

A new Approximate Bayesian Computation framework to distinguish among complex evolutionary models using whole-genome data

Silvia Ghirotto^{*§}, Maria Teresa Vizzari^{*}, Francesca Tassi, Guido Barbujani and
Andrea Benazzo[§]

Department of Life Sciences and Biotechnology, University of Ferrara, 44121 Ferrara, Italy

^{*} these authors contributed equally to this work

[§] correspondence should be addressed to andrea.benazzo@unife.it and silvia.ghirotto@unife.it

Abstract

Inferring past demographic histories is crucial in population genetics, and the amount of complete genomes now available for many species should in principle facilitate this inference. In practice, however, the available inferential methods suffer from severe limitations. Although hundreds complete genomes can be simultaneously analyzed, complex demographic processes can easily exceed computational constraints, and there are no standard procedures to make sure that the estimates obtained are reliable. Here we present ABC-SeS, a new Approximate Bayesian Computation (ABC) framework, based on the Random Forest algorithm, to infer complex past population processes using complete genomes. All possible pairs of populations are compared, and the data are summarized by the full genomic distribution of the four mutually exclusive categories of segregating sites, a set of statistics fast to compute even from unphased genome data. We constructed an efficient ABC pipeline and tested how accurately it allows one to recognize the true model among models of increasing complexity, using simulated data and taking into account different sampling strategies in terms of number of individuals analyzed, number and size of the genetic loci considered. Once assessed the power of ABC-SeS in the comparison of even complex models, we applied it to the analysis of real data, testing models on the dispersal of anatomically modern humans out of Africa and exploring the evolutionary relationships of the three species of

26 Orangutan inhabiting Borneo and Sumatra. The flexibility of our ABC framework, combined with
27 the power provided by the set of statistics proposed pave the way for reliable inference of past
28 population processes for any species for which high coverage genomes are available.

29

30 **Authors' summary**

31 We propose a new method for inferring demographic history from whole-genome data, ABC-SeS.
32 The idea is to analyze the distributions of segregating sites within an Approximate Bayesian
33 Computation framework, using a Random Forest approach. Extensive simulation shows that ABC-
34 SeS can handle very large datasets and identify the model that generated the data, even when
35 population relationships are complicated by factors such as admixture. We also provide guidelines
36 concerning the sampling strategies leading to the most accurate inference. As examples of possible
37 applications, we show: (1) that complex alternative models on human dispersal out of Africa can
38 now be formally tested and compared by ABC-SeS; and (2) that the evolutionary relationships
39 among three orangutan species can now be better assessed by ABC-SeS than by alternative
40 methods.

41

42 **Introduction**

43 A robust reconstruction of the demographic dynamics of a species is important both to
44 improve our knowledge about the past and to disentangle the effects of demography from those of
45 natural selection [1–3]. In recent years, thousands of modern and ancient complete genome
46 sequences have become available, potentially containing vast amounts of information about the
47 evolutionary history of populations [4–10]. However, these genomes do not speak by themselves; to
48 extract the evolutionary information they contain, appropriate inferential statistical methods are
49 required. Some methods based on the Sequential Markovian Coalescent (SMC) model [11], became
50 popular among population geneticists due to their ability to infer population size changes through
51 time (PSMC [12]) and divergence times (MSMC [13]), and to scale well on whole genome

52 sequences. Under these approaches, the local density of heterozygote sites along chromosomes is
53 used to estimate the times of the most recent common ancestor (TMRCA) of genomic regions
54 separated by recombination, thus providing insight into ancestral population sizes and the timing of
55 past divergence processes. These estimates are often used to indirectly support hypotheses
56 regarding the evolution of the studied organisms. Albeit sophisticated, these methods present some
57 limitations; the temporal resolution of the inferred demographic events seems to be strongly
58 dependent on the number of individuals included, with poor performance in the recent past
59 especially when analyzing single individuals. Moreover, these methods assume no gene flow among
60 the investigated populations, which in many case is plainly implausible. The consequences on the
61 inferential process of violation of this assumption are still under investigation [14].

62 Other methods infer demographic parameters via the diffusion approximation [15], or
63 coalescent simulations [16], from the site frequency spectrum (SFS) computed on large genomic
64 datasets. The SFS records the observed number of polymorphisms segregating at different
65 frequencies in a sample of n individuals and it is generally computed over a certain number of
66 genomic regions where no influence of natural selection is assumed. The expectation of the SFS
67 under different evolutionary scenarios could be approximated by the diffusion theory (as
68 implemented e.g. in *dadi*) or directly via coalescent simulations (as in *fastsimcoal*); alternative
69 demographic histories can be explicitly compared via e.g. AIC [17]. Still, there are limits to the
70 complexity of models that can be analyzed, and AIC-like approaches can only be used to
71 understand which modifications significantly improve the model. Therefore, through these
72 approaches, it is not easy to evaluate whether and to what extent the compared models can actually
73 be distinguished from each other, and hence quantify the strength of the support associated to the
74 best model. Indeed, the only available procedure to assess the model identifiability requires the
75 analysis of many datasets simulated under known demographic conditions, which can be
76 computationally prohibitive, in particular for complex evolutionary scenarios [16].

77 Recently, an inferential method that couples the ability of the SMC to deal with whole

78 genome sequences and the population signal gathered from the SFS has been developed (SMC++
79 [18]). Under this inferential framework, both the genomic and the SFS variation are jointly used to
80 estimate population size trajectories through time, as well as the divergence time between pairs of
81 populations. Although this approach seems to scale well on thousands of unphased genomes, it is
82 based on the same assumption of classical SMC methods (with populations evolving
83 independently), which severely limits its use whenever gene flow cannot be ruled out.

84 One powerful and flexible way to quantitatively compare alternative models and estimating
85 model's parameters relies on the Approximate Bayesian Computation (ABC) methods. Under these
86 methods, the likelihood functions do not need be specified, because posterior distributions can be
87 approximated by simulation, even under complex (and hence realistic) population models,
88 incorporating prior information. The genetic data, both observed and simulated, are summarized by
89 the same set of “sufficient” summary statistics, selected to be informative about the genealogic
90 processes under investigation. The ability of the framework to distinguish among the alternative
91 demographic models tested and the quality of the results can be evaluated with rather limited
92 additional effort (for a review see e.g. [19]). Although ABC has the potential to deal with complex
93 and realistic evolutionary scenarios, its application to the analysis of large genomic datasets, such as
94 complete genomes, is still problematic. In its original formulation, indeed, the ABC procedure
95 requires the simulation of millions data sets of the same size of those observed, which becomes
96 computationally very expensive as the dataset increases in size, or when many models need be
97 compared. In addition, there is a problem with the choice of the summary statistics describing both
98 observed and simulated data, as recognized since the first formal introduction of ABC [20,21].
99 Increasing the number of summary statistics, indeed, makes it easier to choose the best model, but
100 inevitably reduces the accuracy of the demographic inference (this problem is referred to as the
101 “curse of dimensionality” [22]). Ideally, the good practice would be to select a set of summary
102 statistics that is both low-dimensional and highly informative on the demographic parameters
103 defining the model. In practice, however, this problem has never really been solved, despite several

104 serious attempts [23].

105 Recently, a new ABC framework has been developed based on a machine-learning tool
106 called Random Forest (ABC-RF [24]). Under ABC-RF, the Bayesian model selection is rephrased
107 as a classification problem. At first, the classifier is constructed from simulations from the prior
108 distribution via a machine learning RF algorithm. Once the classifier is constructed and applied to
109 the observed data, the posterior probability of the resulting model can be approximated through
110 another RF that regresses the selection error over the statistics used to summarize the data. The RF
111 classification algorithm has been shown to be insensitive both to the correlation between the
112 predictors (in case of ABC, the summary statistics) and to the presence of relatively large numbers
113 of noisy variables. This means that even choosing a large collection of summary statistics, the
114 correlation between some of them and others (which may be uninformative about the models
115 tested), have no consequences on the RF performance, and hence on the accuracy of the inference.
116 Moreover, compared to the standard ABC methods, the RF algorithm performs well with a radically
117 lower number of simulations (from millions to tens of thousands per model). These properties make
118 the new ABC-RF algorithm of particular interest for the statistical analysis of massive genetic
119 datasets.

120 In this paper we present ABC-SeS (where SeS stands for Segregating Sites), a new
121 framework for demographic inference via ABC-RF, suitable for the analysis of complete genomes.
122 ABC-SeS compares all possible pairs of populations, summarizing both the observed and the
123 simulated genetic variation by the distribution over the genome of the four mutually exclusive
124 categories of segregating sites (i.e. private polymorphisms in either population, shared
125 polymorphisms and fixed differences) known to be informative about processes of divergence and
126 admixture [25], and not requiring phased data. These statistics have already been successfully used
127 in a standard ABC context [26], but only in the form of the first four moments of the distribution
128 across loci. Here, for the first time, and thanks to the ABC-RF procedure, we analyze the full
129 genomic distribution of each statistic. We performed a power analysis, to evaluate how accurately

130 this ABC pipeline can recognize the true model among models of increasing complexity, using
131 simulated data. We also explored the inferential power of the presented procedure with respect to
132 the experimental conditions, evaluating the consequences of sampling strategies involving different
133 numbers of genomes, different numbers of loci, and different locus lengths. Finally, we applied our
134 method to two case studies. First, we analyzed the demographic history of anatomically modern
135 humans and the dynamics of migration out of the African continent, explicitly comparing the
136 models proposed by [27,28]. Secondly, we reconstructed the past demographic history and the
137 interaction dynamics among the three orangutan species inhabiting Borneo and Sumatra, revising
138 the models presented by [29].

139

140 **Results**

141 **Summary Statistics**

142 We summarize the genetic data by means of the genomic distributions of the four mutually
143 exclusive categories of segregating sites in two populations, namely 1- segregating sites private of
144 the first population, 2- segregating sites private of the second populations, 3- segregating sites that
145 are polymorphic in both populations and 4- segregating sites fixed for different alleles in both
146 populations [25]. To calculate the complete genomic distribution of these statistics we considered
147 the genome as subdivided in a certain number of independent fragments of a certain length, and for
148 each fragment we counted the number of sites belonging to each of the four above-mentioned
149 categories. The final vector of summary statistics is thus composed by the truncated frequency
150 distribution of fragments having from 0 to n segregating sites in each category, for each pair of
151 populations considered. We refer to this set of statistics as FDSS (frequency distribution of
152 segregating sites) hereafter. The maximum number of segregating sites in a locus of a certain length
153 is fixed to n , and hence the last category contains all the observations higher than n . In the one
154 population models, the four distributions described above collapse in a single distribution
155 describing the frequency of loci showing specific levels of within-population segregating sites.

156

157 **Power Analysis**

158 To determine the power of the ABC-SeS procedure in distinguishing among alternative
159 evolutionary trajectories, we performed an extensive power analysis, as described below.

160 We simulated genetic data considering different experimental conditions, thus testing all the
161 possible combinations of locus length (bp) {200; 500; 1,000; 2,000; 5,000}, number of loci {1,000;
162 5,000; 10,000} and number of chromosomes {2, 4, 10, 20}, for a total of 60 combinations of
163 sampling conditions tested. The data were generated using the *msms* simulator according to each of
164 these combinations for three sets of non-nested models of increasing complexity, namely one-
165 population models (four alternative models, Fig 1A), two-population models (three alternative
166 models, Fig 2A) and multi-populations models (two alternative models, Fig 3A); for a more
167 comprehensive description of the models see the Materials and Methods section.

168 For each combination of experimental conditions, we compared alternative models within
169 the three sets tested treating 1,000 simulated datasets for each model as pseudo-observed data
170 (pods). The models were compared through ABC-RF, and we counted the number of times the
171 model selection procedure correctly assigned a pod to the true model, i.e. to the model that actually
172 generated that pseudo-observed dataset. The proportion of true positives (TP) thus calculated is a
173 measure of the power of the whole procedure, considering all its features (model selection
174 approach, alternative models compared, statistics summarizing the data, genomic parameters
175 simulated). The TP results are summarized in Figs 1-3B.

176

177 *One-population models*

178 We designed four one-population models, respectively considering a population of constant
179 size through time, a bottleneck, an exponential growth and a structured population. In each of the
180 50,000 simulations (per model and per combination of experimental parameters) demographic and
181 evolutionary parameter values are drawn from prior distributions, detailed in S1 Table. The four

182 plots of Fig 1B report the results of the power analyses. In each plot, we see the proportion of times
183 each model was correctly recognized as the most likely one. In general, the percentage of true
184 positives is quite high, ranging from almost 80% to 100% depending on the model generating the
185 pod and on the combination of experimental conditions tested. The bottleneck model has the highest
186 rate of identification, with most combinations of experimental conditions yielding nearly 100% true
187 positives. By contrast, the least identifiable model seems the one considering a structured
188 population, with 0.76 to 0.86 true positives, depending on the combination of experimental
189 conditions. However, we observed that the decrease in the power is actually linked to the extent of
190 gene flow among demes, and to the number of demes sampled; as increasing gene flow and
191 decreasing the number of demes sampled, the structured and the panmictic models converge, hence
192 becoming harder to distinguish (S2 Fig). As expected, we observed a general increase in the power
193 with the increase of both the locus length and the number of loci considered. The number of
194 sampled chromosomes does not appear to be directly linked to the increase of the proportion of true
195 positives. For some sampling conditions, we observed instead a decrease in the TP rate going from
196 2 to 20 chromosomes (see Fig 1B). However, we showed that this unexpected behavior can be
197 explained with the overlap of the FDSS generated by the constant and the structured models, which
198 increases as increasing the number of chromosomes sampled (S3 Fig). We repeated the whole
199 analysis considering intra-locus recombination (S4 Fig), and the general pattern did not change. We
200 observed a reduced power for the structured model (TP from 75 to 90%) with respect to the other
201 models tested (TP from 80 to 100%), and a stronger effect of the locus length and of the number of
202 loci, than of the number of chromosomes simulated.

203

204 *Two-populations models*

205 The two-population scenarios include a divergence model with isolation after the divergence
206 (no gene flow); a divergence model with continuous migration from the separation until present
207 times; and a divergence model with a single event of bidirectional migration (i.e. pulse of

208 admixture), for a total of three models compared for each combination of experimental parameters
209 (Fig 2A, prior distributions reported in S2 Table). The plots in Fig 2B clearly show that the most
210 recognizable model is the one including divergence and migration, with proportions of true
211 positives close to 0.95, regardless of the combination of experimental conditions tested. For the
212 other two models, the power was clearly lower than that estimated for the one-population models,
213 with values ranging from 40% to 65% (for the model without migration) and from 60% to 80% (for
214 the model with a pulse of admixture). This low power might be due to the fact that, when the pulse
215 of admixture of the third model occurs right after the divergence, the two models become
216 essentially identical. To verify this prediction, we calculated the proportion of TP for the pulse of
217 admixture model, subdividing the pods in seven categories, depending on the time interval between
218 the divergence and the admixture event (S5 Fig). The proportion of the pods correctly assigned to
219 the pulse of admixture model increases as the intervals between divergence and admixture increase,
220 reaching values of 85% for some combination of parameters. In this case, we did not observe
221 remarkable differences with respect to specific experimental parameters. The performance of the
222 method improved when we included intra-locus recombination in the simulations (S6 Fig). In this
223 case, we observed an increase of the power with increasing locus length, both for the no migration
224 and for the pulse of admixture model, respectively reaching values of 95% and 100%.

225 We also ran many parallel simulations, using either the whole FDSS or its first four
226 moments only. In all cases (a selection of the results is in S3 Table) the FDSS's ability to
227 distinguish among different evolutionary scenarios was greater, and sometimes substantially so,
228 under the former conditions.

229

230 *Multi-populations models*

231 In most realistic cases, interactions among several populations need be considered. Among
232 the many possible scenarios, we chose to initially focus on the alternative hypotheses proposed to
233 explain the expansion of anatomically modern humans out of the African continent. We shall refer

234 to the two models (shown in Fig 3A and detailed in Materials and Methods) as Single Dispersal
235 (proposed e.g. in [27]) and the Multiple Dispersal (proposed e.g. in [28,30,31]). The main
236 demographic events are shared between the two models, and here we shall refer to the description
237 given by [27]: three archaic branches (unknown, Denisova and Neandertal), three modern branches
238 (Africans, Eurasians and Papuans) and a specific pattern of gene flow. The difference between the
239 two alternative models lies in the details of the expansion from Africa. Under the Single Dispersal
240 model (SDM), Eurasians and Papuans derived from the same expansion from Africa, through the
241 Near East. Under the Multiple Dispersal model (MDM), the Papuans derived from an earlier
242 dispersal from Africa, independent from that giving rise to the Eurasian populations.

243 The prior distributions associated with these models are reported in S4 and S5 Table. Fig 3B
244 summarizes the power analysis. The proportion of true positives ranges between 0.65 and 0.70 for
245 the SDM, regardless of the combination of experimental parameters tested, and between 0.6 and 0.8
246 for the MDM, in this case with a slight increase of the power with the size of the fragments
247 simulated. Being the SDM and the MDM quite similar, in particular when the time interval between
248 the first and second exit in the MDM is short, we also evaluated the ABC-SeS ability to identify the
249 correct model as a function of the time span between the divergence time of the African ghost
250 populations and the second exit in the Multiple Dispersal model. To do this, we considered 10,000
251 pods from the MDM. We then subdivided these 10,000 pods in 6 bins of increasing interval
252 between these two events (up to 60,000 years), and, within each bin, we separately calculated the
253 proportion of times in which the MDM is correctly recognized by the ABC-RF procedure. As might
254 be expected, the proportion of true positives increases with increasing intervals between the
255 divergence of the two African ghost populations and the time of the second exit (S7 Fig), reaching
256 values of 90% for some combinations of experimental parameters.

257 For the multi-populations models we tested the effect of intra-locus recombination only for
258 4 key sampling schemes that combine few (1,000) and many (10,000) genetic loci with either short
259 (500 bp) or long (2,000 bp) DNA fragments. We selected these combinations because they might

260 represent extremely favorable (10,000 loci, 2,000 bp length) or unfavorable (1,000 loci 500bp
261 length) conditions that can be encountered when analyzing real data. Compared with the results
262 obtained for the same experimental parameters without considering intra-locus recombination, we
263 observed a slight increase in the TP rate when analyzing high number of fragments, or when
264 considering longer locus lengths (S6 Table).

265

266 **Real Case: out of Africa dynamics**

267 Simulations in the previous section basically show that alternative models can be
268 distinguished by ABC-SeS, except when they differ for parameters approaching 0, thus creating
269 overlaps between their expected consequences. We then moved to applying the ABC-SeS method to
270 estimate posterior probabilities of alternative models about early human expansion from Africa.
271 Whether human demographic history is better understood assuming one [5,27] or two [28,30,31]
272 major episodes of African dispersal is still an open question. While concluding that indigenous
273 Australians and Papuans seem to derive their ancestry from the same African wave of dispersal as
274 most Eurasians, Mallick et al. [5] admitted that these inferences change depending on the
275 computational method used for phasing haplotypes. Therefore, it made sense to compare the SDM
276 and the MDM by ABC-SeS. For this purpose, we analyzed the high coverage Neandertal genome
277 [7], the high coverage Denisova genome [6], and a large set of modern individuals published in [28]
278 (S7 Table). To make sure that the choice of the genomes does not affect the results, we repeated the
279 model selection analysis replacing the Papuan individuals with those published by [27] (S7 Table).
280 We extracted from all these genomes 10,000 shared independent fragments of 500 base pairs length
281 (see Material and Methods for details). Since we knew that the number of chromosomes has a
282 minor impact on the analysis, we considered two chromosomes per population, so that the sample
283 size would be the same in ancient and modern samples. The proportion of true positives for the
284 combination of experimental parameters here considered (i.e. 10,000 loci of 500 bp length and 2
285 chromosomes per population) was 0.66 for the SDM, and between 0.69 and 0.85 for the MDM (Fig

286 3A and S7 Fig). Considering recombination, we observed an increase of the proportion of TP for
287 the MDM, which (averaged over the whole range of possible divergence times between the African
288 ghost populations and the second exit) reached a value of 0.75.

289 As the SDM and the MDM mainly differ for the origin of the Papuan population (whether or
290 not descending from the same ancestor of Eurasians), to test for the robustness of the results we
291 used the genomic information of all the six Papuan individuals sequenced by [28]. To this purpose,
292 we repeated the ABC-RF model selection procedure six times, each time considering a different
293 Papuan individual. Fig 4 and S8-S11 Tables presents the results of the ABC-RF model selection,
294 showing that, in all the comparisons, the results supported the MDM, with posterior probabilities
295 ranging from 0.67 to 0.69. We repeated the whole analysis considering the 25 Papuan individuals in
296 [27], this time repeating the ABC-RF model selection procedure 25 times, and always found results
297 favoring the MDM over the SDM. The posterior probabilities estimated for the MDM were
298 comparable to those estimated with the Pagani dataset [28] (Fig 4). Considering intra-locus
299 recombination, MDM received additional support, with posterior probabilities ranging from 0.70 to
300 0.76 for both datasets (Fig 4).

301

302 **Real Case: Orangutan evolutionary history**

303 As a second application of our framework to real data, we investigated the past demographic
304 and evolutionary dynamics of the three known orangutan species. In addition to the orangutan
305 species previously recognized in Borneo (*Pongo pygmeus*) and in Sumatra, North of Lake Toba
306 (*Pongo abelii*), [29] described a new species of Sumatran orangutan, *Pongo tapanuliensis*, South of
307 Lake Toba. To reduce the otherwise excessive computational effort in their ABC analysis, [29] had
308 to compare simplified versions of the alternative models. Factors such as bottlenecks and population
309 structure were considered only when estimating the parameters of the best-performing model,
310 namely the one with an earlier split between the two Sumatran species, and an origin of the Bornean
311 species from *Pongo tapanuliensis* ancestors. The ABC-SeS method does not suffer from this

312 shortcoming. We compared four complex scenarios, designed following the models presented in
313 [29] (Fig 5A). While considering the same substructure within the three orangutan species, the four
314 models differ in the genealogical relationships assumed among species (Fig 5A and Materials and
315 Methods for details). From the whole genomic dataset presented in [29], we selected seven
316 individuals, one from each of the populations emerged from the study (S12 Table, see Materials and
317 Methods for details). Following the procedure described in Materials and Methods, we extracted
318 9,000 independent sites 1kb length shared among the 7 selected individuals, on which we calculated
319 the FDSS. We generated 50,000 simulations per model considering the above mentioned genomic
320 structure (9,000 loci, 1kb length and 2 chromosomes per population), with prior distributions
321 detailed in S13-S16 Tables. We first assessed the ability to correctly recognize the four alternative
322 models tested through a power analysis, considering 1,000 pods per model. The results are
323 presented in Fig 5A. The most identifiable model (TP=0.81) appeared to be the model 2b, under
324 which there is a first separation of South Toba from Borneo Orangutan, followed by the divergence
325 of North Toba from South Toba. The model assuming an early separation of South Toba from
326 North Toba, followed by the separation of Borneo from South Toba, actually showed the lowest
327 proportion of true positives (0.45). The application to real data favored the model 1a, (also
328 associated with the highest posterior probability in [29]), with a posterior probability of 0.49. Under
329 the most supported model both the North Toba (first) and Borneo (later) separated from *Pongo*
330 *tapanuliensis* Fig 5B.

331

332 **Discussion**

333 The cost of typing large genomic datasets has dramatically decreased lately, making
334 population-scale genomic data available for a large set of organisms [4,9,10,32,33]. The main
335 challenge now is how to extract as much information as possible from these data, developing
336 flexible and robust statistical methods of analysis [12,13,16]. Approximate Bayesian Computation,
337 explicitly comparing alternative demographic models and estimating the models' probabilities,

338 represents a powerful inferential tool for about past population dynamics [34]. Another advantage
339 of such a simulation-based approach is the possibility to easily check whether the models being
340 compared are actually distinguishable, hence quantifying the reliability of the estimates produced
341 [35]. Nevertheless, only recently, with the development of the Random Forest procedure for ABC
342 model selection [24], it has become possible to apply ABC to large genomic datasets. With this
343 work, we took advantage of this newly proposed algorithm to develop ABC-SeS, an ABC-based
344 flexible framework to compare demographic models. The novelty of this framework mainly lies in
345 the summary statistics used to summarize genomic data, namely the complete genomic distribution
346 of the four mutually exclusive categories of segregating sites for pairs of populations [25].

347

348 **Power Analysis**

349 Initially, we analyzed sets of models with increasing levels of complexity, simulating
350 genetic data under a broad spectrum of experimental conditions. This extensive power analysis
351 showed that ABC-SeS can often identify the model under which the data were generated, with some
352 uncertainties only when two models are just marginally different. This was the case for both simple
353 (one or two-population scenarios, Figs 1 and 2) and complex (multi-populations scenarios, Fig 3)
354 demographies. When we compared one-population scenarios, to summarize the data we necessarily
355 considered only one of the four distributions of polymorphic sites. Nonetheless the model
356 identifiability, calculated as the proportion on TPs over 1,000 pods, reached values between 80%
357 and 100%, with slightly lower values only for the structured model. This reduction in the power was
358 always due to the levels of gene flow among demes; when it is high, the structured model becomes
359 hardly distinguishable from the panmictic model (S2 Fig), as has already been known since
360 Wright's times [36]. We also showed that the power depends on the number of demes; indeed, the
361 proportion of TPs increases in parallel with the number of demes considered in the structured model
362 (S2 Fig). Among the two-populations demographies, the model with bi-directional migration at a
363 constant rate proved easiest to identify, with almost 100% TPs, regardless of the combination of

364 experimental parameters tested. Predictably (see e.g. [37]), the pulse of admixture model produced
365 the most accurate results when the two populations were highly genetically divergent and the time
366 interval between admixture and divergence was large (S5 Fig). Our results showed that models are
367 difficult to discriminate only when the divergence is recent, and/or close the admixture event. Pods
368 generated under these conditions, indeed, are often erroneously assigned to the other models.

369 Even when more complicated scenarios were compared (e.g., the multi-populations models),
370 the ABC-SeS framework recognizes the true model, with about 70% TPs. It is worth noting that in
371 the MDM the first and the second exit from Africa could sometimes occur at very close times. In
372 these cases, the SDM and the MDM models become extremely similar. Indeed, we observed an
373 increase in the power of the test at increasing intervals between the African divergence and the
374 second exit (S7 Fig), reaching values close to 90%.

375 In the first round of comparisons, we assumed complete linkage among all sites within the
376 same fragment. Even if acceptable for some experimental conditions (i.e. for short locus lengths),
377 this is generally simplistic. We thus repeated the power tests, this time generating data that could
378 undergo intra-locus recombination. In all the comparisons, we found that, with recombination, the
379 results are generally better, regardless of the experimental parameters considered, and particularly
380 when the loci are long. In particular, we observed a substantial increase of the proportion of TPs for
381 the MDM, but not for the SDM (S6 Table). One possible explanation for this finding is, in a tree
382 generated under SDM, all loci of the African, European, Asian and Papuan populations, belong to a
383 branch that has a unique topology (T1 in S8 Fig) and recombination will not alter this condition. By
384 contrast, under MDM, the topologies are three (T1, T2 and T3 in S8 Fig) a recombination event will
385 generate a transition from one of these baseline topologies to the recombinant topology (denoted as
386 R in S8 Fig). Thus, recombination under MDM reduces the fraction of loci showing the T1
387 topology, i.e. the topology shared with the alternative SDM, thus increasing the possibility that the
388 two models be discriminated. As could be expected, under most conditions tested, the model
389 identifiability increases increasing the locus length, and the number of loci considered. Indeed, by

390 analyzing more loci and longer loci, the probability of observing recombination increases, resulting
391 in a higher fraction of the R topologies. Thus, when one is analyzing many loci, or long stretches of
392 DNA, recombination should be considered, although this is not necessarily leading to an increase in
393 TPs, because the specific features of the demographic models compared are also a factor.

394 As it is intuitively reasonable, and how it has been already pointed out e.g. by [38], the
395 accuracy of the model selection seems to be more dependent on the number of loci considered and
396 on the locus length rather than on the number of individuals sampled per population. In absence of
397 recombination, the locus length appeared to be crucial to correctly identify simple demographies,
398 such as those generated under one-population models, whereas its effect was minor when complex
399 scenarios were compared (Figs 1 and 3). When recombination was included in the modeling, the
400 locus length seemed to become the main factor leading to a good discrimination between models. In
401 the two-populations scenarios we observed a significant increase of the proportion of TPs at
402 increasing lengths of the fragments simulated, reaching values close to 100% (S6 Fig). It is anyway
403 worth noting that the power to correctly identify the true model was quite good even when we
404 simulated short fragments, as well as in the comparison of complex demographies (Fig 3). This
405 finding means that the ABC-SeS procedure is particularly suitable for the analysis of ancient data
406 [6] and of RAD sequencing data [39], where short DNA fragments are more the rule than the
407 exception.

408

409 **Applications to real datasets**

410 We compared two demographic models about the anatomically modern humans expansion
411 out of Africa, combining ancient (Neandertal and Denisova) and modern genomes data. Because we
412 were including ancient data, we restricted the analysis to short fragments (500 bp) to maximize the
413 number of independent loci retrievable. This combination of experimental parameters (10,000 loci
414 500 bp length, 2 chromosomes per population) showed anyway a good ability to distinguish
415 between models (Fig 3). For the first comparison, we used as modern genomes those published by

416 [28], and repeated the model selection procedure six time, each time considering one of the six
417 Papuan genomes in the dataset. All such comparisons supported the MDM (Fig 4), i.e. a first
418 expansion of the ancestors of the current Austro-Melanesians, followed by a second expansion
419 leading to the peopling of Eurasia. We obtained the same conclusion by using Papuan genomes
420 coming from another dataset [27] (Fig 4). Considering different modern individuals from African,
421 European and Asian populations did not change the support for the MDM. These results raise
422 several questions; indeed, it was the SDM that showed the best fit in [27], whereas the MDM
423 appeared to account for the data only when the analysis was restricted to modern populations.
424 However, our findings are in agreement with those by [28], who estimated that at least 2% of the
425 Papuan genome derive from an earlier, and distinct, dispersal out of Africa. Other genomic studies
426 [31], but not all [5], and phenotypic analyses [30] appear support the MDM, which calls for further
427 research in this area. Be that as it may, in no other study besides the present one (i) the alternative
428 hypotheses are explicitly compared analyzing complete genomes; (ii) posterior probabilities are
429 estimated for each model, and (iii) the accuracy of the estimates is assessed by power analysis.

430 We then moved to investigating the evolutionary history of the three extant Orangutan
431 subspecies, the last of which (located south of the Lake Toba) has been recently described by [29].
432 We basically repeated the ABC analysis performed by [29] applying our ABC-SeS framework.
433 Nater and colleagues [29] started comparing simple evolutionary scenarios, and considered
434 population substructure and gene flow only when estimating parameters, but not in the phase of
435 model choice. ABC-SeS allowed us to analyze complex demographies from the very beginning, i.e.
436 from the model selection phase, thus comparing more realistic evolutionary scenarios. We basically
437 obtained the same results presented by [29]; in both cases the most supported model was the one
438 accounting for a first separation of the North Toba and a later separation of the Borneo species from
439 the newly identified South Toba species (Fig 5B). The main difference was about the strength of the
440 support associated to this model. While Nater and colleagues [29] estimated high posterior
441 probabilities for the best-fitting model (73% when comparing the 4 models and 98% when

442 comparing the two best scenarios), our procedure associated to the same model a posterior
443 probability of 49% (Fig 5A). Moreover, the power analysis that we conducted (absent in [29]),
444 revealed that the ability to correctly distinguish among the four tested models is between 45% and
445 81%, with the selected model that can be erroneously recognized as the most probable one in the
446 38% of cases. These results emphasize (i) the importance of including complex demographic
447 histories in the model selection step, so as to evaluate the real posterior probability associated to the
448 best model, on which the parameter estimation will be performed and (ii) the importance of
449 performing a power analysis of the models tested, so as to be aware of the level of uncertainty about
450 the conclusions of the study.

451

452 **Perspectives**

453 In this paper we presented a new framework based on ABC-RF to explicitly compare
454 complex demographic models exploiting whole-genome data. We proposed to summarize the data
455 with a set of statistics that were already known to be informative about past population dynamics
456 [25], but for the first time, and thanks to the Random Forest algorithm, we used the entire genomic
457 distribution of the four categories of segregating sites between pairs of populations in an ABC
458 context. It is clear that considering the whole FDSS, rather than its first four moments, significantly
459 improves our ability to distinguish among different evolutionary scenarios. How large this
460 improvement, depends on the amount of genetic loci available, and, of course, on how large are the
461 differences between the alternative models. At present, the ABC-SeS is optimized for high-
462 coverage genotypes [5,10,33]. A natural extension of this work will thus be to implement the use of
463 low coverage data, developing an approach able to retrieve the FDSS taking into account the
464 genotype uncertainty and sequencing errors, for instance through the use of the genotype likelihood
465 (as, e.g., in ANGSD [40]).

466 The flexibility of the ABC-RF model selection approach, combined with the inferential
467 power proven by the summary statistics that we proposed to calculate on genomic data, pave the

468 way for a detailed and comprehensive study of complex demographic dynamics for any species for
469 which few high coverage genomes are available.

470

471 **Materials and Methods**

472 **ABC Random Forest**

473 Approximate Bayesian computation is a flexible framework, based on coalescent
474 simulations, to compare different evolutionary hypotheses and identify the one that with the highest
475 probability generated the observed data [20,34]. Despite its potential, the application of the classical
476 ABC approach to the analysis of large genomic datasets and complex evolutionary models, has
477 been quite limited. Possible reasons are (i) the levels of arbitrariness in the choice of the sufficient set
478 of summary statistics and (ii) the high number of simulations required to obtain good quality
479 posterior estimates, resulting in an unacceptable increase of computational time when multiple
480 complex models are simultaneously analyzed. Both these issues appear to have been solved with the
481 development of a new ABC approach, based on a machine learning procedure (Random Forest
482 [24]). Random Forest uses the simulated datasets for each model in a reference table to predict the
483 best fitting model at each possible value of a set of covariates. After selecting it, another Random
484 Forest obtained from regressing the probability of error of the same covariates estimates the
485 posterior probability. This procedure allows one to overcome the difficulties traditionally associated
486 with the choice of summary statistics, while gaining a larger discrimination power among the
487 competing models. Moreover, a satisfying level of accuracy in the estimates is achieved with about
488 30-50 thousand simulations per model [24], significantly reducing the computational cost of the
489 analysis of complex demographic histories. All the ABC-RF estimates have been obtained using the
490 function *abcrf* from the package *abcrf* and employing a forest of 500 trees, a number suggested to
491 provide the best trade-off between computational efficiency and statistical precision [24]. We
492 compared all models and obtained the posterior probabilities using the function *predict* from the
493 same package.

494

495 **Power Analysis**

496 For the power analysis, we generated data under different combinations of experimental
497 parameters, varying the number of loci (calculating FDSS on 1,000, 5,000 and 10,000 loci), the
498 locus length (200; 500; 1,000; 2,000; 5,000 bp) and the number of chromosomes sampled per
499 population (2, 4, 10, 20), for a total of 60 combinations per model. We evaluated the power
500 considering three sets of models of increasing complexity, detailed below.

501

502 **One-population models.**

503 We started by considering four demographic models (Fig 1). The first model represents a
504 constantly evolving population with an effective population size NI , drawn from a uniform prior
505 distribution (S1 Table). Under the second model, the population experienced a bottleneck of
506 intensity i , T generations ago. The intensity and the time of the bottleneck, and the ancient effective
507 population size Na are drawn from uniform prior distributions, showed in S1 Table. The third
508 model represents an expanding population. The expansion (of intensity i) is exponential and starts T
509 generations ago, with the effective population size increasing from NI/i to NI (prior distributions in
510 S1 Table). Under the last model, the population is structured in different demes, exchanging
511 migrants at a certain rate. The actual number of demes d , the migration rate m and the effective
512 population size NI are drawn from prior distributions (S1 Table).

513 **Two-populations models.**

514 We then moved to considering three demographic models with two populations (Fig 2). The
515 first one is a simple split model without gene flow after the divergence. Under this model, an
516 ancestral population of size $Nanc$ splits $Tsep$ generation ago into two populations. These two
517 derived populations evolve with a constant population size ($N1$ and $N2$) until the present time
518 (priors for these free parameters are shown in S2 Table). The second model also includes a
519 continuous and bidirectional migration, all the way from the divergence moment to the present. The

520 per generation migration rates $m12$ and $m21$ are drawn from priors defined in S2 Table. The third
521 and last model assumes a single pulse of bidirectional admixture at time $Tadm$ after divergence.
522 Admixture rates $adm12$ $adm21$, and the time of admixture are drawn from priors (S2 Table).

523 **Multi-populations models.**

524 Finally, we compared two demographic models, representing the two alternative hypotheses
525 that have been proposed to explain the dispersal of anatomically modern humans out of the African
526 continent (i.e single and multiple dispersals, Fig 3). To design the models we followed the
527 parametrization proposed by [27], with some minor modifications. Both models share the main
528 demographic structure: on the left the archaic groups (i.e. Neandertal, Denisova and an unknown
529 archaic source), and on the right the anatomically modern humans (with a first separation between
530 Africans and non-Africans and subsequent separations among population that left Africa). Given
531 the evidence for admixture of Neandertals and Denisovans with non-African modern human
532 populations [6,7], we allowed for genetic exchanges from archaic to modern species, indicated in
533 Fig 3 by the colored arrows. It has to be noted that the archaic populations actually sending
534 migrants to modern humans are two ghost populations that diverged from the Denisovan and the
535 Neandertal Altai samples 393 kya and 110 kya, respectively [27]. This way, we took into account
536 that the archaic contributions to the modern gene pool did not necessarily come from the archaic
537 populations that have been genotyped so far. We modeled bidirectional migration between modern
538 populations along a stepping-stone, thus allowing for gene flow only between geographically
539 neighboring populations. Under the Single Dispersal model (SDM) a single wave of migration
540 outside Africa gave rise to both Eurasian and Austromelanesian populations, whereas under the
541 Multiple Dispersal model (MDM) there are two waves of migration out of Africa, the first giving
542 rise to Austromelanesians and the second to Eurasians. We took into account the presence of
543 genetic structure within Africa modeling the expansion from a single unsampled “ghost” population
544 under the SD model, and from two separated unsampled “ghost” populations for the MD model.
545 The prior distributions for all the parameters considered in these models are in S4 and S5 Tables.

546 We simulated both demographic models under all possible combinations of experimental
547 parameters. We ran 50,000 simulations per model and combination of experimental parameters,
548 using the *ms/msms* software [41,42]. To evaluate the quality of the estimates of the model selection
549 procedure we calculated the proportion of True Positives (TP) as follows. We randomly sampled
550 1,000 datasets from each simulated combination of demographic model and experimental
551 parameters. We treated each of these sampled datasets as pseudo-observed data (pod) and
552 performed a models comparison within the correspondent set of models tested. For each model and
553 combination of experimental parameters, we calculated the proportion of TPs as the proportion of
554 times the true model was the most supported by the model selection procedure.

555

556 **Real Case: out of Africa dynamics**

557 We considered the ancient high-coverage genomes of Denisova and Neandertal [6,7],
558 together with modern human samples from [28]. A detailed description of the samples is in S7
559 Table. All the individuals were mapped against the human reference genome hg19 build 37. To
560 calculate the observed FDSS we only considered autosomal regions found outside known and
561 predicted genes +/- 10,000 bp and outside CpG islands and repeated regions (as defined on the
562 UCSC platform, [43]). We also took into consideration only the fragments found in genomic
563 regions that passed a set of minimal quality filters used for the analysis of the ancient genomes
564 (map35_50%, [6,7]). From the resulting regions we extracted 10,000 independent fragments of 500
565 bp length, separated by at least 10,000 bps. All comparisons involved a single individual (i.e. two
566 chromosomes) per population, and so each run of the analysis took into account the Denisova, the
567 Neandertal, one African, one European one Asian and one Papuan individual (detailed in S7 Table).
568 We calculated the observed FDSS six times, each time considering a different Papuan individual
569 present in the Pagani dataset [28]. We repeated the whole procedure substituting the Papuan
570 individuals with those published by [27]. To do this, we downloaded the corresponding alignments
571 in CRAM format from <https://www.ebi.ac.uk/ega/datasets/EGAD00001001634>. The *mpileup* and

572 *call* commands from *samtools-1.6* [44], were used to call all variants within the 10,000 neutral
573 genomic fragments, using the *--consensus-caller* flag, without considering indels. We then filtered
574 the initial call set according to the filters reported in [27] using *vcflib* and *bcftools* [44]. We
575 calculated the FDSS for each of these Papuan individuals, hence obtaining 25 observed FDSS, that
576 have been separately analyzed through the ABC-RF model selection procedure.

577

578 **Real Case: Orangutan evolutionary history**

579 We selected seven Orangutan individuals, one from each of the populations defined by [29],
580 choosing the genomes presenting the highest coverage (S12 Table). We downloaded the FASTQ
581 files from <https://www.ncbi.nlm.nih.gov/sra/PRJEB19688>, and mapped the reads to the ponAbe2
582 reference genome (<http://genome.wustl.edu/genomes/detail/pongo-abelii/>) using the BWA-MEM
583 v0.7.15 [45]. We used *picard-tools-1.98* (<http://picard.sourceforge.net/>) to add read groups and to
584 filtered out duplicated reads from the BAM alignments. We performed local realignment around
585 indels by the Genome Analysis Toolkit (GATK) v2.7-2 [46]. To obtain genomic fragments suitable
586 to calculate the FDSS, we generated a mappability mask (identified with the *GEM-mappability*
587 module from the *GEM* library build [47]) so as to consider only genomic positions within a
588 uniquely mappable 100-mer (up to 4 mismatches allowed). We then excluded from this mask all the
589 exonic regions +/- 10,000 bp, repeated regions (as defined in the *Pongo abelii* Ensembl gene
590 annotation release 78), as well as loci on the X chromosome and in the mitochondrial genome. We
591 then generated the final mask calculating the number of fragments separated by at least 10 kb, thus
592 obtaining 9,000 fragments of 1,000 bp length. We called the SNPs within these fragments using the
593 *UnifiedGenotyper* algorithm from *GATK*; the filtering step has been performed as reported in [29]
594 through *vcflib*. We finally calculated the observed FDSS from the quality filtered VCF file.

595 To investigate past population dynamics of the three Orangutan species, we designed
596 competitive scenarios following the demographic models reported in [29]. Nater [29] first compared
597 models assuming simple demographies, and then estimated the parameters from a more complicated

598 model. By contrast, we directly compared complex demographies, designing the within-species
599 substructure as described by [29], (Fig 5A). The four competing models indeed share the same
600 within species features (four populations for the Bornean group, two Sumatran populations north of
601 Lake Toba, and a single population south of Lake Toba), while differing for the topology, i.e. for
602 the evolutionary relationships among the three species, as reported in Fig 5A. We modeled
603 bidirectional migration both among populations within a species and between neighboring species.
604 A detailed description of the models' parameters and of the priors are in S13-S16 Tables. We ran
605 50,000 simulations per model using the *ms* software [41], generating two chromosomes per
606 population (4 Bornean, 1 south of Lake Toba and 2 north of Lake Toba), and 9,000 independent
607 fragments of 1kb length per chromosome. We first assessed the power to distinguish among the four
608 models calculating the proportion of TPs as described above, and then explicitly compared the
609 simulated variation with the FDSS calculated on the observed data (Fig 5B).

610

611 **Software and data availability**

612 All the scripts of the ABC-SeS framework can be found at <https://github.com/anbena/ABC-SeS>.

613

614 **Acknowledgments**

615 We would like to thank the DFG Center for Advanced Studies, “Words, Bones, Genes, Tools,” at
616 University of Tübingen, that hosted AB and SG during the first phase of the project. FT was
617 supported by ERC Advanced Grant 295733, ‘LanGeLin’.

618

619 **Author Contributions**

620 AB conceived the study; AB and SG designed the experiments; MTV, AB, SG and FT analyzed the
621 data; SG, MTV, FT, GB and AB discussed the results; SG, GB and AB wrote the paper with inputs
622 from all coauthors.

623

624 **References**

- 625 1. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, et al. Population
626 history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.*
627 2004;
- 628 2. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history
629 and natural selection in the human major histocompatibility complex loci. *Genetics.* 2006;
- 630 3. Lohmueller KE. The Impact of Population Demography and Selection on the Genetic
631 Architecture of Complex Traits. *PLoS Genet.* 2014;
- 632 4. 1000 Genomes Project Consortium. An integrated map of genetic variation. *Nature.* 2012;
- 633 5. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome
634 Diversity Project: 300 genomes from 142 diverse populations. *Nature.* 2016;
- 635 6. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage
636 genome sequence from an archaic Denisovan individual. *Science* (80-). 2012;
- 637 7. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete
638 genome sequence of a Neanderthal from the Altai Mountains. *Nature.* 2014;
- 639 8. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, et
640 al. Early human dispersals within the Americas. *Science* (80-) [Internet]. 2018 Dec
641 7;362(6419):eaav2621. Available from:
642 <http://science.sciencemag.org/content/362/6419/eaav2621.abstract>
- 643 9. Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, et al.
644 Butterfly genome reveals promiscuous exchange of mimicry adaptations among species.
645 *Nature.* 2012;
- 646 10. De Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al.
647 Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* (80-).
648 2016;

- 649 11. McVean GAT, Cardin NJ. Approximating the coalescent with recombination. *Philos Trans R*
650 *Soc B Biol Sci.* 2005;
- 651 12. Li H, Durbin R. Inference of human population history from individual whole-genome
652 sequences. *Nature.* 2011;
- 653 13. Schiffels S, Durbin R. Inferring human population size and separation history from multiple
654 genome sequences. *Nat Genet.* 2014;
- 655 14. Heller R, Chikhi L, Siegmund HR. The Confounding Effect of Population Structure on
656 Bayesian Skyline Plot Inferences of Demographic History. *PLoS One.* 2013;
- 657 15. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Diffusion Approximations
658 for Demographic Inference : DaDi. *Nat Preced.* 2010;
- 659 16. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust Demographic
660 Inference from Genomic and SNP Data. *PLoS Genet.* 2013;
- 661 17. Akaike H. A New Look at the Statistical Model Identification. *IEEE Trans Automat Contr.*
662 1974;
- 663 18. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from
664 hundreds of unphased whole genomes. *Nat Genet.* 2017;
- 665 19. Bertorelle G, Benazzo A, Mona S. ABC as a flexible framework to estimate demography
666 over space and time: Some cons, many pros. *Molecular Ecology.* 2010.
- 667 20. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population
668 genetics. *Genetics.* 2002;
- 669 21. Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods.
670 *Proc Natl Acad Sci.* 2003;
- 671 22. Blum MGB, François O. Non-linear regression models for Approximate Bayesian
672 Computation. *Stat Comput.* 2010;
- 673 23. Blum MGB, Nunes MA, Prangle D, Sisson SA. A Comparative Review of Dimension
674 Reduction Methods in Approximate Bayesian Computation. *Stat Sci.* 2013;

- 675 24. Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP. Reliable ABC model
676 choice via random forests. *Bioinformatics*. 2015;
- 677 25. Wakeley J, Hey J. Estimating ancestral population parameters. *Genetics*. 1997;
- 678 26. Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ. ABC inference of multi-
679 population divergence with admixture from unphased population genomic data. *Mol Ecol*.
680 2014;
- 681 27. Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. A genomic
682 history of Aboriginal Australia. *Nature*. 2016.
- 683 28. Pagani L, Lawson DJ, Jagoda E, Mörseburg A, Eriksson A, Mitt M, et al. Genomic analyses
684 inform on migration events during the peopling of Eurasia. *Nature*. 2016;
- 685 29. Nater A, Mattle-Greminger MP, Nurcahyo A, Nowak MG, de Manuel M, Desai T, et al.
686 Morphometric, Behavioral, and Genomic Evidence for a New Orangutan Species. *Curr Biol*.
687 2017;
- 688 30. Reyes-Centeno H, Ghirotto S, Detroit F, Grimaud-Herve D, Barbujani G, Harvati K.
689 Genomic and cranial phenotype data support multiple modern human dispersals from Africa
690 and a southern route into Asia. *Proc Natl Acad Sci*. 2014;
- 691 31. Tassi F, Ghirotto S, Mezzavilla M, Vilaça ST, De Santi L, Barbujani G. Early modern human
692 dispersal from Africa: Genomic evidence for multiple waves of migration. *Investig Genet*.
693 2015;
- 694 32. Benazzo A, Trucchi E, Cahill JA, Maisano Delser P, Mona S, Fumagalli M, et al. Survival
695 and divergence in a small group: The extraordinary genomic history of the endangered
696 Apennine brown bear stragglers. *Proc Natl Acad Sci*. 2017;
- 697 33. Miller W, Schuster SC, Welch AJ, Ratan A, Bedoya-Reina OC, Zhao F, et al. Polar and
698 brown bear genomes reveal ancient admixture and demographic footprints of past climate
699 change. *Proc Natl Acad Sci*. 2012;
- 700 34. Beaumont MA. Approximate Bayesian Computation in Evolution and Ecology. *Annu Rev*

- 701 Ecol Evol Syst. 2010;
- 702 35. Csilléry K, Blum MGB, Gaggiotti OE, François O. Approximate Bayesian Computation
703 (ABC) in practice. Trends in Ecology and Evolution. 2010.
- 704 36. Wright S. Evolution in Mendelian Populations. Genetics [Internet]. 1931 Mar;16(2):97–159.
705 Available from: <https://www.ncbi.nlm.nih.gov/pubmed/17246615>
- 706 37. Medina P, Thornlow B, Nielsen R, Corbett-Detig R. Estimating the timing of multiple
707 admixture pulses during local ancestry inference. Genetics. 2018;
- 708 38. Felsenstein J. Accuracy of coalescent likelihood estimates: Do we need more sites, more
709 sequences, or more loci? Mol Biol Evol. 2006;
- 710 39. Rowe HC, Renaut S, Guggisberg A. RAD in the realm of next-generation sequencing
711 technologies. Molecular Ecology. 2011.
- 712 40. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation
713 Sequencing Data. BMC Bioinformatics. 2014;
- 714 41. Hudson RR. ms: a Program for Generating Samples Under Neutral Models. Bioinformatics.
715 2002;
- 716 42. Ewing G, Hermisson J. MSMS: A coalescent simulation program including recombination,
717 demographic structure and selection at a single locus. Bioinformatics. 2010;
- 718 43. Hinrichs AS, Raney BJ, Speir ML, Rhead B, Casper J, Karolchik D, et al. UCSC Data
719 Integrator and Variant Annotation Integrator. Bioinformatics. 2016;
- 720 44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
721 Alignment/Map format and SAMtools. Bioinformatics. 2009;
- 722 45. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
723 Bioinformatics. 2010;
- 724 46. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et
725 al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best
726 practices pipeline. Curr Protoc Bioinforma. 2013;

727 47. Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, et al. Fast computation and
728 applications of genome mappability. PLoS One. 2012;

729

730 **Figures Legend:**

731 **Fig 1. One-population models and proportion of True Positives.** A) Demographic models
732 compared: Constant, Bottleneck, Expansion, Structured population. N_i is the effective population
733 size, i the intensity of the bottleneck or of the expansion, T the time of the bottleneck or of the start
734 of the expansion, m is the migration rate. B) True Positives rates. The plot below each of the four
735 models represents the proportion of TPs obtained analyzing pods coming from the above model
736 under 60 combinations of experimental parameters. Different locus lengths are in the x-axes,
737 number of loci is represented by different colors and the number of chromosomes is represented by
738 different symbols.

739

740 **Fig 2. Two-populations models and proportion of True Positives.** A) Demographic models
741 compared: Divergence with isolation, Divergence with migration, Divergence with a single pulse of
742 admixture. N_{anc} is the effective population size of the ancestral population, N_1 and N_2 are the
743 effective population sizes of the diverged populations, T_{sep} is the time of the split, m_{12} and m_{21} the
744 migration rates, T_{adm} is the time of the single pulse of admixture, adm_{12} and adm_{21} the proportions
745 of admixture. B) True Positives rates. The plots have the same features of Fig 1.

746

747 **Fig 3. Multi-populations models and proportion of True Positives.** A) Demographic models
748 compared: Single Dispersal and Multiple Dispersals. The populations sampled are indicated in bold.
749 B) True Positives rates. The plots have the same features of Fig 1.

750

751 **Fig 4. Posterior Probabilities for the MDM.** Left panel: posterior probabilities obtained analyzing
752 6 Papuan individuals from [28], simulating data without (PWR) or considering (PR) intra-locus

753 recombination. Right panel: posterior probabilities obtained analyzing 25 Papuan individuals from
754 [27], simulating data without (MWR) or considering (MR) intra-locus recombination.

755

756 **Fig 5. Demographic models tested to study the evolutionary history of Orangutan species.** A-

757 Four demographic models compared. The numbers in the black boxes indicate the proportion of TP

758 calculated analyzing 1,000 pods coming from that demographic model. NT, Sumatran populations

759 north of Lake Toba; ST, the Sumatran population south of Lake Toba; BO, Bornean populations. B-

760 Number of votes associated to each model by ABC-RF and posterior probability of the most

761 supported model (model 1a).

762

763 **Supporting Informations**

764 **Supplementary Figures:**

765 **S1 Fig. Example of a frequency distribution of segregating sites.** On the x-axes: the number of

766 segregating sites (here from 0 to 20). On the y-axes: the number of genomic loci showing a certain

767 number of segregating sites. We calculated four FDSS for each pair of populations compared

768 (representing the genomic distribution of segregating sites private of the first population, private of

769 the second population, shared, and fixed for different alleles). In the one-population models we used

770 a single FDSS.

771

772 **S2 Fig. Proportion of True positives for the one-population structured model as a function of**

773 **the migration rate (A) and the number of demes considered (B).** (A) Each plot represents the

774 proportion of pods from the structured model assigned to each of the four one-population models

775 with the migration rates among demes in the structured model constrained at ranges of increasing

776 values (from $1 \cdot 10^{-5}$ to $1 \cdot 10^{-1}$). All the plots consider two chromosomes and a specific combination

777 of locus length and number of loci; the number of demes in the structured model is fixed to four. In

778 general, the TP rate (in dark blue) decreases as increasing the migration rate among demes, with the

779 constant model erroneously recognize as the true model for higher migration rates. (B) Proportion
780 of pods from the structured model assigned to each of the four one-population models as a function
781 of the number of demes (from 2 to 10). The TP rate increase with the number of demes, regardless
782 of the level of migration among demes.

783

784 **S3 Fig. FDSS generated under the one-population models for each number of chromosomes**
785 **tested.** Each plot represents the FDSS simulated under each of the four one-population models
786 considering 1,000 fragments of 1,000 base pair length, for a specific number of chromosomes
787 sampled. Going from two to twenty chromosomes, we observe an increase of the overlapping
788 between the FDSS generated under the Constant and the Structured model, thus possibly explaining
789 the decrease in the models' identifiability as increasing the number of chromosomes considered.

790

791 **S4 Fig. Proportion of True Positives for the one-population models with recombination.** The
792 plots have the same features of Fig 1, we considered a per nucleotide per generation recombination
793 rate of $1 \cdot 10^{-8}$.

794

795 **S5 Fig. Proportion of True Positives for the Pulse of Admixture model as a function of the**
796 **time span between the divergence and the admixture.** We expressed the x-axes as $(T_{\text{sep}} -$
797 $T_{\text{adm}})/T_{\text{sep}}$, so as to normalize the results respect to the age of the divergence. Each plot represents
798 the results for a different locus length.

799

800 **S6 Fig. Proportion of True Positives for the two-population models with recombination.** The
801 plots have the same features of Fig 1, we considered a per nucleotide per generation recombination
802 rate of $1 \cdot 10^{-8}$

803

804 **S7 Fig. Proportion of True Positives for the MDM a function of the time span between the**

805 **divergence time of the African ghost populations and the second exit (Delta τ YG-OOA2).** The
806 time difference between the divergence time of the African ghost populations and the second exit
807 from Africa is on the x-axis and it is expressed in years (considering a generation time of 29 years).
808 Each plot reports the results for a different locus length.

809

810 **S8 Fig. Expected topologies under the SDM and MDM.** T1, T2 and T3 are baseline topologies,
811 of which T1 is shared between SDM and MDM, whereas T2 and T3 are exclusively originated by
812 MDM. R is the recombinant topology expected under MDM.

813

814 **Supplementary Tables:**

815 **S1 Table. Demographic parameters and prior distributions of one-population models.**
816 Mutation and Recombination rates are expressed per nucleotide per generation.

817

818 **S2 Table. Demographic parameters and prior distributions of two-population models.**
819 Mutation and Recombination rates are expressed per nucleotide per generation. Time is in
820 generations.

821

822 **S3 Table. RF Prior Error Rates using the complete FDSS or its first four moments.** We
823 calculated the Prior Error Rates under the complete SDFD and under a summary of the distribution
824 (i.e. the first four moments). We analyzed ten combinations of experimental parameters (indicated
825 in the "ID" column), comparing the three two-population divergence models. l = locus length, n =
826 number of loci, r = recombination rate, n_c = number of chromosomes. The Prior Error Rate is a
827 measure of the global quality of the RF classifier (the lower the better), which depends on the
828 summary statistics used. The complete FDSS always generate lower values of Prior Error Rates.

829

830 **S4 Table. Demographic parameters and prior distributions of multi-population models:**

831 **Single Dispersal model.** Migration and admixture rates are expressed per generation, times in
832 years. We considered a generation time of 29 years as in [27]. Per nucleotide per generation
833 mutation and recombination rates are fixed as in [27].

834

835 **S5 Table. Demographic parameters and prior distributions of multi-population models:**

836 **Multiple Dispersals model.** Migration and admixture rates are expressed per generation, times in
837 years. We considered a generation time of 29 years as in [27]. Per nucleotide per generation
838 mutation and recombination rates are fixed as in [27].

839

840 **S6 Table. Power Analysis of the Multi-Populations models. Comparison of the results with**

841 **and without intra-locus recombination for the four combinations of experimental parameters**

842 **tested.** The first four rows present the results without considering intra-locus recombination,

843 whereas the last four rows refer to the same combinations of experimental parameters simulated

844 considering a per nucleotide per generation recombination rate of $1.12 \cdot 10^{-8}$. TP% columns

845 represent the percentage of True Positives calculated analyzing 1,000 pods coming from SDM and

846 MDM. SDM mpp and MDM mpp represent the mean posterior probability value obtained for the

847 datasets that have been correctly assigned to the true model.

848

849 **S7 Table. Genomes used for the comparison of SDM and MDM using real data.**

850

851 **S8 Table. Model selection results using Papuan individuals from [28] without intra-locus**

852 **recombination.** The first column represents the id of the Papuan sample used in that comparison as

853 reported in the dataset. The second column shows the model selected by the ABC-SeS procedure.

854 The third and the fourth columns represent the proportion of votes assigned to SDM and MDM by

855 the RF algorithm. The last column is the posterior probability of the most supported model.

856

857 **S9 Table. Model selection results using Papuan individuals from [28] considering intra-locus**
858 **recombination.** The first column represents the id of the Papuan sample used in that comparison as
859 reported in the dataset. The second column shows the model selected by the ABC-SeS procedure.
860 The third and the fourth columns represent the proportion of votes assigned to SDM and MDM by
861 the RF algorithm. The last column is the posterior probability of the most supported model.

862

863 **S10 Table. Model selection results using Papuan individuals from [27] without intra-locus**
864 **recombination.** The first column represents the id of the Papuan sample used in that comparison as
865 reported in the dataset. The second column shows the model selected by the ABC-SeS procedure.
866 The third and the fourth columns represent the proportion of votes assigned to SDM and MDM by
867 the RF algorithm. The last column is the posterior probability of the most supported model.

868

869 **S11 Table. Model selection results using Papuan individuals from [27] considering intra-locus**
870 **recombination.** The first column represents the id of the Papuan sample used in that comparison as
871 reported in the dataset. The second column shows the model selected by the ABC-SeS procedure.
872 The third and the fourth columns represent the proportion of votes assigned to SDM and MDM by
873 the RF algorithm. The last column is the posterior probability of the most supported model.

874

875 **S12 Table. Genomes used for the comparison of the four Orangutan evolutionary scenarios.**

876

877 **S13 Table. Demographic parameters and prior distributions for Model 1a.** Migration rates are
878 expressed per generation, times in years. We used a generation time of 25 years as in [29]. The per
879 nucleotide per generation mutation rate is fixed as in [29].

880

881 **S14 Table. Demographic parameters and prior distributions for Model 2a.** Migration rates are
882 expressed per generation, times in years. We used a generation time of 25 years as in [29]. The per

883 nucleotide per generation mutation rate is fixed as in [29].

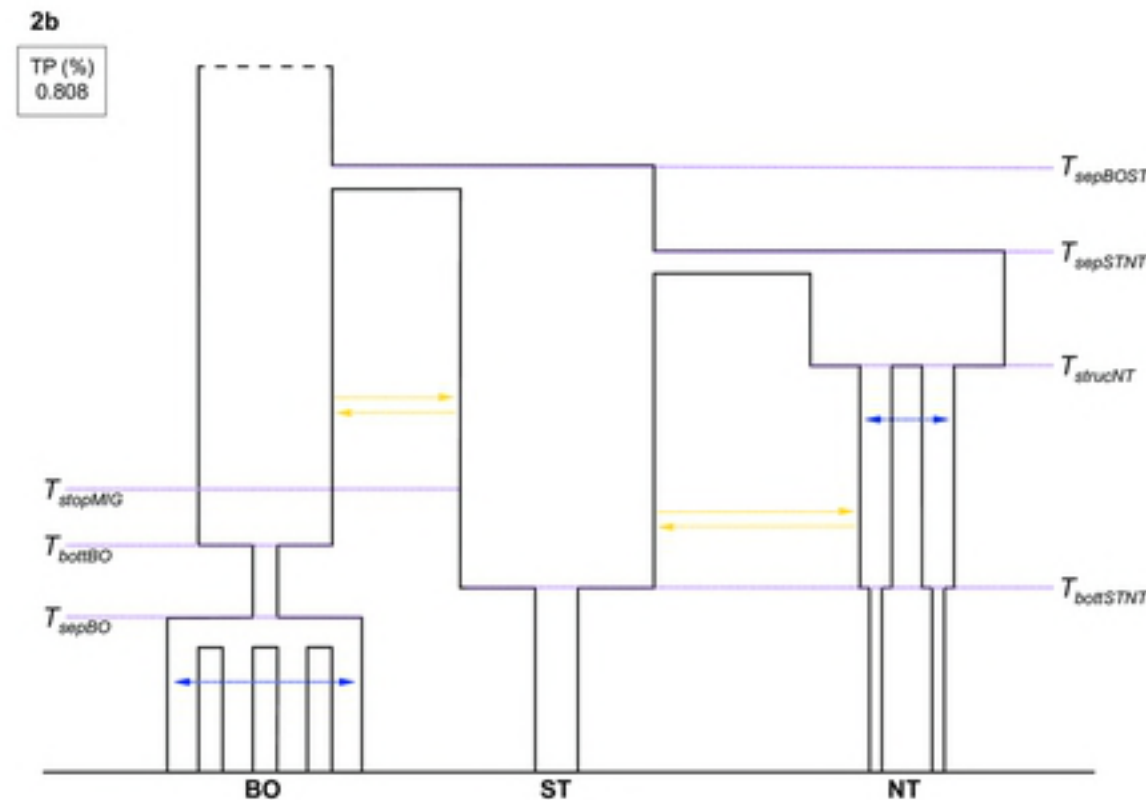
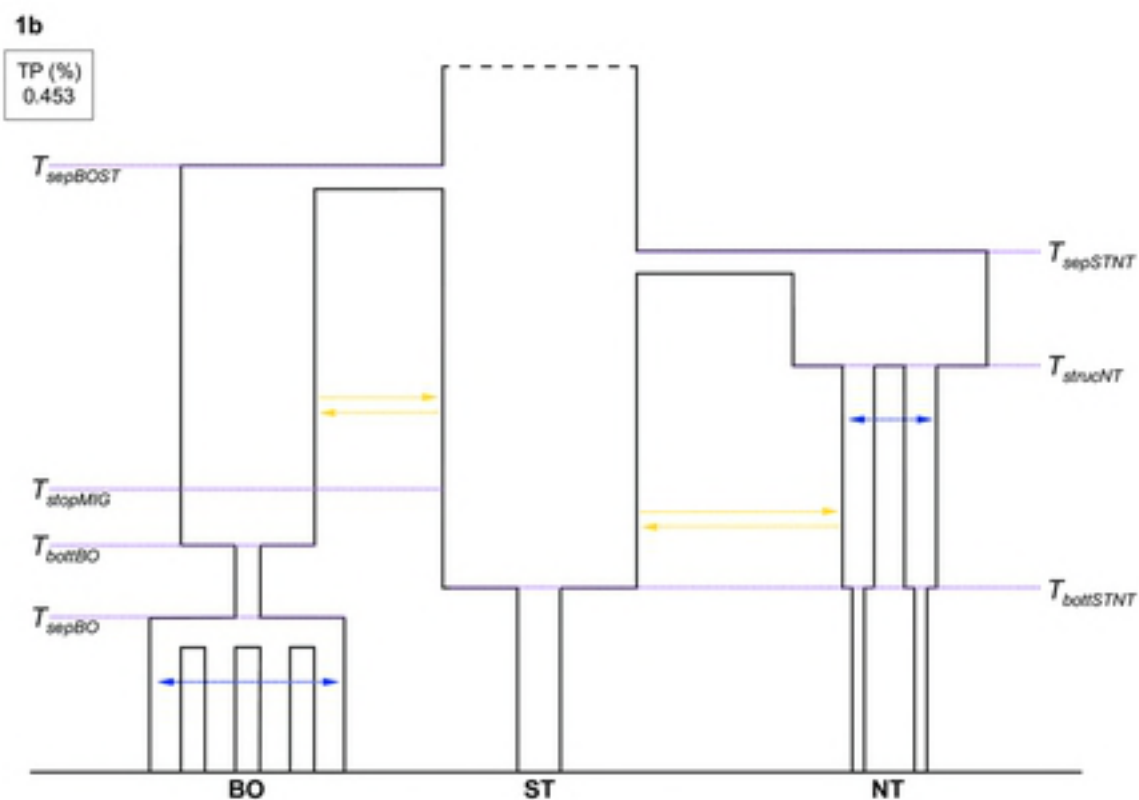
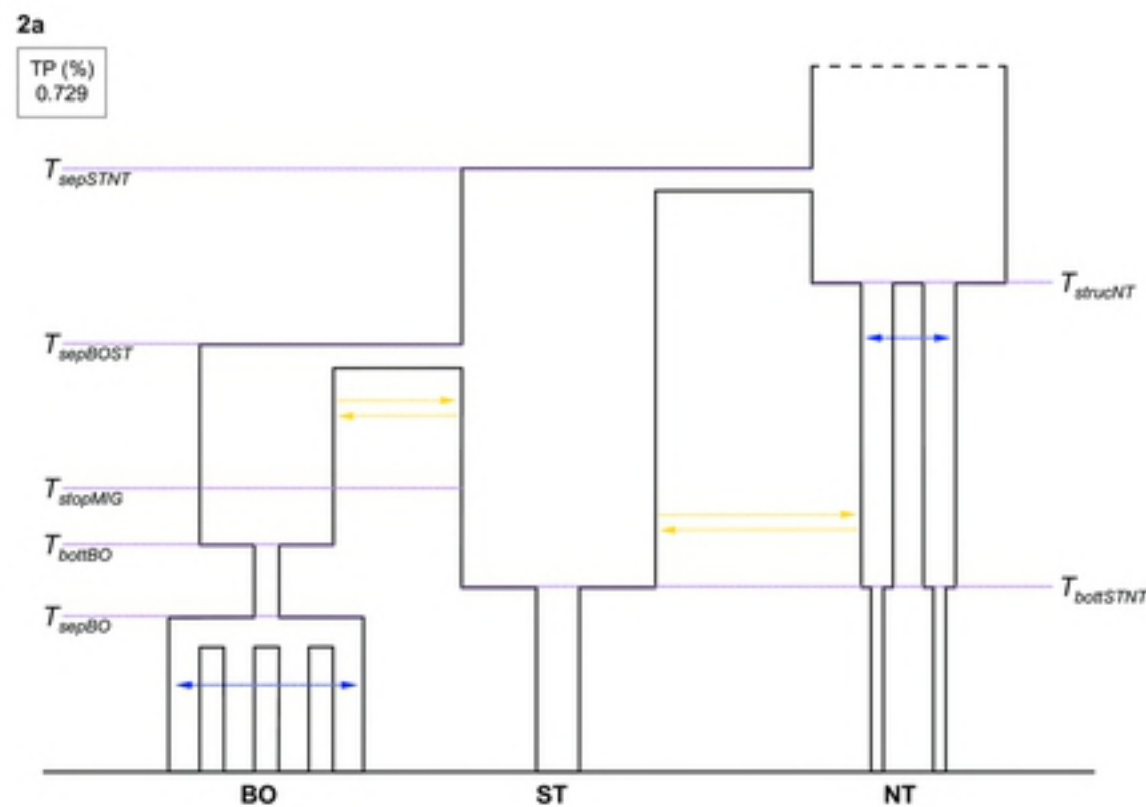
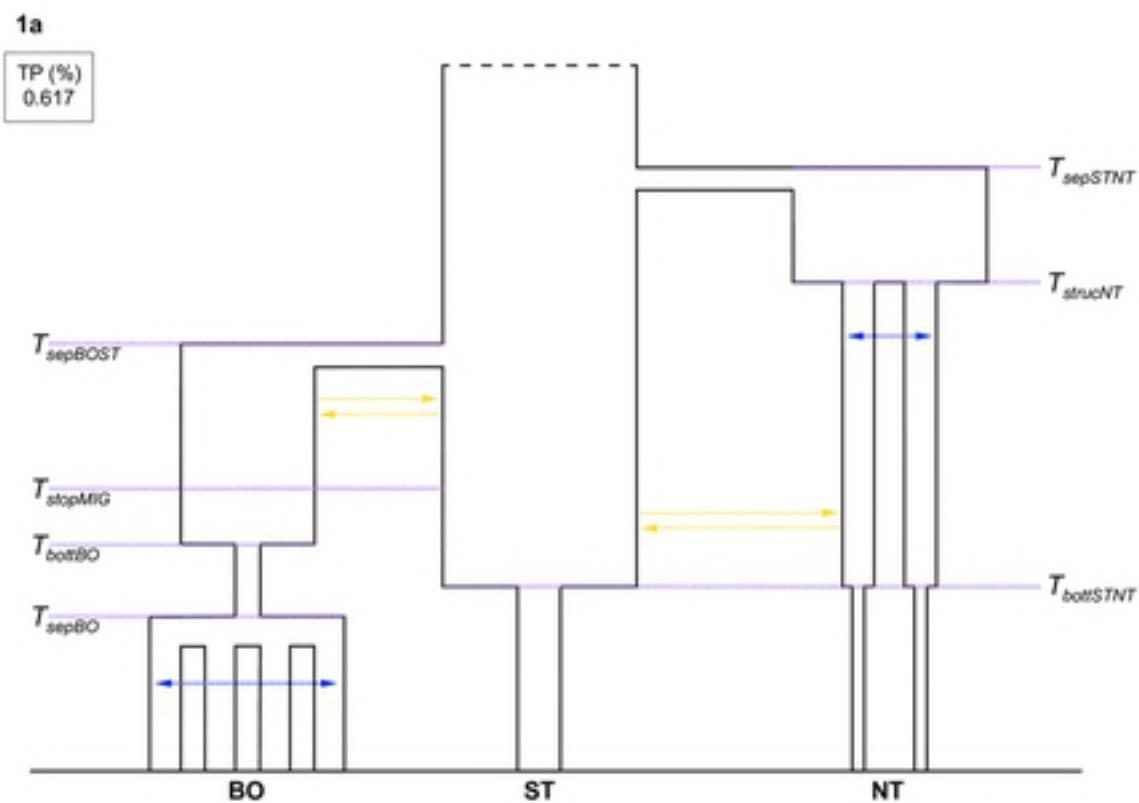
884

885 **S15 Table. Demographic parameters and prior distributions for Model 1b.** Migration rates are
886 expressed per generation, times in years. We used a generation time of 25 years as in [29]. The per
887 nucleotide per generation mutation rate is fixed as in [29].

888

889 **S16 Table. Demographic parameters and prior distributions for Model 2b.** Migration rates are
890 expressed per generation, times in years. We used a generation time of 25 years as in [29]. The per
891 nucleotide per generation mutation rate is fixed as in [29].

892



B

Selected Model	Votes model 1A	Votes model 2A	Votes model 1B	Votes model 2B	PP
1A	0.398	0.190	0.292	0.120	0.489

Figure 5

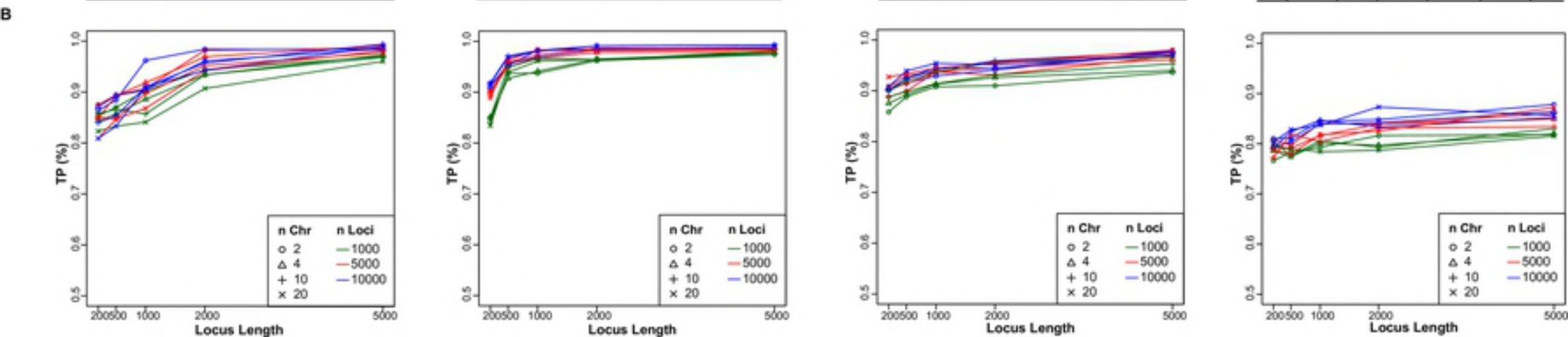
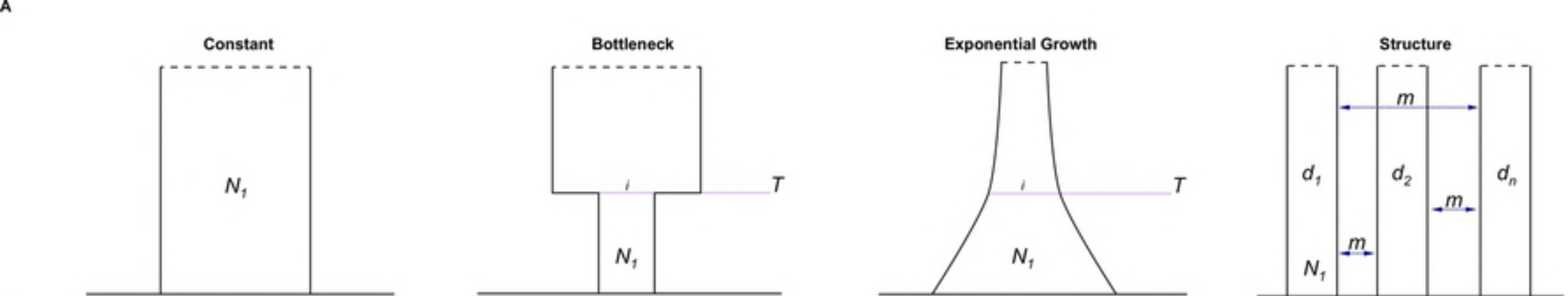


Figure 1

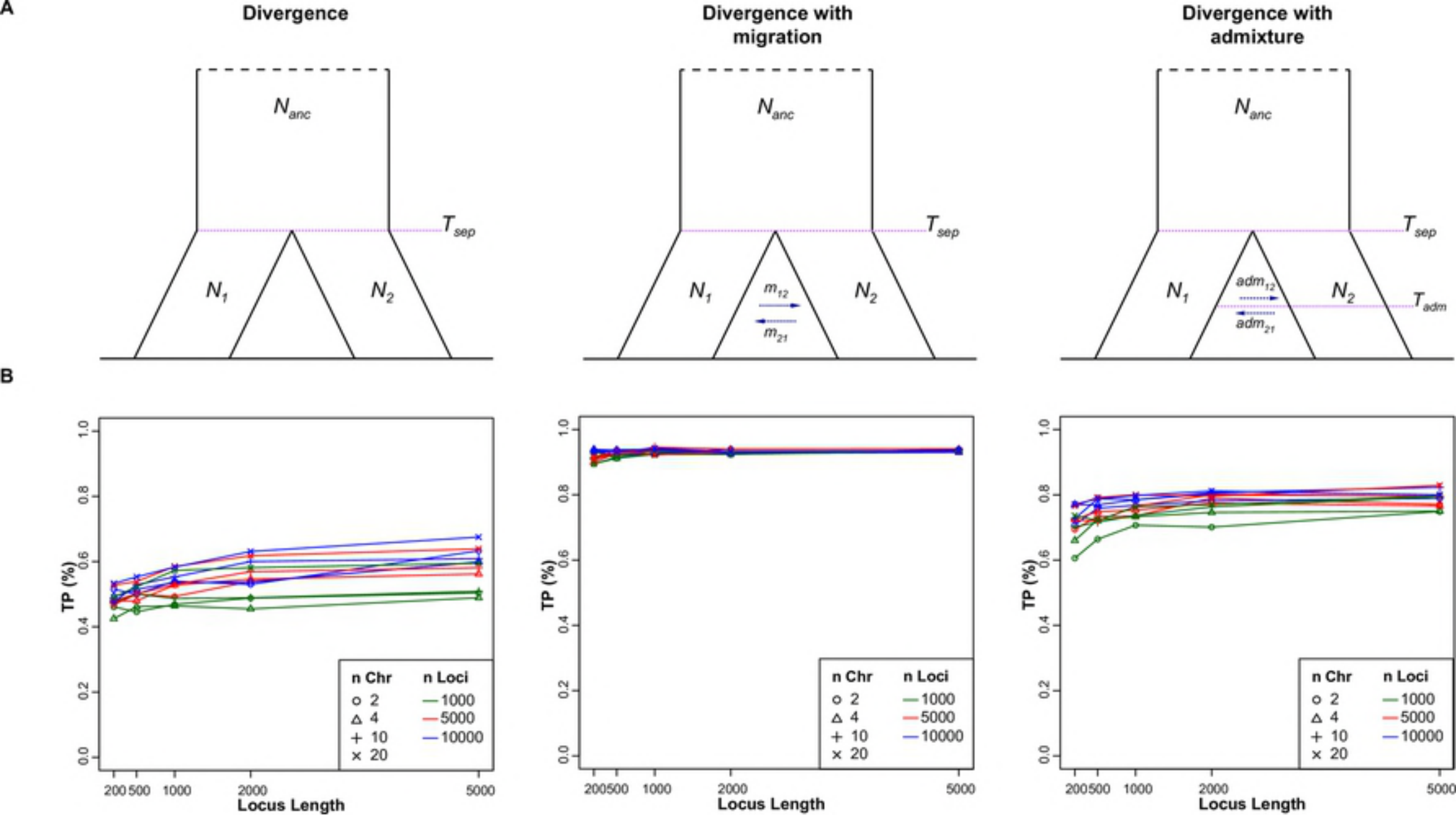


Figure 2

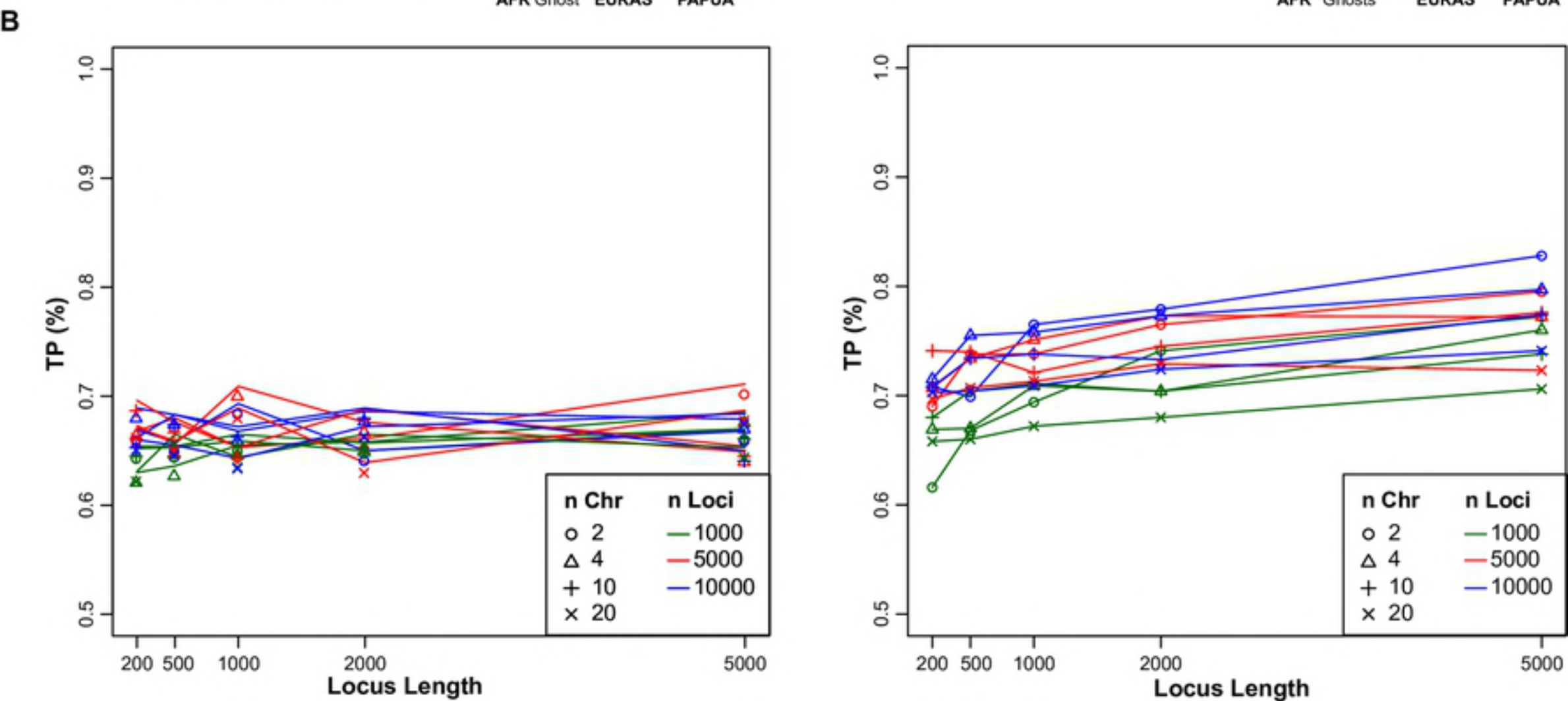
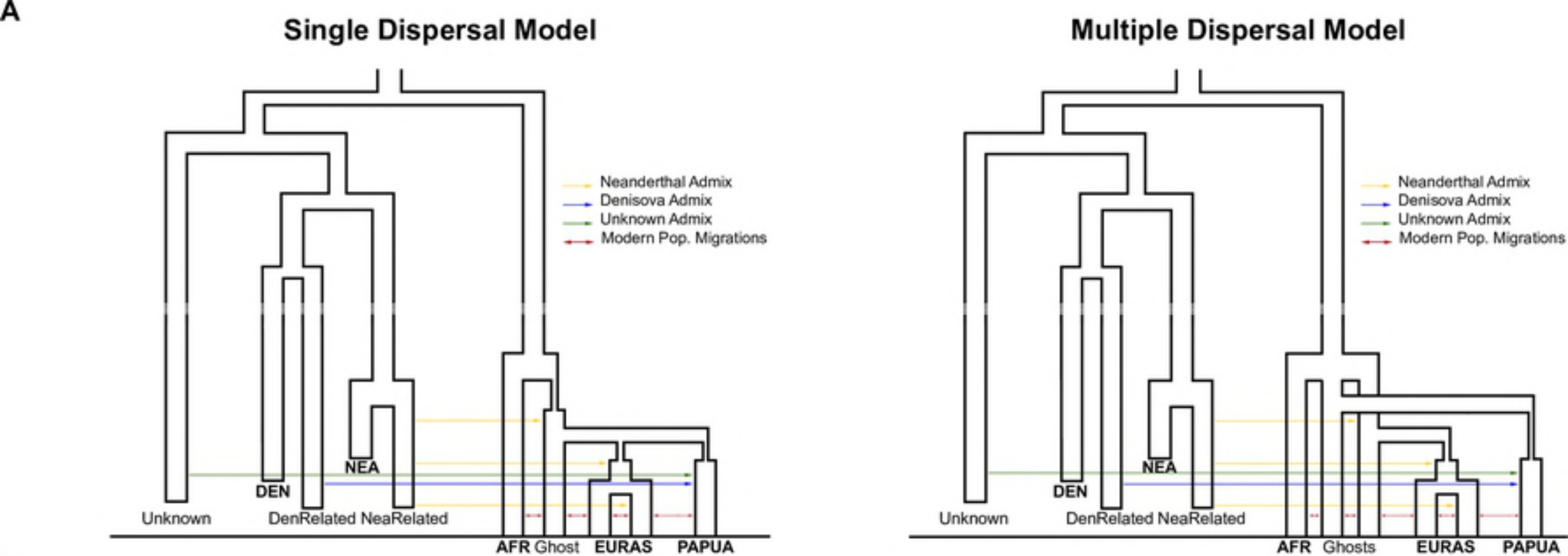


Figure 3

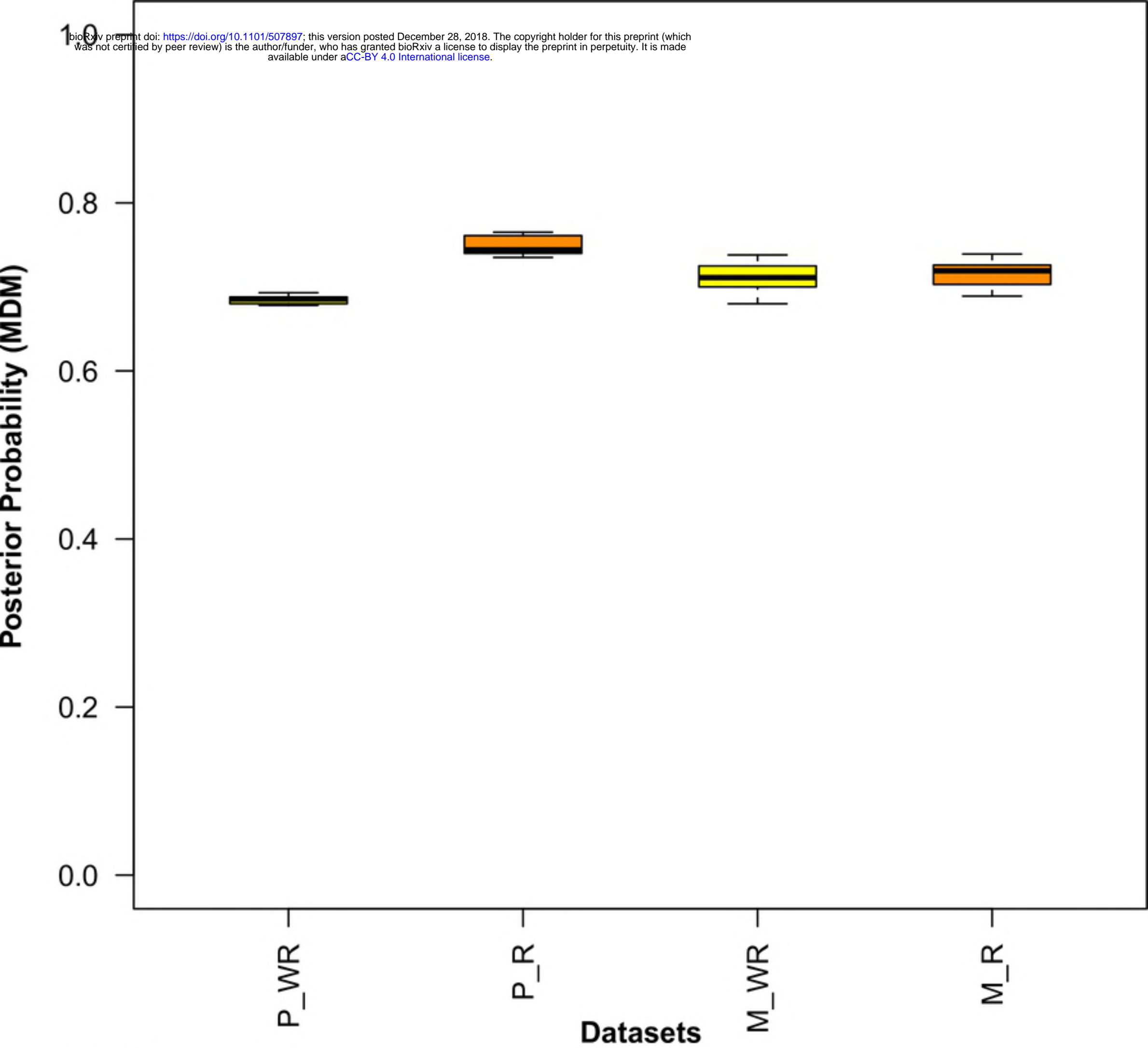


Figure 4