Genetic variation in *Loudetia simplex* supports the presence of ancient grasslands in Madagascar

George P. Tiley[1,†], Andrew A. Crowl[2], Tchana O. M. Almary[3], W. R. Quentin Luke[4], Cédrique L. Solofondranohatra[5], Guillaume Besnard[6], Caroline E.R. Lehmann[7,8], Anne D. Yoder[2], Maria S. Vorontsova[1]

[1]Royal Botanic Gardens Kew, Richmond, UK

[2]Department of Biology, Duke University, Durham, North Carolina, 27708 USA

[3]Kew Madagascar Conservation Centre, Antananarivo 101, Madagascar

[4]East African Herbarium, National Museums of Kenya, P.O. Box 45166, Nairobi, 00100, Kenya

[5]Madagascar Biodiversity Center, Parc Botanique et Zoologique de Tsimbazaza, Antananarivo 101, Madagascar

[6]CNRS-UPS-IRD, UMR5174, Laboratoire Évolution & Diversité Biologique, Université Paul Sabatier, Toulouse, France

[7]School of GeoSciences, University of Edinburgh, Edinburgh, UK

[8]Royal Botanic Garden Edinburgh, Edinburgh, UK

[†]Author for Correspondence: g.tiley@kew.org

Summary

1) Research Aims — The extent of Madagascar's grasslands prior to human colonization is unresolved. We used population genetic analyses of a broadly dominant C_4 fire-adapted grass, *Loudetia simplex*, as a proxy for estimating grassland change through time. In the absence of population genomic resources, we used target-enrichment data. We carefully examined the utility of these data for population genetics to make recommendations on cost-effective strategies for conservation genetics. We explored the potential of estimating individual ploidy levels from target-enrichment data and how assumptions about ploidy could affect analyses.
2) Methods — We developed a novel bioinformatic pipeline based to estimate ploidy and genotypes from target-enrichment data. We estimated standard population genetic summary statistics in addition to species trees and population structure. Extended Bayesian skyline plots provided estimates of population size through time for empirical and simulated data.
3) Key Result — All Malagasy *Loudetia simplex* individuals sampled in this study formed a clade and possibly indicated an ancestral high-altitude distribution near the present-day Itremo protected area. Demographic models suggested grassland expansions occurred prior to the Last Interglacial Period and supported extensive grasslands prior to human colonization. Though there are limitations to target-enrichment data for population genetic studies, we find that analyses of population structure are reliable.
4) Key Point —Genetic variation in *Loudetia simplex* supports widespread grasslands in Madagascar prior to the more recent periods of notable paleoclimatic change. However, the methods explored here could not differentiate between paleoclimatic change near the Last Glacial Maximum and anthropogenic effects. Target-enrichment data can be a valuable tool for analyses of population structure in the absence of reference genomic resources.

Keywords

Effective Population Size, Target-Enrichment, HybSeq, Polyploidy, Anthropogenic Climate Change, Conservation Genetics

Societal Impact Statement

Recognizing *Loudetia* dominated grasslands were widespread prior to human colonization highlights that open ecosystems were and continue to be an important component to Madagascar's biodiversity. Urgently required are biodiversity inventories and risk assessments for unique grassland flora and fauna under present day environmental conditions to recognize and quantify modern human impacts within ecosystems historically regarded as wastelands. Substantial financial and logistic barriers exist to implementing conservation studies using contemporary genomic tools that we seek to ameliorate by developing computational resources to leverage a cost-effective data generation strategy that requires no prior genetic knowledge of the target species.

Introduction

The degree to which anthropogenic activities have transformed the ancestral vegetative composition of Madagascar's ecosystems is heavily debated. While the narrative that most of central Madagascar was once covered in closed-canopy forests has been rejected (Bond et al., 2008; Vorontsova et al., 2016; Hackel et al., 2018 Solofondranohatra et al., 2020; Bond et al., 2022; Lehmann et al., 2022), the degree to which a former mosaic of grassland, shrubland, and forests has become more dominated by grasses is contentious (reviewed in Joseph et al., 2021). Understanding population dynamics of individual ecologically important species and plant communities will have profound effects on conservation policy decisions and the economic productivity of grasslands for local communities. However, it is difficult to determine where grasslands were in the past and how extensive they were. Stratigraphic sediment records from central Madagascar indicate an increased relative abundance of grasses and fire through the Holocene (Burney, 1987; Gasse & Van Campo, 1998; Virah-Swamy et al., 2010), especially near the expected age of human colonization [c. two thousand years ago (ka); Dewar & Wright, 1993; Dewar et al., 2013; Pierron et al., 2017]. Megafaunal isotope records also support Holocene expansions of C_4 grasslands but show a vegetative mosaic of C_4 grasses and C_3 vegetation was at least present (Crowley et al., 2021) near the last glacial maximum (LGM; c. 26.5-19 ka). Reconstructing spatio-temporal vegetative dynamics beyond the LGM begin to approach limitations of paleoecological data, but molecular evolutionary approaches can fill the gaps by providing probabilistic inferences through time. Population genetic approaches for reconstruct demographic histories has been successfully applied to lemurs to reconstruct patterns of past forest connectivity (Quéméré et al., 2012; Yoder et al., 2016; Salmona et al., 2017; Teixeira et al., 2021; Tiley et al., 2022) and may similarly provide insights into the Late Pleistocene dynamics of grasses between the LGM and Last Interglacial period (LIG; c. 132-112 ka) or older. This strategy has been used to reconstruct vegetative dynamics of temperate Japanese grasslands from forb species (Yamaura et al., 2019).

An array of tools now exists to estimate the effective population size (N_e) of species through time. As generating genomic data for non-model groups has become more accessible, population genomic methods that utilize precisely estimated site frequency spectra (SFS; Nielsen et al., 2000; Gutenkunst et al., 2009; Excoffier et al., 2013; Liu & Fu, 2020; Blischack et al., 2022) have become a valuable tool for reconstructing the demographic history of natural populations. There are still barriers to generating population genomic data though, either due to the lack of suitable tissue, an appropriate reference genome, or costs. Target-enrichment

libraries have been transformative for the analysis of plant phylogenetic relationships (e.g. Breinholt et al., 2021; Baker et al., 2022), but the population genetic utility of such data are not well-explored. A survey of taxa with low sample sizes suggested population genetic parameters estimated from target-enrichment data could be unreliable, in addition to violating assumptions of the neutral coalescent model that underpins most tests of selection and demographic change (Slimp et al., 2021). However, the relatively long loci recovered by target-enrichment (or HybSeq) methods could be appropriate for a demographic modeling approach, the extended Bayesian skyline plot (EBSP; Heled & Drummond, 2008), which has been successfully applied to a range of conservation genetic questions despite some model violations. We are interested in exploring the potential of target-enrichment data for population or conservation genetic insights from plant groups that may be lacking genomic resources.

Here, we examined the population genetics of the grass species *Loudetia simplex* (Tristachyideae: Panicoideae: Poaceae). It is largely dominant in Madagascar's Central Highlands (Koechlin, 1993; Hagl et al., 2021; Fig. 1a) and is notable for its fire- and grazing-adapted functional traits (Solofondranohatra et al., 2018). Population genetic analyses of microsatellites suggest Malagasy *L. simplex* has been isolated from mainland African populations with no detectable level of gene flow, and that there is additional population structure between the northern and southern extents of the species range across the Central Highlands of Madagascar (Hagl et al., 2021). Here, we use a subset of individuals analyzed by Hagl et al. (2021) along with a new sample from the center of the *L. simplex* distribution to explore the population genetic utility of target-enrichment data and its implications for the natural history of Madagascar's grasslands. Because Malagasy *L. simplex* are putative polyploids (tetraploids and hexaploids; Hagl et al., 2021), we developed novel bioinformatic tools for the processing and analysis of polyploid data, integrated into the PATÉ allele phasing pipeline (Tiley et al., 2021).

Materials and Methods

Sampling and Sequencing

Target-enrichment data was generated for 24 individuals of *Loudetia simplex* and one outgroup, *Tristachya nodiglumus* (Supplementary Table S1). Twenty individuals were selected to best represent Madagascar from available collections as a single unstructured population and avoid complications of population substructure and low sample sizes in downstream analyses (Fig. 1b). Genomic DNA samples were used from Hagl et al. (2021) while new extracts used a

modified CTAB protocol (Doyle & Doyle, 1987) from silica-dried leaf tissue. Genomic DNA was then sent to Rapid Genomics (Gainesville, FL, USA) where sample enrichment and sequencing was performed with a bait kit developed for angiosperms following Breinholt et al. (2021). All sample metadata is provided in Supplementary Table S1.

Sequence Assembly and Processing

Initial contigs were assembled with HybPiper (Johnson et al., 2016) with default settings. We then used HybPiper's supercontigs as reference sequences for PATÉ (Tiley et al., 2021; Crowl et al., 2022) to genotype individuals, filter low-quality variants with GATK v4.2.0 (McKenna et al., 2010; DePristo et al., 2011), and phase haplotype sequences using HPOP-G (Xie et al., 2016). Ploidy information for individuals (Supplementary Table S1) was based on previous microsatellite allele count data (Hagl et al., 2021); however, we also implemented a statistical test to infer ploidy level in PATÉ based on mixtures of distributions from allele balance data, similar to other approaches that use the number of reads supporting an alternate allele over total reads to infer ploidy (Weiß et al., 2018). We extended the mixture models of Tiley et al. (2018) to determine the ploidy of a sample with the model weights as a selection criterion (Supplementary Methods: *Ploidy Test*). PATÉ control files with all filtering options for GATK are provided in the Supplementary Data (Dryad X).

To investigate nucleotide variation and the potential population genetic value of non-coding versus coding regions of target-enrichment loci, we identified the following regions:

1) *Left Flank* – Non-coding sequence upstream of the 5' start codon or splice site
2) *Core* – The coding region targeted by probes for sample enrichment
3) *Right Flank* – Non-coding sequence downstream of the 3' stop codon or splice site

Identification of individual regions was automated (https://github.com/gtiley/Locus-Splitter) based on the tblasx program from BLAST v2.10.0 (Camacho et al., 2009) to find in-frame alignments between our assembled sequences and the original transcriptome loci used for probe development. Individual regions were aligned with MAFFT v7.471 (Katoh & Standley, 2013). The aligned regions were then concatenated together to make whole-locus alignments.

Because we are interested in investigating the analytic impact of ploidy in sequence data, we created three datasets from the phased ploidy-aware data:

1) *One Haplotype (OH)* – Only one phased haplotype sequence was randomly selected from each individual for each locus.

2) *Ambiguous Genotype (AG)* – The genotypic information for an individual is represented as a single sequence with IUPAC ambiguity codes for biallelic sites. Analyses that treat IUPAC codes as missing data were randomly resolved into two sequences.
3) *Phased Haplotypes (PH)* – All phased haplotype sequences for an individual are used.

We expect the OH data to be representative of majority consensus sequences (e.g. HybPiper supercontigs) since our target-enrichment loci are anchored in single exons and we anticipate that any risk of assembling chimeric sequences from such data without paralog warnings is low. Any site genotyped as polyallelic was assumed to be an error and treated as missing sequence across all data sets. All expected copies of a locus are represented even if they are invariant. For example, some individuals in the PH data may have six identical sequences at a locus if they are hexaploids and every site is homozygous. While the ploidy at an individual locus may vary, this is a simplifying assumption to preserve expected allele frequency information.

Phylogeny estimation

A species tree was estimated for all *L. simplex* individuals using whole-locus sequences for the OH, AG, and PH datasets with BPP v4.6.2 (Flouri et al., 2020). We chose a full-likelihood multispecies coalescent (MSC; Rannala & Yang, 2003) model since it is robust to low-information gene tree estimates (Xu & Yang, 2016). Markov chain Monte Carlo (MCMC) analyses collected 10,000 posterior samples, saving every 10 samples after a 10,000 generation burnin. Model and prior specifications are provided in the supplementary information (Supplementary Methods: *BPP Prior Specification*). The starting tree for our Bayesian species tree analyses was based on a concatenated maximum likelihood (ML) estimate from the OH data with IQTREE v2.1.2 (Minh et al., 2020). We used unpartitioned analyses with ModelFinderPlus (Kalyaanamoorthy et al., 2017) for the OH, AG, and PH data to estimate ML evolutionary distances between individuals, with random concatenation of haplotypes within individuals.

Population structure

Potential population structure within our sample of *L. simplex* from Madagascar was explored with STRUCTURE v2.3.4 (Pritchard et al., 2000). For each dataset, we sampled only one biallelic SNP per locus, if present, and required a minimum minor allele count of three with no more than 25% missing data. To explore uncertainty due to sampling error of sites in a target-enrichment dataset that is relatively small compared to contemporary population genomic data,

199 we generated 10 datasets with random SNP sampling. The number of clusters that best satisfy
200 Hardy-Weinberg assumptions was determined with the ΔK method (Evanno et al., 2005).
201 Principal component analyses (PCAs) were conducted on the OH, AG, and PH datasets with
202 adegenet v2.1.3 (Jombart & Ahmed, 2011) by compressing allele counts to the same dimension
203 as the lowest ploidy level present. For example, a hexaploid with three reference and three
204 alternate alleles would be scored as a 2 on a scale from 0 to 4, to be comparable to tetraploids
205 in the dataset. Additional details on allelic imbalances would be lost as, for example, hexaploids
206 with either one or two alternate alleles would be represented as a 1 on the tetraploid scale.
207 Scripts for generating PCA or other clustering-ready matrices from fasta files are available on
208 GPT's GitHub (https://github.com/gtiley/fasta-conversion). Additional details about compression
209 of PCA data matrices can be found in the Supplementary Material (Supplementary Methods:
210 *PCA Input Data*). The presence of isolation-by-distance as an explanation for observed patterns
211 of genetic structure and ordination was tested with the patristic distances from the ML analyses
212 or the uncorrected pairwise distances (p-distances) regressed against the natural logarithm of
213 the great circle distance between individuals. A Mantel test was performed for isolation-by-
214 distance analyses in the ape v5.6.2 R package (Paradis & Schliep, 2019).
215
216 *Population genetic summary statistics*
217 The pairwise nucleotide diversity (π), number of segregating sites (*S*), and Tajima's *D* were
218 calculated across locus regions and datasets with the PopGenome v2.7.5 R package (Pfeifer et
219 al., 2014). The significance of Tajima's *D* statistics was determined by 100 standard coalescent
220 simulations with ms (Hudson, 2002). To estimate 95% confidence intervals on summary
221 statistics, we generated 200 bootstrap replicates of each alignment. The relative estimation
222 error between regions and datasets was compared using the root squared mean error (RSME)
223 from the bootstrap sample. The scaling and interpretation of parameters are described in the
224 supplementary material (Supplementary Methods: *Summary Statistics*). We tested for
225 differences in the distribution of summary statistics among locus regions (left flanks, cores, and
226 right flanks) and data types (OH, AG, and PH) with Wilcoxon Rank-Sum tests. We tested for
227 biases in Tajima's *D* results among locus regions and data types with X^2 tests.
228
229 *Demographic History*
230 We chose the extended Bayesian skyline plot (EBSP, Heled & Drummond, 2008) as the most
231 appropriate model of N_e change through time for our sample of target enrichment data. The
232 model assumes N_e changes coincide with coalescent events, and although an unrealistic

assumption (Yang, 2014), is common. Here we assumed a strict molecular clock underlying all lineages but allowed rates of evolution to vary among loci. One locus was fixed to the clock rate and all other locus rates were relative. Rates of molecular evolution among plants are especially complicated, as there can be life histories with variable reproductive strategies, unknown generation times, and overlapping generations. Thus, we relied on an estimate of the neutral substitution rate at third-position synonymous sites from grass genomes (Christin et al., 2014) and constructed a vague prior around this mean rate, such that any biologically plausible uncertainty in the rate of evolution is reflected in our posterior sample. It is a tenuous rate estimate at best, since it remains unknown whether polyploid *L. simplex* largely propagates clonally, through apomixis, or through sexual reproduction. The MCMC analyses for the OH, AG, and PH data collected 20,000 samples after a burnin of 50,000,000 and sample interval of 10,000. Two replicates were run for each analysis to evaluate convergence and mixing was assessed with Tracer v1.7 (Rambaut et al., 2017). All model and prior specifications are provided in the supplementary information (Supplementary Methods: *EBSP Prior Specification*) and are replicable from XML files on Dryad (X).

Adequacy of Demographic Models

To explore if EBSP models could reliably differentiate between Late Pleistocene versus Holocene population size increases, we evaluated the accuracy of EBSP inferences with simulations that captured some features of our target-enrichment data. We simulated 57 loci that were 1000 bp in length for 20 haplotype individuals 100 times with ms (Hudson, 2002) and seq-gen (Rambaut & Grass, 1997) for four scenarios: 1) a constant N_e of 100,000, 2) an N_e increase from 10,000 to 100,000 individuals roughly coinciding with the LGM at 20,000 years in the past, 3) the same ten-fold increase but at 2,000 years in the past to reflect anthropogenic expansion, and 4) a scenario combining both LGM and anthropogenic effects by exponentially increasing N_e from 10,000 to 100,000 between 20,000 years ago and the present. Each simulated dataset was then evaluated with the same EBSP model using a fixed clock used for our empirical analyses, except the MCMC algorithm collected 10,000 posterior samples with a sample interval of 1,000 after a 100,000 sample burnin. Commands for simulations are given in the supplementary information (Supplementary Methods: *Simulating Demographic Histories*).

Results

Sequencing statistics

266 The probe set targeted 408 loci from which an average of 327 loci were assembled by HybPiper
267 (Supplementary Table S2). These were supported by an average of 217 read pairs per locus.
268 Although there was variable read depth across loci and individuals, the number of assembled
269 loci was significantly correlated with read depth (Pearson's correlation coefficient = 0.57, *p*-
270 value = 0.002). After joint-genotyping all individuals as diploids, normal mixture models
271 suggested all individuals were hexaploids, with the exception of two individuals from South
272 Africa (Supplementary Table S3). Some individuals were inferred as pentaploids, but since this
273 was unexpected based on previous observations from microsatellites (Supplementary Table S1)
274 these cases were assumed to be hexaploids. Although available evidence leaves the ploidy of
275 some individuals ambiguous and raises skepticism of the inference methods, genotyping and
276 phasing of individuals proceeded with the ploidy suggested by our mixture models. We inferred
277 an average of 247 loci with allelic variation, which had a heterozygosity of 0.012 variants per
278 base. After reducing the number of loci with BLAST clustering to avoid tight linkage, a final 301
279 loci were retained from the OH, AG, and PH datasets for downstream analyses.
280
281 *Phylogeny Estimation*
282 Species trees revealed all Malagasy *L. simplex* form a well-supported clade (Supplementary
283 Fig. S1), but there was disagreement in the biogeographic history within Madagascar based on
284 the treatment of ploidy. Individuals in northern Madagascar formed a clade with a posterior
285 probability of 0.9 or greater across the OH, AG, and PH analyses. However, the individuals from
286 central and southern Madagascar were paraphyletic in the OH and AG analyses
287 (Supplementary Figs. 1a and 1b). A clade of central and southern individuals was recovered in
288 the PH analysis, but with a weak posterior probability of 0.64 (Fig. 1c; Supplementary Fig. 1c).
289 Individuals from the center of the *L. simplex* distribution appear to be the earliest diverging
290 lineages in the OH and AG analyses (Supplementary Figs. 1a and 1b), or with respect to the
291 sample of central and southern individuals in the PH analysis (Fig. 1c; Supplementary Fig. 1c).
292
293 *Population sub-structure in Madagascar*
294 STRUCTURE analyses recovered a variable number of Hardy-Weinberg groups across the ten
295 datasets that randomly sampled one non-singleton biallelic SNP per locus (Supplementary
296 Table S4). For the OH data, we found a *K* of two, three, or four in eight, one, and one of the
297 respective replicates. For AG data, the optimal number of clusters was more ambiguous with a
298 *K* of two, three, and four recovered for four, five, and one replicates. PH data found *K* of two and
299 three for eight and two replicates. We note that the method of Evanno et al. (2005) cannot infer

300 if one panmictic population is the best model, and while there is reasonable geographic
301 structure at *K* of two weighted across biological and technical replicates (Fig. 2a) the non-zero
302 admixture between individuals at the most extreme ends could indicate a single population with
303 a cline of genetic variation. This is further supported by the lack of distinct cluster assignments
304 at higher values of *K* despite some evidence for additional population structure (Supplementary
305 Fig. S2). A weak signal of isolation-by-distance was detected when regressing genetic distance
306 against spatial distance (Fig. 2b). Notably, the AG data inferred much higher pairwise distances
307 between individuals; although, the relative differences and overall trend were similar. The same
308 patterns can be observed when using patristic distances from ML phylogenetic estimates,
309 although the disparity between AG versus OH or PH distances is reduced, as is the percent of
310 genetic variation explained by spatial distances (Supplementary Fig. S3).
311
312 PCAs reflected findings from STRUCTURE analyses and reinforce the presence of a genetic
313 cline (Fig. 3). For OH data, where only one reference or alternate allele was observed per
314 individual, Malagasy *L. simplex* is clearly distinguished from mainland African individuals (Fig.
315 3a), but there is no clear separation of individuals based on geographic location (Fig. 3b). PCAs
316 of data that accounts for heterozygosity but not dosage recovered a single Malagasy versus
317 mainland African pattern (Fig. 3c), PCAs reflect a north-to-south gradient of genetic variation
318 regardless of whether allelic dosage was treated as hexaploid (Fig. 3d), tetraploid (Fig. 3e), or
319 diploid (Fig. 3f). The total percent of variance explained by principal component axis 1 (PC1)
320 and axis 2 (PC2) increased as allelic dosage was compressed from a hexaploid to a diploid
321 state.
322
323 *Variation in summary statistics*
324 There were significant differences in the distributions of π, *S*, and Tajima's *D* among locus
325 regions and between the OH, AG, and PH data (Fig. 4). For π, there was always a difference
326 between flanking regions and the exon core (Supplementary Table S5), with the cores
327 surprisingly having more variation. Differences were also observed between the flanking regions
328 and cores for *S* and Tajima's *D*, except when comparing the left flanks and cores for OH data
329 (Supplementary Table S5). The only case of a difference between the left and right flanks was
330 when estimating π for AG data (Supplementary Table S5). The OH and PH data resulted in
331 similar estimates of π; however, the AG and PH data had similar distributions of *S*
332 (Supplementary Table S5). Any differences in Tajima's *D* statistics were largely explained by an

333 over-representation of values less than negative two, indicating positive selection or population
334 size expansion, in the OH data and among core regions across datasets (Table 1).
335
336 We attempted to explore biases in summary statistic estimates due to sampling error of sites
337 through bootstrapping the OH (Supplementary Fig. S4), AG (Supplementary Fig. S5), and PH
338 data (Supplementary Fig. S6). Since mutations are assumed to arise through a Poisson process
339 and exponential distributions of measures of genetic variation, such as π and *S* are often
340 observed, the bootstrap confidence intervals are not a reasonable measure of uncertainty.
341 However, we used the bias, difference between the bootstrap mean and observed estimate, and
342 the RSME of bootstrap replicates with respect to the observed estimate, to gain some general
343 insights. Notably, the bias of π is larger and much noisier in the AG data (Supplementary Fig.
344 S5) compared to the OH (Supplementary Fig. S4) and PH (Supplementary Fig. S6) data. Even if
345 the bias for a whole-locus is low, there can be high variability between regions. For the
346 observed values, estimates are not always proportional among regions either. The RSME of an
347 estimate tends to decrease along with genetic variation, implying some role of genotyping error
348 or rare variants and singletons inflating values at the higher end of the spectrum. Loci with very
349 low levels of genetic variation can also be problematic though, as the RSME for flanking regions
350 increases towards the tail of the distribution, creating opportunity for sampling error to wildly
351 swing estimates regardless of ploidy (Supplementary Figs. S4-S6). The number of segregating
352 sites increased with ploidy of the data, but π was relatively stable. However, differences
353 between π and *S* will affect estimates of Tajima's *D*, and this could affect interpretations of
354 demography or selection acting on loci; although, the majority of loci and their regions were not
355 significantly different from neutral expectations given the simulation-based test statistics
356 (Supplementary Table S6).
357
358 *Population Size Change through time*
359 After filtering for longer loci with low amounts of missing data, 57 loci were retained for EBSP
360 analyses. Reducing our data to a subset of more informative loci was done for computational
361 convenience, because EBSP models are parameter-rich and require long MCMC runs for
362 convergence. Runs converged for each data type for both analyses with a fixed rate of
363 molecular evolution (Fig. 5) and variable rate (Supplementary Fig. S7). All analyses suggested
364 the *L. simplex* N_e increased through the Late Pleistocene, well before the LIG, approximately
365 one million years ago. Between the LGM and period of potential human colonization of
366 Madagascar in the Holocene, N_e had already reached present-day estimates. N_e estimates

through the Holocence were only possible for the PH data, as accounting for the within-individual variation should have captured more recent coalescences. Simulations cast doubt on the seriousness with which EBSP results can be interpreted through the Holocene though. For example, when N_e was truly constant through time, some individual EBSP analyses would suggest recent N_e increases; although, the mean estimate across simulations recovered the constant population scenario (Fig. 6a). This problem was exaggerated in the presence of model over-fitting (Supplementary Fig. S8). In scenarios with recent N_e increases, the contemporary N_e was underestimated but a trend indicating a recent N_e increase was detectable (Figs. 6b-6d). However, simulations indicated it would not be possible to differentiate between N_e increases that occurred during the LGM (Fig. 6b) versus the Late Holocene (Fig. 6c), let alone some combination of paleoclimate and anthropogenic effects (Fig. 6d). Interpreting these recent N_e increases from EBSP analyses becomes impossible with model over-fitting (Supplementary Fig. S8).

Discussion

Robustness of analyses across loci and ploidy

Although transformative for species-level phylogenomic studies (Breinholt et al., 2021; Baker et al., 2022), the robustness of population genetic estimates for a sample of closely-related individuals was largely unexplored (but see Slimp et al., 2021). Some of the most important measures of genetic variation for populations were susceptible to sampling error, especially in the flanking intron regions (Supplementary Figs. S4-S6). While estimates from cores showed less error, they will lead to some violations of population genetic methods that assume evolution consistent with neutral theory, given potential signatures of purifying selection (Table 1). This creates some conflict, since most of the information is contained within the exon cores of target-enrichment loci (Fig. 4), presumably due to read coverage. Because Tajima's *D* alone cannot disentangle demography from selection, some judgment calls and tolerance for downstream model violations are necessary. The majority of loci examined were consistent with neutral expectations based on simulations (Supplementary Table S6), and the proportion of loci with significant Tajima's *D* statistics were exaggerated when ignoring heterozygosity (Table 1).

Predicting the effects of assumptions about genotype on estimates of genetic variation and downstream analyses are not always straightforward. For example, the pairwise distances (Fig. 2b) or evolutionary distances (Supplementary Fig. S3) between individuals was higher when assuming all individuals were diploid (AG) compared to assuming all individuals were haploid

401 (OH) or genotyped at their anticipated ploidy level (PH). If we were willing to ignore the absolute
402 estimates of genetic distance, the interpretation of isolation-by-distance analyses was the same
403 across the OH, AG, and PH data. This implies that biologically meaningful conclusions can arise
404 even with uncertainty in ploidy and phasing. Interestingly, treating all individuals as diploid was
405 sufficient for STRUCTURE to recover northern and southern clusters with some admixed central
406 individuals of varying proportions (Fig. 2a; Supplementary Fig. S2). Similarly, PCAs reflected the
407 geographic distribution of our *L. simplex* samples in the AG and PH data with, surprisingly, the
408 AG data explaining more of the observed genetic variance (Fig. 3f) than the PH data (Fig. 3d).
409 This may be due to genotyping error at higher ploidy levels caused by read stochasticity and a
410 flat genotype likelihood surface (Blischak et al., 2016) or caused by incorrect ploidy inferences.
411 A similar mixture model approach to the one implemented here (Weiß et al., 2018) has been
412 shown to have good performance based on comparisons of allele balance distributions from
413 target-enrichment data and flow cytometry-based ploidy estimates (Viruel et al., 2019). We are
414 cautiously optimistic about the prospects of direct ploidy estimation from sequence data,
415 because likelihood-based criteria can be prone to overfitting (Tiley et al., 2018). However, our
416 ploidy estimates are generally consistent with the previous inferences from microsatellite allele
417 counts (Hagl et al., 2021; Supplementary Table S1; Supplementary Table S3). A positive insight
418 from our comparisons across ploidy levels is that if unknown individuals are genotyped as
419 diploid, there will be sufficient information about population structure, and that ploidy estimation
420 should not be a barrier to investigations of natural plant populations with little *a priori* cytological
421 or genetic knowledge.
422
423 *Limits to differentiating anthropogenic and paleoclimate N_e change*
424 Large population genomic studies in plants that can leverage demographic estimates from SFS
425 (e.g. Excoffier et al., 2013; Liu & Fu, 2020) are commonly focused on temperate regions (e.g.
426 Yamaura et al., 2019) and species exploited for energy or agriculture (e.g. Huang et al., 2014;
427 Wang et al., 2017). Thus, it is tempting to leverage models that use gene trees as data such as
428 the EBSP (Heled & Drummond, 2008) or other extensions of skyline plots (Pybus et al., 2000),
429 since they could be paired with relatively low-cost target-enrichment strategies that yield multi-
430 locus nuclear data while stretching material quality.
431
432 While our analyses provide valuable insights into the ancient origins of *L. simplex* dominated
433 grasslands, both our empirical and simulated analyses show we cannot differentiate between N_e
434 increases that occurred during human colonization or paleoclimate change near the LGM. The

435 change in vegetative composition towards more C_4 grass species in central Madagascar is well-
436 documented (Burney, 1987; Gasse & Van Campo, 1998; Virah-Swamy et al., 2010; Crowley et
437 al., 2021). It is unlikely that these increases in C_4 grasses did not involve *L. simplex*; although,
438 an increase in census size may not be reflected by N_e (Fig. 5; Supplementary Fig. S7), the
439 expected population size under a neutral coalescent model. Our simulations show that detecting
440 these recent increases is difficult, and that we cannot differentiate between a ten-fold increase
441 that occurred 200 generations (2 ka) or 2000 generations (20 ka; Fig. 6). Diagnosing a true N_e
442 change becomes impossible in the presence of model overfitting (Supplementary Fig. S8) and
443 diagnosing these problems in empirical data can be difficult aside from convergence
444 assessments.
445
446 These limitations do not necessarily disagree with an assessment of EBSP models with a larger
447 genomic dataset from endemic palm species in Madagascar (Helmstetter et al., 2021).
448 Helmstetter et al. (2021) found that the estimated N_e trends were valuable predictors of IUCN
449 risk assessments, and that population decreases could be detected at anthropogenic time
450 scales. However, it should be easier to detect population declines than population increases in a
451 system that already had a large degree of standing genetic variation, such as *L. simplex*. This is
452 because the time until the most recent common ancestor of a genealogy will be much more
453 recent in a small population than a large one (e.g. see equation 2.13 from Gillespie 2004).
454 Analyses of demographic models with genomic SFS should be able to increase resolution at
455 recent anthropogenic time scales (Patton et al., 2019) and differentiate anthropogenic from
456 paleoclimatic demographic effects (Tiley et al., 2022), but this would require much more
457 extensive sampling and sequencing effort. Additionally, our analyses treat all individuals as a
458 single population, and since we detected some level of population structure, we cannot rule out
459 their confounding effects on analyses of N_e through time (Heller et al., 2013; Mazet et al., 2016)
460 or other confounding affects such as clonal growth and overlapping populations. Population-
461 level sampling of multiple sites throughout a species range should be used to control for
462 population structure, while accounting for localized effects such as agriculture and grazing
463 versus regional climate effects.
464
465 *New insights into Loudetia simplex evolution and policy implications*
466 Our analyses yielded new findings about the evolution of *L. simplex* that are potentially
467 consequential for the broader understanding of Malagasy grasslands. First, our species trees
468 (Supplementary Fig. S1) indicate that individuals from what we call central Madagascar – a

469 loose term meant only to indicate individuals from the center of the sampled *L. simplex*
470 distribution – represent the ancestral distribution. It is likely that *L. simplex* dominated
471 grasslands have origins in some of the higher elevation regions of the Central Highlands. For
472 example, MSV1564 comes from the Itremo protected area (Fig. 1b; Supplementary Table S1),
473 and further investigations of this region should be productive for characterizing natural
474 grasslands in Madagascar. The hypothesis of central high-altitude origins is also supported by
475 our analyses of population structure (Fig. 2a; Supplementary Fig. S2) and the weak but
476 significant pattern of isolation-by-distance (Fig. 2b; Supplementary Fig. S3).
477
478 Our models of N_e change through suggests that *L. simplex* grasslands were already extensive
479 (Fig. 5; Supplementary Fig. S7) well before any hypothesized human presence in Madagascar
480 (c. 10 ka; Hansford et al., 2018) let alone human colonization (c. 2 ka; Dewar and Wright, 1993;
481 Dewar et al., 2013; Pierron et al., 2017). We found an approximate ten-fold increase between
482 one MA and the LIG which is beyond the limitations of available stratigraphic records. Increases
483 in grass and especially C_4 grass relative abundance through the Holocene are well-
484 characterized (Burney, 1987; Gasse & Van Campo, 1998; Virah-Swamy et al., 2010; Railsback
485 et al., 2020; Crowley et al., 2021); however, we showed that limitations of our model do not
486 permit us to make inferences at these very recent timescales (Fig. 6). Strong assumptions about
487 the mutation rate and generation time to calibrate analyses to absolute time and N_e were
488 necessary, but our interpretations of results are within the biologically plausible uncertainty of
489 these parameters. Even if the per-year substitution rate was overestimated by a factor of ten,
490 the N_e increase would still be constrained to the Late Pleistocene, occurring between the LIG
491 and LGM. These analyses, especially in the context of a previous investigation of population
492 structure (Hagl et al., 2021), suggest that *L. simplex* dominated grasslands have pre-human
493 origins and likely experienced some N_e increase coincident with Pleistocene climate change.
494 Tree planting programs have been deployed to combat the measurable habitat loss and
495 deforestation in Madagascar since the 1950s (Green & Sussman, 1990; Harper et al., 2007;
496 Vieilledent et al., 2018). The results of our study strongly imply that these efforts should be
497 cautious of treating *L. simplex* dominated grasslands as degraded forests, and at least a mosaic
498 of *L. simplex* and native woody species would be more appropriate.
499
500 *Potential use cases of target-enrichment data for conservation genetics*
501 One application of target-enrichment data going forward could be analyses of population
502 structure and population assignment of unknown individuals. For example, many conservation

503 programs operating at scale rely on reduced-SNP panels to determine provenance of an
504 individual or sample (e.g. Bourgeois et al., 2018; Stronen et al., 2022). Developing SNP panels
505 requires at least some *a priori* genomic resources, so target-enrichment libraries could provide
506 feasible approaches in the absence of such references and practical improvements over
507 existing barcoding strategies for identifying, for example, Madagascar's rosewoods (Hassold et
508 al., 2016); although acquiring viable DNA samples from woody tissue would remain a barrier
509 (Jiao et al., 2018). We anticipate limited use cases of target-enrichment data in population
510 genetics going forward as reference genome assembly is becoming more accessible through
511 decreasing long-read costs, and less complicated library preparation protocols for whole-
512 genome resequencing will become preferential to any target-enrichment strategy. However, the
513 population genetic inferences from target-enrichment data should be biologically meaningful
514 when available.
515
516 *A practical need for plant population genomics in Madagascar*
517 There are few studies of plant population genetics from the Central Highlands and those are
518 limited to AFLP (Gardiner et al., 2017) or microsatellite data (Salmona et al., 2020; Hagl et al.,
519 2021). Demographic analyses from the endemic olive tree *Noronhia lowryi* suggested recent
520 population declines of the woody species, but it is unclear if these declines represent
521 anthropogenic or paleoclimatic change near the LGM (Salmona et al., 2020). While our
522 analyses of *L. simplex* are also ambiguous regarding this window of anthropogenic effects on
523 the extent of grasslands, we can at least conclude that the *L. simplex* N_e was quite large.
524 Contemporary population genomic analyses will be necessary to provide statistical power to
525 differentiate between competing demographic hypotheses, not only for grasses such as *L.*
526 *simplex* but also species representing plant diversity across families and life-history. A better
527 reconstruction of the natural history of Madagascar's grasslands is not only necessary for
528 understanding diversity within and outside of the Central Highlands (Yoder et al., 2016; Burbrink
529 et al., 2019; Everson et al., 2020), but also to make effective land management decisions for
530 tree planting programs.
531
532 **Acknowledgements**
533 GPT was supported by a Marie Curie Individual Fellowship (MSCA-IF-EF-ST 101026923). GB is
534 supported by LabEx TULIP (ANR-10-LABX-0041). CERL was supported by a The Royal Society
535 International Collaboration Award. Thank you to MEDD Madagascar, Madagascar National
536 Parks, Isalo MNP, and Itremo NPA for supporting our research and exportation permits in 2014

and 2017, and also to COSTECH Tanzania for their support for the 2014 collection permits for MSV and WRQL. This work would not have been possible without Stuart Cable and everyone at the Kew Madagascar Conservation Centre. We thank Emily Lockwood for DNA extractions, and Lorena Endara and J. Gordon Burleigh for coordinating library preparation and sequencing.

Author Contributions

GPT, ADY, and MSV conceived the study. WRQL, CLS, CERL, GB and MSV collected samples. GPT and TOMA curated data. GPT and AAC analyzed data. GPT wrote the first draft. All authors revised the manuscript and approved the final version.

Data Availability Statement

Scripts for analyses and input files are available through the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.[NNNN]. Additional scripts are available on GPT's GitHub for identifying exon-intron boundaries from target enrichment data (https://github.com/gtiley/Locus-Splitter) generating matrices for PCAs from fasta files (https://github.com/gtiley/fasta-conversion) and the ploidy test implemented through PATÉ (https://github.com/gtiley/Phasing). Raw sequence data is available via the NCBI short read archive (SRA) database, with individual SRA and BioSample identifiers in Supplementary Table S1. Raw sequence data is associated with NCBI BioProject PRJNA952819.

Conflict of Interest Statement

The authors declare no conflict of interest.

References

Baker, W. J., Bailey, P., Barber, V., Barker, A., Bellot, S., Bishop, D., ... & Forest, F. (2022). A comprehensive phylogenomic platform for exploring the angiosperm tree of life. *Systematic Biology*, *71*(2), 301-319. https://doi.org/10.1093/sysbio/syab035

Blischak, P.D., Kubatko, L.S., Wolfe, A.D. (2016_. Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Molecular Ecology Resources, 16*(3), 742-754. https://doi.org/10.1111/1755-0998.12493

Blischak, P.D., Sajan, M., Barker, M.S., Gutenkunst, R.N. (2022). Demographic history inference and the polyploid continuum. *bioRxiv*, https://doi.org/10.1101/2022.09.15.508148

Bond, W. J., Silander Jr, J. A., Ranaivonasy, J., & Ratsirarson, J. (2008). The antiquity of Madagascar's grasslands and the rise of C4 grassy biomes. *Journal of Biogeography*, *35*(10), 1743-1758. https://doi.org/10.1111/j.1365-2699.2008.01923.x

Bond, W. J., Silander, J. A., & Ratsirarson, J. (2023). Madagascar's grassy biomes are ancient and there is much to learn about their ecology and evolution. *Journal of Biogeography*, *50*(3), 614-621. https://doi.org/10.1111/jbi.14494

Bourgeois, S., Senn, H., Kaden, J., Taggart, J. B., Ogden, R., Jeffery, K. J., ... & McEwing, R. (2018). Single-nucleotide polymorphism discovery and panel characterization in the African forest elephant. *Ecology and Evolution*, *8*(4), 2207-2217. https://doi.org/10.1002/ece3.3854

Breinholt, J. W., Carey, S. B., Tiley, G. P., Davis, E. C., Endara, L., McDaniel, S. F., ... & Burleigh, J. G. (2021). A target enrichment probe set for resolving the flagellate land plant tree of life. *Applications in Plant Sciences*, *9*(1), e11406. https://doi.org/10.1002/aps3.11406

Burbrink, F. T., Ruane, S., Kuhn, A., Rabibisoa, N., Randriamahatantsoa, B., Raselimanana, A. P., ... & Raxworthy, C. J. (2019). The origins and diversification of the exceptionally rich gemsnakes (Colubroidea: Lamprophiidae: Pseudoxyrhophiinae) in Madagascar. *Systematic Biology*, *68*(6), 918-936. https://doi.org/10.1093/sysbio/syz026

Burney, D. A. (1987). Late Holocene vegetational change in central Madagascar. *Quaternary Research*, *28*(1), 130-143. https://doi.org/10.1016/0033-5894(87)90038-X

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*, 1-9. https://doi.org/10.1186/1471-2105-10-421

Christin, P.-A., Spriggs, E., Osborne, C. P., Strömberg, C. A., Salamin, N., & Edwards, E. J. (2014). Molecular dating, evolutionary rates, and the age of the grasses. *Systematic biology*, *63*(2), 153-165. https://doi.org/10.1093/sysbio/syt072

Crowl, A. A., Fritsch, P. W., Tiley, G. P., Lynch, N. P., Ranney, T. G., Ashrafi, H., & Manos, P. S. (2022). A first complete phylogenomic hypothesis for diploid blueberries (Vaccinium section Cyanococcus). *American Journal of Botany*, *109*(10), 1596-1606. https://doi.org/10.1002/ajb2.16065

Crowley, B. E., Godfrey, L. R., Hansford, J. P., & Samonds, K. E. (2021). Seeing the forest for the trees—and the grasses: revisiting the evidence for grazer-maintained grasslands in Madagascar's Central Highlands. *Proceedings of the Royal Society B*, *288*(1950), 20201785. https://doi.org/10.1098/rspb.2020.1785

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, *43*(5), 491-498. https://doi.org/10.1038/ng.806

Dewar, R. E., Radimilahy, C., Wright, H. T., Jacobs, Z., Kelly, G. O., & Berna, F. (2013). Stone tools and foraging in northern Madagascar challenge Holocene extinction models. *Proceedings of the National Academy of Sciences*, *110*(31), 12583-12588. https://doi.org/10.1073/pnas.1306100110

Dewar, R. E., & Wright, H. T. (1993). The culture history of Madagascar. *Journal of World Prehistory*, *7*, 417-466. https://doi.org/10.1007/BF00997802

Doyle, J.J., Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin, 19*(1), 11-15.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. Molecular Ecology, 14(8), 2611–2620. https://doi.org/10.1111/J.1365-294X.2005.02553.X

Everson, K. M., Jansa, S. A., Goodman, S. M., & Olson, L. E. (2020). Montane regions shape patterns of diversification in small mammals and reptiles from Madagascar's moist evergreen forest. *Journal of Biogeography*, *47*(10), 2059-2072. https://doi.org/10.1111/jbi.13945

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Flouri, T., Jiao, X., Rannala, B., & Yang, Z. (2020). A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Molecular Biology and Evolution*, *37*(4), 1211-1223. https://doi.org/10.1093/molbev/msz296

Gardiner, L. M., Rakotoarinivo, M., Rajaovelona, L. R., & Clubbe, C. (2017). Population genetics data help to guide the conservation of palm species with small population sizes and fragmented habitats in Madagascar. *PeerJ*, *5*, e3248. https://doi.org/10.7717/peerj.3248

Gasse, F., & Van Campo, E. (1998). A 40,000-yr pollen and diatom record from Lake Tritrivakely, Madagascar, in the southern tropics. *Quaternary Research*, *49*(3), 299-311. https://doi.org/10.1006/qres.1998.1967

Gillespie, J.H. (2004). The coalescent. In, *Population Genetics: A Concise Guide* (pp.40-47). Baltimore, MD. John Hopkins University Press.

Green, G. M., & Sussman, R. W. (1990). Deforestation history of the eastern rain forests of Madagascar from satellite images. *Science*, *248*(4952), 212-215. https://doi.org/10.1126/science.248.4952.212

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Hackel, J., Vorontsova, M.S., Nanjarisoa, O.P., Hall, R.C., Razanatsoa, J., Malakasi, P., & Besnard, G. (2018). Grass diversification in Madagascar: In situ radiation of two large C3 shade clades and support for a Miocene to Pliocene origin of C4 grassy biomes. *Journal of Biogeography, 45*(4), 750–761. https://doi.org/10.1111/jbi.13147

Hagl, P. A., Gargiulo, R., Fay, M. F., Solofondranohatra, C., Salmona, J., Suescun, U., ... & Vorontsova, M. S. (2021). Geographical structure of genetic diversity in Loudetia simplex (Poaceae) in Madagascar

and South Africa. *Botanical Journal of the Linnean Society*, *196*(1), 81-99. https://doi.org/10.1093/botlinnean/boaa098

Hansford, J., Wright, P. C., Rasoamiaramanana, A., Pérez, V. R., Godfrey, L. R., Errickson, D., ... & Turvey, S. T. (2018). Early Holocene human presence in Madagascar evidenced by exploitation of avian megafauna. *Science Advances*, *4*(9), eaat6925. https://doi.org/10.1126/sciadv.aat6925

Harper, G. J., Steininger, M. K., Tucker, C. J., Juhn, D., & Hawkins, F. (2007). Fifty years of deforestation and forest fragmentation in Madagascar. *Environmental conservation*, *34*(4), 325-333. https://doi.org/10.1017/S0376892907004262

Heled, J., & Drummond, A. J. (2008). Bayesian inference of population size history from multiple loci. *BMC evolutionary biology*, *8*(1), 1-15. https://doi.org/10.1186/1471-2148-8-289

Heller, R., Chikhi, L., & Siegismund, H. R. (2013). The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PloS one*, *8*(5), e62992. https://doi.org/10.1371/journal.pone.0062992

Helmstetter, A. J., Cable, S., Rakotonasolo, F., Rabarijaona, R., Rakotoarinivo, M., Eiserhardt, W. L., ... & Papadopulos, A. S. (2021). The demographic history of Madagascan micro-endemics: have rare species always been rare?. *Proceedings of the Royal Society B*, *288*(1959), 20210957. https://doi.org/10.1098/rspb.2021.0957

Huang, P., Feldman, M., Schroder, S., Bahri, B. A., Diao, X., Zhi, H., ... & Kellogg, E. A. (2014). Population genetics of Setaria viridis, a new model system. *Molecular Ecology*, *23*(20), 4912-4925. https://doi.org/10.1111/mec.12907

Hudson, R. R. (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337-338. https://doi.org/10.1093/bioinformatics/18.2.337

Jiao, L., Yu, M., Wiedenhoeft, A. C., He, T., Li, J., Liu, B., ... & Yin, Y. (2018). DNA barcode authentication and library development for the wood of six commercial Pterocarpus species: the critical role of xylarium specimens. *Scientific Reports*, *8*(1), 1945. https://doi.org/10.1038/s41598-018-20381-6

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., ... & Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in plant sciences*, *4*(7), 1600016. https://doi.org/10.3732/apps.1600016

Jombart, T., & Ahmed, I. (2011). adegenet 1.3–1: New tools for the analysis of genome-wide SNP data. *Bioinformatics, 27(*21), 3070–3071. https://doi.org/10.1093/bioin forma tics/btr521

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772-780. https://doi.org/10.1093/molbev/mst010

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, *14*(6), 587-589. https://doi.org/10.1038/nmeth.4285

Koechlin, J. (1993). Grasslands of Madagascar. In R. T. Coupland (Ed.), *Natural Grasslands: Eastern Hemisphere and Résumé* (pp. 291-301). Amsterdam, Netherlands. Elsevier.

Lehmann, C. E. R., Solofondranohatra, C., Morton, J. A., Phelps, L. N., Ralimanana, H., Razanatsoa, J., Rakotoarimanana, V., Vorontsova, M.S. (2022). The grassy ecosystems of Madagascar. In S. M.

Goodman (Ed.), *The new natural history of Madagascar* (pp. 152-168). Princetonm NJ. Princeton University Press.

Liu, X., & Fu, Y. X. (2020). Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome biology*, *21*(1), 1-9. https://doi.org/10.1186/s13059-020-02196-9

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference?. *Heredity*, *116*(4), 362-371. https://doi.org/10.1038/hdy.2015.104

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303. https://doi.org/10.1101/gr.107524.110

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530-1534. https://doi.org/10.1093/molbev/msaa015

Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, *154*(2), 931-942. https://doi.org/10.1093/genetics/154.2.931

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526-528. https://doi.org/10.1093/bioinformatics/bty633

Patton, A. H., Margres, M. J., Stahlke, A. R., Hendricks, S., Lewallen, K., Hamede, R. K., ... & Storfer, A. (2019). Contemporary demographic reconstruction methods are robust to genome assembly quality: A case study in Tasmanian devils. *Molecular Biology and Evolution*, *36*(12), 2906-2921. https://doi.org/10.1093/molbev/msz191

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929-1936. https://doi.org/10.1093/molbev/msu136

Pierron, D., Heiske, M., Razafindrazaka, H., Rakoto, I., Rabetokotany, N., Ravololomanga, B., ... & Letellier, T. (2017). Genomic landscape of human diversity across Madagascar. *Proceedings of the National Academy of Sciences*, *114*(32), E6498-E6506. https://doi.org/10.1073/pnas.1704906114

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959. https://doi.org/10.1093/genetics/155.2.945

Pybus, O. G., Rambaut, A., & Harvey, P. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, *155*(3), 1429-1437. https://doi.org/10.1093/genetics/155.3.1429

Quéméré, E., Amelot, X., Pierson, J., Crouau-Roy, B., & Chikhi, L. (2012). Genetic data suggest a natural prehuman origin of open habitats in northern Madagascar and question the deforestation narrative in this region. *Proceedings of the National Academy of Sciences*, *109*(32), 13028-13033. https://doi.org/10.1073/pnas.1200153109

Railsback, L. B., Dupont, L. A., Liang, F., Brook, G. A., Burney, D. A., Cheng, H., & Edwards, R. L. (2020). Relationships between climate change, human environmental impact, and megafaunal extinction inferred from a 4000-year multi-proxy record from a stalagmite from northwestern Madagascar. *Quaternary Science Reviews*, *234*, 106244. https://doi.org/10.1016/j.quascirev.2020.106244

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic biology*, *67*(5), 901-904. https://doi.org/10.1093/sysbio/syy032

Rambaut, A., & Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(3), 235-238. https://doi.org/10.1093/bioinformatics/13.3.235

Rannala, B., & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, *164*(4), 1645-1656. https://doi.org/10.1093/genet ics/164.4.1645

Salmona, J., Heller, R., Quéméré, E., & Chikhi, L. (2017). Climate change and human colonization triggered habitat loss and fragmentation in Madagascar. *Molecular Ecology*, *26*(19), 5203-5222. https://doi.org/10.1111/mec.14173

Salmona, J., Olofsson, J. K., Hong-Wa, C., Razanatsoa, J., Rakotonasolo, F., Ralimanana, H., ... & Besnard, G. (2020). Late Miocene origin and recent population collapse of the Malagasy savanna olive tree (Noronhia lowryi). *Biological Journal of the Linnean Society*, *129*(1), 227-243. https://doi.org/10.1093/biolinnean/blz164

Solofondranohatra, C. L., Vorontsova, M. S., Hackel, J., Besnard, G., Cable, S., Williams, J., ... & Lehmann, C. E. (2018). Grass functional traits differentiate forest and savanna in the Madagascar central highlands. *Frontiers in Ecology and Evolution*, *6*, 184. https://doi.org/10.3389/fevo.2018.00184

Solofondranohatra, C. L., Vorontsova, M. S., Hempson, G. P., Hackel, J., Cable, S., Vololoniaina, J., & Lehmann, C. E. (2020). Fire and grazing determined grasslands of central Madagascar represent ancient assemblages. *Proceedings of the Royal Society B*, *287*(1927), 20200598. https://doi.org/10.1098/rspb.2020.0598

Stronen, A. V., Mattucci, F., Fabbri, E., Galaverni, M., Cocchiararo, B., Nowak, C., ... & Caniglia, R. (2022). A reduced SNP panel to trace gene flow across southern European wolf populations and detect hybridization with other Canis taxa. *Scientific Reports*, *12*(1), 4195. https://doi.org/10.1038/s41598-022-08132-0

Teixeira, H., Montade, V., Salmona, J., Metzger, J., Bremond, L., Kasper, T., ... & Radespiel, U. (2021). Past environmental changes affected lemur population dynamics prior to human impact in Madagascar. *Communications biology*, *4*(1), 1084. https://doi.org/10.1038/s42003-021-02620-1

Tiley, G. P., Barker, M. S., & Burleigh, J. G. (2018). Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome biology and evolution*, *10*(11), 2882-2898. https://doi.org/10.1093/gbe/evy200

Tiley, G.P., Crowl, A.A., Manos, P.S., Sessa, E.B., Solís-Lemus, C., Yoder, A.D., Burleigh, J.G. 2021. Phasing alleles improves network inference with allopolyploids*. bioRxiv* https://doi.org/10.1101.2021.05.04.442457

Tiley, G. P., van Elst, T., Teixeira, H., Schüßler, D., Salmona, J., Blanco, M. B., ... & Yoder, A. D. (2022). Population genomic structure in Goodman's mouse lemur reveals long-standing separation of Madagascar's Central Highlands and eastern rainforests. *Molecular ecology*, *31*(19), 4901-4918. https://doi.org/10.1111/mec.16632

Vieilledent, G., Grinand, C., Rakotomalala, F. A., Ranaivosoa, R., Rakotoarijaona, J. R., Allnutt, T. F., & Achard, F. (2018). Combining global tree cover loss data with historical national forest cover maps to look at six decades of deforestation and forest fragmentation in Madagascar. *Biological Conservation*, *222*, 189-197. https://doi.org/10.1016/j.biocon.2018.04.008

Virah-Sawmy, M., Willis, K. J., & Gillson, L. (2010). Evidence for drought and forest declines during the recent megafaunal extinctions in Madagascar. *Journal of Biogeography*, *37*(3), 506-519. https://doi.org/10.1111/j.1365-2699.2009.02203.x

Viruel, J., Conejero, M., Hidalgo, O., Pokorny, L., Powell, R. F., Forest, F., ... & Leitch, I. J. (2019). A target capture-based method to estimate ploidy from herbarium specimens. *Frontiers in Plant Science*, 937. https://doi.org/10.3389/fpls.2019.00937

Vorontsova, M. S., Besnard, G., Forest, F., Malakasi, P., Moat, J., Clayton, W. D., ... & Linder, H. P. (2016). Madagascar's grasses and grasslands: anthropogenic or natural?. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1823), 20152262. https://doi.org/10.1098/rspb.2015.2262

Wang, L., Beissinger, T. M., Lorant, A., Ross-Ibarra, C., Ross-Ibarra, J., & Hufford, M. B. (2017). The interplay of demography and selection during maize domestication and expansion. *Genome biology*, *18*(1), 1-13. https://doi.org/10.1186/s13059-017-1346-4

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., & Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC bioinformatics*, *19*(1), 1-8. https://doi.org/10.1186/s12859-018-2128-z

Xie, M., Wu, Q., Wang, J., & Jiang, T. (2016). H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics*, *32*(24), 3735-3744. https://doi.org/10.1093/bioinformatics/btw537

Yamaura, Y., Narita, A., Kusumoto, Y., Nagano, A. J., Tezuka, A., Okamoto, T., ... & Lindenmayer, D. (2019). Genomic reconstruction of 100 000-year grassland history in a forested country: population dynamics of specialist forbs. *Biology letters*, *15*(5), 20180577. https://doi.org/10.1098/rsbl.2018.0577

Yang, Z. (2014). Population demographic process. In, *Molecular Evolution: A Statistical Approach* (pp.323-325). Croyden, UK. Oxford University Press.

Yoder, A. D., Campbell, C. R., Blanco, M. B., Dos Reis, M., Ganzhorn, J. U., Goodman, S. M., ... & Weisrock, D. W. (2016). Geogenetic patterns in mouse lemurs (genus Microcebus) reveal the ghosts of Madagascar's forests past. *Proceedings of the National Academy of Sciences*, *113*(29), 8049-8056. https://doi.org/10.1073/pnas.1601081113

Figures

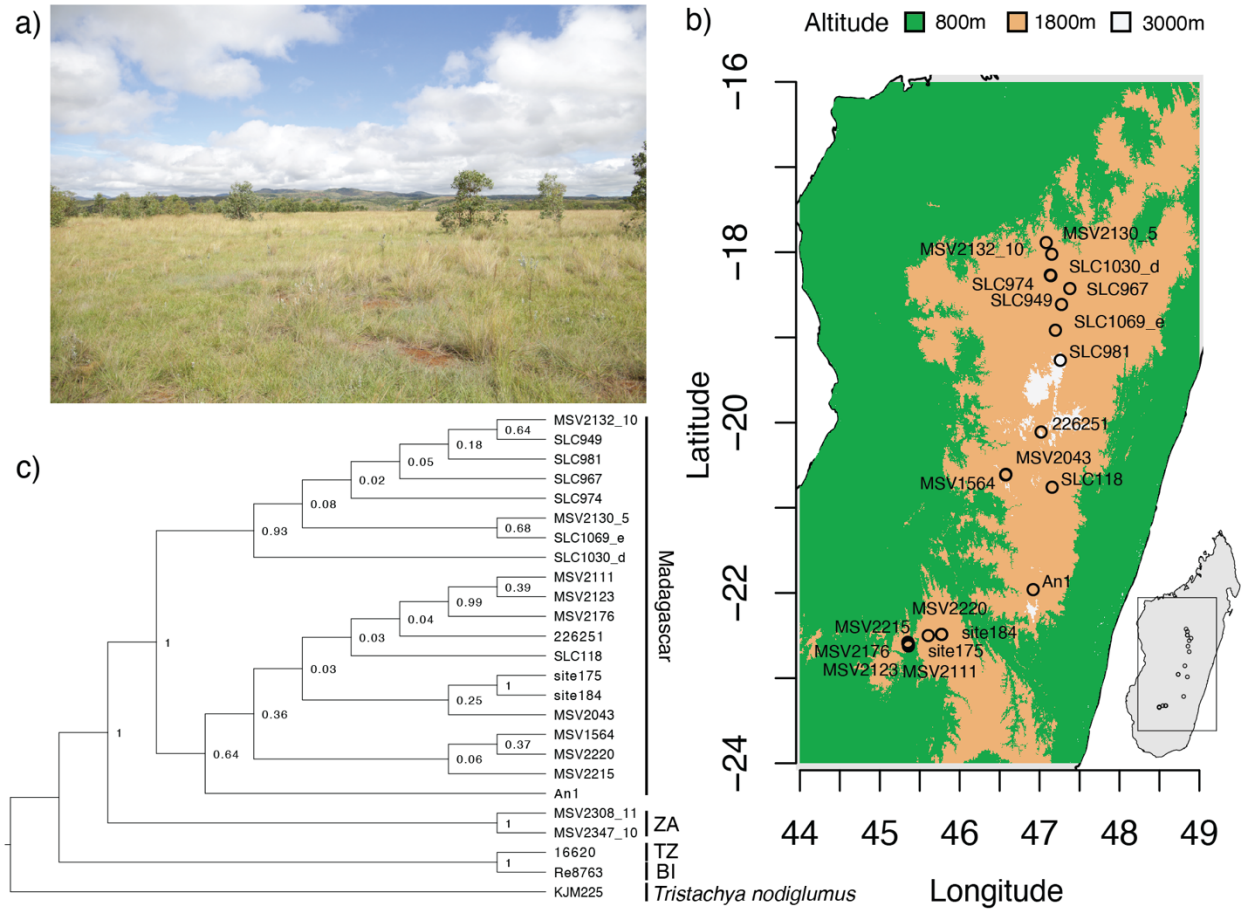

Figure 1 — *Loudetia simplex* distribution and single Madagascar origin. a) A field of *L. simplex* near Antsirabe Madagascar planted with eucalyptus (photo credit: George P. Tiley). b) Our sampling of *L. simplex* was based on collections made throughout Madagascar's Central Highland Plateau, with an elevation between 800 and 1800 meters. c) Phylogenetic relationships of individuals based on the phased (PH) data. Numbers next to nodes are posterior probabilities.

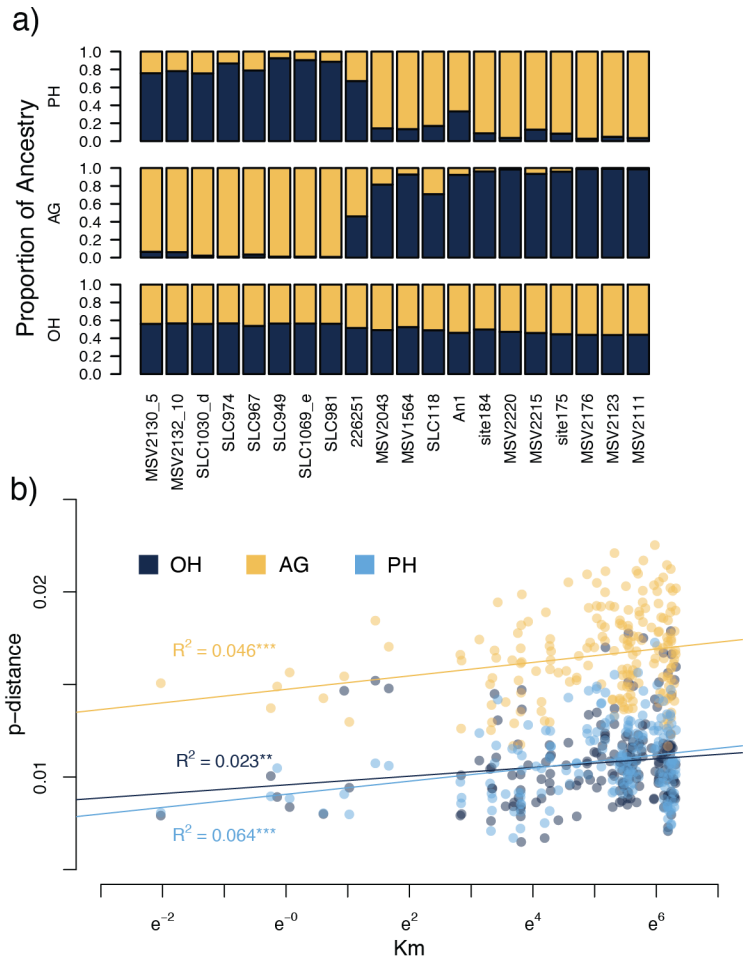

Figure 2 — Patterns of genetic structure and variation in Madagascar. a) Aggregated STRUCTURE results across runs for an *a priori* number of clusters of two (K=2) for the different data types. c) A plot of natural-log-scaled geographic distance against the pairwise distance (p-distance). Lines are from a least-squares linear regression with the coefficient of determination (R^2) shown for each data type. Statistical significance was determined with a Mantel test using 9999 permutations. Significance levels are *p*-value < 0.05 (*), *p*-value < 0.01 (**), and *p*-value < 0.001 (***). Colors represent different data types: one haplotype (OH), ambiguous genotypes (AG), and phased haplotypes (PH).

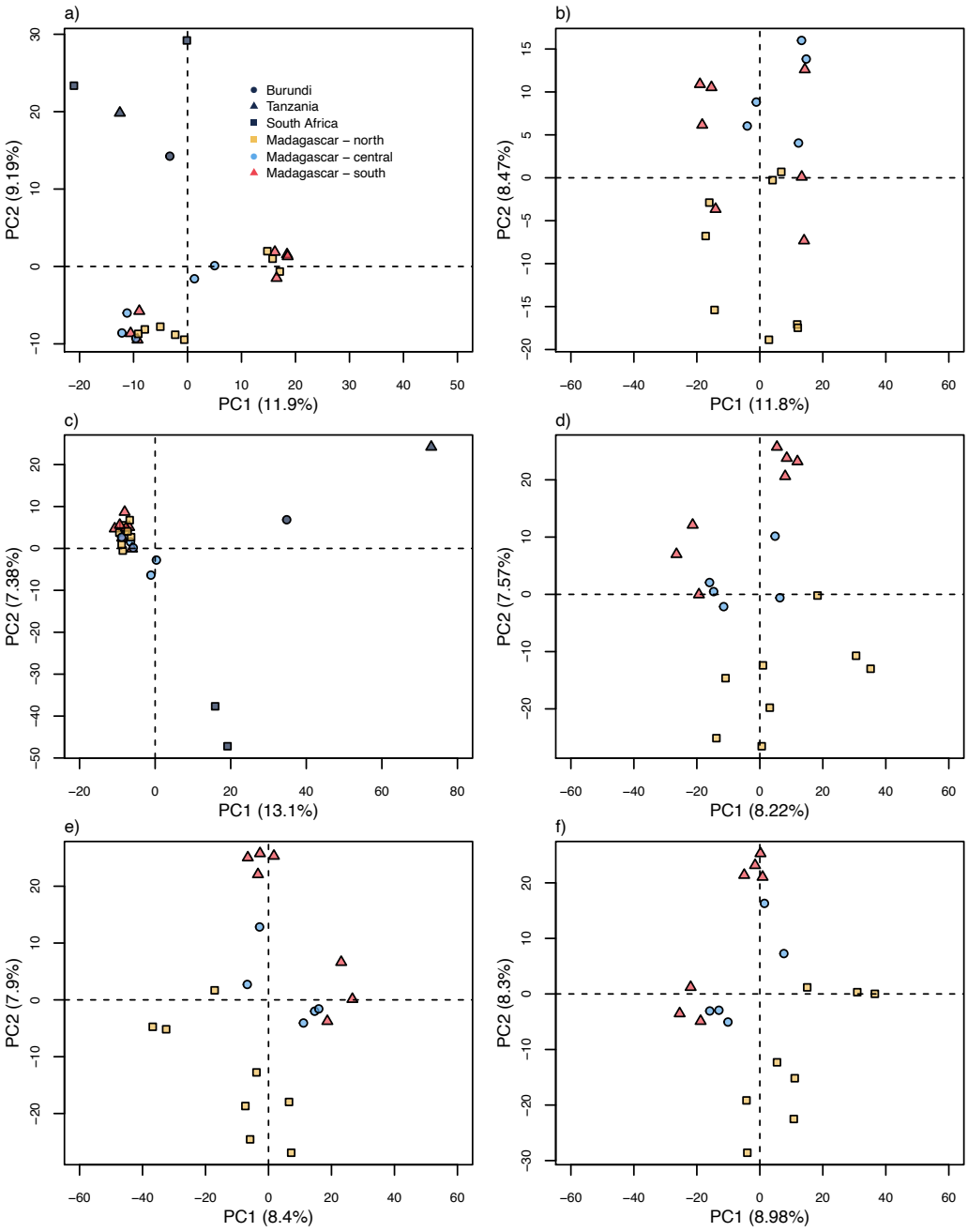

Figure 3 — Principal component analyses across data types. a) All individuals using haploid (OH) data. b) Only Malagasy individuals using OH data. c) All individuals treating all sites as diploid. d) Only Malagasy individuals treating all sites as hexaploid. e) Only Malagasy individuals compressing unbalanced allelic dosage to single categories. f) Only Malagasy individuals treating all sites as diploid.

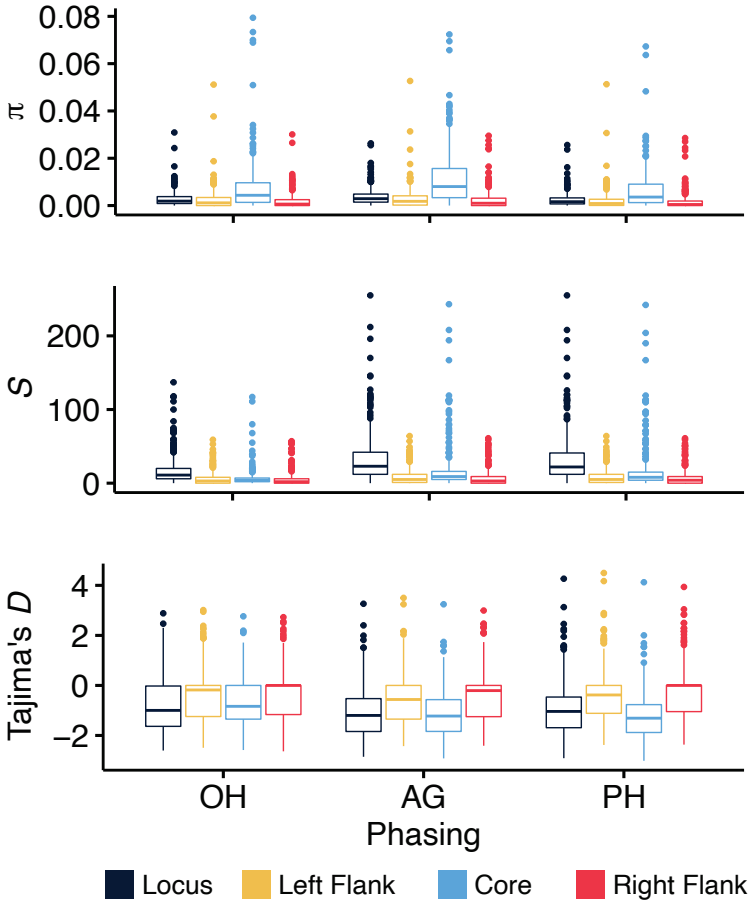

Figure 4 — Distributions of summary statistics across locus regions. Box-and-whisker plots show distribution medians and quartiles for the expected pairwise nucleotide divergence (π), the number of segregating sites (S), and a normalized measure of the difference between two estimators of nucleotide divergence, indicating deviations from neutral expectations (Tajima's D).

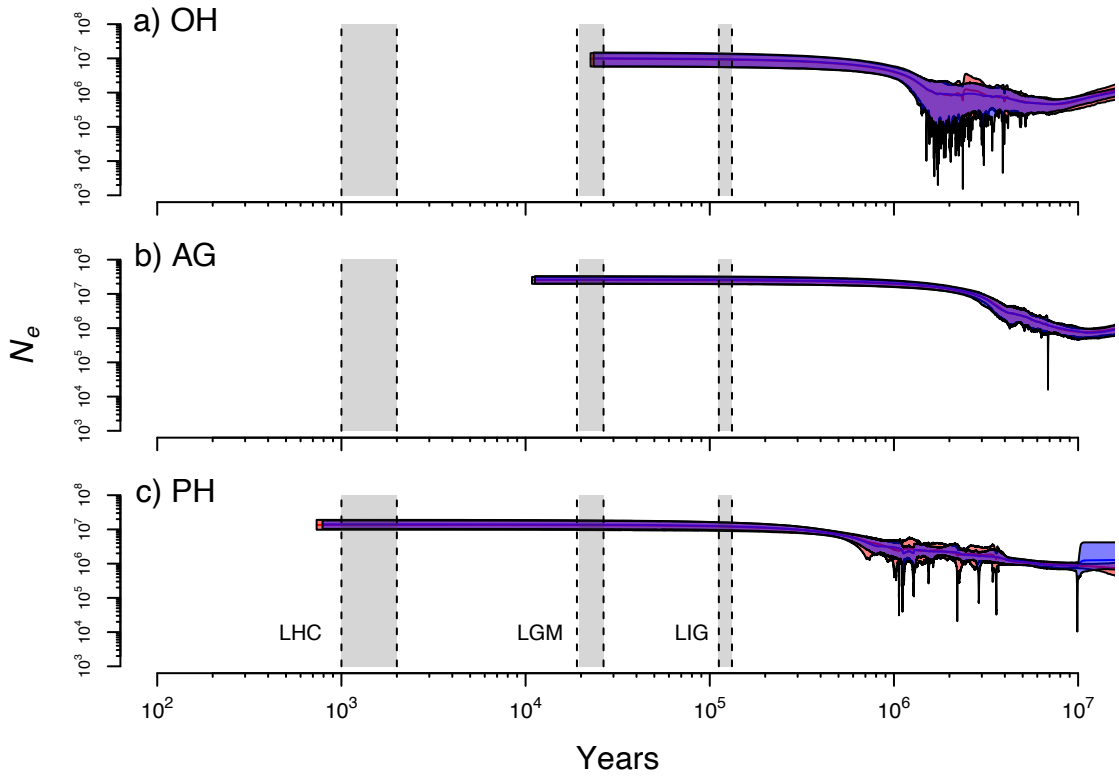

Figure 5 — Extended Bayesian Skyline Plot results with fixed clock rates. Two independent runs are shown with mean population sizes shown by solid lines. Polygons show the 95% HPD intervals with the overlap between runs in purple. Results are shown for the a) OH, b) AG, and c) PH data. Relevant time periods are shown in gray, which includes likely human colonization (LHC), the last glacial maximum (LGM), and last interglacial period (LIG).

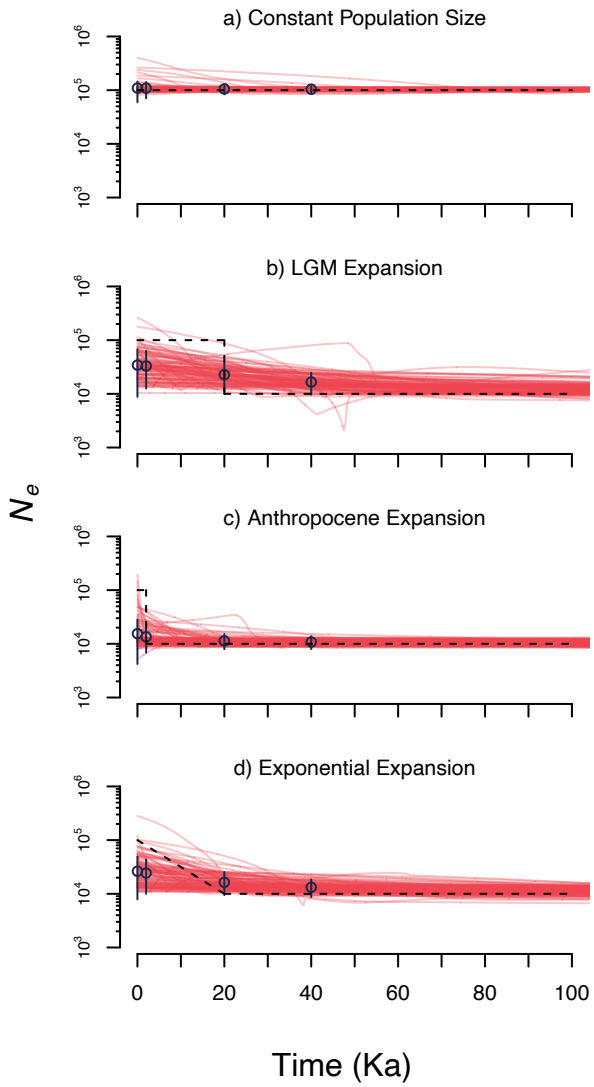

Figure 6 — Simulation results from EBSP analyses for an appropriate number of change points. Individual red lines are the mean N_e estimates from 100 simulated data sets. Blue points and vertical lines are the mean of mean N_e estimates and their 95% HPD intervals. The dashed black line is the true demographic history of the population.

Tables

Table 1 — X^2 test result from Tajima's D statistics across datasets and locus regions

Tested Effect	Fixed Effect	X^2	*p*-value	Largest Residual
Region	OH	7.72	0.27	Under-representation of Tajima's D < -2 in core regions
	AG	37.77	0.0001	Over-representation of Tajima's D < -2 in core regions
	PH	83.47	0.0001	Over-representation of Tajima's D < -2 in core regions
Dataset	Left Flank	12.65	0.046	Over-representation of Tajima's D < -2 in OH dataset
	Right Flank	8.61	0.19	Over-representation of Tajima's D < -2 in OH dataset
	Core	57.15	0.0001	Over-representation of Tajima's D < -2 in OH dataset