MOLECULAR ECOLOGY

Supplemental Information for:

The biogeographic and evolutionary processes shaping population divergence in *Laupala*

Thomas Blankers & Kerry L. Shaw

Table of Contents:

MOLECULAR ECOLOGY

SUPPLEMENTARY METHODS

SNP selection

Two SNP callers were used: FreeBayes 0.9.13 (Garrison & Marth, 2012) and GATK 3.6.0 (DePristo et al., 2011; Van der Auwera et al., 2013). FreeBayes was used using default settings. GATK Haplotypecaller was used in GVCF mode and g.vcf file were then subject to joint genotyping. FreeBayes SNPs were filtered using vcflib (Garrison, Kronenberg, Dawson, Pedersen, & Prins, 2021) with the following instructions: "TYPE = snp & QUAL > 30 & DP > 5 & SAP > 0.0001 & SRP > 0.0001". GATK SNPs were filtered using GATK VariantFiltration with the following expressions: --filterExpression "QD < 2.0 " --filterName "LowQD" --filterExpression "FS > 60.0" --filterName "StrandBiasFishers" --filterExpression "MQ < 35.0" --filterName "LowMapQual" --filterExpression "MQRankSum < -12.5". Both vcf files were then processed in VCFtools (Danecek et al., 2011) to select only bi-allelic SNPs with fewer than 10 individuals missing. The two vcf files were then checked for correspondence of genotypes using a custom script. The GATK vcf file was filtered to contain only those positions with 95% identical genotypes between the GATK and FreeBayes calls and then further reduced to have only one SNP each 50,000 bp. We always kept the SNP with the highest quality score if we had to select from two or more SNPs in a 50,000 bp window.

REFERENCES

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Group, 1000 Genomes Project Analysis. (2011). The variant call format and VCFtools. *Bioinformatics* , *27*(15), 2156–2158. Retrieved from http://bioinformatics.oxfordjournals.org/content/27/15/2156.abstract

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. Retrieved from http://dx.doi.org/10.1038/ng.806

Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2021). Vcflib and tools for processing the VCF variant call format. *BioRxiv*. doi: https://doi.org/10.1101/2021.05.21.445151

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv Preprint ArXiv:1207.3907*.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*, 1–11. doi: 10.1002/0471250953.bi1110s43

MOLECULAR ECOLOGY

TABLES (in separate excel sheet)

Table S1. Sampling locations/information. The sequencing ID, phenotypic sample name, sex, sampling population, GPS coordinates and sampling date are shown.

Table S2. Two-dimensional dadi demographic analysis model selection and parameter estimates. For each two-population model, the model statistics (number of variable sites over which the site frequency spectrum was calculated, number of parameters, composite likelihood), the model selection statistics (D_adj is two times the difference between composite likelihoods adjusted for the difference in parameters, p is the associated p-value for $D \leq 0$), and the inferred parameters for the model with the highest likelihood (mean and standard deviation shown for selected model) are shown.

Table S3. Three-dimensional dadi demographic analysis model selection and parameter estimates. For each three-population model the model statistics (number of variable sites over which the site frequency spectrum was calculated, number of parameters, composite likelihood), the model selection statistics (D is two times the difference between composite likelihoods, D_adj is that difference adjusted for the difference in parameters, p is the associated p-value for $D \leq 0$), and the inferred parameters for the model with the highest likelihood (mean and standard deviation shown for selected model) are shown.

Table S4. The relationship between genetic variation among populations (FST and dxy) and features of the genome (recombination rate and proximity to a QTL for song divergence) for all population pairs.

Table S5. The relationship between genetic diversity and features of the genome (recombination rate and proximity to a QTL for song divergence) for all populations.

Table S6. Phylogenetic independent contrasts for analyzing the relationship between time since divergence and the strength of the association between among-population genetic variation and features of the genome.

MOLECULAR ECOLOGY

FIGURES

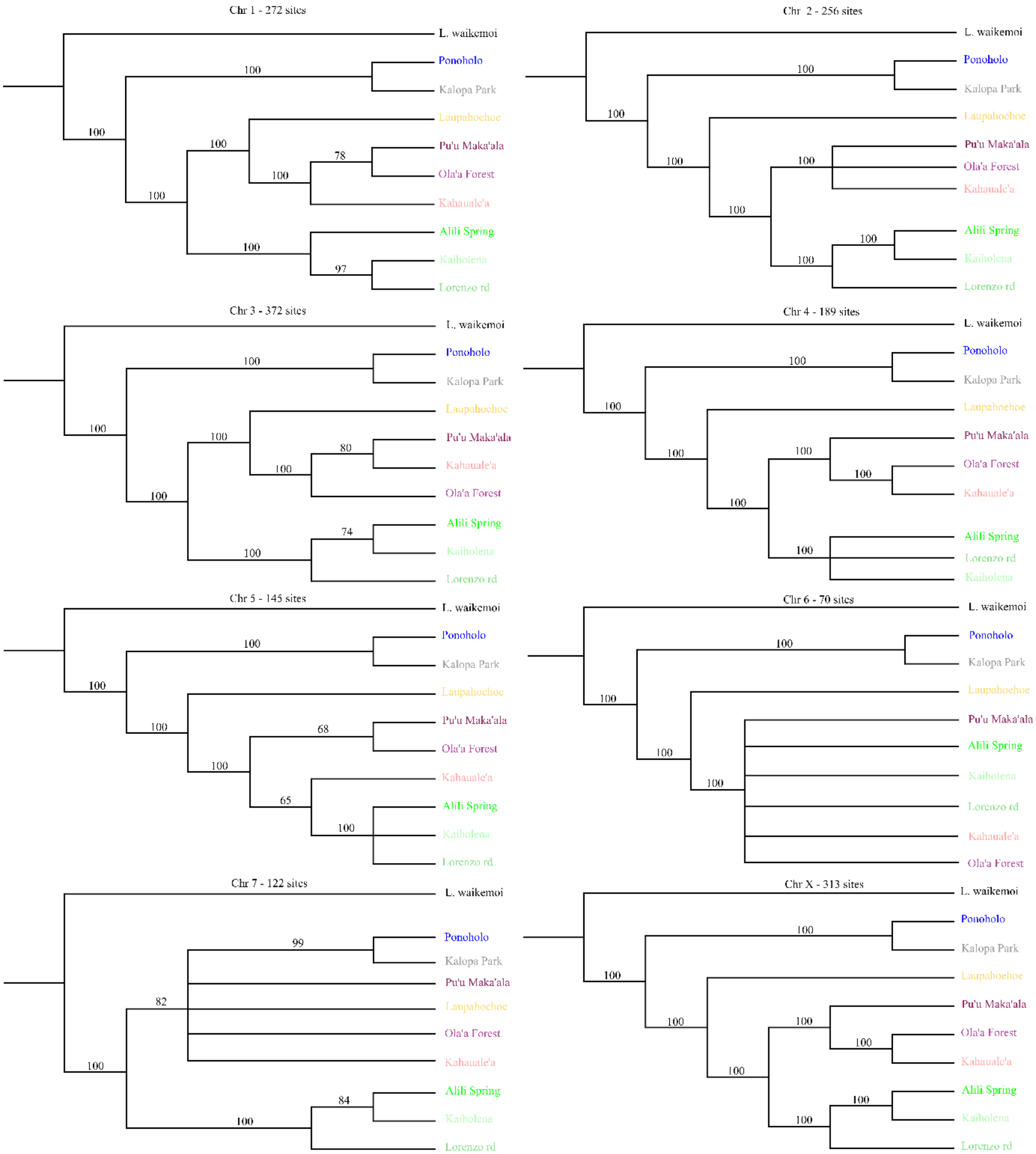


Figure S1. Chromosome specific topologies. For each of the seven autosomes and the X chromosome, a MSC tree produced in PAUP4.0 using the SVDquartets method is shown. The number of variable sites available for analysis for each chromosome is shown in the label above each tree. Bootstrap support values for branches are shown. Distal nodes of branches with bootstrap support < 50 are collapsed.

MOLECULAR ECOLOGY

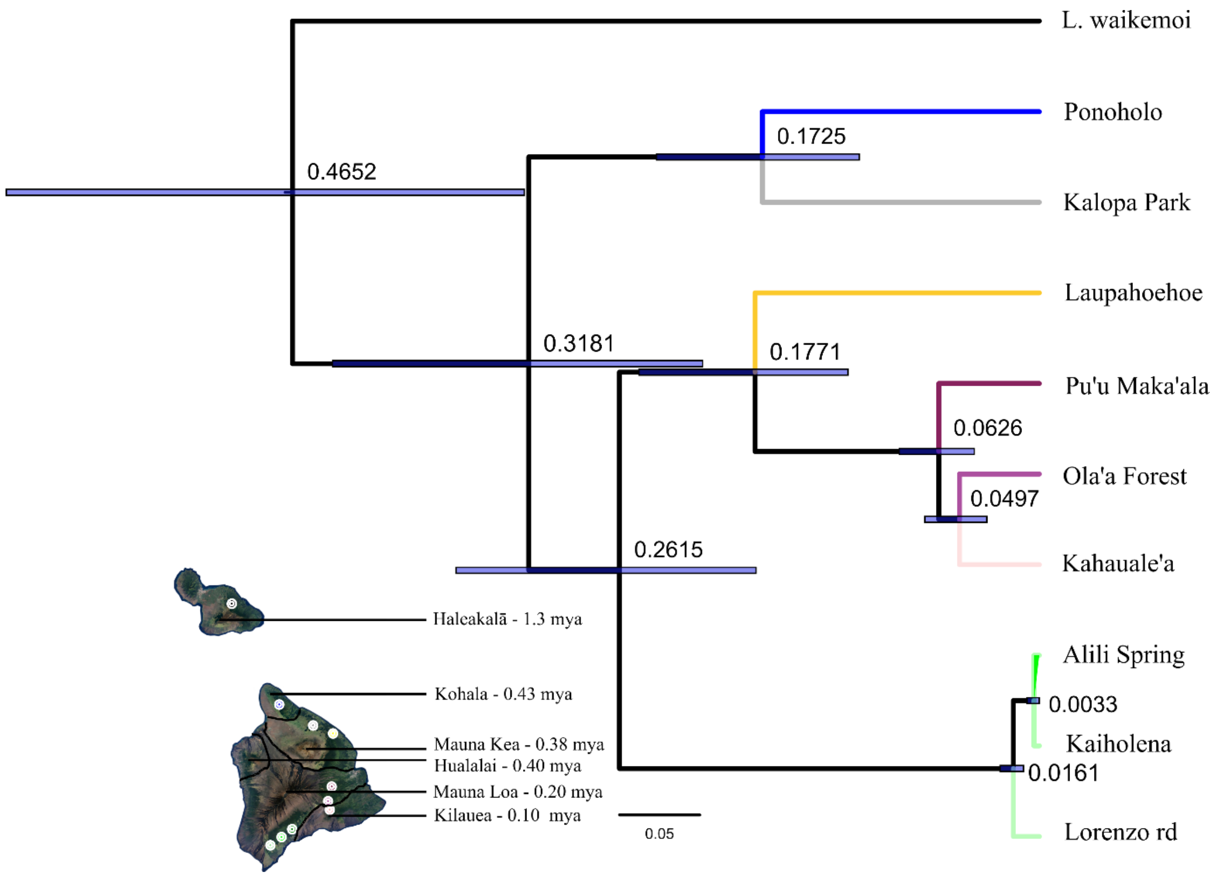
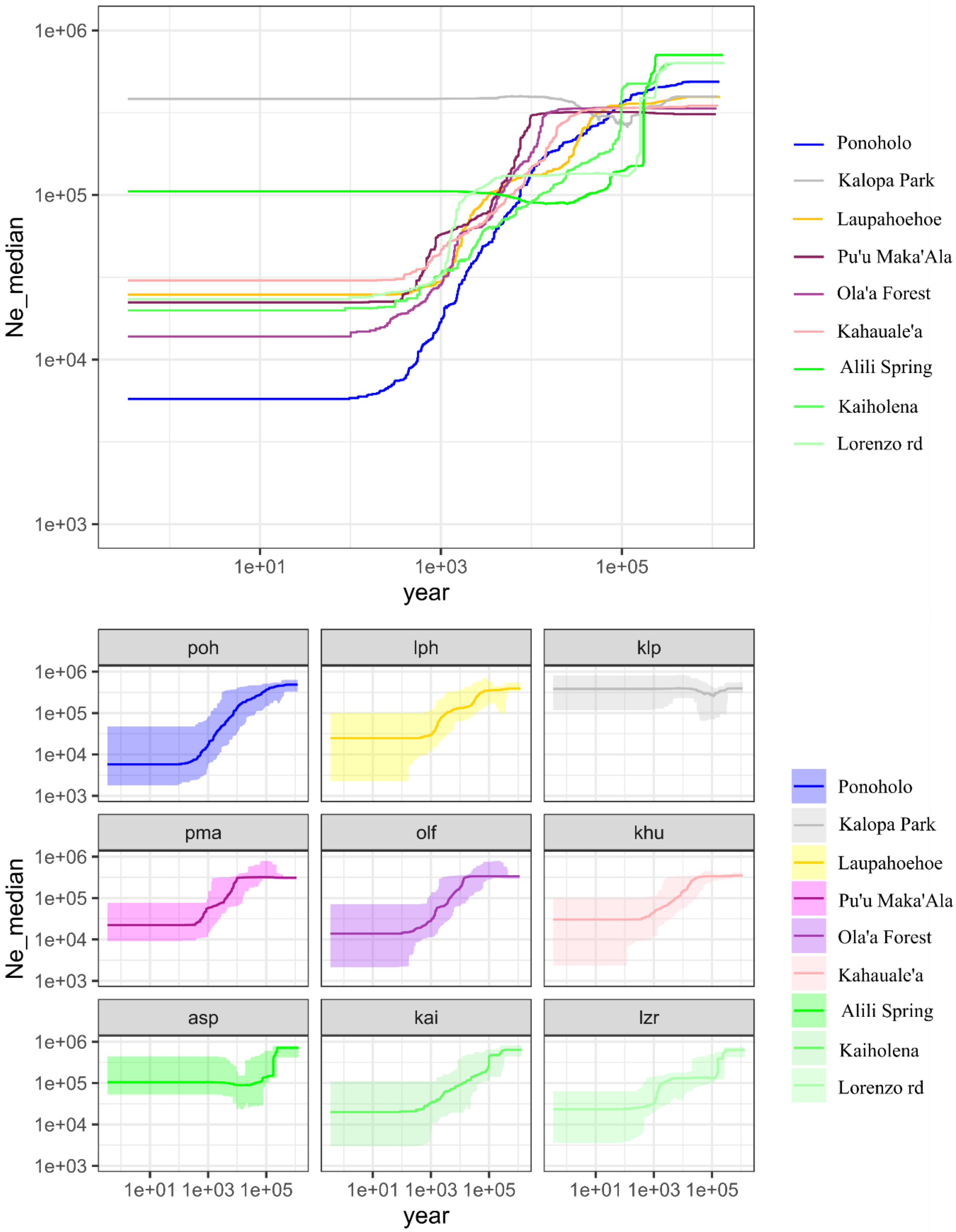


Figure S2. Time-calibrated population divergence with alternative topology (placement of Laupahoehoe relative to derived populations). Node values represent divergence time estimates in million years. The only calibration point used was the surfacing age of Hawai'i (around 0.5 mya) which translated to a log-normal prior with mean 0.5 and standard deviation 0.2 on the *L. cerasina* stem. All branches are supported by Bayes factor = 1.0. Numbers next to the node indicate mean age (in million years). Blue bars indicate 95% Highest Posterior Density. The inset map on the bottom left shows sampling locations, approximate present-day boundaries of the volcanoes, and estimated time of emergence above sea level (in million years ago [mya]]).



1e+06
1e+05
1e+04
1e+03
Ne_median
1e+01
1e+03
1e+05
year
Ponoholo
Kalopa Park
Laupahoehoe
Pu'u Maka'Ala
Ola'a Forest
Kahauale'a
Alili Spring
Kaiholena
Lorenzo rd
poh
lph
klp
pma
olf
khu
asp
kai
lzr

Figure S3. Stairway plots for effective population size (Ne) through time. Top panel shows median Ne (log-scaled y-axis) through time (log-scaled x-axis) for all nine *L. cerasina* populations. The bottom panel shows median Ne along with 95% confidence intervals for each population separately.

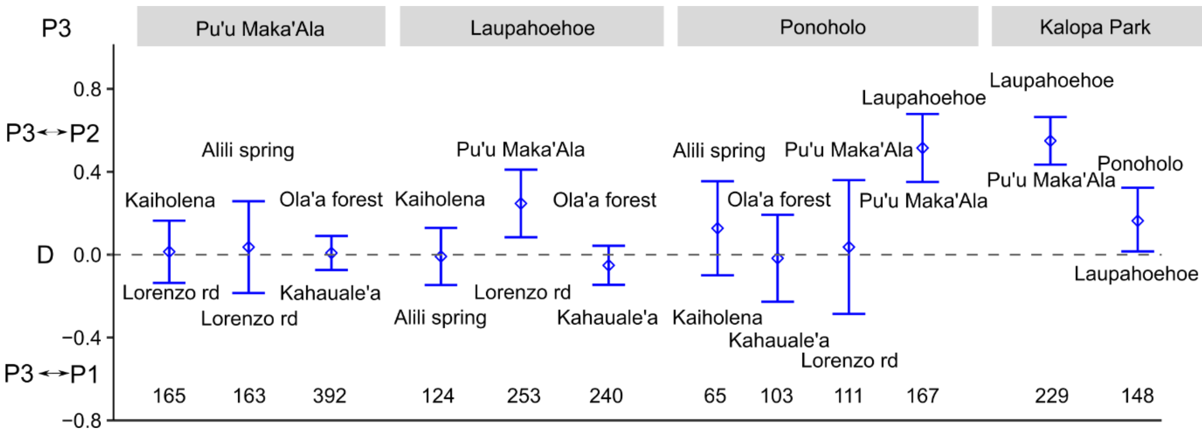


Figure S4. ABBA-BABA patterns across the phylogeny. Diamonds show whether D was positive, indicating gene flow between P3 and P2, or negative, indicating gene flow between P3 and P1. The bars show the 95% confidence interval (1.96 times the standard deviation), revealing only three significantly positive D values. Number of variable sites included in the analysis for each trio is printed along the x-axis.

Figure S4. Model fit for the four models in dadi. See also separate PDF.

MOLECULAR ECOLOGY

Model fit for the Ponoholo - Laupahoehoe - KalopaPark model with gene flow

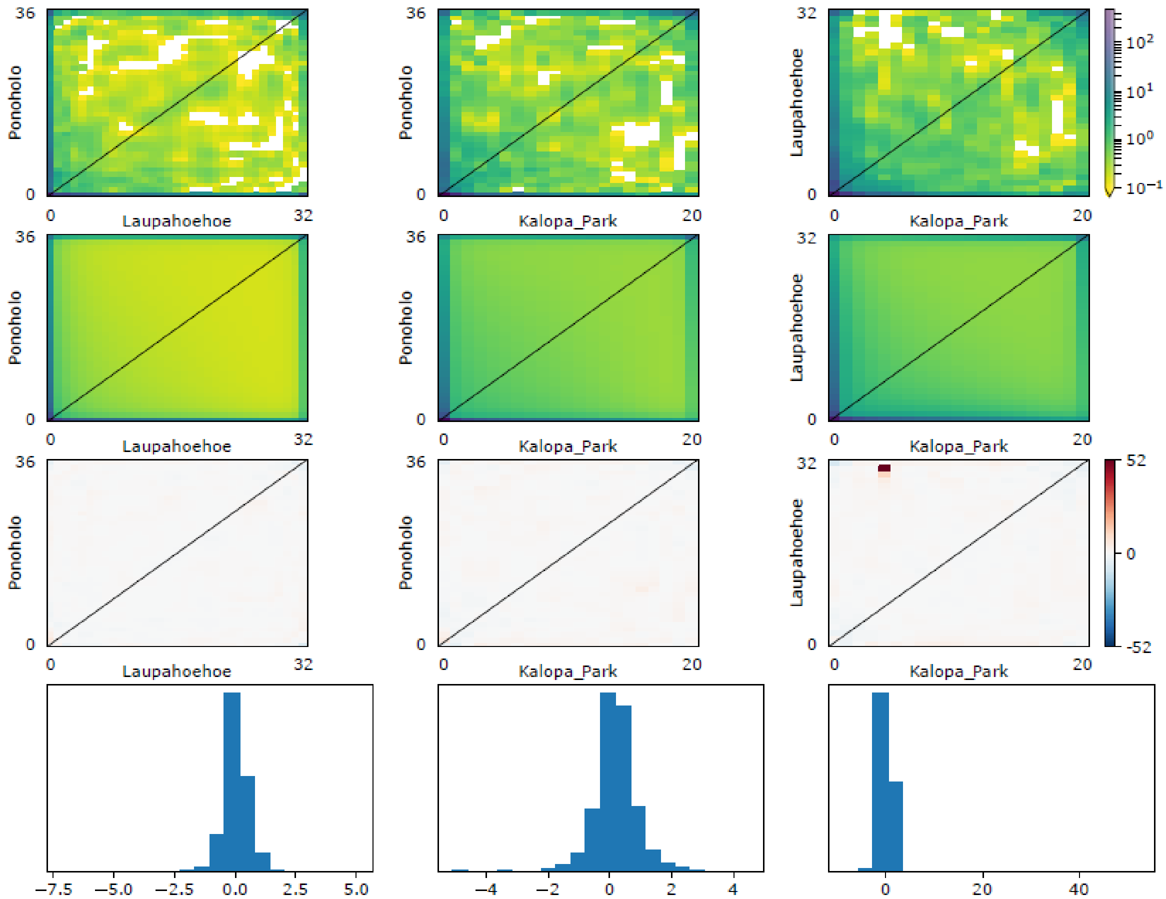


Figure S5. Ponoholo – Laupahoehoe – Kalopa Park. The two-dimensional Site Frequency Spectrum (SFS) is shown as a heatmap (top panels). Each cell represents the logarithm of the number of variants of a given frequency in either of the two populations. More purple hues indicate higher values for the logarithm of the number of variants. White cells represent counts of zero or close to zero (minimum value to plot is the logarithm of 0.01, as the logarithm of zero is minus infinity and cannot be plotted). The three panels below show the same for the model SFS with the highest fit. The bottom six panels show the deviation between the model and the data as the residuals for each cell and histograms for the residuals.

MOLECULAR ECOLOGY

Model fit for the Pu'u Maka'ala - Ola'a Forest - Kahaualeia model with no gene flow

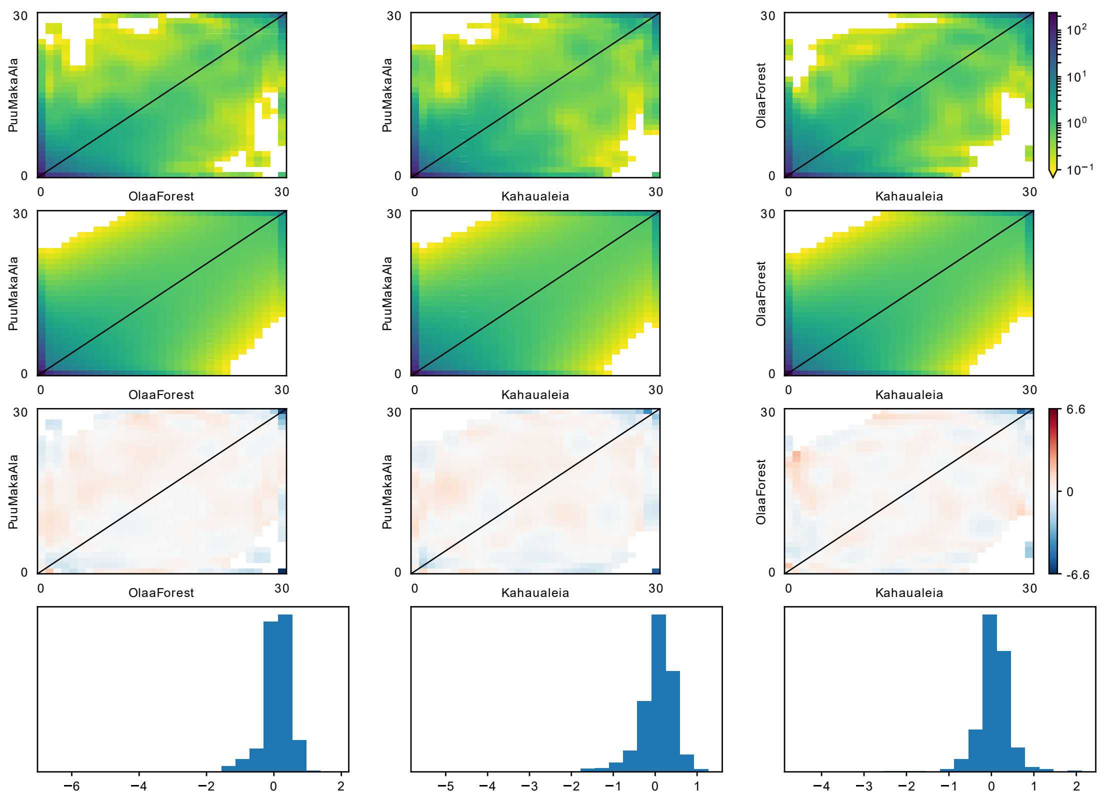


Figure S5 – continued. Pu'u Maka'ala – Ola'a Forest – Kahauale'a. The two-dimensional Site Frequency Spectrum (SFS) is shown as a heatmap (top panels). Each cell represents the logarithm of the number of variants of a given frequency in either of the two populations. More purple hues indicate higher values for the logarithm of the number of variants. White cells represent counts of zero or close to zero (minimum value to plot is the logarithm of 0.01, as the logarithm of zero is minus infinity and cannot be plotted). The three panels below show the same for the model SFS with the highest fit. The bottom six panels show the deviation between the model and the data as the residuals for each cell and histograms for the residuals.

MOLECULAR ECOLOGY

Model fit for the Laupahoehoe - Pu'u Maka'ala - Lorenzo Road model with gene flow

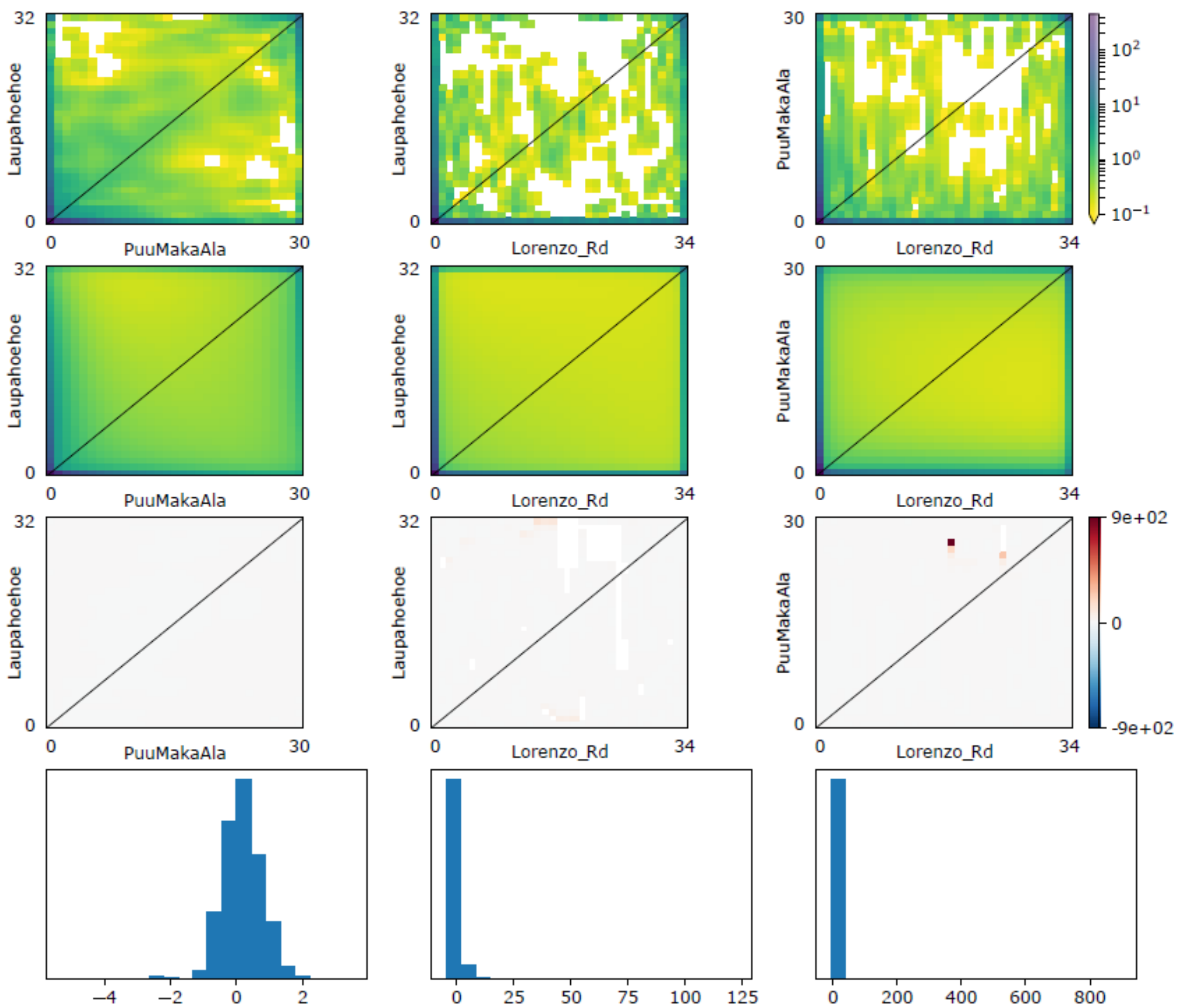


Figure S5 – continued. Laupahoehoe - Pu'u Maka'ala – Lorenzo Road. The two-dimensional Site Frequency Spectrum (SFS) is shown as a heatmap (top panels). Each cell represents the logarithm of the number of variants of a given frequency in either of the two populations. More purple hues indicate higher values for the logarithm of the number of variants. White cells represent counts of zero or close to zero (minimum value to plot is the logarithm of 0.01, as the logarithm of zero is minus infinity and cannot be plotted). The three panels below show the same for the model SFS with the highest fit. The bottom six panels show the deviation between the model and the data as the residuals for each cell and histograms for the residuals.

MOLECULAR ECOLOGY

Model fit for the Lorenzo Road - Alili Springs - Kaiholena model with no gene flow

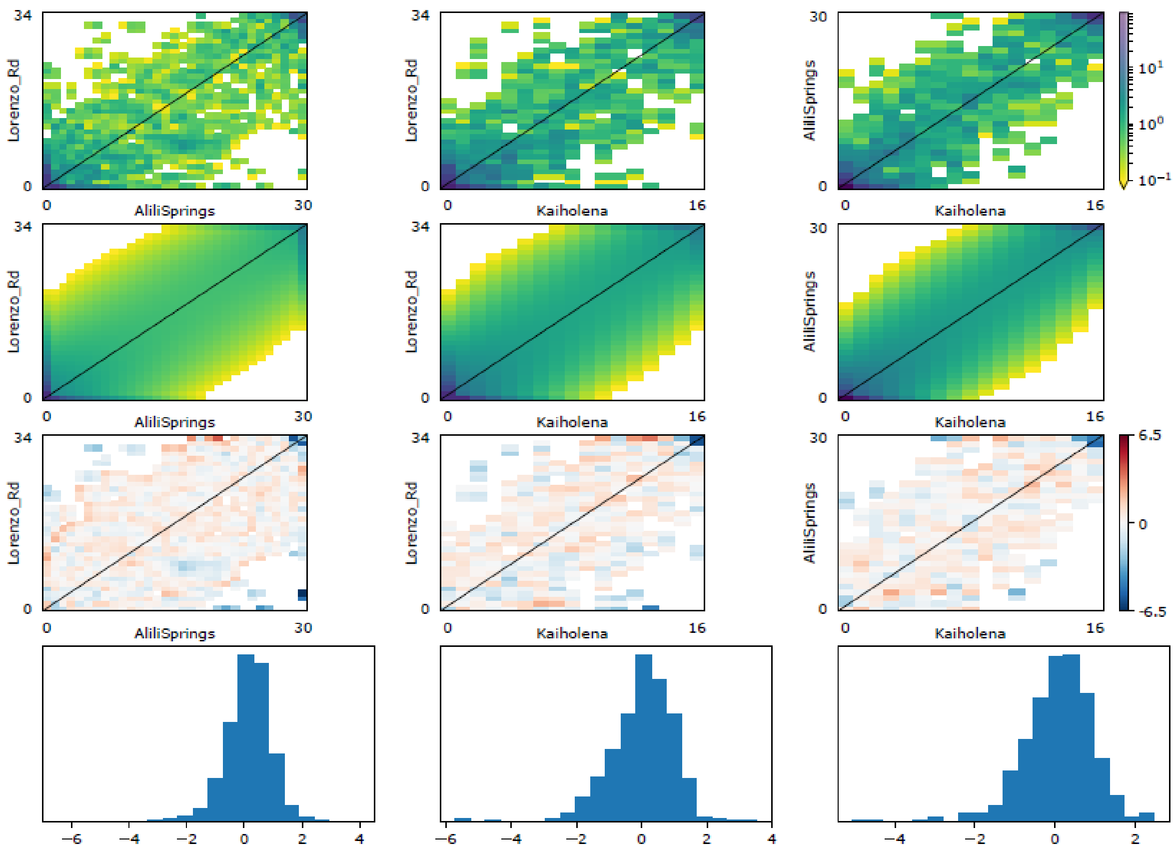


Figure S5 – continued. Lorenzo Road – Alili Springs – Kaiholena. The two-dimensional Site Frequency Spectrum (SFS) is shown as a heatmap (top panels). Each cell represents the logarithm of the number of variants of a given frequency in either of the two populations. More purple hues indicate higher values for the logarithm of the number of variants. White cells represent counts of zero or close to zero (minimum value to plot is the logarithm of 0.01, as the logarithm of zero is minus infinity and cannot be plotted). The three panels below show the same for the model SFS with the highest fit. The bottom six panels show the deviation between the model and the data as the residuals for each cell and histograms for the residuals.