Identifying loci under selection via explicit demographic models

Hirzi Luqman[1*], Alex Widmer[1], Simone Fior[1†], Daniel Wegmann[2,3†*]

[1] Institute of Integrative Biology, ETH Zurich, Universitätstrasse 16, 8092 Zürich, Switzerland

[2] Department of Biology, University of Fribourg, Fribourg, Switzerland

[3] Swiss Institute of Bioinformatics, Fribourg, Switzerland

[†] denotes joint senior authors

Corresponding authors: hirzi.luqman@env.ethz.ch and daniel.wegmann@unifr.ch

Abstract

Adaptive genetic variation is a function of both selective and neutral forces. In order to accurately identify adaptive loci, it is hence critical to account for demographic history. Theory suggests that signatures of selection can be inferred using the coalescent, following the premise that the genealogies of selected loci deviate from neutral expectations. Here, we build on this theory to develop an analytical framework to identify Loci under Selection via explicit Demographic models (LSD). Under this framework, signatures of selection are inferred by demographic parameters, rather than through isolated summary statistics, and demographic history is accounted for explicitly. Given that demographic models can incorporate directionality, we show that LSD can provide information on the environment in which selection acts on a population. This can prove useful in dissecting the genomics of local adaptation, by characterising genetic trade-offs and extending the concepts of antagonistic pleiotropy and conditional neutrality from ecological theory to practical application in genomic data. We implement LSD via Approximate Bayesian Computation and demonstrate, via simulations, that LSD has high power to identify selected loci across a large range of demographic-selection regimes, including complex demographies, and that the directionality of selection can be inferred accurately for identified candidates. Using the same simulations, we further characterise the behaviour of isolation-with-migration models conducive to the study of local adaptation under regimes of selection. Finally, we apply LSD to the

detection and characterisation of loci underlying floral guides in *Antirrhinum majus*, and find consistent results with previous studies.

Keywords

Approximate Bayesian computation, demography, genetic trade-offs, genome scan, local adaptation, selection

1 INTRODUCTION

Elucidating the genetic basis of adaptation and identifying genetic determinants of population and species divergence are key foci in evolutionary biology. In natural systems, genetic variation is shaped by the demographic history, driven by the neutral processes of mutation, migration and drift, together with natural selection on loci underlying adaptive traits. Conceptually, while all gene genealogies are constrained by the demographic history of the population, the genealogies of loci affected by selection are perturbed and may differ in key characteristics compared to those evolving under neutrality, though converging patterns can arise (Bierne, Welch, Loire, Bonhomme, & David, 2011; Edmonds, Lillie, & Cavalli-Sforza, 2004; Excoffier, Foll, & Petit, 2009; J. Li et al., 2012; Montgomery Slatkin & Excoffier, 2012). Disentangling the genomic signatures of these two processes, e.g. correctly identifying adaptive loci, remains a prevailing challenge in the field of population

56 genetics (Biswas & Akey, 2006; Horscroft, Ennis, Pengelly, Sluckin, & Collins, 2019;
57 Luikart, England, Tallmon, Jordan, & Taberlet, 2003).
58
59 A multitude of methods have been developed that identify loci under selection as
60 those whose summary statistics deviate from the genome-wide distribution. These
61 "outlier" approaches can generally be grouped into three classes: those that 1)
62 highlight regions of elevated differentiation between populations (via e.g. F_{ST}-related
63 statistics), 2) highlight regions of perturbed site frequency spectrum (SFS) via
64 diversity or diversity-related estimators (e.g. π, Tajima's D) and 3) highlight regions
65 of extensive linkage disequilibrium (LD) via haplotype statistics (e.g. EHH, iHS)
66 (Beaumont & Nichols, 1996; Biswas & Akey, 2006; Luikart et al., 2003; Oleksyk,
67 Smith, & O'Brien, 2010; Vitti, Grossman, & Sabeti, 2013). While in empirical studies
68 inference of selection is often achieved through corroborating evidence from multiple
69 measures, each approach generally builds upon a single statistic that individually
70 captures only a partial aspect of the effect of selection on the underlying genealogies.
71 Additionally, reliance on statistics that describe the sample rather than on
72 parameters that define the population imply that these approaches provide an
73 incomplete description of the system. Combining the information from multiple
74 statistics, either indirectly to inform underlying parameters via demographic
75 modelling or directly via composite statistics, has the potential to provide insight and

increase power in the localisation of targets of selection (Grossman et al., 2010; Sugden et al., 2018; Vitti et al., 2013; Zeng, Shi, & Wu, 2007).

Under the premise that adaptive genetic variation is a function of both selective and neutral forces, accounting for the demographic history of the study system is critical for the correct identification of selected loci (François, Martins, Caye, & Schoville, 2016; Hoban et al., 2016). To achieve this, empirical studies generally employ a demographic null model which describes the neutral distribution of the statistics used to infer selection (e.g. F_{ST}; Beaumont & Nichols, 1996; Eckert et al., 2010; Excoffier et al., 2009; Hofer, Ray, Wegmann, & Excoffier, 2009), or estimates of sample relatedness or covariance to correct for neutral population structure (Bonhomme et al., 2010; Engelhardt & Stephens, 2010; Gautier, 2015; Günther & Coop, 2013; Price et al., 2006). These approaches can be successful in controlling the confounding influence of demography, however they generally suffer from one or more of the following caveats; relying on post-hoc treatment (Eckert et al., 2010; Excoffier et al., 2009; Hofer et al., 2009), not providing direct estimates for underlying demographic parameters (sample covariance approaches) or assuming a set of rather simple demographic models (Foll & Gaggioti 2008, Beaumont & Nichols 1996). Alternatively, a (complex) demographic model might be inferred from putatively neutral sites and loci under selection may be identified as those for which the an additional selection parameter is required (e.g. Williamson et al., 2005) or as those for

97 which locus-specific estimates of demographic parameters differ from genome-wide
98 estimates (e.g. Sousa, Carneiro, Ferrand, & Hey, 2013). Such an approaches may not
99 only account for demography explicitly and concurrently with the inference of
100 selection, but may also directly estimate the probability of a locus being under
101 selection if performed in a Bayesian framework, and hence avoid arbitrary
102 thresholds.
103
104 Coalescent theory provides a powerful framework to infer the demographic
105 processes that drive observed genetic variation (Kingman, 1982; Wakeley, 2001).
106 While the coalescent was derived to model neutral variation, theory suggests that it
107 can be exploited to infer signatures of selection following the premise that the
108 genealogies of selected loci are expected to deviate from neutral expectations (Barton
109 & Bengtsson, 1986; Charlesworth, 2009; Charlesworth, Nordborg, & Charlesworth,
110 1997; Fusco & Uyenoyama, 2011; Galtier, Depaulis, & Barton, 2000; Gossmann,
111 Woolfit, & Eyre-Walker, 2011; Petry, 1983; Sousa, Carneiro, Ferrand, & Hey, 2013).
112 Under coalescent theory, demographic models are parametrised by effective
113 population size(s) (N_E) and in the case of multiple populations additionally by
114 effective migration rate(s) (m_E), which respectively describe the level of drift and
115 gene flow within and between populations (Charlesworth, 2009; Petry, 1983).
116 Importantly, both N_E and m_E may change through time. Different modes of selection
117 and adaptive processes can be expected to alter these demographic parameters in

different ways. In a single population, imprints of selection may be reflected in variations in N_E, which is expected to be reduced in regions undergoing selective sweeps and increased in regions undergoing diversifying selection (Galtier et al., 2000; Gossmann et al., 2011). In the case of two or more populations connected by gene flow, divergent selection is expected to reduce m_E at selected and linked sites (Petry, 1983), while balancing selection and adaptive introgression may be expected to increase m_E at affected regions. Such theory implies that information can be acquired not just on the strength (via the magnitude of deviation of N_E or m_E from neutral expectations) and mode of selection (i.e. whether these deviations in N_E or m_E are characterised by a reduction or elevation relative to neutral expectations), but also potentially of the population (environment) in which selection acts (i.e. whether the reduction in m_E is in a particular direction, in a multi-population model). This paradigm presents a unique opportunity to address the genetic basis of local adaptation, a key concept in ecological genetics.

The framework to describe the genetic basis of local adaptation derives from ecological theory that relates the fitness advantage of alternate alleles under different environmental conditions (Savolainen, Lascoux, & Merilä, 2013). Specifically, alternate alleles may confer higher fitness in their respective local environment but reduced fitness in the foreign environment, i.e. antagonistic pleiotropy (AP), or alleles may confer higher fitness in their local environment but have no differential

139 effect relative to the alternative allele in the foreign environment, i.e. conditional
140 neutrality (CN) (Anderson, Lee, Rushworth, Colautti, & Mitchell-Olds, 2013;
141 Kawecki & Ebert, 2004; Savolainen et al., 2013). Such genetic trade-offs are generally
142 investigated via estimates of fitness in genetic crosses grown in reciprocal transplant
143 experiments (Kawecki & Ebert, 2004; Savolainen et al., 2013). In such experiments, as
144 well as in natural populations inhabiting different environments and connected by
145 gene flow, directional selection acting on phenotypic traits modulates fitness and
146 purges individuals that carry maladaptive alleles, e.g. via hybrid (Naisbit, Jiggins, &
147 Mallet, 2001; Rundle & Whitlock, 2001; Schluter, 2000) or immigrant (Nosil, Vines, &
148 Funk, 2005) inviability, or lower fecundity. This effectively reduces m_E at these
149 (selected) loci between these populations, with the reduction (from neutral
150 expectations) proportional to the strength of selection imposed on the alternative
151 alleles in each environment. Such genetic trade-offs have long been assumed to
152 underlie phenotypic trade-offs in populations connected by gene flow and the
153 modelling of selection in terms of m_E via the coalescent provides a means to explicitly
154 describe and disentangle such processes.
155
156 In this paper, we outline a conceptual and methodological framework for
157 identifying Loci under Selection via explicit Demographic models called LSD, that
158 scales to genomic data. Our approach explicitly accounts for demography in the
159 identification of candidate loci, avoids reliance on singular summary statistics, and

elucidates the driving parameters underlying differentiation at putative selected loci. Furthermore, applied in a probabilistic Bayesian framework, our approach does not rely on arbitrary thresholds to delimit candidate loci, addressing an inherent limitation of many outlier approaches. While LSD is flexible regarding the choice of demographic model and can in principle accommodate any discrete population model (including single population and stepping-stone models) as well as detect different modes of selection, we here demonstrate LSD's utility in studies of local adaptation by focusing on the detection of loci under divergent selection under isolation-with-migration (IM) models. We validate and assess the performance of LSD via extensive simulations, and apply the method to the detection of functionally validated loci underlying floral guides in two parapatric subspecies of *Antirrhinum majus* (common snapdragon) (Schwinn et al., 2006; Tavares et al., 2018).

2 MATERIALS AND METHODS

2.1 Model

We begin by outlining the conceptual framework underlying LSD. Consider a demographic model $\mathcal{M}$, parametrised by demographic parameters $\boldsymbol{\theta}$, that generates genetic data $\boldsymbol{D}$. To quantify deviations from neutrality, LSD first estimates the demographic parameters $\widehat{\boldsymbol{\theta}}$ from a collection of loci assumed to be neutral (Figure S1). In a second step, LSD performs demographic inference on all loci and

180 determines the posterior distribution $\pi_l = \pi(\boldsymbol{\theta}|\boldsymbol{D_l})$ for each locus. Finally, LSD
181 assesses the concordance of $\widehat{\boldsymbol{\theta}}$ with π_l by determining p_l, the highest posterior
182 density interval (HPDI) of π_l that contains $\widehat{\boldsymbol{\theta}}$. Here, $q_l = 1 - p_l$ corresponds to the
183 false-discovery-rate (FDR) of identifying locus l as incompatible with $\widehat{\boldsymbol{\theta}}$. The joint
184 posterior distribution π_l may further provide information on the magnitude and
185 directionality of selection.
186
187 Given that the evaluation of the likelihood is non-trivial and may be
188 intractable under more complex models, we resort to an approximate approach
189 (Marjoram & Tavaré, 2006) (Figure 1). Under an Approximate Bayesian Computation
190 (ABC) framework, the likelihood is approximated by simulations, the outcomes of
191 which are compared with observed data in terms of summary statistics. That is, we
192 find the set of parameters $\boldsymbol{\theta}$ that minimise the distance between the observed data $\boldsymbol{D}$
193 and the simulated data $\boldsymbol{D}'$. To efficiently evaluate this, we reduce the dimensionality
194 of the data via summarising them into a set of lower-dimensional summary statistics
195 $\boldsymbol{S}$ and $\boldsymbol{S}'$, which are selected to capture the relevant information in $\boldsymbol{D}$ and $\boldsymbol{D}'$,
196 respectively (Joyce & Marjoram, 2008; Peter, Huerta-Sanchez, & Nielsen, 2012).
197
198 An appropriate model for generating simulated genetic data is provided by
199 coalescent theory (Kingman, 1982; Wakeley, 2001), parametrised by population
200 demographic parameters $\boldsymbol{\theta} = \{\boldsymbol{N}_E, \boldsymbol{M}_E, \mu\}$, where $\boldsymbol{N}_E$ refers to the vector of effective

population sizes, $\boldsymbol{M}_E$ to the vector of effective migration rates, and μ to the mutation rate. We stress that population sizes and migration rates may vary through time. To minimise the distance $\boldsymbol{D}$ from $\boldsymbol{D}'$, model space in addition to parameter space should be explored.

2.2 Implementation

The analytical framework described above is practically implemented as follows (Figure 1). Coalescent simulations are generated by a coalescent simulator, e.g. msms (Ewing & Hermisson, 2010), under a user-defined demographic model. Definition and choice of the demographic model should i) be informed by knowledge of the study system, ii) be motivated by the model's capacity to provide a useful approximation of a biological process of interest, and iii) be sufficiently simple to remain computationally tractable. Additionally, given that we condition the inference of selection on demographic parameters, the model should be formulated according to whether deviation in $\boldsymbol{N}_E$ or $\boldsymbol{M}_E$ is desired for the inference of selection. Importantly, the model should be validated by demonstrating that the observed data can be accurately and sufficiently captured (Figure S7).

The processing, format and final output of observed genetic data will often differ from that of raw coalescent simulations, given that observed genetic data may be subject to various pre-sequencing (e.g. pooling), sequencing (e.g. sequencing

222 errors, stochastic sampling of reads) and post-sequencing (e.g. filters) events that
223 perturb and reformat the data from the original source. We thus implemented two
224 complimentary programs that interface with coalescent simulators to replicate
225 observed sequencing pipelines and generate simulated sequencing data: *LSD-High*
226 can accommodate and simulate both individual and pooled data and assumes mid to
227 high coverage (>10x) data, while *LSD-Low* accepts individual data and can
228 additionally accommodate low coverage (>2x) data by utilising genotype likelihoods
229 via msToGLF and ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014). A suite of
230 summary statistics is then calculated for the simulated and observed data via the
231 same programs. Summary statistics currently implemented include the number of
232 segregating sites (S), private S, nucleotide diversity (π), Watterson's estimator (θ_W),
233 Tajima's D (θ_D), relative divergence (F_{ST}), absolute divergence (D_{XY}), and site
234 frequencies, though in principle any summary statistic can be included, contingent
235 on the data and appropriate additions to the programs' scripts. To account for
236 potential correlation between summary statistics and to retain only their informative
237 components, we apply a Partial Least Squares transformation (Wegmann,
238 Leuenberger, & Excoffier, 2009).
239
240 The estimation of demographic parameters is performed using ABCtoolbox
241 (Wegmann, Leuenberger, Neuenschwander, & Excoffier, 2010), via the ABC-GLM
242 algorithm using the subset of n simulations closest to the observed summary

243 statistics, separately for the putative neutral regions and for a sliding window across
244 the genome. Here, windows may be interpreted as loci, and for modelling simplicity
245 we assume that recombination is free between loci and fixed within. To acquire a set
246 of putative neutral regions, we assume that sites belonging to a particular structural
247 or functional class are selectively neutral (Williamson et al., 2005). We may for
248 instance rely on genomic regions outside all structural annotations unlinked to all
249 structural genomic elements (i.e. with a conservative flanking distance, informed by
250 the linkage disequilibrium decay distance). Alternatively, a more naïve
251 approximation may rely on the whole genome to provide genome-wide expectations,
252 which may in some cases, but not always (Begun et al., 2007; Fay, Wyckoff, & Wu,
253 2002; H. Li & Stephan, 2006), reflect the neutral case. In order to remain
254 computationally tractable, departure from neutrality is evaluated only for that subset
255 of demographic parameters informative of selection, while assuming the others to be
256 shared among neutral and selected loci.
257

258 ## 2.3 Simulations

259 **Demography** To test the performance of the LSD implementation, we
260 simulated pseudo-observed genomes using the program msms under different
261 demographic and selection parameter values (regimes), focusing on isolation-with-
262 migration (IM) models relevant for the characterisation of local adaptation. We

263 simulated three models representing different levels of complexity in terms of
264 population structure and evolutionary history (Figure 2): i) A simple 2-deme IM
265 model (model $\mathcal{M}_1$), with effective population sizes $N_1 = N_2 = 10{,}000$ and symmetric
266 migration rates $M_{12} = M_{21} = M$. ii) A 6-deme IM model comprising 2 contrasting
267 environments of 3 demes each (model $\mathcal{M}_2$), structured as 'islands' with effective
268 population size $N_i = 1{,}000$ connected via migration $M_{ic} = 100$ to meta-population
269 'continents' of size $N_1 = N_2 = 100{,}000$. These continents exchange migrants at
270 symmetric rates $M_{12} = M_{21} = M$. iii) A 2-deme divergence with bottleneck and
271 exponential growth model (model $\mathcal{M}_3$). Under this model, two demes split from an
272 ancestral population of size $N_A = 10{,}000$ at $T_D = 200{,}000$ generations ago. Following
273 divergence, deme 2 stays constant at $N_2 = 10{,}000$, while deme 1 undergoes a sudden
274 bottleneck, immediately followed by exponential growth with rate $\alpha = 2$ until
275 $T_G = 160{,}000$ generations ago, at which $N_1 = 10{,}000$ is reached and thereafter
276 remains constant. Demes 1 and 2 are initially separated with no gene flow between
277 demes (isolation period), after which secondary contact is established at $T_c = 20{,}000$
278 generations ago with symmetric migration rates $M_{12} = M_{21} = M$. In all models, we
279 used neutral migration rates $M = 0.5, 5$ and 50 migrants per generation and inferred
280 selection as deviations from these rates. We use model $\mathcal{M}_1$ to represent a simplified,
281 generalised model of local adaptation, model $\mathcal{M}_2$ to represent a more complex case of
282 local adaptation comprising multiple, structured populations and model $\mathcal{M}_3$ to
283 reflect a scenario typical of glacial-induced secondary contact population dynamics.

284

285 **Selection** Each simulated pseudo-genome comprised $n_n = 1{,}000$ neutral loci
286 and $n_s = 50$ selected loci of 5kb length, for a total (pseudo-genome) size of 5.25Mb.
287 We assumed a diploid system and all loci to be biallelic with ancestral allele *a* and
288 derived allele *A*. In order to generate a certain fraction of segregating sites per locus
289 (~2%), the mutation μ varied from 2.5 x 10-3 (model $\mathcal{M}_1$) to 2.5 x 10-4 (model $\mathcal{M}_3$) and
290 5 x 10-5 (model $\mathcal{M}_2$) per locus per generation.

291

292 To simulate genetic trade-offs, selection was simulated on alternate alleles in
293 the contrasting environments. Specifically, we assumed the beneficial alleles to be
294 dominant such that the relative fitness was $1 + s_1, 1, 1$ and $1, 1, 1 + s_2$ for the three
295 genotypes *AA*, *Aa* and *aa* in the demes or meta-populations occupying the two
296 environments, respectively. For the selection coefficients s_1 and s_2, we used all
297 combinations of coefficient values 0, 0.001, 0.01 and 0.1 and thus included cases of
298 conditional neutrality (CN) in which either $s_1 > 0, s_2 = 0$ or $s_1 = 0, s_2 > 0$ as well as
299 cases of antagonistic pleiotropy (AP) with $s_1 > 0, s_2 > 0$ with both symmetric ($s_1 =$
300 s_2) and asymmetric ($s_1 \neq s_2$) regimes. CN regimes are by definition always
301 asymmetric and AP regimes are defined such that both alleles confer higher fitness in
302 their respective local environments but reduced fitness in the other, both with respect
303 to the alternate allele and to the fitness conferred in their home environments. We

304 further varied the time of the onset of selection from $T_S = 400$, 4,000, 40,000 and
305 400,000 generations ago.
306
307 For all models, we considered selection on standing variation with the initial
308 frequency of the derived allele at $f_1 = f_2 = 0.1$ in all demes. For model $\mathcal{M}_1$, we
309 additionally investigated the case of *de-novo* mutations with initial frequencies
310 $f_1 = 1/2N_1$ and $f_2 = 0$. These two cases represent the often-considered starting points
311 for local adaptation (Peter et al., 2012). Dependent on the selection regime and due to
312 the stochasticity of drift, the derived allele A may sometimes be lost and hence be
313 absent in the simulation of selected loci (especially in the *de-novo* case). Because such
314 a scenario contains no signal for detection of selection, we excluded such simulations
315 (via the -SFC parameter in msms).
316
317 **Assessing accuracy** We inferred selection by contrasting the locus-specific
318 migration rates m_{12} and m_{21} against their neutral estimates $\hat{m}_{12}$ and $\hat{m}_{21}$ (Figure 3).
319 We evaluated the performance of our LSD implementation at identifying selected
320 loci under these simulations by plotting the true positive rate (TPR) against the false
321 positive rate (FPR) under the choice of HDPI thresholds from 0 to 1, and reporting
322 the area under the curve (AUC) of the resultant receiver operating characteristic
323 (ROC) curve. To evaluate the accuracy of the inferred symmetry of the joint posterior
324 (of reciprocal migration rates), we compared this to the true underlying selection

coefficients, under the expectation that deviations from symmetry in the joint posterior should reflect asymmetry in selection regimes. Specifically, we determined for each locus l the posterior mass

$$\sigma_l = \int I\left(\frac{m_{21}}{m_{12}} < \frac{\hat{m}_{21}}{\hat{m}_{12}}\right)\pi_l d_{\boldsymbol{\theta}},$$

where the indicator function $I(\cdot)$ limits the integral to cases in which the deviation of one of the migration rates has reduced more than the reciprocal migration rate compared to a proportional deviation of both migration rates from their neutral estimates $\hat{m}_{12}$ and $\hat{m}_{21}$. From this, we calculate the asymmetry as

$$a = \log\frac{\sigma}{1-\sigma},$$

where $\sigma = \frac{1}{n_s}\sum\sigma_l$ across loci simulated under selection.

2.4 Case study

To evaluate the performance of LSD on real data, we applied it to the detection of loci underlying floral guides in two parapatric subspecies of *Antirrhinum majus*. *A.majus* is an herbaceous, perennial, flowering plant native to the western Mediterranean. Owing to its diploid inheritance, relatively short generation time, ability for both self- and cross-pollination and rich and varied flower morphology, *A.majus* has lent itself as a model organism for over a century, with several key floral genes being first identified within this genus (Schwarz-Sommer, Davies, & Hudson, 2003; Schwinn et al., 2006). Two subspecies, *A.m.striatum* and *A.m.pseudomajus*, differ

343 in the flower colouration that signposts the pollinator entry point, and form a natural
344 hybrid zone in the Pyrenees that constitutes a benchmark example of divergent
345 selection driven by assortative mating (Whibley et al., 2006). Several genetic loci have
346 been shown to control the differences in these floral patterns (Bradley et al., 2017;
347 Schwinn et al., 2006), and recently, Tavares et al. (2018) produced evidence of
348 genomic signatures of selection at the ROS and EL loci, of which the former has been
349 further functionally validated. Here, we apply LSD to sequencing data from this
350 study to isolate the ROS and EL loci and to characterise their underlying selection
351 signal.
352
353 We modelled this study system via a simple representation (model $\mathcal{M}_1$) of one
354 population on either side of the hybrid zone (YP1 (*A.m.striatum*) vs MP2
355 (*A.m.pseudomajus*); populations 2.5km apart) and via a more inclusive island-
356 continent model (model $\mathcal{M}_2$) comprising three (distant) populations each per
357 subspecies (CAM, ML, YP1 (*A.m.striatum*) vs MP2, CHI, CIN (*A.m.pseudomajus*);
358 Figure S5), allowing all N_E and m_E parameters to be free. Given that these populations
359 were previously sequenced using pool-seq (Tavares et al., 2018), we simulated
360 pooling of individuals in-silico by pooling twice the amount of msms coalescent
361 (haploid) samples as (diploid) individuals in the pooled populations via *LSD-High*.
362 Samples were drawn from a parametric (negative binomial) distribution fitted to the
363 empirical coverage distribution using the '*fitdistrplus*' package (Delignette-Muller &

364 Dutang, 2015) and *LSD-High*. We focused our analysis on chromosome 6 on which
365 the ROS and EL loci lie. To acquire empirical estimates of neutral demographic
366 parameters, we excluded all genomic regions present in the structural annotation
367 plus 10kb flanking regions to generate a subset of putatively neutral regions on that
368 chromosome. The inference of selection was then performed in sliding windows of
369 size 10kb with a 1kb step-size. The mutation rate for the simulations (1.7*10-8 per site
370 per generation) and the filtering of the empirical and simulated data followed those
371 reported in the original study, though we mapped on a more recent and complete
372 version of the *A.majus* reference (version 3.0; M. Li et al., 2019).
373

374 # 3 RESULTS

375 ## 3.1 Two-deme IM case (model $\mathcal{M}_1$)

376 **Power to identify selected loci** Our LSD implementation demonstrated a high
377 diagnostic ability to discriminate between neutral and selected loci (AUC >> 0.9)
378 across a large range of migration-selection regimes (Figure 4). Notably, our results
379 point towards an optimal, intermediate rate of migration (M=5) at which selection is
380 best detectable with high AUC values across a large set of selection coefficients. As
381 migration rates increase (M=50), migration from the foreign deme where selection
382 acts on the alternate allele increasingly inhibits the build-up of beneficial
383 polymorphisms in the local deme, in which case the power to detect selected loci

384 becomes limited to scenarios under longer regimes of strong selection. At lower
385 migration rates (M=0.5), long regimes of selection permit the detection of loci under
386 the lowest selection coefficients, but power decreases for younger times compared to
387 scenarios simulated under intermediate migration rates. This owes to LSD relying on
388 the reduction of effective migration relative to neutral or genome-wide expectations,
389 which in this case is already at a low level.
390
391 The power to detect selection increased with increasing selection coefficients
392 in the case of AP if selection coefficients were similar ($s_1 \approx s_2$, cells along diagonal of
393 sub-panels in Figure 4). In such cases, stronger selection coefficients on alternate
394 alleles increasingly polarise and ultimately maintain larger allele frequency
395 differences between the two environments. In tandem, the power to detect selection
396 also generally increased with the time since the onset of selection T_S. However, in
397 most cases of CN or when $s_1 \gg s_2$ or $s_1 \ll s_2$, one of the two alleles may proceed to
398 fixation, in which case the power to detect selection decays or is lost (e.g. cells along
399 bottom row and left-most column of sub-panels in Figure 4). This is particularly
400 evident when the onset of selection is more distant in the past and implies that cases
401 of CN may be harder to detect than AP as their signatures of selection are often more
402 transient and decay more rapidly.
403

404 To evaluate the false discovery rate (FDR) of LSD at different migration rates,
405 we also conducted simulations in the absence of selection. For intermediate
406 migration rates (M=5), LSD performed better than FDR expectations (Figure S2B); at
407 a defined threshold of 0.95, LSD resulted in a false positive rate of 2%. For the low
408 and high migration rates (M=0.5, 50) however, LSD performed worse than FDR
409 expectations, producing false positive rates higher than expected with false positive
410 rates of 44% and 8% respectively at a threshold of 0.95 (Figures S2A and S2C). This
411 optimal, intermediate rate of migration for minimising FDR is consistent with the
412 optimal migration rate for detecting selection.
413
414 **Power to characterise (a)symmetry** A benefit of LSD over classic outlier
415 approaches is that it can provide insight into trade-offs underlying selection, by
416 identifying cases in which selection acts at equal strength in the two demes or
417 metapopulations (symmetric AP), or whether selection coefficients differ
418 considerably (CN or asymmetric AP). As shown in Figure 5, the inferred
419 (a)symmetry generally reflected the true (a)symmetry of the underlying selection
420 coefficients well, particularly for regimes with high power to correctly identify
421 selected loci (AUC > 0.95). In lower powered regimes, we observe a few cases where
422 the inferred asymmetry does not reflect the underlying asymmetry of the selection
423 coefficients accurately (e.g. blue cells along diagonals in sub-panels at T_S = 4,000;
424 Figures 5 and S3). We treat these exceptional cases further in the discussion.

425

426 **Standing variation vs *de-novo*** A lower initial frequency of the derived allele
427 may be expected to affect LSD's power to identify selected loci and its power to
428 capture the underlying (a)symmetry of selection coefficients. However, we find that
429 results for simulations building on selection from the *de-novo* and standing variation
430 cases showed generally very similar patterns (Figures 4, 5 and S3). One notable
431 exception however was the inaccurate inference of (a)symmetry in a few regimes
432 with high power (AUC > 0.95) in the *de-novo* case (e.g. blue cells along diagonals in
433 sub-panels at $T_S = 4{,}000$ in Figure S3B). This we attribute to the lower initial
434 frequency of the derived allele *A* and consequently longer time needed to reach drift-
435 migration-selection equilibrium for the *de-novo* cases. This is explored further in the
436 discussion.

437

438 ## 3.2 More complex cases (models $\mathcal{M}_2$ and $\mathcal{M}_3$)

439 A key feature of LSD is its potential for explicit accommodation of complex
440 demographies, which when not properly accounted for can lead to an inflation in
441 false positives (De Villemereuil, Frichot, Bazin, François, & Gaggiotti, 2014; Foll &
442 Gaggiotti, 2008; Lotterhos & Whitlock, 2014). Despite the added complexity of
443 models $\mathcal{M}_2$ and $\mathcal{M}_3$, results were generally very similar to that of model $\mathcal{M}_1$, with
444 high power to identify selected loci (AUC >> 0.9) across a large range of migration-
445 selection regimes, an optimal migration rate at an intermediate value (M=5), a similar

446 dependence of power to detect selection on s_1, s_2 and $T_{S'}$ and inferences of
447 (a)symmetry that reflected well the underlying (a)symmetry of selection coefficients
448 (Figures 6 and S4). A notable difference compared to model $\mathcal{M}_1$ was the longer
449 amount of time needed to generate a high power to detect selection across a range of
450 migration values in models $\mathcal{M}_2$ (Figures 6A) and especially $\mathcal{M}_3$ (Figure S4A), which
451 we attribute to the lower mutation rates used in these models compared to $\mathcal{M}_1$.
452

453 ## 3.3 Case study results

454 We identified a region of reduced effective migration between 52.9-53.2MB on
455 chromosome 6 (Figure 7), consistent with the location of the ROS and EL loci
456 (Tavares et al., 2018). Under model $\mathcal{M}_1$, this region is characterised by a set of
457 smaller, multiple peaks (posterior probability of being divergent from neutral
458 expectations > 99.9%) reflecting signatures identified by previous authors, with the
459 left-most peaks (shaded in red) corresponding to ROS1 and ROS2 and the right peaks
460 (shaded in green) to EL (Figure 7B). The joint posterior probability distributions
461 reveal symmetric selection acting on both regions, implying that selection acts with
462 similar strength in the two populations. Under Model $\mathcal{M}_2$, we find fewer outliers in
463 the ROS-EL region than in model $\mathcal{M}_1$, with the left-most peak in this region
464 corresponding to ROS2 and the right peaks consistent with EL. ROS1 appears to be
465 less of an outlier than in model $\mathcal{M}_1$ (posterior probability of being divergent ~99.8%).

In contrast to model $\mathcal{M}_1$, the ROS2 and EL peaks in model $\mathcal{M}_2$ are characterised by asymmetry, specifically with stronger selection acting in the populations of *A.m.pseudomajus* than in the populations of *A.m.striatum*.

4 DISCUSSION

The trajectory of selected loci depends on demographic and selection parameters that define the system, namely the effective population sizes, effective migration rates and selection coefficients, as well as the intrinsic properties of mutation and recombination. Despite well-developed theory which relates the effect of population parameters on the trajectory of selected alleles, few methods or empirical studies have combined estimates of differential selection with explicit quantification of migration rates and effective population sizes to examine the conditions under which local adaptation can arise. In this study, we condition the identification of candidate loci on divergent population parameters using explicit demographic models, and demonstrate that under certain demographic-selection regimes, we can both detect and elucidate the processes underlying signatures of selection. While LSD is flexible regarding the choice of demographic models employed, we focus here specifically on those processes that are expected to lead to selection against gene flow, namely local adaptation and extrinsic reproductive barriers, that can be inferred via their expectation to reduce effective migration rates.

4.1 Identifying selection

In our simulations, we demonstrate that LSD has high diagnostic power (AUC >> 0.9) to identify selected loci across a large range of demographic-selection regimes. This power relies upon two fundamental aspects that contribute to generating observable patterns. First, selection must effectively be realised, i.e. result in a frequency shift of the beneficial allele. This requires that the strength of selection and initial frequency of the beneficial allele be sufficient to both counter the homogenising effect of migration (Felsenstein, 1976; Haldane, 1930; Lenormand, 2002; M. Slatkin, 1973; Yeaman, 2015) and the eroding effect of drift (Wright, 1931). Secondly, the genomic data must contain signatures of selection that can be detected. In the case of LSD, this requires that the signatures of selection are discernible from the underlying noise (drift and migration) that characterises the system, as well as requires sufficient time for said signatures to be reflected in the employed statistics and hence in the inferred parameters N_E or m_E. A lack of power in LSD must be interpreted considering these two conceptually different perspectives. Notably, the lack of discrimination power for high migration rates and low selection coefficients can be attributed to selection failing to realise as a consequence of local, beneficial alleles being swamped by immigrant, maladaptive alleles. In contrast, the lack of signal under low migration rates constitutes a methodological limitation of our implemented model, as it becomes increasingly difficult to detect reductions in effective migration when neutral or genome-wide migration rates are already at a

508 low level; even when selection is effectively being realised in the demes. This is
509 analogous in effect to the loss of power to detect selection in highly differentiated
510 populations in F_{ST} outlier tests (Hoban et al., 2016). Under the same principle, we
511 argue that the converse expectation can be assumed to hold for loci underlying
512 adaptive introgression or balancing selection. That is, we expect power to detect such
513 loci to be low when populations are minimally differentiated, as the detection of
514 candidate loci in these cases is informed by increased effective migration.
515
516 The power of LSD to correctly identify selected loci generally increased with
517 stronger selection coefficients and longer time since the onset of selection, though
518 with exceptions related to differential selection on alternate alleles in multiple
519 populations. Specifically, when selection is of similar or equal strength in both demes
520 or meta-populations, we observed a strong correlation between the power to detect
521 selection and the true underlying selection coefficients. This follows theory which
522 states that the reduction in effective migration is proportional to the strength of
523 selection (Petry, 1983). However, we defer from translating these changes to explicit
524 selection coefficients because in addition to the strength of selection, changes in
525 effective migration are also a function of the recombination rate between linked and
526 selected loci (Cutter & Payseur, 2013; Lotterhos, 2019; Petry, 1983). On the other
527 hand, if selection is highly divergent in strength ($s_i >> s_j$) between the demes or meta-
528 populations or when the onset of selection is sufficiently distant in the past, one of

529 the two selected alleles may have fixed in the system. In such a case, the signal to
530 detect selection rapidly decays (Huber, DeGiorgio, Hellmann, & Nielsen, 2016;
531 Przeworski, 2002). Finally, we observed little power to detect very recent selection,
532 intrinsically related to our choice of summary statistics (Hohenlohe, Phillips, &
533 Cresko, 2010). While a signal of selection necessarily requires time to build up, we
534 note that extending LSD to include additional statistics sensitive to linkage
535 disequilibrium such as extended haplotype homozygosity (EEH) (Szpiech &
536 Hernandez, 2014) or single density score (SDS) (Field et al., 2016) may increase the
537 power to detect more recent selection.
538
539 From our simulations, we find that the effect of the tested selection regimes on
540 the power to detect selection is similar between the *de-novo* and standing genetic
541 variation cases (model $\mathcal{M}_1$; Figures 4 and S3A). This result relies on the fact that we
542 only kept simulations of selected loci (comprising the pseudo-genomes) in which
543 derived allele *A* was not lost. This particularly affected the *de-novo* case, where most
544 simulations of selected loci were observed to result in the loss of allele *A*. This
545 implies firstly, that in most observed cases, signals of selection are most likely to
546 arise from standing variation (Jones et al., 2012; Lai et al., 2019; Reid et al., 2016), and
547 secondly, that to distinguish selection arising *de-novo* to that from standing variation,
548 additional information such as allele age, mutation rate or supplementary
549 phylogenetic information is likely required (Peter et al., 2012).

4.2 Revealing trade-offs underlying selection

A major benefit of genome scans performed under a demographic framework is the capacity to infer the directionality of selection. In our simulations, we observe that the (a)symmetry in the reduction of reciprocal migration rates between demes or meta-populations as inferred by LSD reflects the (a)symmetry of underlying selection coefficients accurately for older onsets of selection ($T_S > 4{,}000$, red cells inhabiting left-upper triangle, blue in right-lower triangle; Figures 5, 6B, S3B, S4B), but less so for more recent onsets of selection ($T_S \leq 4{,}000$; Figures 5, 6B, S3B, S4B). This is because prior to reaching drift-migration-selection equilibrium, estimated asymmetries in effective migration rates are also affected by asymmetry in allele frequencies (Figure 8). This is highlighted when contrasting the standing variation and *de-novo* simulation results for model $\boldsymbol{\mathcal{M}_1}$, where incorrectly inferred (a)symmetries are more evident in the *de-novo* case due to the lower initial frequency of the derived allele ($T_S \leq 4{,}000$; Figures 5 and S3B). A direct link between the (a)symmetry in inferred migration rates and selection coefficients is only established through time as the beneficial allele increases in frequency towards an equilibrium. From this, we deduce that if a strong asymmetry between effective migration rates is inferred, the system may still be in the process of evolving towards an equilibrium state, which may include cases in which one allele will ultimately be lost. If symmetry between effective migration rates is inferred on the other hand, the system

571 is likely near-equilibrium and we can expect both alleles to be maintained in the
572 system. We note that in practice, however, the interpretation of the results is
573 straightforward. This is because the inference of directionality is only relevant if the
574 targets of selection can be detected accurately (e.g. in regimes with AUC > 0.9). For
575 these cases, we generally find the inferred (a)symmetries in migration rates to reflect
576 the true (a)symmetry in selection coefficients (Figure 5, 6B, S3B, S4B) accurately,
577 mostly because the power to detect selection is generally low if selection started
578 acting only very recently.
579
580 The ability of LSD to infer the directionality of selection directly from genomic
581 data can greatly facilitate investigations of genetic trade-offs underlying adaptation,
582 which are seldom performed due to the considerable effort required to set up field
583 trials of recombinant lines. As shown above, the inference of symmetry in LSD-
584 identified candidates accurately reflects cases of AP with equal strength of selection
585 on alternate alleles in the contrasting environments. The inference of asymmetry on
586 the other hand can either indicate AP with stronger selection in one environment
587 than the other, or CN. From our simulations, we find that scenarios reflecting AP are
588 generally more readily detected than those reflecting CN. Given that selection acts
589 only upon one of the two alleles in the latter case, fixation becomes likely and the
590 ability to detect selection is transient. This implies that there may be an observation
591 bias between AP and CN; such that the inference of CN may be comparatively

592 under-represented. This bias appears to contrast with that reported in ecological
593 literature, where instances of AP are more rarely detected compared to CN due to
594 the additional power required to detect variance in fitness concurrently in two
595 environments (Anderson et al., 2013). LSD may further complement field trials as
596 such experiments typically test genetic trade-offs under contemporary selective
597 environments, which may not reflect past conditions driving the observed adaptive
598 responses, but whose signature may still be inferred from genomic data. Using LSD
599 to formulate expectations about fitness effects and to inform the choice of
600 environmental conditions under which to validate identified candidate genes can
601 thus greatly aid such experiments.
602

603 ## 4.3 Real-world application

604 We demonstrate a real-world application of LSD by successfully isolating and
605 characterising the selection signal of loci underlying an extrinsic reproductive barrier
606 in *A.majus*. Our results from contrasting a single population (model $\mathcal{M}_1$) and three
607 populations (model $\mathcal{M}_2$) per subspecies both identified the ROS and EL loci which
608 were previously reported to underlie differences in floral patterns between these
609 subspecies (Tavares et al., 2018). Interestingly however, our results show a different
610 signal of (a)symmetry between the tested models $\mathcal{M}_1$ and $\mathcal{M}_2$, with the peaks
611 corresponding to the ROS and EL loci characterised by symmetry under model $\mathcal{M}_1$,
612 in contrast to a pattern of asymmetry under model $\mathcal{M}_2$, specifically with stronger

613 selection inferred to act in the populations of *A.m.pseudomajus* than in the
614 populations of *A.m.striatum*. We stress that the interpretation of LSD genome scans is
615 conditional on the model and populations used, such that in our example model $\mathcal{M}_1$
616 uncovers population-pair specific differences at the contact zone (YP1 vs MP2) while
617 model $\mathcal{M}_2$ reveals common (global) differences between the two subspecies. We do
618 not necessarily expect these two signals to be identical, and indeed, Tavares et al.
619 (2018) also found differences between distant and close *A.m.striatum*-*A.m.pseudomajus*
620 population pairs in terms of observed θ_W and F_{ST} summary statistics. Given that there
621 is no evident difference in environment or pollinators on opposite sides of the hybrid
622 zone, reproductive barriers in this system have often been proposed to be maintained
623 through assortative mating mediated by pollinator preference for the dominant
624 (most common) flower phenotype in a given area and the subspecies' distinct flower
625 colouration. However, whether selection on alternate alleles follows the same
626 positive frequency-dependence across the broader scale including more distant
627 populations is currently unknown. The difference in signal between local pairs at the
628 contact zone ($\mathcal{M}_1$) and the global set ($\mathcal{M}_2$) may be generated by different frequency-
629 dependent selection curves for the alternate alleles and potentially loss of AP away
630 from the contact zone (Figure S6).
631

632 ## 5 CONCLUSION

When selection occurs in the presence of gene flow, selected sites are predicted to exhibit gene genealogies with demographic parameters divergent from those of neutral non-linked sites, leading to heterogeneity in demography across the genome. In this study, we condition the identification of candidate loci on divergent population parameters using explicit demographic models, and demonstrate that under certain conditions of migration, selection strength and onset time, we can both detect and elucidate the underlying processes driving signatures of selection. Incorporating and utilising the inference of demographic parameters in the identification of candidate loci address some key issues and assumptions that prevail in the discrimination of selected variants, namely 1) the explicit consideration of demography, 2) heterogeneity in drift and gene flow across the genome, 3) information synthesis of multiple, complementary summary statistics, and 4) transparency towards underlying driving mechanisms.

Our power analysis using simulations shows that LSD, and our implementation of it, represents a powerful method for detecting selection that is robust to different and complex demographies. Furthermore, given that certain demographic parameters e.g. migration are not inherently commutative, we show that the directionality or population-specificity in selection can be inferred. This can facilitate identifying in which environment selection acts and hence elucidate genetic trade-offs; bridging an analytical divide between experimental ecology and

654 population genomics. Importantly, the proposed approach as well as our
655 implementation is not limited to the demographic models investigated here, nor the
656 explicit choice of simulation programs or summary statistics used. This flexibility
657 and customisability of LSD can facilitate e.g. more realistic accommodation of
658 recombination (via different coalescent simulators), improved detection of more
659 recent selection (via linkage-informative statistics), and inference of other modes of
660 selection (e.g. balancing selection) and adaptive introgression by conditioning the
661 detection of selection on e.g. increase (rather than reduction) of m_E or changes in N_E
662 relative to neutral expectations.
663

664 ## ACKNOWLEDGEMENTS

665 We would like to thank Niklaus Zemp for providing IT support. This work
666 was supported by the Swiss National Science Foundation (SNSF) grants
667 31003A_160123 and 31003A_182675 awarded to AW and grant 31003A_173062
668 awarded to DW.
669

REFERENCES

Anderson, J. T., Lee, C. R., Rushworth, C., Colautti, R., & Mitchell-Olds, T. (2013). Genetic tradeoffs and conditional neutrality contribute to local adaptation. *Molecular Ecology*, *22*(3), 699–708. https://doi.org/10.1038/jid.2014.371

Barton, N., & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity*, *57*(3), 357–376. https://doi.org/10.1038/hdy.1986.135

Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B: Biological Sciences*, *263*(1377), 1619–1626. https://doi.org/10.1098/rspb.1996.0237

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y. P., Hahn, M. W., … Langley, C. H. (2007). Population genomics: Whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biology*, *5*(11), 2534–2559. https://doi.org/10.1371/journal.pbio.0050310

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Biswas, S., & Akey, J. M. (2006). Genomic insights into positive selection. *Trends in Genetics*, *22*(8), 437–446. https://doi.org/10.1016/j.tig.2006.06.005

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, *186*(1), 241–262. https://doi.org/10.1534/genetics.110.117275

Bradley, D., Xu, P., Mohorianu, I. I., Whibley, A., Field, D., Tavares, H., … Coen, E. (2017). Evolution of flower color pattern through selection on regulatory small RNAs. *Science*, *358*(6365), 925–928. https://doi.org/10.1126/science.aao3526

Charlesworth, B. (2009). Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Charlesworth, B., Nordborg, M., & Charlesworth, D. (1997). The effects of background and interference selection on patterns of genetic variation in subdivided populations. *Genetics Research*, *70*, 155–174. https://doi.org/10.1534/genetics.115.178558

Cutter, A. D., & Payseur, B. A. (2013). Genomic signatures of selection at linked sites: Unifying the disparity among species. *Nature Reviews Genetics*, *14*(4), 262–274. https://doi.org/10.1038/nrg3425

De Villemereuil, P., Frichot, É., Bazin, É., François, O., & Gaggiotti, O. E. (2014). Genome scan methods against more complex models: When and how much should we trust them? *Molecular Ecology*, *23*(8), 2006–2019.

711 https://doi.org/10.1111/mec.12705
712 Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting
713 distributions. *Journal of Statistical Software*, *64*(4), 1–34.
714 https://doi.org/10.18637/jss.v064.i04
715 Eckert, A. J., Van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J.,
716 González-Martínez, S. C., & Neale, D. B. (2010). Patterns of population structure
717 and environmental associations to aridity across the range of loblolly pine (Pinus
718 taeda L., Pinaceae). *Genetics*, *185*(3), 969–982.
719 https://doi.org/10.1534/genetics.110.115543
720 Edmonds, C. a, Lillie, A. S., & Cavalli-Sforza, L. L. (2004). Mutations arising in the
721 wave front of an expanding population. *Proceedings of the National Academy of*
722 *Sciences of the United States of America*, *101*(4), 975–979.
723 https://doi.org/10.1073/pnas.0308064100
724 Engelhardt, B. E., & Stephens, M. (2010). Analysis of population structure: A unifying
725 framework and novel methods based on sparse factor analysis. *PLoS Genetics*,
726 *6*(9). https://doi.org/10.1371/journal.pgen.1001117
727 Ewing, G., & Hermisson, J. (2010). MSMS: A coalescent simulation program
728 including recombination, demographic structure and selection at a single locus.
729 *Bioinformatics*, *26*(16), 2064–2065. https://doi.org/10.1093/bioinformatics/btq322
730 Excoffier, L., Foll, M., & Petit, R. J. (2009). Genetic Consequences of Range
731 Expansions. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 481–501.
732 https://doi.org/10.1146/annurev.ecolsys.39.110707.173414
733 Fay, J. C., Wyckoff, G. J., & Wu, C. I. (2002). Testing the neutral theory of molecular
734 evolution with genomic data from Drosophila. *Nature*, *415*(6875), 1024–1026.
735 https://doi.org/10.1038/4151024a
736 Felsenstein, J. (1976). The theoretical population genetics of variable selection and
737 migration. *Annu. Rev. Genet*, *10*, 253–280.
738 Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Yengo, L., … Pritchard, J. K.
739 (2016). Detection of human adaptation during the past 2000 years, *354*(6313),
740 760–764. https://doi.org/10.1126/science.aag0776.Detection
741 Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci
742 appropriate for both dominant and codominant markers: A Bayesian
743 perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221
744 François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false
745 discoveries in genome scans for selection. *Molecular Ecology*, *25*(2), 454–469.
746 https://doi.org/10.1111/mec.13513
747 Fusco, D., & Uyenoyama, M. K. (2011). Sex-specific incompatibility generates locus-
748 specific rates of introgression between species. *Genetics*, *189*(1), 267–288.
749 https://doi.org/10.1534/genetics.111.130732
750 Galtier, N., Depaulis, F., & Barton, N. H. (2000). Detecting bottlenecks and selective
751 sweeps from DNA sequence polymorphism. *Genetics*, *155*(2), 981–987.

752 Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with
753 population-specific covariates. *Genetics*, *201*(4), 1555–1579.
754 https://doi.org/10.1534/genetics.115.181453
755 Gossmann, T. I., Woolfit, M., & Eyre-Walker, A. (2011). Quantifying the variation in
756 the effective population size within a genome. *Genetics*, *189*(4), 1389–1402.
757 https://doi.org/10.1534/genetics.111.132654
758 Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G.,
759 … Sabeti, P. C. (2010). A Composite of Multiple Signals Distinguishes Causal
760 Variants in Regions of Positive Selection. *Science*, *327*, 883–886.
761 Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele
762 frequencies. *Genetics*, *195*(1), 205–220. https://doi.org/10.1534/genetics.113.152462
763 Haldane, J. B. S. (1930). A mathematical theory of natural and artificial selection.
764 (Part VI, Isolation.). *Math. Proc. Cambridge Philos. Soc.*, *26*, 220–230.
765 Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., …
766 Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls,
767 practical solutions, and future directions. *American Naturalist*, *188*(4), 379–397.
768 https://doi.org/10.1086/688018
769 Hofer, T., Ray, N., Wegmann, D., & Excoffier, L. (2009). Large allele frequency
770 differences between human continental groups are more likely to have occurred
771 by drift during range expansions than by selection. *Annals of Human Genetics*,
772 *73*(1), 95–108. https://doi.org/10.1111/j.1469-1809.2008.00489.x
773 Hohenlohe, P. A., Phillips, P. C., & Cresko, W. A. (2010). Using population genomics
774 to detect selection in natural populations: Key concepts and methodological
775 considerations. *International Journal of Plant Sciences*, *171*(9), 1059–1071.
776 https://doi.org/10.1086/656306
777 Horscroft, C., Ennis, S., Pengelly, R. J., Sluckin, T. J., & Collins, A. (2019). Sequencing
778 era methods for identifying signatures of selection in the genome. *Briefings in*
779 *Bioinformatics*, *20*(6), 1997–2008. https://doi.org/10.1093/bib/bby064
780 Huber, C. D., DeGiorgio, M., Hellmann, I., & Nielsen, R. (2016). Detecting recent
781 selective sweeps while controlling for mutation rate and background selection.
782 *Molecular Ecology*, *25*(1), 142–156. https://doi.org/10.1111/mec.13351
783 Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., …
784 Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine
785 sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944
786 Joyce, P., & Marjoram, P. (2008). Approximately sufficient statistics and bayesian
787 computation. *Statistical Applications in Genetics and Molecular Biology*, *7*(1).
788 https://doi.org/10.2202/1544-6115.1389
789 Kawecki, T. J., & Ebert, D. (2004). Conceptual issues in local adaptation. *Ecology*
790 *Letters*, *7*(12), 1225–1241. https://doi.org/10.1111/j.1461-0248.2004.00684.x
791 Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*,
792 *13*(3), 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

793 Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next
794 Generation Sequencing Data. *BMC Bioinformatics*, *15*(1), 356.
795 https://doi.org/10.1186/s12859-014-0356-4
796 Lai, Y. T., Yeung, C. K. L., Omland, K. E., Pang, E. L., Hao, Y., Liao, B. Y., … Li, S. H.
797 (2019). Standing genetic variation as the predominant source for adaptation of a
798 songbird. *Proceedings of the National Academy of Sciences of the United States of*
799 *America*, *116*(6), 2152–2157. https://doi.org/10.1073/pnas.1813597116
800 Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology*
801 *& Evolution*, *17*(4), 183–189. https://doi.org/10.1111/j.1365-2184.1976.tb01248.x
802 Li, H., & Stephan, W. (2006). Inferring the demographic history and rate of adaptive
803 substitution in Drosophila. *PLoS Genetics*, *2*(10), 1580–1589.
804 https://doi.org/10.1371/journal.pgen.0020166
805 Li, J., Li, H., Jakobsson, M., Li, S., SjÖdin, P., & Lascoux, M. (2012). Joint analysis of
806 demography and selection in population genetics: Where do we stand and
807 where could we go? *Molecular Ecology*, *21*(1), 28–44.
808 https://doi.org/10.1111/j.1365-294X.2011.05308.x
809 Li, M., Zhang, D., Gao, Q., Luo, Y., Zhang, H., Ma, B., … Xue, Y. (2019). Genome
810 structure and evolution of Antirrhinum majus L. *Nature Plants*, *5*(2), 174–183.
811 https://doi.org/10.1038/s41477-018-0349-9
812 Lotterhos, K. E. (2019). The effect of neutral recombination variation on genome
813 scans for selection. *G3: Genes, Genomes, Genetics*, *9*(6), 1851–1867.
814 https://doi.org/10.1534/g3.119.400088
815 Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and
816 neutral parameterization on the performance of FST outlier tests. *Molecular*
817 *Ecology*, *23*(9), 2178–2192. https://doi.org/10.1111/mec.12725
818 Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power
819 and promise of population genomics: From genotyping to genome typing.
820 *Nature Reviews Genetics*, *4*(12), 981–994. https://doi.org/10.1038/nrg1226
821 Marjoram, P., & Tavaré, S. (2006). Modern computational approaches for analysing
822 molecular genetic variation data. *Nature Reviews Genetics*, *7*(10), 759–770.
823 https://doi.org/10.1038/nrg1961
824 Naisbit, R. E., Jiggins, C. D., & Mallet, J. (2001). Disruptive sexual selection against
825 hybrids contributes to speciation between Heliconius cydno and Heliconius
826 melpomene. *Proceedings of the Royal Society B: Biological Sciences*, *268*(1478), 1849–
827 1854. https://doi.org/10.1098/rspb.2001.1753
828 Nosil, P., Vines, T. H., & Funk, D. J. (2005). Perspective: Reproductive Isolation
829 Caused By Natural Selection Against Immigrants From Divergent Habitats.
830 *Evolution*, *59*(4), 705. https://doi.org/10.1554/04-428
831 Oleksyk, T. K., Smith, M. W., & O'Brien, S. J. (2010). Genome-wide scans for
832 footprints of natural selection. *Philosophical Transactions of the Royal Society B:*
833 *Biological Sciences*, *365*(1537), 185–205. https://doi.org/10.1098/rstb.2009.0219

834 Peter, B. M., Huerta-Sanchez, E., & Nielsen, R. (2012). Distinguishing between
835 Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLoS*
836 *Genetics*, *8*(10). https://doi.org/10.1371/journal.pgen.1003011
837 Petry, D. (1983). The effect on neutral gene flow of selection at a linked locus.
838 *Theoretical Population Biology*, *23*(3), 300–313. https://doi.org/10.1016/0040-
839 5809(83)90020-5
840 Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich,
841 D. (2006). Principal components analysis corrects for stratification in genome-
842 wide association studies. *Nature Genetics*, *38*(8), 904–909.
843 https://doi.org/10.1038/ng1847
844 Przeworski, M. (2002). The Signature of Positive Selection at Randomly Chosen Loci.
845 *Genetics*, *162*(4), 1179–1189.
846 Reid, N. M., Proestou, D. A., Clark, B. W., Warren, W. C., Colbourne, J. K., Shaw, J.
847 R., … Whitehead, A. (2016). The genomic landscape of rapid repeated
848 evolutionary adaptation to toxic pollution in wild fish. *Science*, *354*(6317), 1305–
849 1308. https://doi.org/10.1126/science.aah4993
850 Rundle, H. D., & Whitlock, M. C. (2001). A genetic interpretation of ecologically
851 dependent isolation. *Evolution*, *55*(1), 198–201.
852 Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local
853 adaptation. *Nature Reviews Genetics*, *14*(11), 807–820.
854 https://doi.org/10.1038/nrg3522
855 Schluter, D. (2000). *The Ecology of Adaptive Radiation*. Oxford University Press, Oxford,
856 U.K. Retrieved from https://global.oup.com/academic/product/the-ecology-of-
857 adaptive-radiation-9780198505228?cc=ch&lang=en&
858 Schwarz-Sommer, Z., Davies, B., & Hudson, A. (2003). An everlasting pioneer: The
859 story of Antirrhinum research. *Nature Reviews Genetics*, *4*(8), 655–664.
860 https://doi.org/10.1038/nrg1127
861 Schwinn, K., Venail, J., Shang, Y., Mackay, S., Alm, V., Butelli, E., … Martin, C.
862 (2006). A Small Family of MYB-Regulatory Genes Controls Floral Pigmentation
863 Intensity and Patterning in the Genus Antirrhinum. *The Plant Cell*, *18*(April),
864 831–851. https://doi.org/10.1105/tpc.105.039255.1
865 Slatkin, M. (1973). Gene flow and selection in a cline. *Genetics*, *75*(4), 733–756.
866 Slatkin, Montgomery, & Excoffier, L. (2012). Serial founder effects during range
867 expansion: A spatial analog of genetic drift. *Genetics*, *191*(1), 171–181.
868 https://doi.org/10.1534/genetics.112.139022
869 Sousa, V. C., Carneiro, M., Ferrand, N., & Hey, J. (2013). Identifying loci under
870 selection against gene flow in isolation-with-migration models. *Genetics*, *194*(1),
871 211–233. https://doi.org/10.1534/genetics.113.149211
872 Sugden, L. A., Atkinson, E. G., Fischer, A. P., Rong, S., Henn, B. M., &
873 Ramachandran, S. (2018). Localization of adaptive variants in human genomes
874 using averaged one-dependence estimation. *Nature Communications*, *9*(1).

https://doi.org/10.1038/s41467-018-03100-7

Szpiech, Z. A., & Hernandez, R. D. (2014). Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, *31*(10), 2824–2827. https://doi.org/10.1093/molbev/msu211

Tavares, H., Whibley, A., Field, D. L., Bradley, D., Couchman, M., & Copsey, L. (2018). Selection and gene flow shape genomic islands that control floral guides, 1–6. https://doi.org/10.1073/pnas.1801832115

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Wakeley, J. (2001). The coalescent in an island model of population subdivision with variation among demes. *Theoretical Population Biology*, *59*(2), 133–144. https://doi.org/10.1006/tpbi.2000.1495

Wegmann, D., Leuenberger, C., & Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, *182*(4), 1207–1218. https://doi.org/10.1534/genetics.109.102509

Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, *11*, 116. https://doi.org/10.1186/1471-2105-11-116

Whibley, A. C., Langlade, N. B., Andalo, C., Hanna, A. I., Bangham, A., Thébaud, C., & Coen, E. (2006). Evolutionary paths underlying flower color variation in Antirrhinum. *Science*, *313*(5789), 963–966. https://doi.org/10.1126/science.1129161

Williamson, S. H., Hernandez, R., Fledel-alon, A., Zhu, L., Nielsen, R., & Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *PNAS*, *102*(22), 7882–7887. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1142382&tool=pmcentrez&rendertype=abstract

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97–159.

Yeaman, S. (2015). Local adaptation by alleles of small effect. *American Naturalist*, *186*(october), S74–S89. https://doi.org/10.1086/682405

Zeng, K., Shi, S., & Wu, C. I. (2007). Compound tests for the detection of hitchhiking under positive selection. *Molecular Biology and Evolution*, *24*(8), 1898–1908. https://doi.org/10.1093/molbev/msm119

DATA ACCESSIBILITY

We provide scripts to perform LSD genome scans at the GitHub repository: https://github.com/hirzi/LSD.

916
917 ## AUTHOR CONTRIBUTIONS
918
919 HL, DW, AW and SF designed the study. HL wrote the LSD scripts and performed the
920 simulations and analyses. HL and DW developed the methods. HL wrote the manuscript,
921 which all authors critically revised.
922
923

FIGURES

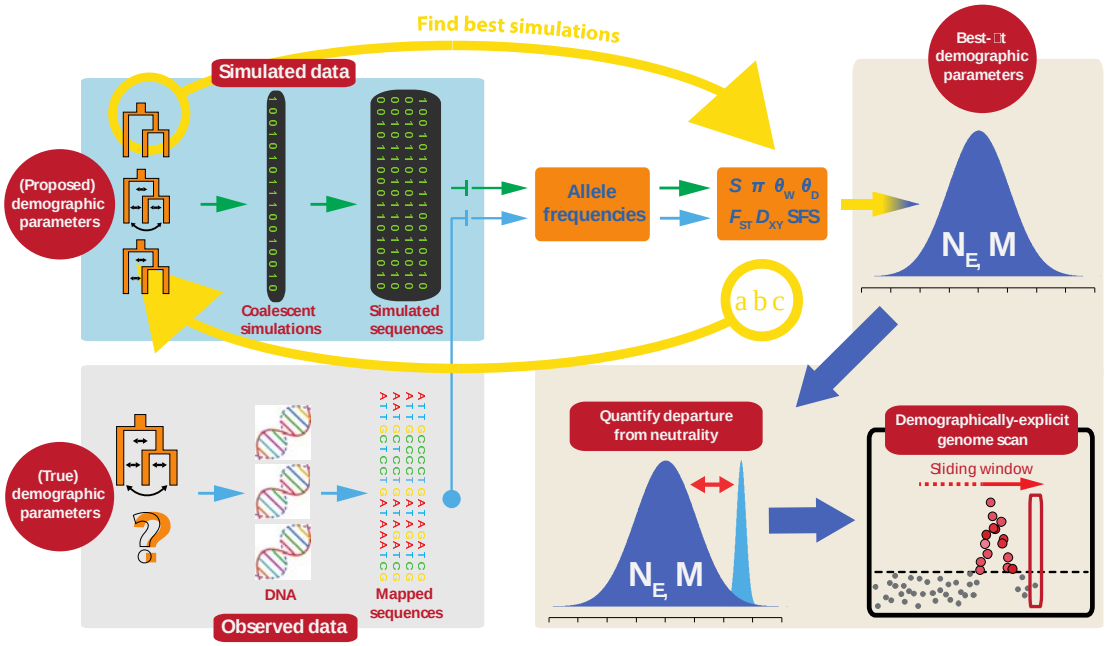

Figure 1. Identifying Loci under Selection via explicit Demographic models (LSD). LSD identifies loci under selection by first estimating demographic parameters and then quantifying the departure of these parameters from neutral expectations. Our specific implementation of LSD employs Approximate Bayesian Computation (ABC) for parameter estimation, and is performed in a genome scan approach.

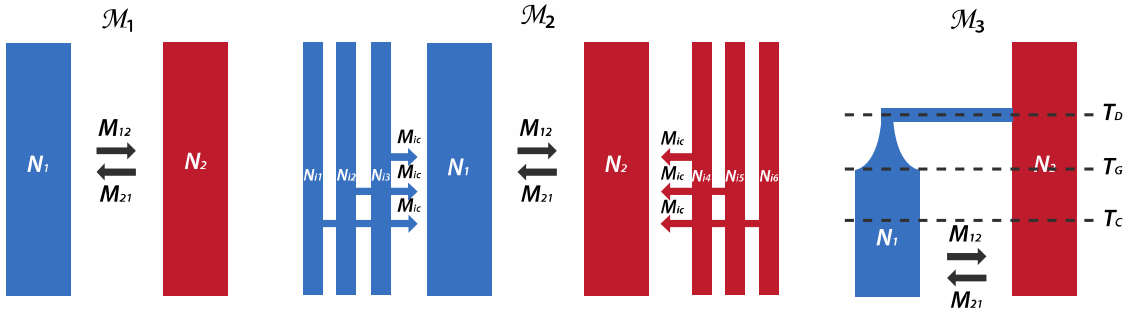

Figure 2. Models used in the simulations and case study. Model $\mathcal{M}_1$ represents a simple 2-deme isolation-with-migration (IM) model with reciprocal migration. Model $\mathcal{M}_2$ represents a 6-deme island-continent model where common differences between environments are modelled by connecting the sampled demes (i.e. islands) to respective meta-population continents via gene flow. Model $\mathcal{M}_3$ represents a 2-deme divergence with bottleneck and exponential growth model. Different deme colours reflect contrasting environments. In all models, selection is inferred from the

deviation from neutrality of the reciprocal migration rates between the two contrasting environments.

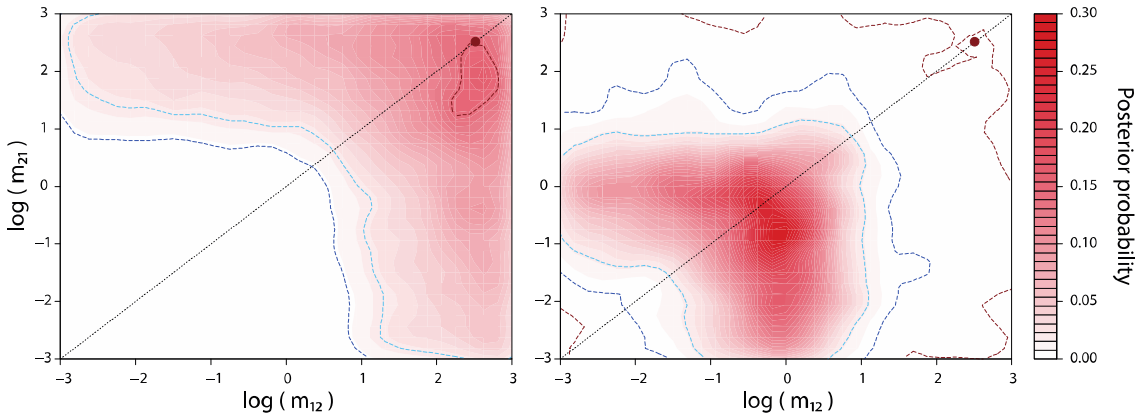

Figure 3. Joint posterior distribution of the reciprocal migration parameters, m_{12} and m_{21}. The neutral joint parameter estimate, as informed by the global posterior distribution of all neutral regions (Fig. S1), is indicated by the red dot in the top right corner. The red contours represent the joint posterior distribution of a genomic region (i.e. window), with the blue contours representing the 95% (light blue) and 99% (dark blue) highest density region (HDR) credible intervals. Left – a window not significantly divergent from the neutral estimate; right – a window significantly divergent from the neutral estimate, with slightly higher relative reduction in m_{12} than in m_{21}.

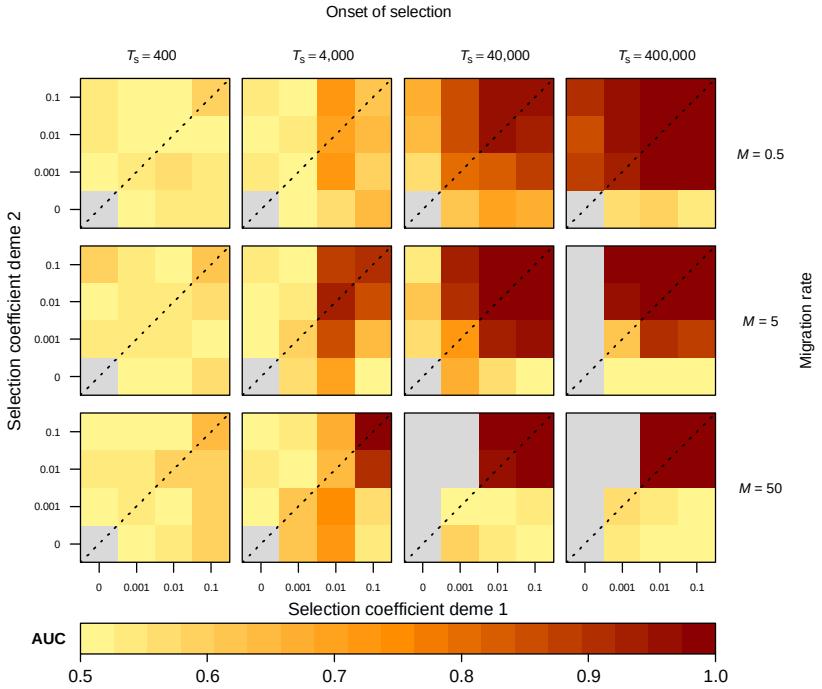

Figure 4. Simulation results showing the effect of migration rate, time of onset of selection and deme-specific selection coefficients on LSD diagnostic performance (AUC), for the 2-deme IM model (model $\mathcal{M}_1$; standing genetic variation case). Each cell represents a pseudo-genome simulated under a specific selection regime. The cell colours reflect the AUC calculated by the correct discrimination of 1000 neutral loci and 50 selected loci in the 1050 loci simulated pseudo-genomes. An AUC value of 0.5 reflects random assignment while that of 1 reflects perfect classification (i.e. TPR=1, FPR=0). Grey cells indicate selection regimes where the derived allele is always lost.

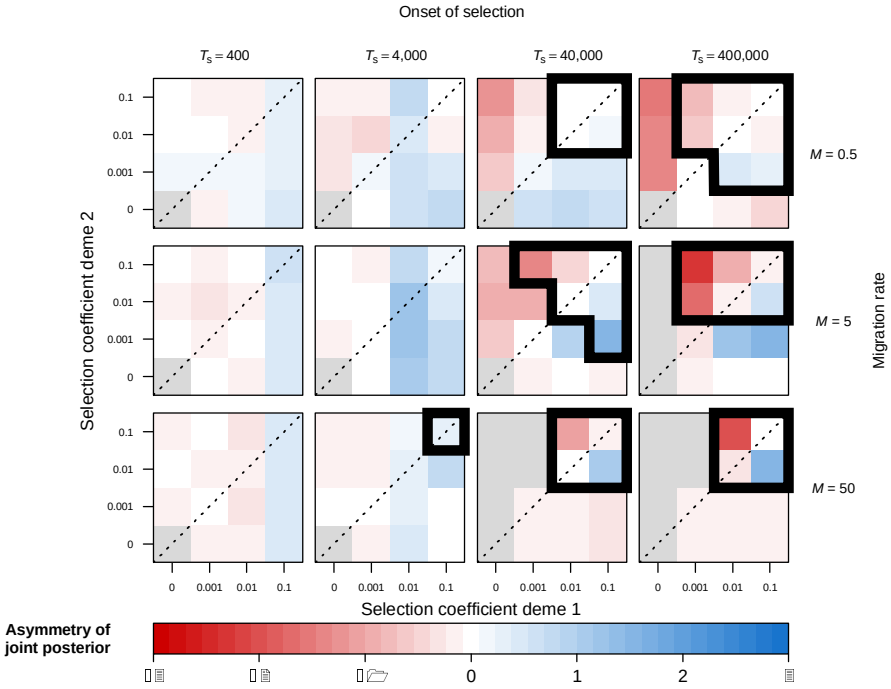

Figure 5. Simulation results showing the effect of migration rate, time of onset of selection and deme-specific selection coefficients on LSD inferred (a)symmetry of selection, for the 2-deme IM model (model $\mathcal{M}_1$; standing genetic variation case). Each cell represents a pseudo-genome simulated under a specific selection regime. The cell colours reflect the (a)symmetry values inferred by LSD, where a value of 0 reflects perfect symmetry of the joint posterior while values divergent from this reflect asymmetry. Cells surrounded by thick lines indicate the values of (a)symmetry for regimes expected to generate meaningful signal (i.e. AUC>0.95). Grey cells indicate selection regimes where the derived allele is always lost.

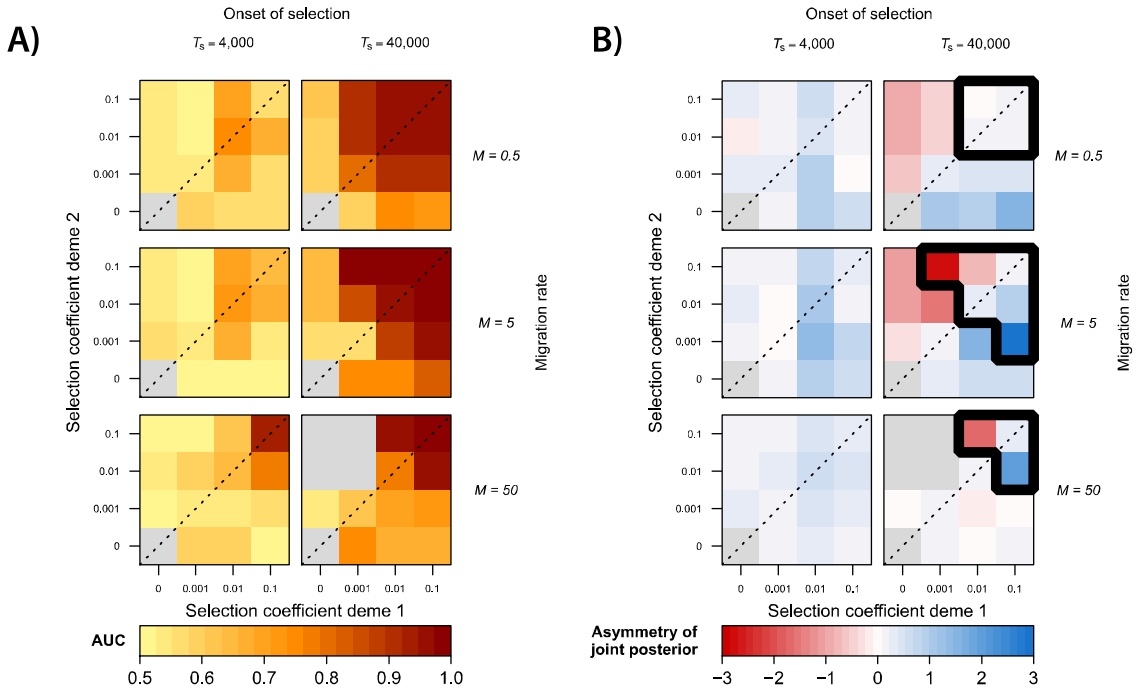

Figure 6. Simulation results for a 2-deme divergence with bottleneck and exponential growth model (model $\mathcal{M}_3$; standing genetic variation case) showing the effect of migration rate, time of onset of selection and deme-specific selection coefficients on A) LSD diagnostic performance (AUC) and B) LSD inferred (a)symmetry of selection. Divergence time of the two populations, T_D, is 200,00 generations ago. Each coloured cell represents a pseudo-genome simulated under a specific selection regime. Grey cells indicate selection regimes where the derived allele is always lost. B) Cells surrounded by thick lines indicate the values of (a)symmetry for regimes expected to generate meaningful signal (i.e. AUC>0.95).

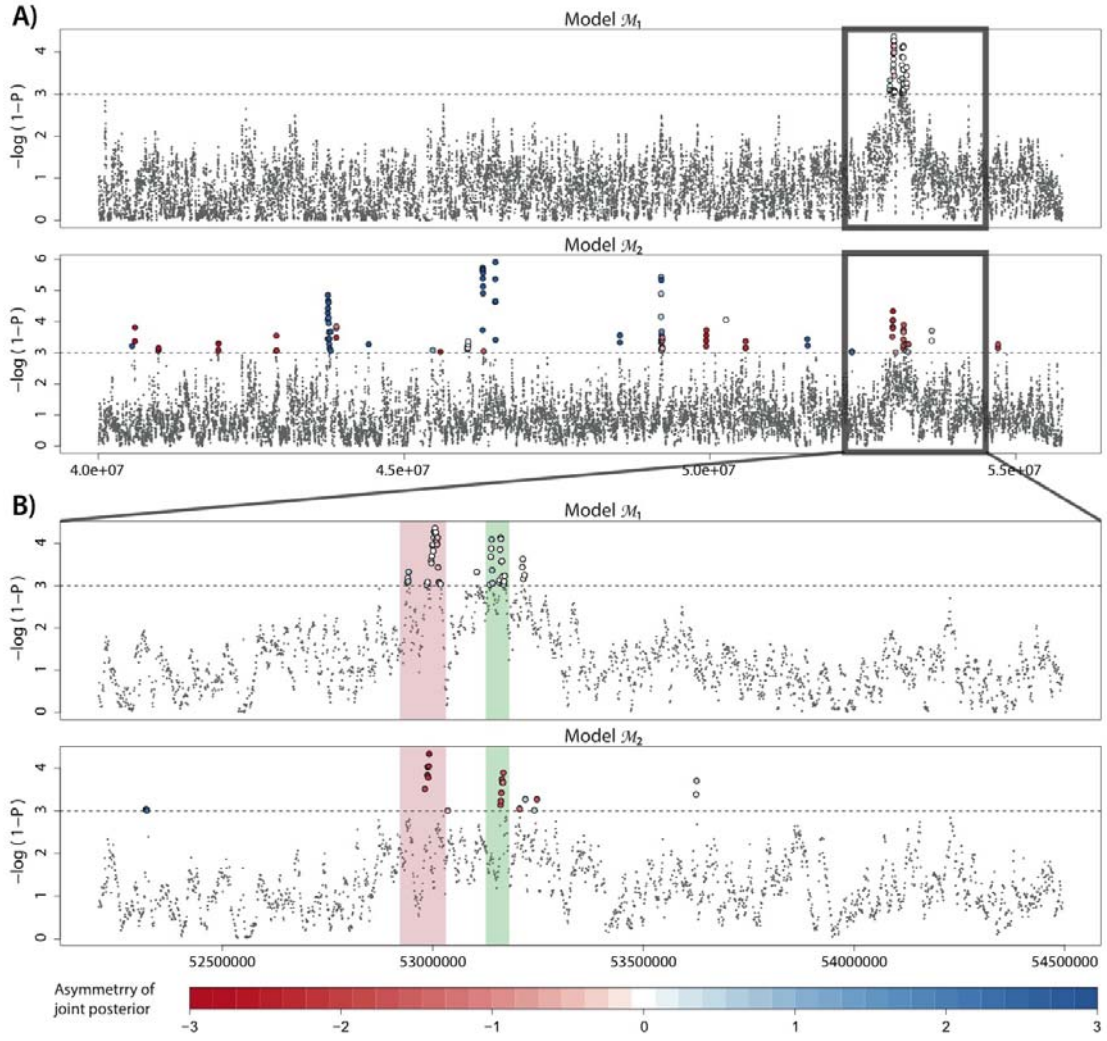

Figure 7. Manhattan plot for the LSD scan of the *A.m.striatum-A.m.pseudomajus* system. The posterior probability of observing the neutral estimate for 10kb windows (1kb step-size) is plotted for A) a 16Mb region of chromosome 6, under models $\boldsymbol{\mathcal{M}_1}$ and $\boldsymbol{\mathcal{M}_2}$, and B) a 2Mb zoomed-in region of chromosome 6 focusing on the ROS-EL region, under the same two models. The horizontal dashed line indicates a 99.9% posterior probability of deviating from neutral expectations. Colour for loci above this threshold denotes the joint (m_{12}-m_{21}) posterior (a)symmetry, and reflects the relative strengths of selection in the two divergent demes or subspecies. A large divergent peak centered around the ROS-EL region (A) is composed of a set of smaller peaks (B), consistent with the ROS (red) and EL (green) loci.

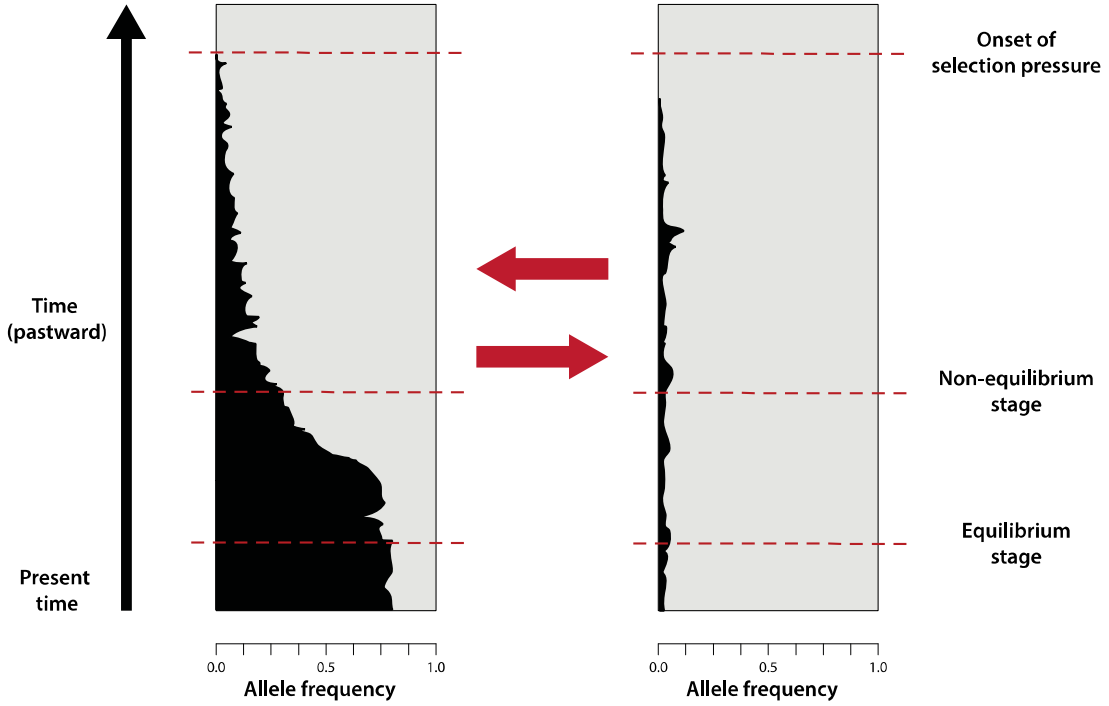

Figure 8. Conceptual illustration of allele frequency trajectories over time in a 2-deme IM model ($\mathcal{M}_1$), for an example *de-novo* case and antagonistic pleiotropic selection regime. The frequency of derived allele *A* is indicated in black and that of ancestral allele *a* in grey. Red arrows represent migration. Prior to reaching drift-migration-selection equilibrium, estimated asymmetries in effective migration rates are also affected by asymmetry in allele frequencies.