is robust to many
832 deviations from the assumed model, admixture time estimates are sensitive to a variety of
833 potential confounding factors and examining the resulting ancestry tract distributions after

834 LAI may be necessary to confirm that the assumed demographic model provides a
835 reasonable fit to the data.
836
837 To our knowledge, this is the first method that attempts to simultaneously link LAI and
838 population genetic parameter estimation directly, and we can envision many extensions of
839 this approach that could expand the utility of this method to a broad variety of applications.
840 For example, it is straightforward to accommodate additional reference populations (*e.g.*
841 by assuming multinomial rather than binomial read sampling). Alternatively, any
842 demographic or selective model that can be approximated as a Markov process could be
843 incorporated—in particular, it is feasible to accommodate two-pulse admixture models and
844 possibly models including ancestry tracts that are linked to positively selected sites. Such
845 methods can be used to construct likelihood ratio tests of evolutionary models and for
846 providing improved parameter estimates.
847

848 # Methods

849 **Constructing Emissions Probabilities**

850 We model the ancestry using an HMM $\{H_v\}$ with state space $S = \{0,1,\ldots,n\}$, where $H_v = i$,
851 $i \in S$, indicates that in the vth position i chromosomes are from population 0 and $n - i$
852 chromosomes are from population 1. In the following, to simplify the notation and without
853 loss of generality, we will omit the indicator for the position in the genome as the structure
854 of the model is the same for all positions of equivalent ploidy. We assume each variant site
855 is biallelic, with two alleles A and a, and the availability of reference panels from source
856 populations 0 and 1 with total allelic counts C_{0a}, C_{1a}, C_{0A}, and C_{1A}, where the two subscripts

refer to population identity and allele, respectively. Also, $C_0 = C_{0A} + C_{0a}$ and $C_1 = C_{1A} + C_{1a}$. Finally, we also assume we observe a pileup of r reads from the focal population, with r_A and r_a reads for alleles A and a respectively ($r = r_A + r_a$). The emission probability of state $i \in S$ 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 ε is an error rate. This probability can be calculated by summing over all possible genotypes in the admixed sample and over all possible population identities of the reads, as explained in the following section.

The probability of obtaining r_0 ($=r - r_1$) reads, in the admixed population, from chromosomes of ancestry 0, given r and the hidden state $H = i$, and assuming no mapping or sequencing biases, is binomial,

$$(1)\quad r_0 \mid H = i, n, r \sim \mathrm{Bin}(r, i/n)$$

These probabilities are pre-computed in our implementation for all possible values of $i \in S$ and r_0, $0 \le r_0 \le r$. Similarly, for the reference populations, for j=0,1,

$$(2)\quad C_{jA} \mid C_j, f_j \sim \mathrm{Bin}(C_j, f_j)$$

where f_j is the allele frequency of allele A in population j. The analogous allelic counts in the admixed population, denoted C_{M0a}, C_{M1a}, C_{M0A}, and C_{M1A}, are unobserved (only reads are observed for the admixed population), but are also conditionally binomially distributed, *i.e.*:

$$(3)\ C_{M0A} \mid H = i, f_0 \sim \mathrm{Bin}(i, f_0) \text{ and } C_{M1A} \mid H = i, n, f_1 \sim \mathrm{Bin}(n - i, f_1)$$

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

$$(4)\ r_{0A} \mid H = i, r_0, C_{M0A} \sim \mathrm{Bin}(r_0, C_{M0A} / i)$$

It should be noted that because we are explicitly modeling the process of sampling alleles from the population (Equation 3) and the process of sampling reads conditional on the sample allele frequencies (Equation 4), that this approach corrects for the increased variance associated with two rounds of binomial sampling in poolseq applications that has been reported previously (*e.g.*, in [52]).

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

$$(5)\ \square_{0\square} \mid \square = \square, \square_{\square}, \square_{\square 0\square}, \square \sim \mathrm{Bin}(\square_0, (1 - \square)\square_{\square 0\square}/\square + \square(1 - \square_{\square 0\square}/\square)).$$

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

$$(6)\ \Pr\left(\square_{0\square}, \square_{0\square,} \middle| \square_0, \square_0, \square, \square = \square, \varepsilon\right) =$$

$$\int_0^1 \sum_{\square=0}^{\square} \Pr(\square_{0\square} | \square = \square, \square_0, \square_{\square 0\square} = \square, \square) \Pr(\square_{\square 0\square} = \square | \square = \square, \square_0) \square(\square_0) \square\square_0 \quad =$$

$$(7)\ \frac{\square_0!\square!}{(\square_0 - \square_{0\square})!\square_{0\square}!(\square_0 + \square + 1)!} \sum_{\square=0}^{\square} \Pr(\square_{0\square} | \square = \square, \square_0, \square_{\square 0\square} = \square, \square) \frac{(\square_0 - \square_{0\square} + \square - \square)!(\square_{0\square} + \square)!}{(\square - \square)!\square!}$$

assuming a uniform [0, 1] distribution for f_0. A similar expression is obtained for $\Pr\left(\square_{1\square}, \square_{1\square,} \middle| \square_1, \square_1, \square, \square = \square, \varepsilon\right)$, assuming $f_1 \sim \mathrm{U}[0,1]$, and these expressions combine multiplicatively to give

$$(8)\ \Pr\left(\square_{\square}, \square_{1\square,}, \square_{0\square,} \middle| \square_0, \square_0, \square_1, \square_1, \square, \square = \square, \varepsilon\right) =$$

$$\sum_{\square_{0\square}=\max\{0, \square_{\square} - \square_1\}}^{\min\{\square_0, \square_{\square}\}} \Pr\left(\square_{0\square}, \square_{0\square,} \middle| \square_0, \square_0, \square, \square = \square, \varepsilon\right) \Pr\left(\square_{1\square} = \right.$$

$$\left. \square_{\square} - \square_{0\square}, \square_{1\square,} \middle| \square_1, \square_1, \square, \square = \square, \varepsilon\right),$$

and the emission probabilities become

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

$$\sum_{\square_0=0}^{\square} \Pr(\square_0 | \square = \square, \square, \square) \Pr\left(\square_{\square}, \square_{1\square,}, \square_{0\square,} \middle| \square_0, \square_0, \square_1 = \square - \square_0, \square_1, \square, \square = \square, \varepsilon\right)$$

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

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

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

$$= \sum_{k=\max\{C_{MA}-i,0\}}^{\min\{n-i,C_{MA}\}} \frac{C_0!C_1!i!(n-i)!(C_{MA}+C_{1A}-k)!(C_{0A}+k)!(C_1-C_{MA}-C_{1A}+i+k)!(C_0-C_{0A}-i-k+n)!}{(C_0-C_{0A})!C_{0A}!(C_1-C_{1A})!C_{1A}!(C_{MA}-k)!k!(k+i-C_{MA})!(n-k-i)!(n-i+C_0+1)!(i+C_1+1)!} \quad (10)$$

These emissions probabilities are sometimes substantially faster to compute than those for short read pileups, especially when sequencing depths are high. However, the genotypes must be estimated with high accuracy for this approach to be valid. For applications with low read coverage, or with ploidy >2 for which many standard genotype callers are not applicable, it is usually preferable to use the pileup-based approach described above.

Constructing Transition Probabilities

We assume an admixed population, of constant size, with N diploid individuals, in which a proportion m of the individuals in the population where replaced with migrants t generations before the time of sampling. Given these assumptions, and an SMC' model of the ancestral recombination graph [87], the rate of transition from ancestry 0 to 1, along the length of a single chromosome, is

$$\lambda_0 = 2Nm\left(1-e^{\frac{-t}{2N}}\right) \quad (11)$$

per Morgan [57]. Similarly, the rate of transition from ancestry 1 to 0 on a single chromosome is

$$\lambda_1 = 2N(1-m)\left(1-e^{\frac{-t}{2N}}\right) \quad (12)$$

per Morgan. Importantly, because these expressions are based on a coalescence model, they account for the possibility that a recombination event occurs between two tracts of the same ancestry type and the probability that the novel marginal genealogy will back-coalesce with the previous genealogy [57]. Both events are expected to decrease the number of ancestry switches along a chromosome and ignoring their contribution will cause overestimation of the rate of change between ancestry types between adjacent markers.

The transition rates are in units per Morgan, but can be converted to rates per bp, by multiplying with the recombination rate in Morgans/bp, r_{bp} within a segment. The transition probabilities of the HMM for a single chromosome, $\mathbf{P}(l) = \{P_{ij}(l)\}, i, j \in S,$ between two markers with a distance l between each other, is then approximately

$$\mathbf{P}(l) = \begin{bmatrix} 1-\lambda_0 r_{bp} & \lambda_0 r_{bp} \\ \lambda_1 r_{bp} & 1-\lambda_1 r_{bp} \end{bmatrix}^l \quad (13)$$

using discrete distances, or

$$\mathbf{P}(l)=\begin{bmatrix} \frac{\lambda_1}{\lambda_0+\lambda_1}+\frac{\lambda_0}{\lambda_0+\lambda_1}e^{-r_{bp}l(\lambda_0+\lambda_1)} & \frac{\lambda_0}{\lambda_0+\lambda_1}-\frac{\lambda_0}{\lambda_0+\lambda_1}e^{-r_{bp}l(\lambda_0+\lambda_1)} \\ \frac{\lambda_0}{\lambda_0+\lambda_1}+\frac{\lambda_1}{\lambda_0+\lambda_1}e^{-r_{bp}l(\lambda_0+\lambda_1)} & \frac{\lambda_1}{\lambda_0+\lambda_1}-\frac{\lambda_1}{\lambda_0+\lambda_1}e^{-r_{bp}l(\lambda_0+\lambda_1)} \end{bmatrix} \quad (14)$$

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 [57], 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 [57]. 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

979 state 0 in the marginal processes and $j - i + l$ transitions from state 0 to state 1, summed

980 over all admissible values of l, i.e.,

981

982 $$\Pr\left(H_{v+k} = j \mid H_v = i\right) = \sum_{l=\max\{0,i-j\}}^{\min\{n-j,i\}} \binom{n-i}{j-i+l} \left(P_{01}(k)\right)^{j-i+l} \left(1-P_{01}(k)\right)^{n-j+i-l} \binom{i}{l} \left(P_{10}(k)\right)^{l} \left(1-P_{10}(k)\right)^{i-l} \quad (15)$$

983

984 Although this procedure can be computationally expensive when there are many markers,

985 read depths are high, and especially when n is large, in our implementation, we reduce the

986 compute time by pre-calculating and storing all binomial coefficients.

987

988 **Estimating Time Since Admixture**

989 A parameter of central biological interest, that is often unknown in practice, is the time

990 since the initial admixture event (t). We therefore use the HMM representation to provide

991 maximum likelihood estimates of t using the forward algorithm to calculate the likelihood

992 function. As this is a single parameter optimization problem for a likelihood function with a

993 single mode, optimization can be performed using a simple golden section search [88].

994 Default settings for this optimization in our software, including the search range maxima

995 defaults, t_{max} and $t_{min,}$ are documented in the C++ HMM source code provided at

996 https://github.com/russcd/Ancestry_HMM.

997

998 **Posterior Decoding**

999 After either estimating or providing a fixed value of the time since admixture to the HMM,

1000 we obtained the posterior distribution for all variable sites considered in our analysis using

1001 the forward-backward algorithm, and we report the full posterior distribution for each
1002 marker along the chromosome.
1003
1004 **Simulating Ancestral Polymorphism**
1005 To validate our HMM, we generated sequence data for each of two ancestral populations
1006 using the coalescent simulator MACS [55]. We sought to generate data that could be
1007 consistent with that observed in Cosmopolitan and African populations of *D. melanogaster*,
1008 which has been studied previously in a wide variety of contexts [2,11,35-37]. We used the
1009 command line "macs 400 10000000 -i 1 -h 1000 -t 0.0376 -r 0.171 -c 5 86.5 -I 2 200 200 0 -
1010 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
1011 generate genotype data. This will produce 200 samples of ancestry 0 and 200 samples of
1012 ancestry 1 on a 10mb chromosome—*i.e.* this should resemble genotype data for about half
1013 of an autosomal chromosome arm in *D. melanogaster*. Unless otherwise stated below, we
1014 then sampled the first 50 chromosomes from each ancestral population as the ancestral
1015 population reference panel, whose genotypes are assumed to be known with low error
1016 rates. The sample size was chosen because it is close to the size of the reference panel that
1017 we obtained in our application of this approach to *D. melanogaster* (below).
1018
1019 To evaluate the performance of our method on data consistent with human populations, we
1020 simulated data that could be consistent with that observed for modern European and
1021 African human populations. Specifically, we simulated the model of [89] using the
1022 command line "macs 200 1e8 -I 3 100 100 0 -n 1 1.682020 -n 2 3.736830 -n 3 7.292050 -eg
1023 0 2 116.010723 -eg 1e-12 3 160.246047 -ma x 0.881098 0.561966 0.881098 x 2.797460

1024 0.561966 2.797460 x -ej 0.028985 3 2 -en 0.028986 2 0.287184 -ema 0.028987 3 x
1025 7.293140 x 7.293140 x x x x x -ej 0.197963 2 1 -en 0.303501 1 1 -t 0.00069372 -r
1026 0.00069372". Admixture between ancestral populations was then simulated as described
1027 below.
1028
1029 **Simulating Admixed Populations**
1030 Although it is commonly assumed that admixture tract lengths can be modeled as
1031 independent and identically distributed exponential random variables (e.g. [26,29] and in
1032 this work, above), this assumption is known to be incorrect as ancestry tracts are neither
1033 exponentially distributed, independent across individuals, nor identically distributed along
1034 chromosomes [57]. We therefore aim to determine what bias violations of this assumption
1035 will have on inferences obtained from this model. Towards this, we used SELAM [56] to
1036 simulate admixed populations under the biological model described above. Because this
1037 program simulated admixture in forward time, it generates the full pedigree-based
1038 ancestral recombination graph, and is therefore a conservative test of our approach
1039 relative to the coalescent which is known to produce incorrect ancestry tract distributions
1040 for short times [57]. Briefly, we initialized each admixed population simulation with a
1041 proportion, m, of ancestry from ancestral population 1, and a proportion $1-m$ ancestry from
1042 ancestral population 0. Unless otherwise stated, all simulations were conducted with
1043 neutral admixture and a hermaphroditic diploid population of size 10,000.
1044
1045 We then assigned the additional, non-reference chromosomes from the coalescent
1046 simulations, to each ancestry tract produced in SELAM simulations according to their local

1047 ancestry along the chromosome. In this way, each chromosome is a mosaic of the two
1048 ancestral subpopulations. See, *e.g.* [2], for a related approach for simulating genotype data
1049 of admixed chromosomes.
1050
1051 **Pruning Ancestral Linkage Disequilibrium**
1052 Correlations induced by LD between markers within ancestral populations violates a
1053 central assumption of the Markov model framework. Although it may be feasible to
1054 explicitly model linkage within ancestral populations (*e.g.*, [24,25]), when ancestral
1055 populations have relatively little LD, such as those of *D. melanogaster*, another effective
1056 approach is to discard sites that are in strong LD in the ancestral populations. Hence, to
1057 avoid this potential confounding aspect of the data, we first computed LD between all pairs
1058 of markers within each reference panel that are within 0.01 centimorgans of one another.
1059 We then discarded one of each pair of sites where $|r|$ in either reference panel exceeded a
1060 particular threshold, and we decreased this threshold until we obtained an approximately
1061 unbiased estimate of the time since admixture estimates of the HMM. This approach differs
1062 from a previous method, WinPop [34], where LD is pruned from within admixed samples
1063 (see also below).
1064
1065 **Simulating Sequence Data**
1066 We first identified all sites where the allele frequencies of the ancestral populations differ
1067 by at least 20% within the reference panels. We excluded weakly differentiated sites to
1068 decrease runtime and because these markers carry relatively little information about the
1069 LA at a given site. Then, to generate data similar to what would be produced using Illumina

1070 sequencing platforms, we simulated allele counts for each sample, by first drawing the
1071 depth at a given site from a Poisson distribution. In most cases and unless otherwise stated,
1072 the mean of this distribution is set to be equal to the sample ploidy. We did this to ensure
1073 equivalent sequencing depth per chromosome regardless of pooling strategy, and because
1074 this depth is sufficiently low that high quality genotypes cannot be determined. We then
1075 generated set of simulated aligned bases via binomial sampling from the sample allele
1076 frequency and included a uniform error rate of 1% for both alleles at each site.
1077
1078 Unless otherwise stated, we simulated a total of 40 admixed chromosomes. The HMM can
1079 perform LAI on more than one sample at a time, and we therefore included all samples
1080 when running it. Hence, we used 40 haploid, 20 diploid, 4 pools of 10 chromosomes, and 2
1081 pools of 20 chromosomes for most comparisons of accuracy reported below, unless
1082 otherwise stated. It is worth noting that it is possible to jointly analyze distinct samples
1083 from the same subpopulation that have been sequenced at different ploidies.
1084
1085 **Simulating Divergent Ancestral Populations**
1086 To investigate the effects of allele frequency differences between reference populations and
1087 admixed populations, we performed coalescent simulations using the software MACS [55],
1088 using the command line "./macs 500 10000000 -i 1 -h 1000 -t 0.0376 -r 0.171 -c 5 86.5 -I 8
1089 100 100 50 50 50 50 50 50 0 -en 0 2 0.183 -en 0.0037281 2 0.000377 -en 0.00381 2 1 -ej
1090 0.00382 2 1 -eN 0.0145 0.2 -ej 0.0005 3 2 -ej 0.000500001 4 1 -ej 0.001 5 2 -ej
1091 0.001000001 6 1 -ej 0.002 7 2 -ej 0.002000001 8 1". This might be expected to produce
1092 populations that are differentiated similarly to how populations of *D. melanogaster* would

be across European populations or between populations in Central Africa. We then substituted the increasingly divergent populations for the reference panel. All allele frequency differences and LD pruning were performed as described above on each of the substitute reference panels.

Accuracy Statistics

To evaluate the performance of the HMM, we computed four statistics. First, we compute the proportion of sites where the true state is within the 95% posterior credible interval, where ideally, this proportion would be equal to or greater than 0.95. As this HMM has discrete states, there are many ways the 95% credible interval could be defined. In light of the fact that the credible interval tends to be narrow (Results), we defined the interval to include all states that are overlapped, by any amount, in the 95% confidence interval of the posterior distribution. Second, we compute the mean posterior error, the average distance between the posterior distribution of hidden states and the true state

$$E = \frac{\sum_{v=0}^{S}\sum_{i=0}^{n} p(H_v = i \mid \mathbf{r})\left|i - I_v\right|}{Sn}$$

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 data. Third, we also report the proportion of sites where the maximum likelihood estimate of the hidden state is equal to the true ancestry state. Finally, as an indicator of the specificity of our approach, we also report the average width of the 95% credible interval.

1115 **Deviations from the Assumed Neutral Demographic Model**

1116 A potential issue with this framework is that the assumptions underlying the transition
1117 matrixes and related time of admixture estimation procedure is likely to be violated in a
1118 number of biologically relevant circumstances. We therefore simulated populations
1119 wherein individuals of ancestral population 1 began entering a population entirely
1120 composed of individuals from ancestral population 0, at a time *t* generations before the
1121 present, at a constant rate that is sustained across all subsequent generations until the time
1122 of sampling. That is, additional unadmixed individuals of ancestry 1 migrate each
1123 generation from *t* until the present.

1124

1125 Natural selection acting on admixed genetic regions has been inferred in a wide variety of
1126 systems (*e.g.* [5,7,13,17,18]), and is expected to have pronounced effects on the distribution
1127 of LA among individuals within admixed populations. Here again, this aspect of biologically
1128 realistic populations will tend to violate central underlying assumptions of the model
1129 assumed in this work. Towards this, we simulated admixed populations with a single pulse
1130 of admixture *t* generations prior to the time of sampling. We then incorporated selection at
1131 2,5,10, and 20 loci at locations uniformly distributed along the length of the chromosome
1132 arm. All selected loci were assumed to be fixed within each ancestral population. Selection
1133 was additive and selective coefficients were assigned based on a uniform [0.005, 0.05]
1134 distribution to either ancestry 0 or 1 alleles with equal probability. As above, these
1135 simulations were conducted using SELAM [56].

1136

1137 For both selected and continuous migration simulations, we then performed the genotype
1138 and read data simulation procedure, and reran our HMM as described above. We
1139 performed 10 simulations for each treatment.
1140

1141 **Comparisons to WinPop**

1142 We next sought to compare our method to a commonly used local ancestry inference
1143 method, WinPop [34]. Towards this, we again simulated data using MACS and SELAM as
1144 described above. For these comparisons, the initial ancestry contribution was 0.5 and the
1145 number of generations since admixture varied between 5 and 1000. For comparison, we
1146 supplied WinPop and our program the correct time since admixture and ancestry
1147 proportions, as these are required parameters for WinPop. We also supplied the program
1148 with genotypes rather than read counts, another requirement of WinPop, whereas we
1149 supplied our HMM with read data simulated as described above. We then ran WinPop
1150 under default parameters, and we also reran WinPop using LD pruning within the
1151 reference panels, as we do in our method, instead of the default LD pruning implemented in
1152 WinPop.
1153

1154 **Analysis of Inuit Genotype Data**

1155 To demonstrate that LAI methods can be biased by the arbitrary selection of the time since
1156 admixture, we analyzed a dataset of SNP-array genotype data from Greenlandic Inuits.
1157 These data are described in detail elsewhere [60,61]. This population has received some
1158 admixture from a European source population, and the authors had previously used RFMix
1159 [24] to perform LAI, and found some sensitivity to the assumed time since admixture (J.

1160 Crawford *pers. Comm.*). We analyzed data from chromosome 10 using RFMix v1.5.4 [24] as
1161 described in Moltke *et al.* [61] assuming admixture occurred either 5 or 20 generations ago.
1162 We subsequently analyzed chromosome 10 using our HMM including the genotype-
1163 analysis emissions probabilities and assuming a genotype error rate of 0.2%. For our
1164 analysis we identified the LD cutoff that is appropriate for these data as described above.
1165

1166 **Generating *D. melanogaster* Reference Populations**

1167 To generate reference panels, we used a subset of the high quality *D. melenaogaster*
1168 assemblies that have been described previously in Pool *et al.* (2012) and Lack *et al.* (2015).
1169 As in the local ancestry analysis of Pool (2015), we used the French population. For our
1170 African reference panel, we selected a subset of the Eastern and Western African
1171 populations (CO, RG, RC, NG, UG, GA, GU) and grouped them into a single population for the
1172 purposes of our analysis. We elected to combine populations so that we would have a
1173 larger reference panel of African populations for this analysis, this solution may be justified
1174 because these *D. melanogaster* populations are only weakly genetically differentiated
1175 [2,21,90], particularly after common inversion-bearing chromosomes are removed from
1176 analyses. Specific individuals were selected for inclusion in the African reference panel if
1177 previous work found they have relatively little cosmopolitan ancestry (*i.e.,* below 0.2
1178 genome-wide in [2]).
1179

1180 Because of their powerful effects on recombination, chromosomal inversions are known to
1181 have substantial impacts on the distribution of genetic variants on chromosomes
1182 containing chromosomal inversions in *D. melanogaster* [2,63]. For this reason, we removed

1183 all common inversion-bearing chromosome arms from the reference populations [91].
1184 Nonetheless, it is clear that chromosomal inversions are present in the pool-seq samples
1185 [64]. Although the inversions certainly violate key assumptions of our model—particularly
1186 the transmission probabilities—given that our approach is robust to a many perturbations,
1187 we expect the LA within inverted haplotypes can be estimated with reasonable confidence,
1188 and the overall LAI procedure will still perform adequately with low frequencies of
1189 chromosomal-inversion bearing chromosomes present within these samples.
1190
1191 Although these reference populations are believed to have relatively little admixture, some
1192 admixture is likely to remain within these samples [2]. To mitigate this potential issue, we
1193 first applied our HMM to each reference population using the genotype-based emissions
1194 probabilities (above). Calculated across all individuals, we found that our maximum
1195 likelihood ancestry estimates were identical with those of Pool *et al.* (2012) at 96.2% of
1196 markers considered in our analysis. The differences between the results of these methods
1197 may reflect differences in the methodology of LAI or differences in the reference panels.
1198 Nonetheless, the broad concordance suggests the two methods are yielding similar overall
1199 results. We masked all sites where the posterior probability of non-native ancestry was
1200 greater than 0.5 within each reference individual's genome. These masked sequences were
1201 then used as the reference panel for the analyses of pool-seq data below.
1202
1203 **Ancestry Cline Sequence Data Analysis**
1204 We acquired pooled sequencing data from six populations from the east coast of the United
1205 States. The generation of these samples, sequencing data, and accession numbers are

1206 described in detail in [21,40]. Briefly, the samples are comprised of individuals drawn from
1207 natural populations and sequenced in relatively large pools of 66-232 chromosomes. We
1208 aligned all data using BWA v0.7.9a-r786 [92] using the 'MEM' function and the default
1209 program parameters. For all alignments, we used version 5 of the *D. melanogaster*
1210 reference genome [93] in order to make our analysis and coordinates compatible with the
1211 *Drosophila* genome nexus [91]. We then realigned all reads using the indelrealigner tool
1212 within the GATK package [84], and we extracted the sequence pileup using samtools
1213 mpileup v1.1 [94] using the program's default parameters.
1214
1215 We extracted sites at ancestry informative positions within the reference panels, where we
1216 required that the reference panel have a minimum of 50% of individuals with a high quality
1217 genotype call in both Cosmopolitan and African reference populations. As above, ancestry
1218 informative sites were defined as those with a minimum of 20% difference in allele
1219 frequencies between the reference panels used, and we retained only ancestry informative
1220 sites for our analyses. We then produced global ancestry estimates for each chromosome
1221 arm separately for each sample using the method of Bergland *et al.* (2016). We ran our
1222 HMM for each chromosome arm and each population, and we provided the program this
1223 estimate of the ancestry proportion and the time since admixture, 1593 generations [17].
1224 We elected to provide the time since admixture because we have found that this parameter
1225 is difficult to estimate in relatively large pools (see Results). However, the program can
1226 accurately estimate LA in high ploidy samples even when the time since admixture cannot
1227 be estimated correctly (see Results).
1228

1229 **Correlation with Local Recombination Rates**

1230 To assess the correlation between local recombination rates and LA in the genome, we
1231 computed Spearman's rank sum correlation between the proportions African ancestry and
1232 the local recombination rates in windows of 100 ancestry informative markers. As above,
1233 we used the recombination rate estimates of [59]. We estimated confidence intervals using
1234 1000 block-bootstrap samples using window sizes of 100 SNPs.

1235

1236 **Robustness of LAI to Genomic Heterogeneity**

1237 To determine if there are systematic biases in LAI across the genome, we computed the
1238 mean difference in genomic windows between LA estimates for two samples form Maine
1239 and between two samples from Florida. We assessed evidence for systematic biases
1240 through the correlation between local recombination rates and differences in local ancestry
1241 inference using Spearman's rank sum correlation.

1242

1243 **Identifying LA Cline Outliers**

1244 To detect loci that show evidence for steeper ancestry clines than the genomic average, we
1245 first computed the Spearman's rank correlation between mean ancestry proportions and
1246 latitude for each chromosome arm separately. Then, for each site for which we obtained a
1247 posterior ancestry distribution for all samples, we computed the partial Spearman's rank
1248 correlation between the posterior ancestry mean and latitude while correcting for the
1249 correlation between latitude and the overall ancestry proportion. We then computed the
1250 probability of obtaining the observed partial correlation in R, which implements the
1251 approach of [95], and we retained those sites where the probability of the partial

1252 correlation between local ancestry and latitude was less than 0.005 as significant in our
1253 analysis. Although this cutoff is arbitrary, given the strong evidence for local adaptation
1254 and reproductive isolation in these populations [46,47,96], the tail of the LA cline
1255 distribution will likely be enriched for sites experiencing selection on this ancestry
1256 gradient. Due to linkage, adjacent sites show strong autocorrelation. We therefore selected
1257 the local optima for a given clinally significant LA segment (*i.e.* a tract where all positions
1258 are significantly correlated with latitude at our threshold) and retained these for analyses
1259 of outlier loci. Finally, to further reduce the effect of autocorrelation, we retained only
1260 those local optima for which no other optimum had a stronger correlation with latitude
1261 within 100,000bp on either side on the site.
1262

1263 **Identifying Low African Ancestry Outlier Loci**

1264 To identify loci with a disproportionately low proportion of African ancestry across this
1265 ancestry cline, we computed the mean African ancestry across all populations. We then
1266 selected those sites in the lowest 5% tail on each chromosome arm and selected only the
1267 local minima within 100kb windows on either side of a selected locus.
1268

1269 **Gene Ontology Analyses**

1270 We performed Gene-ontology (GO) analyses on outlier SNPs using Gowinda [97], where the
1271 background set of SNPs was all positions at which we obtained a posterior distribution in
1272 all samples (i.e. the set on which we obtained estimates of the posterior probability of
1273 African ancestry). We ran the program using default parameters, except that we included
1274 all genes within 10000bp of a focal SNP, and we performed 1e6 total GO simulations.

Seasonality of LA in the Pennsylvania Populations

To identify recurrent seasonal changes in the local ancestry, we followed an approach similar to [40]. Specifically, we fit a generalized linear model of the form

$$Mean\,Posterior\,Ancestry \sim Season + \varepsilon$$

We then recorded the estimated effect size, and probability of the observed correlation for each site in the genome at which we obtained a posterior ancestry distribution in all samples considered. To correct for multiple testing, we applied a false discovery rate correction [81] to the resulting p-value distribution.

Acknowledgements

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

References

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

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

1301 3. Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J.
1302 The genomic signature of crop-wild introgression in maize. PLoS Genet. 2013;9:
1303 e1003477–13. doi:10.1371/journal.pgen.1003477

1304 4. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and introgression
1305 between Mimulus nasutus and Mimulus guttatus. PLoS Genet. 2014;10: e1004410–
1306 15. doi:10.1371/journal.pgen.1004410

1307 5. Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, et al.
1308 Major ecological transitions in wild sunflowers facilitated by hybridization. Science.
1309 2003;301: 1211–1216. doi:10.1126/science.1086949

1310 6. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence
1311 of the neandertal genome. Science. 2010;328: 710–722.
1312 doi:10.1126/science.1188021

1313 7. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The
1314 genomic landscape of Neanderthal ancestry in present-day humans. Nature.
1315 2014;507: 354–357. doi:10.1038/nature12961

1316 8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal
1317 components analysis corrects for stratification in genome-wide association studies.
1318 Nat Genet. 2006;38: 904–909. doi:10.1038/ng1847

1319 9. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic
1320 association studies supports a contribution of common variants to susceptibility to
1321 common disease. Nat Genet. 2003;33: 177–182. doi:10.1038/ng1071

1322 10. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al.
1323 Assessing the impact of population stratification on genetic association studies. Nat
1324 Genet. 2004;36: 388–393. doi:10.1038/ng1333

1325 11. Caracristi G, Schlotterer C. Genetic differentiation between American and European
1326 Drosophila melanogaster populations could be attributed to admixture of African
1327 alleles. Mol Biol Evol. 2003;20: 792–799. doi:10.1093/molbev/msg091

1328 12. Kolbe JJ, Glor RE, Schettino LR, Lara AC, Larson A. Genetic variation increases during
1329 biological invasion by a Cuban lizard. Nature. 2004;431: 177–181.

1330 13. Consortium THG, Consortium G. Butterfly genome reveals promiscuous exchange of
1331 mimicry adaptations among species. Nature. 2012;487: 94–98.
1332 doi:10.1038/nature11041

1333 14. Racimo F, Sankararaman S, Nielsen R, Huerta-Sanchez E. Evidence for archaic
1334 adaptive introgression in humans. Nat Genet. 2015;16: 359–371.
1335 doi:10.1038/nrg3936

1336 15. Juric I, Aeschbacher S, Coop G. The strength of selection against Neanderthal
1337 introgression. bioRxiv. 2015 Oct pp. 1–24. doi:10.1101/030148

1338 16. Harris K, Nielsen R. The genetic cost of neanderthal introgression. Genetics.
1339 2016;203: 881–891. doi:10.1534/genetics.116.186890

1340 17. Pool JE. The mosaic ancestry of the Drosophila Genetic Reference Panel and the D.
1341 melanogaster reference genome reveals a network of epistatic fitness interactions.
1342 Mol Biol Evol. 2015;: msv194–16. doi:10.1093/molbev/msv194

1343 18. Schumer M, Cui R, Powell DL, Dresner R, Rosenthal GG, Andolfatto P. High-resolution
1344 mapping reveals hundreds of genetic incompatibilities in hybridizing fish species.
1345 eLife. 2014;3: 610–21. doi:10.7554/eLife.02535

1346 19. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using
1347 multilocus genotype data. Genetics. 2000;155: 945–959.

1348 20. Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture
1349 proportions from next generation sequencing data. Genetics. 2013;195: 693–702.
1350 doi:10.1534/genetics.113.154138/-/DC1

1351 21. Bergland AO, Tobler R, González J, Schmidt P, Petrov D. Secondary contact and local
1352 adaptation contribute to genome-wide patterns of clinal variation in Drosophila
1353 melanogaster. Mol Ecol. 2016;25: 1157–1174. doi:10.1111/mec.13455

1354 22. Falush D, Stephens M, Pritchard JK. Inference of population structure using
1355 multilocus genotype data: linked loci and correlated allele frequencies. Genetics.
1356 2003;164: 1567–1587.

1357 23. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in
1358 admixed populations. Am J Hum Genet. 2008;82: 290–303.
1359 doi:10.1016/j.ajhg.2007.09.022

1360 24. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling
1361 approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93:
1362 278–288. doi:10.1016/j.ajhg.2013.06.020

1363 25. Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, et al. Fast and
1364 accurate inference of local ancestry in Latino populations. Bioinformatics. 2012;28:
1365 1359–1367. doi:10.1093/bioinformatics/bts144

1366 26. Pool JE, Nielsen R. Inference of historical changes in migration rate from the lengths
1367 of migrant tracts. Genetics. 2009;181: 711–719. doi:10.1534/genetics.108.098095

1368 27. Koopman W, Li Y, Coart E. Linked vs. unlinked markers: multilocus microsatellite
1369 haplotype-sharing as a tool to estimate gene flow and introgression. Mol Ecol.
1370 2007;16: 243–256. doi:10.1111/j.1365-294X.2006.03137.x

1371 28. Patterson N, Hattangadi N, Lane B. Methods for high-density admixture mapping of
1372 disease genes. Am J Hum Genet. 2004;74: 979–1000.

1373 29. Gravel S. Population Genetics Models of Local Ancestry. Genetics. 2012;191: 607–
1374 619. doi:10.1534/genetics.112.139808

1375 30. Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The history
1376 of African gene flow into southern Europeans, Levantines, and Jews. PLoS Genet.
1377 2011;7: e1001373–13. doi:10.1371/journal.pgen.1001373

1378 31. Loh PR, Lipson M, Patterson N, Moorjani P. Inferring admixture histories of human
1379 populations using linkage disequilibrium. Genetics. 2013;193: 1233–1254.
1380 doi:10.1534/genetics.112.147330/-/DC1

1381 32. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A Genetic Atlas
1382 of Human Admixture History. Science. 2014;343: 747–751.
1383 doi:10.1126/science.1243518

1384 33. Consortium 1GP, BGI-Shenzhen, European Bioinformatics Institute, Illumina,
1385 Hospital BAW, College B, et al. An integrated map of genetic variation from 1,092
1386 human genomes. Nature. 2012. doi:10.1038/nature11632

1387 34. Pasaniuc B, Sankararaman S, Kimmel G, Halperin E. Inference of locus-specific
1388 ancestry in closely related populations. Bioinformatics. 2009;25: i213–i221.
1389 doi:10.1093/bioinformatics/btp197

1390 35. Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. Multilocus patterns of
1391 nucleotide variability and the demographic and selection history of Drosophila
1392 melanogaster populations. Genome Res. 2005;15: 790–799. doi:10.1101/gr.3541005

1393 36. Thornton K, Andolfatto P. Approximate Bayesian inference reveals evidence for a
1394 recent, severe bottleneck in a Netherlands population of Drosophila melanogaster.
1395 Genetics. 2006;172: 1607–1619. doi:10.1534/genetics.105.048223

1396 37. Li H, Stephan W. Inferring the demographic history and rate of adaptive substitution
1397 in Drosophila. PLoS Genet. 2006;2: e166. doi:10.1371/journal.pgen

1398 38. Duchen P, Živković D, Hutter S, Stephan W. Demographic inference reveals African
1399 and European admixture in the North American Drosophila melanogaster
1400 population. Genetics. 2013;193: 291–301. doi:10.1534/genetics.112.145912/-/DC1

1401 39. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, et al. Genomic
1402 variation in natural populations of Drosophila melanogaster. Genetics. 2012;192:
1403 533–598. doi:10.1534/genetics.112.142018

1404 40. Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA. Genomic evidence of
1405 rapid and stable adaptive oscillations over seasonal time scales in Drosophila. PLoS

1406 Genet. 2014;10: e1004775–19. doi:10.1371/journal.pgen.1004775

1407 41. Schrider DR, Hahn MW, Begun DJ. Parallel evolution of copy-number variation across
1408 continents in Drosophila melanogaster. Mol Biol Evol. 2016;33: 1308–1316.
1409 doi:10.1093/molbev/msw014

1410 42. Reinhardt JA, Kolaczkowski B, Jones CD, Begun DJ, Kern AD. Parallel geographic
1411 variation in Drosophila melanogaster. Genetics. 2014;197: 361–373.
1412 doi:10.1534/genetics.114.161463/-/DC1

1413 43. Fabian DK, Kapun M, Nolte V, Kofler R, Schmidt PS, Schlotterer C, et al. Genome-wide
1414 patterns of latitudinal differentiation among populations of Drosophila
1415 melanogasterfrom North America. Mol Ecol. 2012;21: 4748–4769.
1416 doi:10.1111/j.1365-294X.2012.05731.x

1417 44. Oakeshott JG, Gibson JB, Anderson PR, Knibb WR. Alcohol dehydrogenase and
1418 glycerol-3-phosphate dehydrogenase clines in Drosophila melanogaster on different
1419 continents. Evolution. 1982;36: 86–96.

1420 45. Berry A, Kreitman M. Molecular analysis of an allozyme cline: alcohol dehydrogenase
1421 in Drosophila melanogaster on the east coast of North America. Genetics. 1993;134:
1422 869–893.

1423 46. Yukilevich R, True JR. African morphology, behavior, and phermones underlie
1424 incipient sexual isolation between US and Caribbean Drosophila melanogaster.
1425 Evolution. 2008;62: 2807–2828. doi:10.1111/j.1558-5646.2008.00488.x

1426 47. Lachance J, True JR. X-autosome incompatibilities in Drosophila melanogaster: test of
1427 Haldane's rue and geographic patterns within species. Evolution. 2010;64: 3035–
1428 3046. doi:10.1111/j.1558-5646.2010.01028.x

1429 48. Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF. Genetic incompatibilities are
1430 widespread within species. Nature. 2013;504: 135–137. doi:10.1038/nature12678

1431 49. Langley CH, Crepeau M, Cardeno C, Corbett-Detig R, Stevens K. Circumventing
1432 heterozygosity: sequencing the amplified genome of a single haploid Drosophila
1433 melanogaster embryo. Genetics. 2011;188: 239–246.
1434 doi:10.1534/genetics.111.127530

1435 50. Kofler R, Betancourt AJ, Schlotterer C. Sequencing of Pooled DNA Samples (Pool-Seq)
1436 Uncovers Complex Dynamics of Transposable Element Insertions in Drosophila
1437 melanogaster. PLoS Genet. 2012;8: e1002487–16.
1438 doi:10.1371/journal.pgen.1002487

1439 51. terWengel PO, Kapun M, Nolte V, Kofler R, Flatt T, Schlotterer C. Adaptation of
1440 *Drosophila* to a novel laboratory environment reveals temporally heterogeneous
1441 trajectories of selected alleles. Mol Ecol. 2012;21: 4931–4941. doi:10.1111/j.1365-

1442 294X.2012.05673.x

1443 52. Kolaczkowski B, Kern AD, Holloway AK, Begun DJ. Genomic differentiation between
1444 temperate and tropical Australian populations of Drosophila melanogaster. Genetics.
1445 2011;187: 245–260. doi:10.1534/genetics.110.123059

1446 53. Kapun M, Schalkwyk H, McAllister B, Flatt T, Schlotterer C. Inference of chromosomal
1447 inversion dynamics from Pool-Seq data in natural and laboratory populations of
1448 Drosophila melanogaster. Mol Ecol. 2014;23: 1813–1827. doi:10.1111/mec.12594

1449 54. Zhu Y, Bergland AO, González J, Petrov DA. Empirical validation of pooled whole
1450 genome population re-sequencing in Drosophila melanogaster. PLoS ONE. 2012;7:
1451 e41901. doi:10.1371/journal.pone.0041901

1452 55. Chen GK, Marjoram P, Wall JD. Fast and flexible simulation of DNA sequence data.
1453 Genome Res. 2008;19: 136–142. doi:10.1101/gr.083634.108

1454 56. Corbett-Detig R, Jones M. SELAM: simulation of epistasis and local adaptation during
1455 admixture with mate choice. Bioinformatics. 2016;btw365.

1456 57. Liang M, Nielsen R. The lengths of admixture tracts. Genetics. 2014;197: 953–967.
1457 doi:10.1534/genetics.114.162362

1458 58. Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. The Date of Interbreeding
1459 between Neandertals and Modern Humans. PLoS Genet. 2012;8: e1002947–9.
1460 doi:10.1371/journal.pgen.1002947

1461 59. Comeron JM, Ratnappan R, Bailin S. The many landscapes of recombination in
1462 Drosophila melanogaster. PLoS Genet. 2012;8: e1002905.
1463 doi:10.1371/journal.pgen.1002905

1464 60. Moltke I, Grarup N, Jørgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, et al. A
1465 common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2
1466 diabetes. Nature. 2014;512: 190–193. doi:10.1038/nature13425

1467 61. Moltke I, Fumagalli M, Korneliussen TS. Uncovering the genetic history of the
1468 present-day Greenlandic population. Am J Hum Genet. 2015;96: 54–69.
1469 doi:10.1016/j.ajhg.2014.11.012

1470 62. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a
1471 custom genotyping array for genetic studies of metabolic, cardiovascular, and
1472 anthropometric traits. PLoS Genet. 2012;8: e1002793–12.
1473 doi:10.1371/journal.pgen.1002793

1474 63. Corbett-Detig RB, Hartl DL. Population genomics of inversion polymorphisms in
1475 Drosophila melanogaster. 2012;8: e1003056–15. doi:10.1371/journal.pgen.1003056

1476 64. Kapun M, Fabian DK, Goudet J, Flatt T. Genomic evidence for adaptive inversion
1477 clines in Drosophila melanogaster. Mol Biol Evol. 2016;33: 1317–1336.
1478 doi:10.1093/molbev/msw016

1479 65. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The
1480 Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482: 173–178.
1481 doi:10.1038/nature10811

1482 66. Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, et al. Natural
1483 variation in genome architecture among 205 Drosophila melanogaster Genetic
1484 Reference Panel lines. Genome Res. 2014;24: 1193–1208.
1485 doi:10.1101/gr.171546.113

1486 67. Corbett-Detig RB, Cardeno C, Langley CH. Sequence-based detection and breakpoint
1487 assembly of polymorphic inversions. Genetics. 2012;192: 131–137.
1488 doi:10.1534/genetics.112.141622

1489 68. Kulathinal RJ, Stevison LS, Noor MAF. The genomics of speciation in Drosophila:
1490 diversity, divergence, and introgression estimated using low-coverage genome
1491 sequencing. PLoS Genet. 2009;5: e1000550–7. doi:10.1371/journal.pgen.1000550

1492 69. Krimbas CB, Powell JR. Drosophila Inversion Polymorphism. CRC Press; 1992.

1493 70. Aulard S, David JR, Lemenieur F. Chromosomal inversion polymorphism in
1494 Afrotropical populations of *Drosophila melanogaster*. Genet Res. 2002;79: 49–63.
1495 doi:10.1017/S0016672301005407

1496 71. Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Zaitlen N, Eng C, et al.
1497 Analysis of Latino populations from GALA and MEC studies reveals genomic loci with
1498 biased local ancestry estimation. Bioinformatics. 2013;29: 1407–1415.

1499 72. Ting CT, Takahashi A, Wu CI. Incipient speciation by sexual isolation in Drosophila:
1500 concurrent evolution at multiple loci. PNAS. 2001;98: 6709–6713.

1501 73. Hollocher H, Ting CT, Wu ML, Wu CI. Incipient speciation by sexual isolation in
1502 Drosophila melanogaster: extensive genetic divergence without reinforcement.
1503 Genetics. 1997;147: 1191–1201.

1504 74. Hollocher H, Ting CT, Pollack F, Wu CI. Incipient speciation by sexual isolation in
1505 Drosophila melanogaster: variation in mating preference and correlation between
1506 sexes. Evolution. 1997;51: 1175–1181.

1507 75. Coyne JA. Genetics and speciation. Nature. 1992;335: 511–515.

1508 76. Coyne JA, Orr HA. Patterns of speciation in Drosophila. Evolution. 1989;43: 362–381.

1509 77. Charlesworth B, Coyne JA, Barton NH. The relative rates of evolution of sex

1510 chromosomes and autosomes. American Naturalist. 1987;130: 113–146.

1511 78. Levine MT, Begun DJ. Evidence of Spatially Varying Selection Acting on Four
1512 Chromatin-Remodeling Loci in Drosophila melanogaster. Genetics. 2008;179: 475–
1513 485. doi:10.1534/genetics.107.085423

1514 79. Ellis LL, Carney GE. Socially-Responsive Gene Expression in Male Drosophila
1515 melanogaster Is Influenced by the Sex of the Interacting Partner. Genetics. 2011;187:
1516 157–169. doi:10.1534/genetics.110.122754

1517 80. Edwards AC, Zwarts L, Yamamoto A, Callaerts P, Mackay TF. Mutations in many genes
1518 affect aggressive behavior in Drosophila melanogaster. BMC Biol. 2009;7: 29–13.
1519 doi:10.1186/1741-7007-7-29

1520 81. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and
1521 powerful approach to multiple testing. Journal of the Royal Statistical Society.
1522 1995;57: 289–300.

1523 82. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. SNP calling, genotype calling,
1524 and sample allele frequency estimation from new-generation sequencing data. PLoS
1525 ONE. 2012;7: e37558–11. doi:10.1371/journal.pone.0037558

1526 83. Fumagalli M, Vieira FG, Linderoth T, Nielsen R. ngsTools: methods for population
1527 genetics analyses from next-generation sequencing data. Bioinformatics. 2014;30:
1528 1486–1487.

1529 84. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework
1530 for variation discovery and genotyping using next-generation DNA sequencing data.
1531 Nat Genet. 2011;43: 491–498. doi:10.1038/ng.806

1532 85. Baird SJE, Barton NH, Etheridge AM. The distribution of surviving blocks of an
1533 ancestral genome. Theoretical Population Biology. 2003;64: 451–471.
1534 doi:10.1016/S0040-5809(03)00098-4

1535 86. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic ancestry
1536 of north Africans supports back-to-Africa migrations. PLoS Genet. 2012;8:
1537 e1002397–11. doi:10.1371/journal.pgen.1002397

1538 87. Marjoram P, Wall JD. Fast "coalecsent" simulation. BMC Genet. 2006;7: 16–9.
1539 doi:10.1186/1471-2156-7-16

1540 88. Kiefer J. Sequential minimax search for a maximum. Proceedings of the American
1541 Mathematical Society. 1953;4: 502–506.

1542 89. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint
1543 demographic history of multiple populations from multidimensional SNP frequency
1544 data. PLoS Genet. 2009;5: e1000695–11. doi:10.1371/journal.pgen.1000695

1545 90. Pool JE, Aquadro CF. History and Structure of Sub-Saharan Populations of Drosophila
1546 melanogaster. Genetics. 2006;174: 915–929. doi:10.1534/genetics.106.058693

1547 91. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, et al. The
1548 Drosophila Genome Nexus: a population genomic resource of 623 Drosophila
1549 melanogaster genomes, including 197 from a single ancestral range population.
1550 Genetics. 2015;199: 1229–1241. doi:10.1534/genetics.115.174664/-/DC1

1551 92. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
1552 MEM. arXiv preprint arXiv:13033997. 2013.

1553 93. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG. The genome
1554 sequence of Drosophila melanogaster. Science. 2000;287: 2185–2195.

1555 94. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
1556 Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
1557 doi:10.1093/bioinformatics/btp352

1558 95. Best DJ, Roberts DE. Algorithm AS 89: the upper tail probabilities of Spearman's rho.
1559 Journal of the Royal Statistical Society Series C. 1975;24: 377–379.

1560 96. Kao JY, Lymer S, Hwang SH, Sung A, Nuzhdin SV. Postmating reproductive barriers
1561 contribute to the incipient sexual isolation of the United States and Caribbean
1562 Drosophila melanogaster. Ecol Evol. 2015;5: 3171–3182. doi:10.1002/ece3.1596

1563 97. Kofler R, Schlötterer C. Gowinda: unbiased analysis of gene set enrichment for
1564 genome-wide association studies. Bioinformatics. 2012;28: 2084–2085.
1565 doi:10.1093/bioinformatics/bts315

1566

1567 **Figure Legends**

1568 **Figure 1.** The effect of increasing stringency with ancestral LD pruning. From left to right,
1569 ancestry proportions are 0.1, 0.25, 0.5, 0.75 and 0.9. |r| cutoffs are: none (red), 1.0 (orange),
1570 0.9 (yellow), 0.8 (green), 0.7 (dark blue), 0.6 (cyan), 0.5 (indigo), and 0.4 (violet). The solid
1571 line indicates the expectation for unbiased time estimation. All read data were simulated
1572 with ploidy = 1. True admixture time was drawn from a uniform (0, 2000) distribution.

1573

1574 **Figure 2.** Time estimates and accuracy statistics for samples of varying ploidies. From left
1575 to right, ancestry proportions are 0.1, 0.25, 0.5, 0.75 and 0.9. Each sample ploidy is
1576 represented by one point color with ploidy one (black), two (red), ten (blue) and twenty
1577 (green). From top to bottom, each row is the estimated time in generations, the proportion
1578 of sites where the true state is within the 95% credible interval, the width of the 95%
1579 credible interval, the mean posterior error, and the proportion of sites where the maximum
1580 likelihood estimate is equal to the true state.
1581
1582 **Figure 3.** Accuracy of the HMM for samples of high ploidy. The 95% credible interval
1583 (shaded blue region), and the posterior mean (red) contrasted with the true ancestry
1584 frequencies (black). Simulated data were generated with an admixture time of 1500
1585 generations, an ancestry proportion of 0.2, a sample ploidy of 100, and a mean sequencing
1586 depth of 25.
1587
1588 **Figure 4.** Admixture time estimates for simulated data consistent with variation present in
1589 modern European and African populations. From left to right, m = 0.1, m = 0.25, m = 0.5, m
1590 = 0.75, m = 0.9. The top row is completely phased chromosomes and the bottom row is
1591 unphased diploid data.
1592
1593 **Figure 5.** Local ancestry of inversion bearing chromosomes (red) compared with those of
1594 standard arrangement chromosomes (black) for the same chromosome arm. Positions of
1595 inversion breakpoints, as reported in [63,67] are shown as vertical dashed lines.
1596

1597 **Figure 6.** The relationship between the proportion of African ancestry proportion and local
1598 recombination rates in 100 ancestry informative SNP windows within the Raleigh, NC
1599 population (left). The correlation between the proportion of African ancestry proportion
1600 and local recombination rates in 100 ancestry informative SNP windows in all populations
1601 assayed (right). Lines indicate the 95% confidence interval obtained via block bootstrap
1602 replicates (see Methods).
1603
1604 **Figure 7.** The partial correlation between LA and latitude with correction for chromosome-
1605 wide ancestry proportions. Sites for which the probability of the observed clinal
1606 relationship was less than 0.005 were retained as significant (red). Inversion breakpoints
1607 for inversions that are at polymorphic frequencies on this ancestry cline are shown as
1608 dotted blue lines.
1609
1610 **Figure 8.** The mean African ancestry proportion across all populations on the ancestry
1611 cline for chromosome arms 2L, 2R, 3L, 3R, and X (top to bottom). Local minima outlier loci
1612 are shown in red (see Methods).
1613
1614 **Supplemental Figure S1.** Comparison between LAI using the full ancestral recombination
1615 graph via forward-time simulations (red) with those from independent and identically
1616 distributed draws from the SMC' distribution (black). Simulations were conducted using an
1617 ancestry proportion of 0.25 and population size of 10,000 hermaphroditic individuals.
1618

1619 **Supplemental Figure S2.** Effects of unknown admixed population sizes on LAI. All LAI was
1620 conducted assuming the true population size was 10,000. Simulated population sizes were
1621 100 (black), 1,000 (red), 10,000 (blue) and 100,000 (green). Ploidy 1 on the right, ploidy 2
1622 on the left. From top to bottom, rows are the estimated time of admixture, the proportion of
1623 sites where the true state is within the 95% credible interval, the width of the 95% credible
1624 interval, the mean posterior error, and the proportion of times that the maximum
1625 likelihood estimate is equal to the true state. For all simulations, the ancestry proportion
1626 was equal to 0.5.
1627
1628 **Supplemental Figure S3.** LAI accuracy when admixture times are increasingly ancient.
1629 Here, ancestry proportions are 0.5 (black), 0.25 (blue), 0.1 (violet), 0.75 (orange) and 0.9
1630 (red). From top to bottom, statistics plotted are estimated time, the proportion of sites
1631 where the true ancestry frequency is within the 95% credible interval, the mean 95%
1632 credible interval width, mean posterior error, and the proportion of times that the
1633 maximum likelihood estimate is correct.
1634
1635 **Supplemental Figure S4.** The effects of reference panel size on LAI and time estimation
1636 using the HMM. Here, we compare reference panels of size 100 (blue) with reference
1637 panels of size 10 (black). From left to right, ancestry proportions are 0.1, 0.25, 0.5, 0.75 and
1638 0.9. From top to bottom the plotted statistics are estimated time, proportion in the 95%
1639 credible interval, the average width of the 95% credible interval, the mean posterior error,
1640 and the proportion of sites where the maximum likelihood ancestry estimate is correct.
1641

1642 **Supplemental Figure S5.** Accuracy of time estimation and LAI when reference populations
1643 are increasingly divergent from the source of the admixture pulses. In columns are
1644 divergence times between ancestral populations (in units of 4Ne) of 0, 0.0005, 0.001, 0.002.
1645 From top to bottom the plotted statistics are estimated time, proportion in the 95%
1646 credible interval, the average width of the 95% credible interval, the mean posterior error,
1647 and the proportion of sites where the maximum likelihood ancestry estimate is correct.
1648
1649 **Supplemental Figure S6.** Comparison of the proportion of sites where the maximum
1650 likelihood ancestry estimate of local ancestry is correct between WinPop and our method.
1651 WinPop was run with default parameters (black), and with LD pruned in the ancestral
1652 populations, but not in the admixed population (red). Our method was run with default
1653 parameters (blue), but with the time since admixture and correct ancestry proportion
1654 supplied to our program as these parameters are required by WinPop.
1655
1656 **Supplemental Figure S7.** Bias in LAI due to uncertainty in *t*. The posterior probability
1657 estimated using RFMix of European ancestry at a given site in the genome assuming t = 5
1658 (black) and assuming t = 20 (red) for a sample representative of the average difference (top
1659 left) and a more extreme example (top right). The distribution of differences in mean Inuit
1660 ancestry for all samples (bottom left) using RFMix. The log likelihood of each time since
1661 admixture as computed using our method (bottom right), which shows a clear optimum at
1662 6-7 generations since admixture. All analyses were restricted to SNPs on chromosome 10.
1663

1664 **Supplemental Figure S8.** Bias in LAI and time estimation due to incorrect estimation of *m.*
1665 On the left, true m is 0.1and on the right true m is 0.5. Supplied m varies across 0.05 to 0.95.
1666 From top to bottom, the plotted statistics are estimated t, proportion in the 95% confidence
1667 interval, mean 95% confidence interval width, mean posterior error and the proportion of
1668 sites where the maximum likelihood estimate is correct. All plots include ploidy one
1669 (back), ploidy two (red), ploidy ten (blue), and ploidy twenty (green).
1670
1671 **Supplemental Figure S9.** Bias in LAI and time estimation due to incorrect assumptions of
1672 *t.* On the left, true *t* is 100 and on the right true *t* is 1000. Supplied t varies across 100 to
1673 2000 generations. From top to bottom, the plotted statistics are estimated t, proportion in
1674 the 95% confidence interval, mean 95% confidence interval width, mean posterior error
1675 and the proportion of sites where the maximum likelihood estimate is correct. All plots
1676 include ploidy one (back), ploidy two (red), ploidy ten (blue), and ploidy twenty (green).
1677
1678 **Supplemental Figure S10.** Estimates of *t* obtained from block bootstrap replicates for
1679 populations that have admixed for 1000 (top), and 2000 (bottom) generations. From left to
1680 right, sample ploidies are 1, 2, 10, and 20. For both simulations, m = 0.5.
1681
1682 **Supplemental Table S1.** Comparison of run times for various demographic models and
1683 sample ploidies using this method.
1684
1685 **Supplemental Table S2.** LA clinality in the genomic intervals immediately surrounding
1686 breakpoints of known polymorphic inversions.

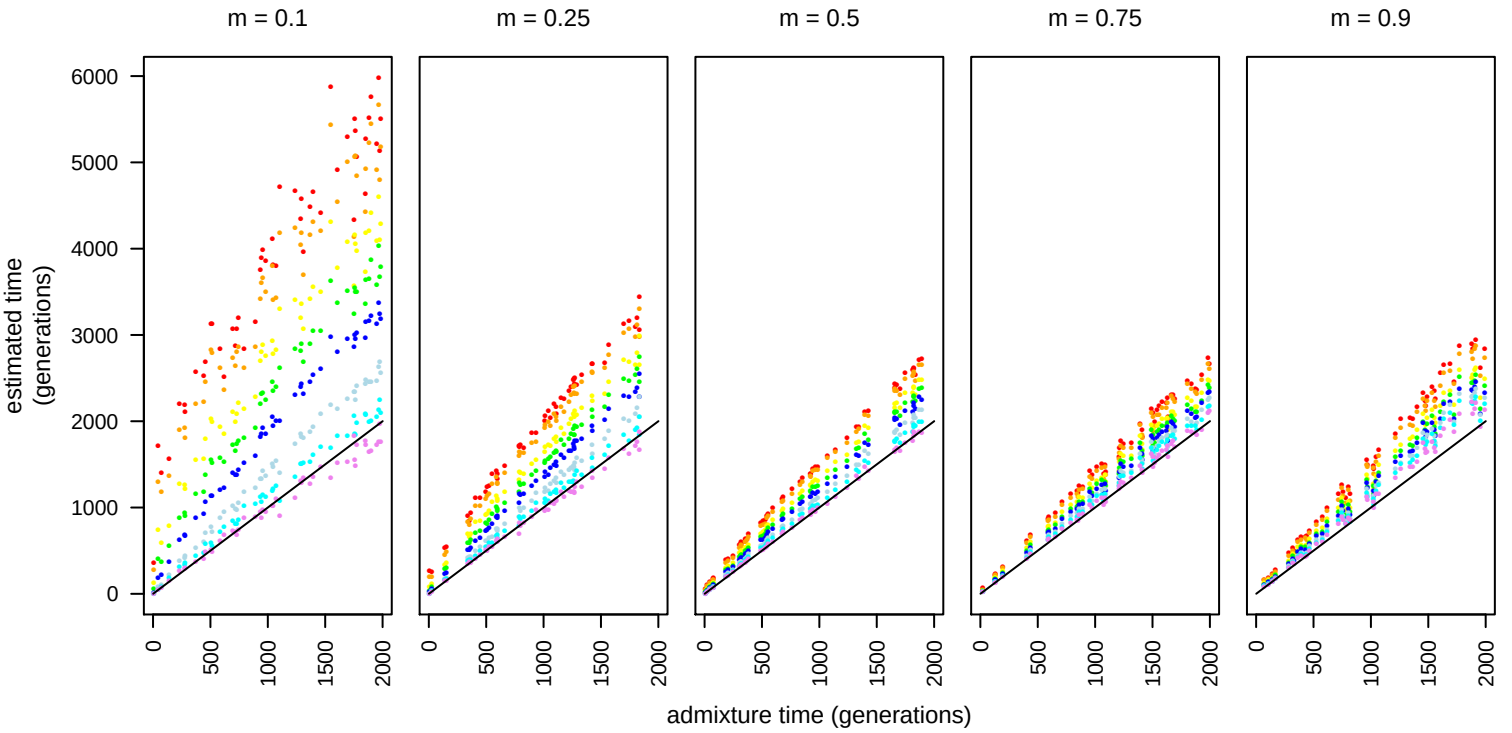
1687

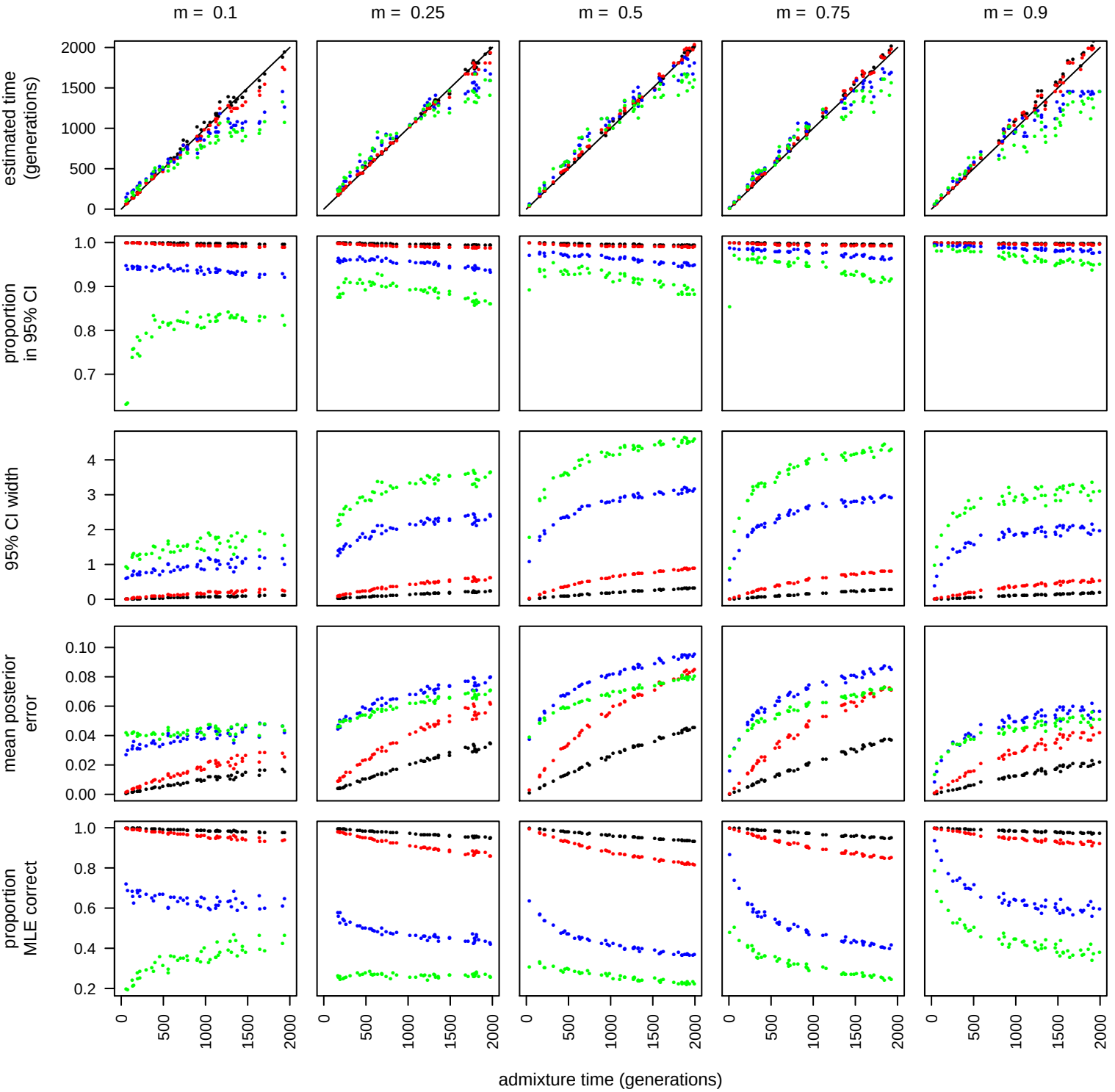
1688 **Supplemental Table S3.** Results of GO analysis of 80 identified LA clinal outlier loci.

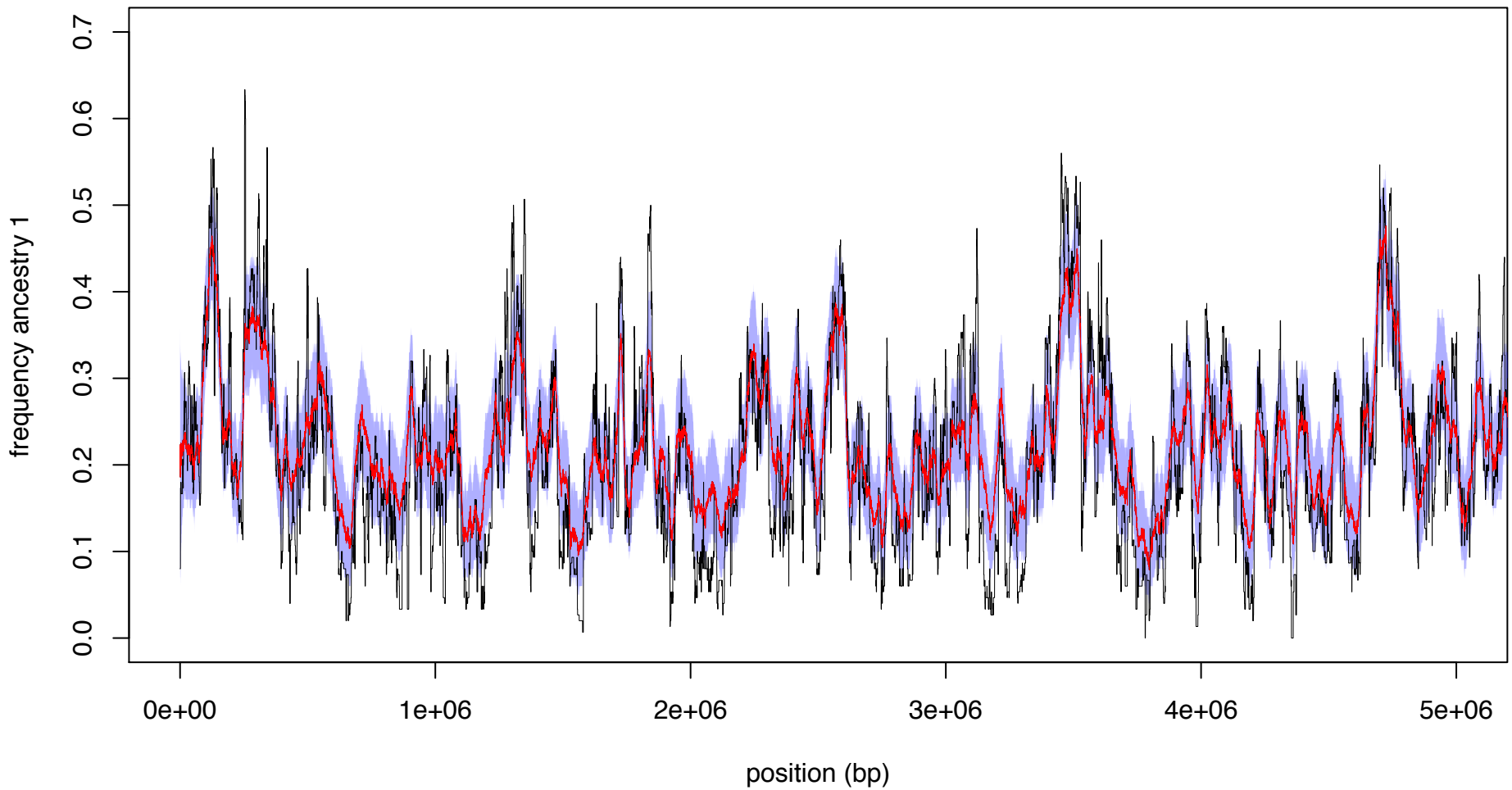
1689

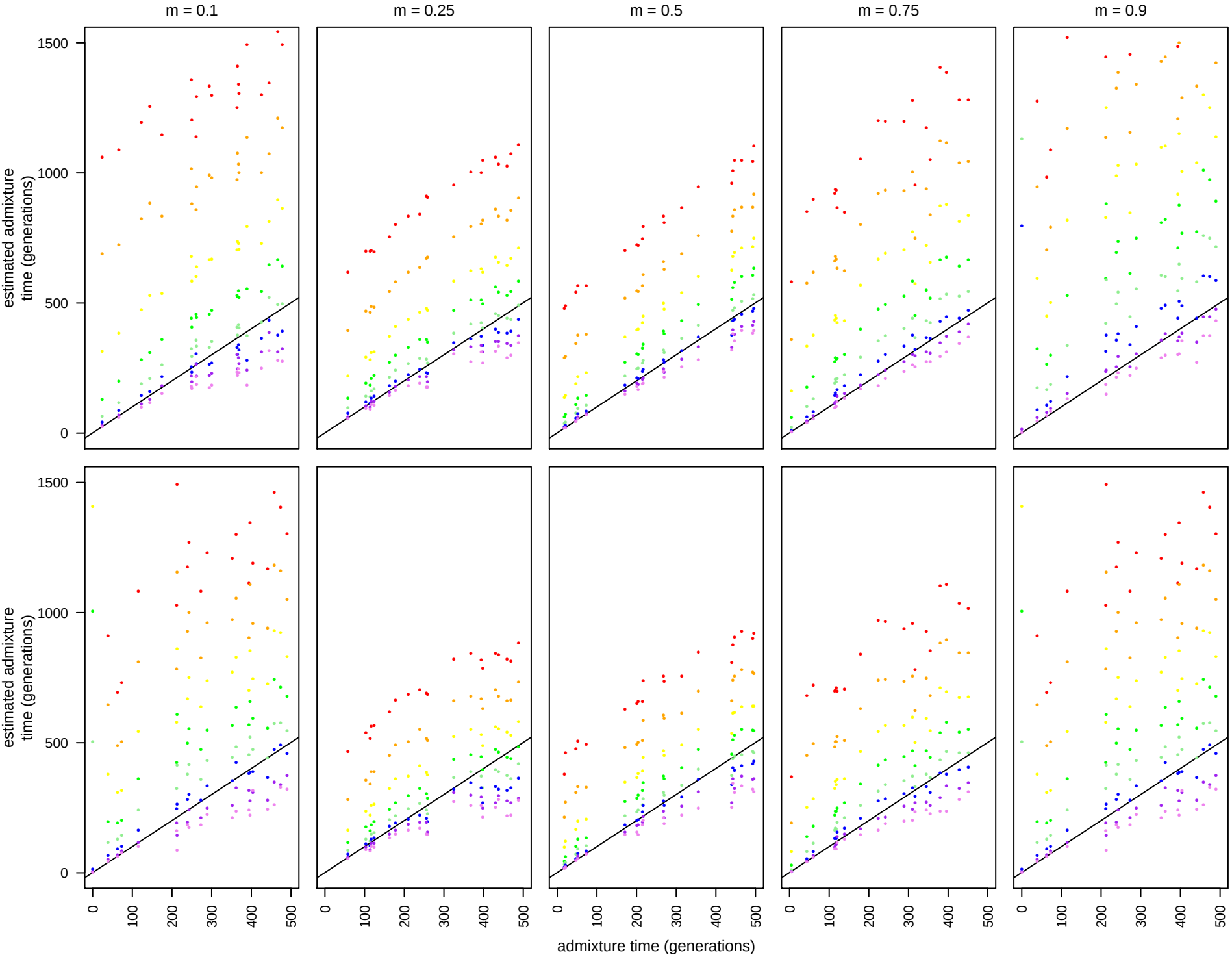
1690 **Supplemental Table S4.** Results of GO analysis of 84 identified low African ancestry

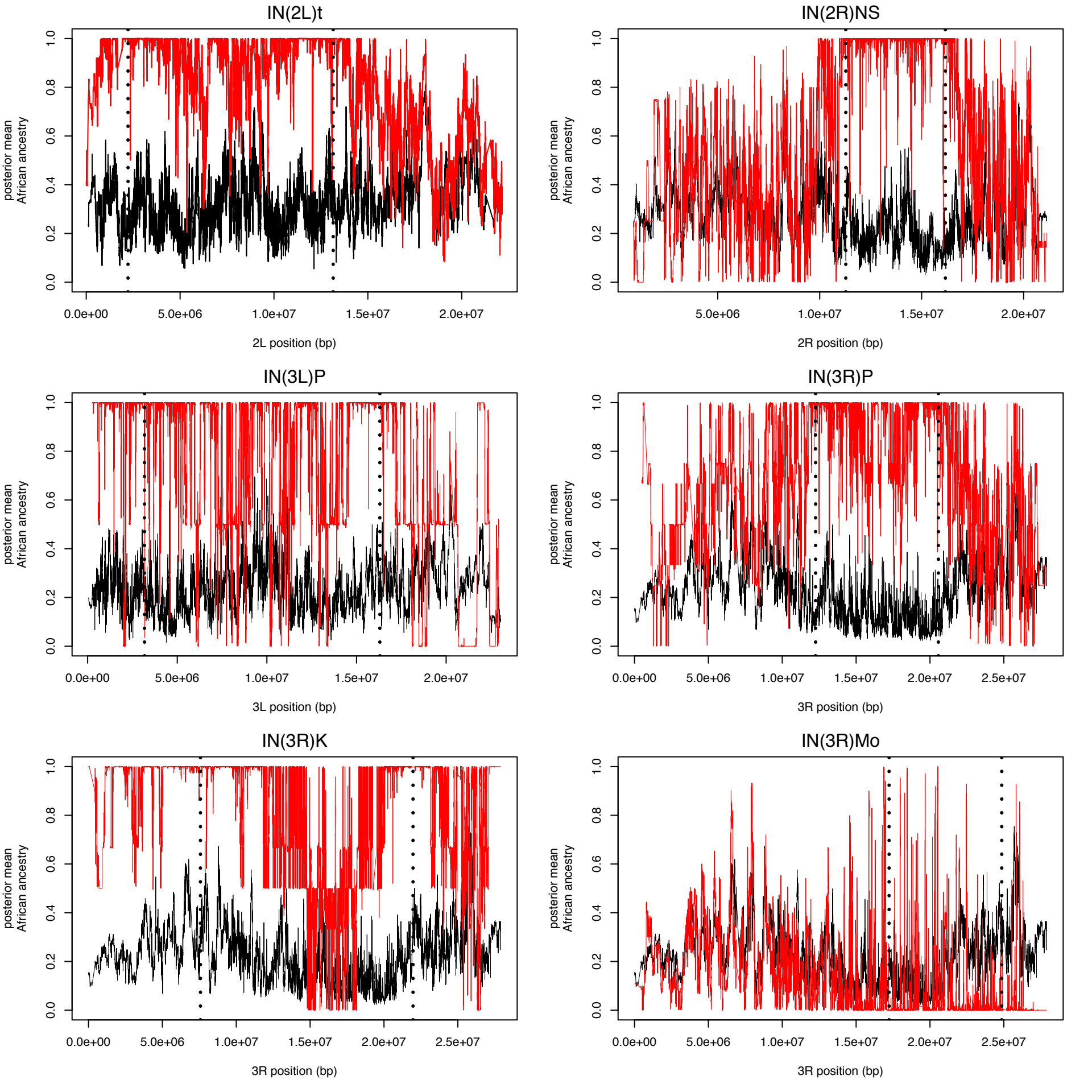
1691 outlier loci.

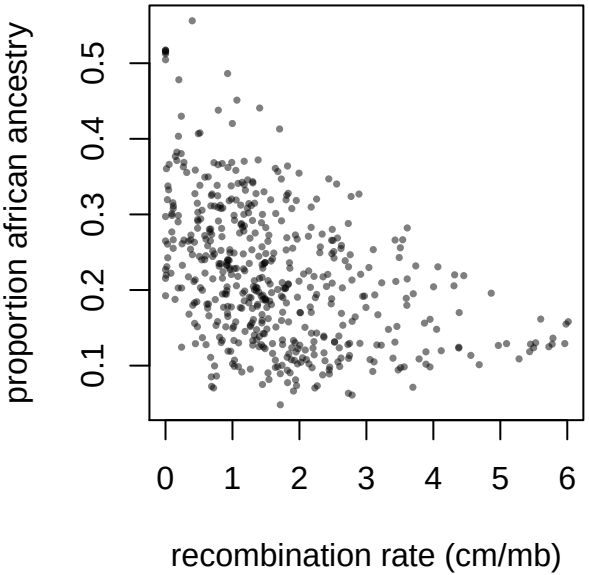

m = 0.1
m = 0.25
m = 0.5
m = 0.75
m = 0.9
estimated time
(generations)
0
1000
2000
3000
4000
5000
6000
0
500
1000
1500
2000
admixture time (generations)

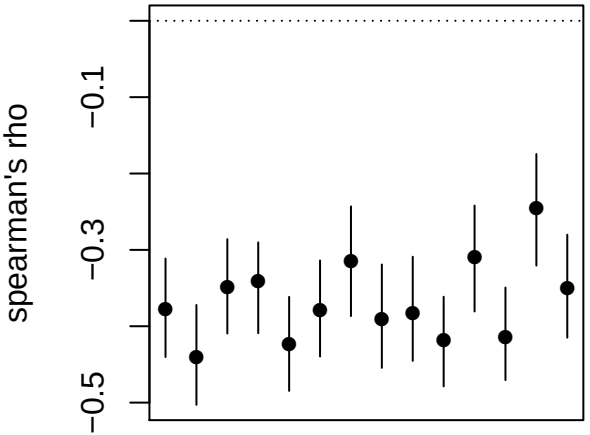

m = 0.1
m = 0.25
m = 0.5
m = 0.75
m = 0.9
estimated time (generations)
proportion in 95% CI
95% CI width
mean posterior error
proportion MLE correct
admixture time (generations)

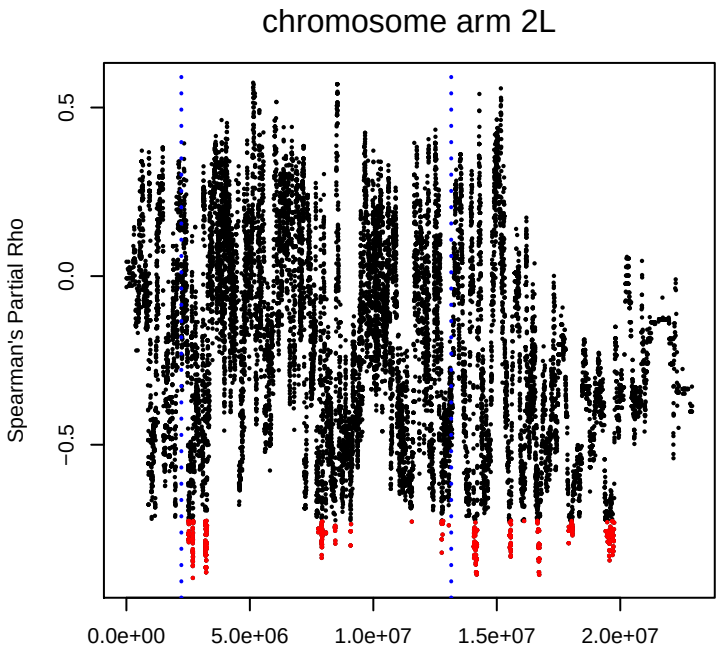

frequency ancestry 1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
position (bp)

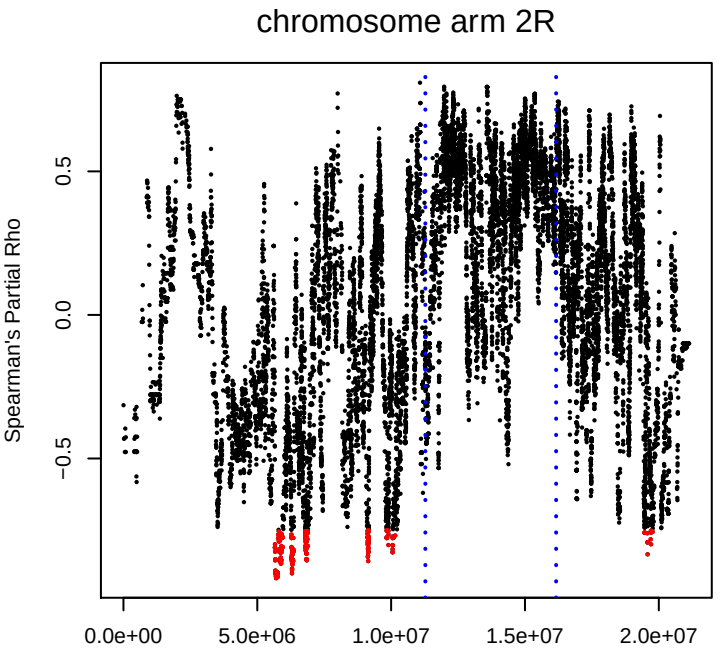

m = 0.1
m = 0.25
m = 0.5
m = 0.75
m = 0.9
estimated admixture time (generations)
0
500
1000
1500
0
100
200
300
400
500
admixture time (generations)

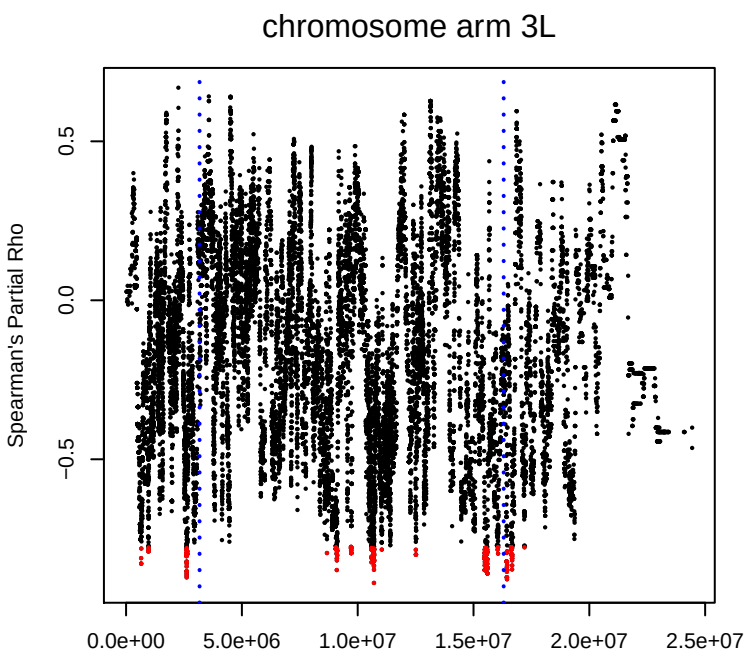

IN(2L)t
IN(2R)NS
IN(3L)P
IN(3R)P
IN(3R)K
IN(3R)Mo
posterior mean
African ancestry
2L position (bp)
2R position (bp)
3L position (bp)
3R position (bp)
0.0
0.2
0.4
0.6
0.8
1.0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07

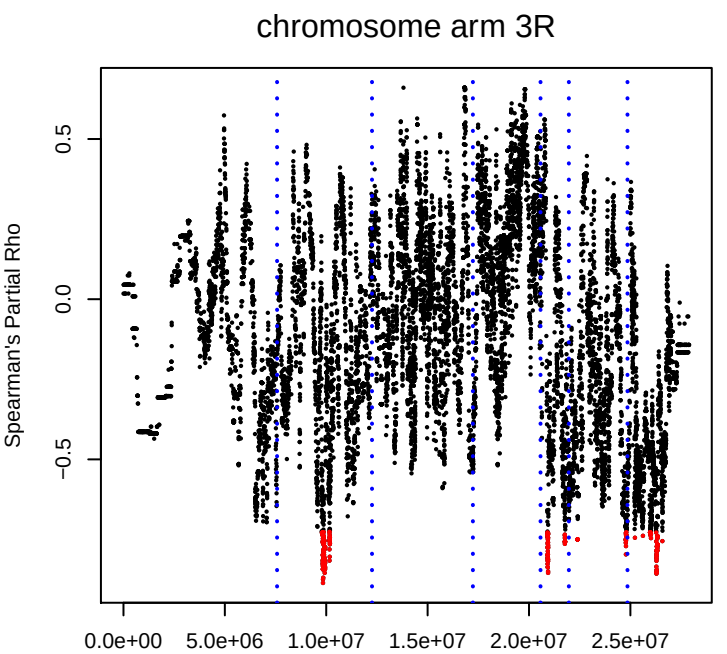

proportion african ancestry
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
recombination rate (cm/mb)

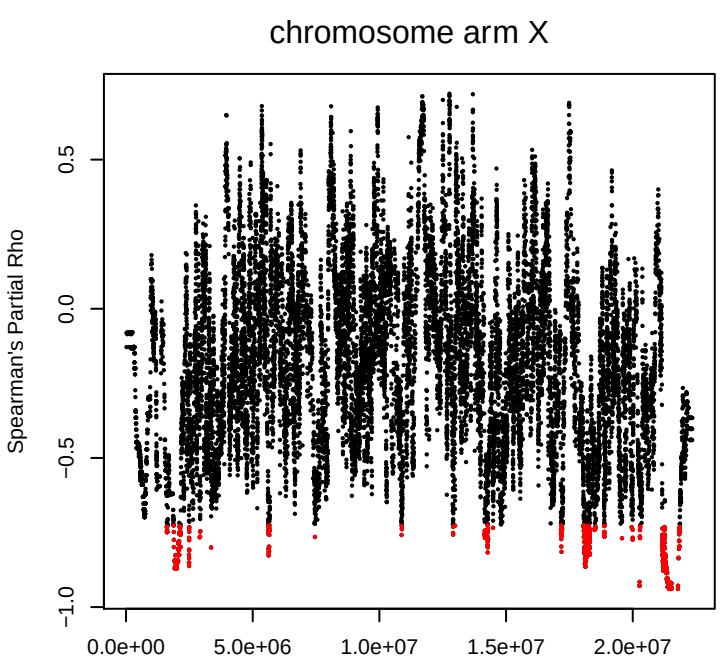

spearman's rho
−0.1
−0.3
−0.5

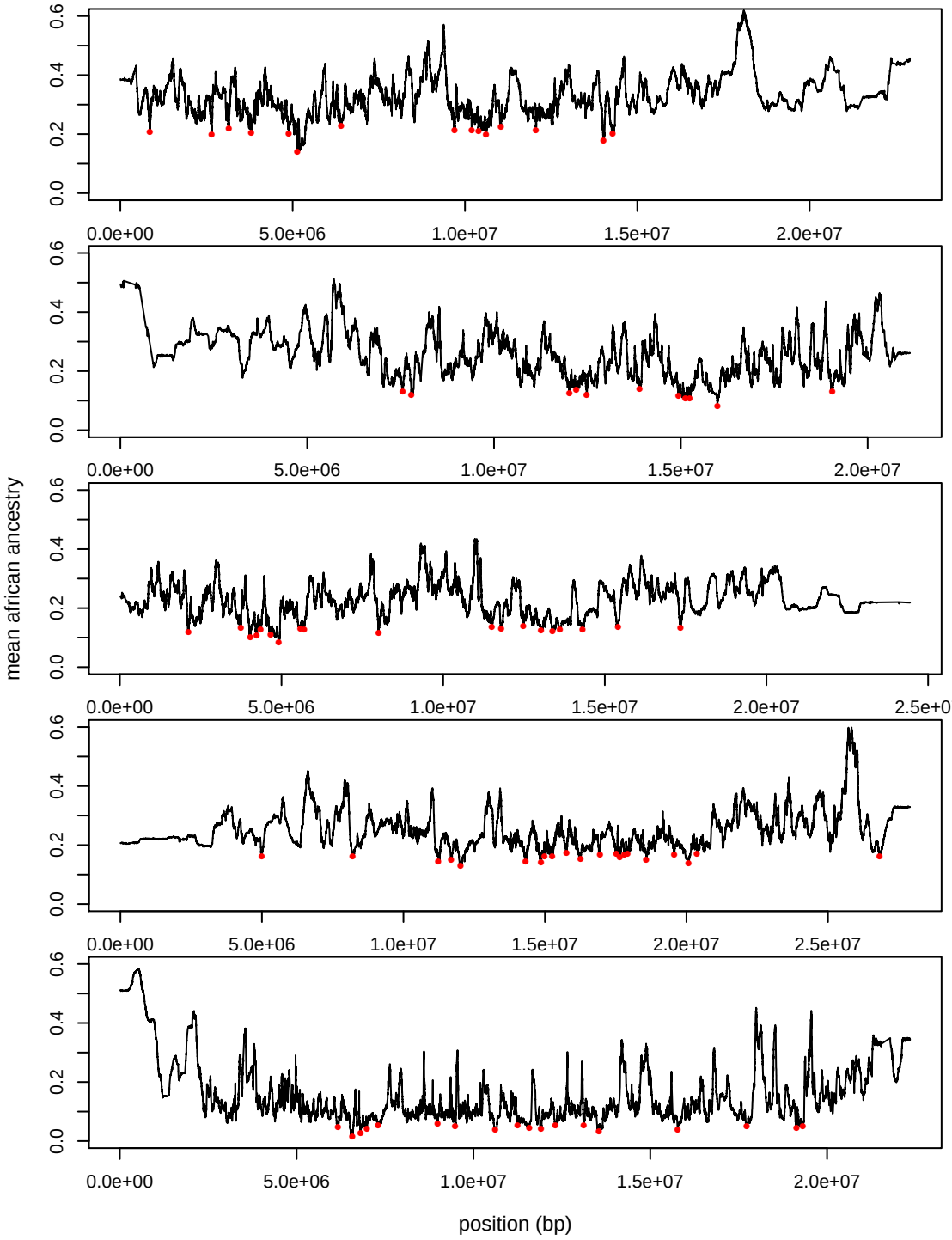

chromosome arm 2L
Spearman's Partial Rho
0.5
0.0
−0.5
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07


chromosome arm 2R
Spearman's Partial Rho
0.5
0.0
−0.5
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07


chromosome arm 3L
Spearman's Partial Rho
0.5
0.0
−0.5
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07


chromosome arm 3R
Spearman's Partial Rho
0.5
0.0
−0.5
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07


chromosome arm X
Spearman's Partial Rho
0.5
0.0
−0.5
−1.0
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07


mean african ancestry
0.0
0.2
0.4
0.6
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
position (bp)