

Divergent selection and primary gene flow shape incipient speciation of a riparian tree on Hawaii Island

Jae Young Choi¹, Michael Purugganan^{1,2}, and Elizabeth A. Stacy^{3*}

¹ Center for Genomics and Systems Biology, Department of Biology, 12 Waverly Place,
New York University, New York, NY USA

² Center for Genomics and Systems Biology, NYU Abu Dhabi Research Institute, New
York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

³ School of Life Sciences, University of Nevada Las Vegas, 4505 S. Maryland Parkway,
Las Vegas, Nevada USA

* Corresponding author: elizabeth.stacy@unlv.edu

1 **Abstract**

2 A long-standing goal of evolutionary biology is to understand the mechanisms
3 underlying the formation of species. Of particular interest is whether or not speciation can
4 occur in the presence of gene flow and without a period of physical isolation. Here, we
5 investigated this process within Hawaiian *Metrosideros*, a hyper-variable and highly
6 dispersible woody species complex that dominates the Hawaiian Islands in continuous
7 stands. Specifically, we investigated the origin of *Metrosideros polymorpha* var. *newellii*
8 (*newellii*), a riparian ecotype endemic to Hawaii Island that is purportedly derived from
9 the archipelago-wide *M. polymorpha* var. *glaberrima* (*glaberrima*). Disruptive selection
10 across a sharp forest-riparian ecotone contributes to the isolation of these varieties and is
11 a likely driver of *newellii*'s origin. We examined genome-wide variation of 42 trees from
12 Hawaii Island and older islands. Results revealed a split between *glaberrima* and *newellii*
13 within the past 0.3-1.2 million years. Admixture was extensive between lineages within
14 Hawaii Island and between islands, but introgression from populations on older islands
15 (*i.e.* secondary gene flow) did not appear to contribute to the emergence of *newellii*. In
16 contrast, recurrent gene flow (*i.e.* primary gene flow) between *glaberrima* and *newellii*
17 contributed to the formation of genomic islands of elevated absolute and relative
18 divergence. These regions were enriched for genes with regulatory functions as well as
19 for signals of positive selection, especially in *newellii*, consistent with divergent selection
20 underlying their formation. In sum, our results support riparian *newellii* as a rare case of
21 incipient ecological speciation with primary gene flow in trees.

23 **Author summary**

A long-standing question in evolution is whether or not new species can arise in the presence of gene flow, which is expected to inhibit the formation of reproductive isolating barriers. We investigated the genomics underlying the origin of a Hawaii Island-endemic riparian tree and purported case of incipient sympatric speciation due to disruptive selection across a sharp forest-riparian ecotone. We find extensive evidence of ongoing gene flow between the riparian tree and its closest relative along with local genomic regions resistant to admixture that likely formed through selection on genes for ecological adaptation and/or reproductive isolation. These results strongly suggest that where disruptive selection is strong, incipient speciation with gene flow is possible even in long-lived, highly dispersible trees.

Introduction

A major aim of evolutionary biology is to understand the mechanisms underlying the formation of species. Of particular interest is the formation of species from within a panmictic population [1] or between populations that co-occur within geographic “cruising range” [2], as gene flow between diverging populations is expected to erode linkage between alleles associated with differential adaptation and reproductive isolation, thus reversing speciation [3]. As a result, the “speciation-with-gene-flow” model is highly controversial [4–6]. Moreover, when diverging populations are connected by ongoing (“primary”) gene flow without any period of physical isolation (i.e., allopatry), speciation is generally not expected [3,7–9]. In fact, recent analyses enabled by the availability of population-level genome sequence data have challenged the few purported

cases of speciation with primary gene flow [10] by revealing historical periods of allopatry [11–13], thus indicating that gene flow had occurred through secondary contact [10,14]. Regardless, despite the rarity of cases, speciation-with-primary-gene-flow (or sympatric speciation [14]) is an intriguing model of speciation for evolutionary biologists, one that requires the integration of both ecology and genetics for its inference [6].

Sympatric speciation occurs predominantly through strong disruptive selection [4,15]. When incipient species inhabit different environments, divergent selection (*i.e.* selection acting in contrasting directions between populations or selection favoring phenotypes of opposite extremes [16,17]) favors locally adaptive alleles that are beneficial in the alternative ecological niches. In the ecological speciation model [18], reproductive isolation evolves between populations as a consequence of ecologically based divergent selection (but see Coyne and Orr [19] for non-ecological mechanisms). For instance, when a pair of incipient species have adapted to contrasting environments, reproductive barriers can arise in the form of immigrant inviability, where selection acts against migrants in each environment [20]. In plants, immigrant inviability is a strong isolating mechanism [21–25], suggesting that ecological speciation may be a principal driver of incipient speciation by restricting gene flow between sympatric populations through a process analogous to reinforcement [20,26]. The ecotypes that arise through divergent selection should represent the earliest stage of speciation [17]. These lineages are likely to be only partially reproductively isolated, which makes them ideal for studying the initial stages of barrier formation [27] before the onset of confounding genetic changes that accumulate once speciation is complete [28].

At the genomic level, under the ecological speciation-with-gene-flow model [27] gene flow will homogenize the genomic landscape and prevent the divergence of regions not directly under divergent selection or linked to such regions [16,29,30]. Intense interest has focused on identifying these genomic islands of divergence [31–33], since these regions should contain the genetic variants involved in speciation [32,34–41]. Any genomic region resistant to gene flow would form localized peaks of differentiation that can be readily detected through genome-wide scans of differentiation (*e.g.* F_{ST}) [29,42]. However, because differential selection rather than differential gene flow can also produce islands of divergence through genetic hitchhiking or background selection [43–45], absolute measures of differentiation such as D_{xy} have been proposed as an alternative for identifying islands of divergence [46,47]. In plants, studies of the role of ecological divergence in the formation of reproductive barriers at the genic/genomic level are few and involve just a handful of “model” species [48]. Reconstructing the past demographic histories of the diverged populations and using the combined measures of F_{ST} and D_{xy} in multiple populations can aid our understanding of the genomic changes that accompany incipient speciation [49].

Metrosideros Gaud. (Myrtaceae) is a young (~3.9 mya; [50]) woody species complex that dominates Hawaii’s native flora, comprising more than 20 vegetatively distinct [51,52] but interfertile taxa that are non-randomly distributed across ecotones and environmental gradients on each of the main islands [51,53]. Differential local adaptation across Hawaii’s heterogeneous landscape [54–61] and partial reproductive isolation [62,63] have evolved among *Metrosideros* taxa despite their long life spans [64] and the high dispersibility of their seeds and pollen [54,65,66], two characters that are expected

to impede diversification [67]. *Metrosideros* appears to be an example of incipient adaptive radiation with gene flow in trees [68].

Metrosideros polymorpha var. *newellii* (hereafter *newellii*) is an incipient species endemic to the waterways of the windward (wet) side of Hawaii Island. *Newellii* is a morphologically distinct, narrow-leaved ecotype (Fig 1A) that appears to derive from the archipelago-wide, broad-leaved *M. polymorpha* var. *glaberrima* (hereafter *glaberrima*; Fig 1A) [68] following a progenitor-derivative speciation model [69]. Riparian zones on Hawaii Island are narrow (typically the width of a single tree on either side of the waterway) and embedded within *glaberrima*-dominated wet forest. Along waterways, mating opportunities among adults of *newellii* and *glaberrima* are presumably random, consistent with a strict definition of sympatry [2,6,70–73], and apparent hybrids are occasionally found. Despite sympatry, *glaberrima* and *newellii* are surprisingly strongly isolated at SSR loci ($F_{ST} = 0.094$ (max = 0.16) relative to an average $F_{ST} = 0.054$ among all Hawaii-Island varieties except *newellii* [68]).

Riparian zones on Hawaii Island are harsh, and disruptive selection across a sharp ecotone results in partial reproductive isolation through significant reciprocal immigrant inviability [20] in adjacent forest and riparian environments [74]. Seedlings of *newellii* are better adapted than those of *glaberrima* to the strong water current and secondarily to the high-light stress of these environments [74]. Whether additional reproductive isolating barriers exist is not yet known. Regardless, the morphological and ecological divergence and strength of genetic differentiation observed between *newellii* and *glaberrima* equal or exceed those of other documented incipient species pairs or radiations (e.g., [75–78]). Although *newellii* is endemic to Hawaii Island (0.5 mya [79]),

an archipelago-wide analysis of SSR variation in Hawaiian *Metrosideros* suggested that it may have arisen from *glaberrima* on the prehistoric island of Maui nui, comprising modern-day Maui, Lanai, Molokai, and Kahoolawe (≤ 1.8 mya [79,80]).

We investigated the demographic history of *M. polymorpha* during its colonization of Hawaii Island, the relationship between *newellii* and *glaberrima*, and the genomic regions that were involved in the origin of *newellii*. We found evidence of extensive admixture between close and divergent lineages, including evidence that incipient speciation between *glaberrima* and *newellii* is consistent with a speciation-with-primary-gene-flow model. Examination of the genomic islands of divergence that formed during incipient speciation revealed evidence of positive selection on genes associated with the regulation of gene expression, consistent with the origin of *newellii* through disruptive selection across a forest-riparian ecotone. We estimate that *newellii* emerged 0.3-1.2 mya.

Results

Forty-two adults were chosen for genome sequencing, including 18 *glaberrima* and 20 *newellii* adults from Hawaii Island, two *glaberrima* adults from Molokai, and two outgroup species (*M. rugosa* and *M. tremuloides*) from Oahu (Fig 1A). Genome coverage was on average $\sim 15\times$, ranging from $1\times$ to $43\times$ per individual (S1 Table). For the majority of the samples more than 95% of the reads could be aligned to the reference *M. polymorpha* genome (S1 Table) [81]. Because the reference genome was still in a draft state with over 55,000 scaffolds, however, we restricted alignments to just scaffolds that

were greater than 100,000 bps (totaling 285 Mbp of the 304-Mbp genome assembly) to avoid analyzing potential repetitive regions in unassembled contigs.

Population structure of glaberrima and newellii on Hawaii Island. Because of the varying sequence depth among individuals, population relationships were examined using a complete probabilistic approach to account for the uncertainty in genotypes [82,83]. Polymorphic sites were pruned randomly to minimize the effect of linkage between physically close sites.

Using the ADMIXTURE method [84,85], ancestry proportions for each individual were estimated by varying the assumed number of ancestral populations (K) from 2 to 7 (S1 Fig). At K=2, individuals were largely divided into two groups representing the two varieties (Fig 1B). Principal components analysis (PCA) corroborated the K=2 ancestry result with the first component dividing the individuals by variety (Fig 1C). The phylogenetic analysis also revealed a near-complete division of individuals by variety (Fig 1D).

Genetic structure was observed within each variety as well (Fig 1). We used monophyly (Fig 1D) to identify three core groups (G_{H1} , G_{H2} , and N). G_{H1} comprised eight glaberrima trees from the Alakahi Bog on Kohala Volcano that formed a separate cluster in the principal component space (Fig 1C) and showed a unique ancestry in the ADMIXTURE analyses with higher K values (Fig 1B). G_{H2} comprised four closely spaced glaberrima trees situated alongside newellii at Kalohehewa Stream. These four individuals, along with three others at the Wailuku_A site (E70, E72, E74) were glaberrima trees sampled immediately adjacent to newellii trees along waterways (Fig

1A). Phylogenetically, G_{H2} was sister to *newellii* (Fig 1D) and in principal component space was closer to *newellii* than was G_{H1} (Fig 1C). The ancestry proportions for G_{H2} were inconsistent across K values, but the results for K=2, 3 and 4 indicated that ancestry of G_{H2} included *newellii* and a *glaberrima* population that was not related to G_{H1} .

Group N comprised 14 individuals of *newellii* that occur along the largest waterway on Hawaii Island, the Wailuku River. While individuals from the N group clustered together in principal component space (Fig 1C), there was evidence of substructure within the group. Trees from the most up-river population (Wailuku_A) were most diverged in the phylogenetic tree and formed their own cluster at higher K values (K=4 and 5; Fig 1B).

Among the individuals excluded from these three monophyletic groups were two *glaberrima* trees (E70 and E74) that were genetically similar to *glaberrima* from Molokai despite occurring alongside the most divergent *newellii* subpopulation (i.e., Wailuku_A; Fig 1). Hence, there may be further genetic structure with both varieties on Hawaii Island, but low sample sizes from these additional subgroups precluded further analysis. Lastly, two trees, H439 and H493, from northern waterways were phenotypically characterized as *newellii*, but were more closely related to the *glaberrima* group, and a single phenotypically defined *glaberrima* (E72) grouped within *newellii* (Fig 1).

Admixture between *glaberrima* and *newellii* on Hawaii Island. We limited

subsequent analyses to the G_{H1} , G_{H2} , and N populations, *glaberrima* from Molokai (individuals X83 and X36, hereafter referred as G_M), and the two outgroup species, *M. rugosa*, and *M. tremuloides*. There were 8,436,114 variable positions across all samples,

7,566,381 variable positions within *M. polymorpha*, and 6,818,011 variable positions among the samples from Hawaii Island.

Consistent with the genotype likelihood-based tree (Fig 1D), the phylogenetic relationships from computationally-called genotypes indicated a tree topology of [(((G_{H2}, N), G_{H1}), G_M), *M. rugosa*, *M. tremuloides*] (S2 Fig). Since admixture events occurring between branches are not detected in a simple bifurcating tree, we used methods that quantify, or test for evidence of, admixture occurring between lineages.

To characterize the relationships among G_{H1}, G_{H2}, and N on Hawaii Island we implemented a topology-weighting method [86]. Using one of G_M, *M. rugosa*, or *M. tremuloides* as the outgroup, we conducted sliding-window analyses and in each window quantified the taxon topology weight, which is a count of all of the unique sub-trees within which a single individual of each taxon is represented. Initially, results showed that regardless of the outgroup used, no single topology dominated the weighting (Fig 2A), suggesting that many variants are shared among G_{H1}, G_{H2}, and N. The choice of outgroup, however, did affect the majority topology. Specifically, with G_M or *M. rugosa* as the outgroup, the topology that grouped N and G_{H2} as sisters had the highest weighting (Fig 2A dotted box). In contrast, with *M. tremuloides* as the outgroup the topology that grouped N and G_{H2} as sisters had the lowest weighting, consistent with admixture between *M. tremuloides* and either G_{H2} or N.

Admixture among the three Hawaii Island populations and each of the outgroup populations was examined using treemix [87] by fitting migration edges on a bifurcating tree (Fig 2B). The initial treemix tree with zero migration recapitulated the population relationships recovered using the topology-weighting method, while the treemix graphs

with increasing numbers of migration edges showed different relationships (S3 Fig). Briefly, the first migration edge was fitted between *M. tremuloides* and the common ancestor of G_{H2} and N, consistent with the topology-weighting results. At two migration edges, there was admixture between G_{H1} and N, whereas in higher migration models the common ancestor of all three Hawaii Island populations admixed into N. At three migration edges, there was admixture between G_M and G_{H2} , while at four migration edges, an unsampled *M. rugosa*-like population was admixed with N. At five migration edges, there was additional gene flow originating from the common ancestor of the Hawaii Island populations into G_{H2} . This suggests that five migration edges may over-fit the data or that Hawaii Island *M. polymorpha* may have originated as a hybrid swarm [88–90].

The treemix results suggested that N and G_{H2} , the two sister groups from Hawaii Island, had extensive evidence of admixture with other populations from both Hawaii Island and one or more older islands. To specifically test for evidence of introgression we implemented the four-population (f_4) test [91]. All significant f_4 test results were consistent with the treemix result with four migration edges (Fig 2C) and revealed two major admixture patterns. (1) The significant f_4 test on the topology $[(G_{H2}, N); (G_{H1}, T \text{ or } G_M)]$ was consistent with evidence of admixture between N and G_{H1} , and admixture between G_{H2} and G_M . The non-significance of the topology $[(G_{H2}, N); (G_{H1}, R)]$ is likely due to N but not G_{H2} sharing alleles with both G_{H1} and *M. rugosa*. (2) The significant f_4 test on topology $[(G_{H2} \text{ or } N, G_{H1}); (R, T)]$ was consistent with *M. tremuloides* admixing with either G_{H2} and N, or their common ancestor. This also indicated that among the three

Hawaii Island groups, G_{HI} was the least affected by admixture from lineages on older islands.

Demographic history of Hawaii Island *glaberrima* and *newellii*. Initially, we focused on the demographic histories of G_{HI} and N to infer the recent splitting event that led to the formation of *newellii*. Levels of polymorphism were similar between the two groups (median $\theta_\pi = 0.0041$ and 0.0043 for G_{HI} and N, respectively), while linkage disequilibrium decreased rapidly to low levels ($r^2 < 0.2$) within 1 kbp for both populations (S4 Fig). Median F_{ST} between G_{HI} and N was 0.043 (S5 Fig).

The demographic histories of G_{HI} and N were estimated using the diffusion approximation method of dadi [92]. Twenty demographic models (S6 Fig) [93] were fit onto the observed two-dimensional joint-site-frequency spectrum (2D-SFS) between G_{HI} and N (see S2 Table for all models with their estimated parameters and log-likelihood). The top three best-fitting models for the 2D-SFS all involved an ongoing or secondary contact-based gene flow between G_{HI} and N. In addition, the demographic model with recent or ongoing gene flow was chosen as the best-fitting model regardless of whether the underlying 2D-SFS was estimated using genotype likelihood or genotype calls (see S3 Table for dadi results using genotype calls). The top three best-fitting models were then further optimized to fit the best-fitting demographic model (S4 Table). In the end the best-fitting model was one that modeled an ongoing asymmetric migration between G_{HI} and N ($\Delta AIC=18.8$ and see S7 Fig for model fit). The model estimated a higher level of gene flow from G_{HI} into N, while the effective population size of G_{HI} was slightly higher than that for N (Fig 3A).

We then tested if different proportions of the ancestral population contributed to G_{H1} and N, specifically testing whether N was derived from a smaller ancestral proportion than G_{H1} (*i.e.* an island model [94]). Seven different scenarios (S8 Fig) were tested, and all models indicated almost equal proportions of the ancestral population contributed to G_{H1} and N (S5 Table). Further, the best-fitting scenario was still less fit than the model of a simple split with continuous asymmetric gene flow ($\Delta AIC=5.42$ and S6 Table), suggesting no evidence of a vicariance-like event between G_{H1} and N.

We also used *dadi* to estimate the demography between G_{H2} and N by testing the same 20 demographic models (see S6 Fig for models and S7 Table for initial optimization results). The best-fitting model was a two-epoch model of asymmetric migration (S8 Table), and the resulting unscaled divergence time (T) almost overlapped the bootstrap confidence interval of the divergence time for G_{H1} and N.

The T between G_{H1} and N can be converted to years (*i.e.* scaled divergence time (τ)) with a mutation rate and generation time, both of which are unknown for *Metrosideros*. Hence, we estimated τ assuming a range of mutation rates (*Arabidopsis* mutation rate = 7×10^{-9} [95], *Prunus* mutation rate = 9.5×10^{-9} [96]) and generation times reasonable for these trees (Fig 3B). τ varied the most at lower mutation rates, but within previously known plant mutation rates ($> 7 \times 10^{-9}$ mutations per bp per generation) all estimated divergence times were within 1.25 million years regardless of generation time.

Genomic islands of divergence between G_{H1} and N. The genomic landscape of differentiation between G_{H1} and N was estimated through local genomic windows of F_{ST} values. F_{ST} was Z-transformed (zF_{ST}), and genomic regions with $zF_{ST} > 4$ were considered

significant outliers [35] (Fig 4A). Genomic islands with increased levels of relative divergence (F_{ST}) were identified in 237 out of 27,508 windows (merged into 147 non-overlapping windows) for a total of 2.37 Mbp. Further, compared to the genomic background, these islands had significantly elevated levels of absolute divergence (D_{xy}) (Fig 4B; Mann Whitney U [MWU] p-value = $<1e-10$). In addition, when the regions of genomic islands identified between G_{H1} and N were examined between G_{H1} and G_{H2} , D_{xy} values were also significantly elevated compared to the genomic background, indicating that islands of divergence are shared between G_{H2} and N. Between G_{H2} and N, however, D_{xy} was significantly lower in the genomic islands identified between G_{H1} and N (S5 Table).

Levels of polymorphism were significantly lower in genomic islands relative to background in both G_{H1} and N (S9 Fig; MWU p-value = $<1e-15$), suggesting that the increased F_{ST} and D_{xy} were partly caused by those regions being frequent targets of selection. We measured haplotype homozygosity (H_{12}) [97] and haplotype length (H) [98], and performed an LD-based test of selection (ω_{max}) [99], and found all selective sweep statistics were significantly elevated in genomic islands (Fig 4C; MWU p-value = $<1e-10$) for both G_{H1} and N. In addition, N seemed to have more genomic islands with evidence of a selective sweep, shown by the overall higher levels of selective sweep statistics for N relative to G_{H1} . We note that recurrent sweeps from a common ancestor are expected to lead to a decrease in D_{xy} [46]. Our observations of increased D_{xy} and evidence of selection, in contrast, suggest that islands of divergence were shaped by positive selection on divergent ancestral haplotypes predating the split between G_{H1} and N [100].

We then examined if the secondary contact between G_{HI} and/or N and outgroup populations on older Hawaiian Islands (G_M , *M. rugosa*, and *M. tremuloides*) had contributed to the formation of the genomic islands of divergence. If secondary gene flow was recurrent, the genomic islands of divergence identified between G_{HI} and N would also have elevated D_{xy} levels between G_{HI} or N and the relevant outgroup. However, no such effect was observed (S9 Table), suggesting that sufficient levels of genome-wide divergence had eroded the genomic islands in allopatric comparisons [30]. We then further examined the N group because of its extensive evidence of admixture with outgroup species, in particular *M. tremuloides*, and its incipient speciation status. We tested whether genomic islands had origins relating to *M. tremuloides* by estimating localized windows of f_4 statistics for the topology [G_{HI} ,N; G_M ,*tremuloides*]. Because G_{HI} was least affected by admixture with any of the outgroups (Fig 2), an excess of positive f_4 statistics across local windows would indicate increased admixture between N and *M. tremuloides* in the region. There was no significant difference in f_4 statistics between the genomic islands of divergence and the genomic background. In addition, we calculated the genic proportion of each window to examine if the admixture had affected coding sequences more than non-coding regions, suggesting a role for selection maintaining the introgression. No significant correlation was seen between the proportion of gene sequence per window and either positive or negative f_4 statistics (S10 Fig). These results, in the end, suggested that secondary contact with allopatric lineages did not contribute to the incipient speciation of N.

Gene ontology (GO)-enrichment analysis was conducted for the 341 genes (S10 Table) that overlapped with genomic islands of divergence. GO could be assigned to 147

genes, which were significantly enriched (hypergeometric test p-values < 0.05) for functions relating to two major categories: 1) DNA binding and transcriptional regulation (GO:0003700, GO:0140110, and GO:0003677) and 2) acetyltransferase activity (GO:0016747 and GO:0016407).

Discussion

Newellii on Hawaii Island appears to be a case of early-stage incipient speciation in a highly dispersible and long-lived tree. Based on our evidence, we argue that divergence between newellii and glaberrima is consistent with a speciation-with-primary-gene-flow model, where barriers to gene flow are developing through disruptive selection. The incipient status of newellii notwithstanding, our results place this riparian tree in the small but growing number of cases of speciation occurring despite the homogenizing effects of gene flow. This pattern is consistent with the observation of differential adaptation of newellii to the mechanical stress of rushing water and high light [74] and the greater shade tolerance of wet-forest-dominant glaberrima [55] resulting from disruptive selection across the forest-riparian ecotone.

Evidence of admixture was extensive among the sampled populations, occurring within and between different island lineages, and may be common throughout the species complex [101]. Our results are in line with the recent view of sympatric speciation in the genomic era, in which gene flow is almost ubiquitous between recently diverged sister species [14,102]. Clarifying the relative roles of primary versus secondary gene flow during speciation, however, is critical for understanding the evolutionary processes

underlying the divergence between sympatric sister lineages [14]. For instance, our genomic analyses suggested a close genetic relationship between G_{H2} and N, yet one that did not result from divergence in sympatry. N and G_{H2} shared islands of divergence, but these regions were not elevated in D_{xy} between N and G_{H2} . This result suggests that the shared islands of divergence between N and G_{H2} arose through recent shared selective sweeps [46], excluding the possibility of primary gene flow (and subsequent divergent selection) underlying their formation. Given the close proximity of N and G_{H2} , the appearance of locally adaptive alleles from N in G_{H2} may result from introgression between varieties where the forest-riparian ecotone is more diffuse and disruptive selection is less intense [74]. The recent admixture between G_{H2} and G_M is also consistent with the formation of G_{H2} through secondary contact between a G_M -like lineage and N. Our results also indicated gene flow between N and an allopatric lineage (*M. tremuloides* on Oahu). The genomic regions resulting from this secondary contact, however, were not enriched in the islands of divergence between G_{H1} and N. This result suggests that primary gene flow between G_{H1} and N, but not secondary gene flow between N and *M. tremuloides*, contributed to the formation of genomic islands of divergence. While this secondary gene flow may have been neutral during speciation, it is possible that some of the introgressed regions were utilized during local adaptation unrelated to incipient speciation [34,103–107].

The divergence time between *glaberrima* and *newellii* was estimated to be between 347-695 kya (assuming a mutation rate of 7×10^{-9} bp per site per generation and generation time of between 10 to 20 years). This time frame coincides roughly with the geological age of Hawaii Island, which formed ~500 kya [79], and is consistent with an

in-situ origin of *newellii*. Our more conservative estimates of divergence time, however, extend back to ~1.2 mya and suggest that divergence between *glaberrima* and *newellii* may have begun on the next youngest island in the chain, Maui Nui, before the colonization of Hawaii Island. While the site of origin of *newellii* remains unresolved, our results suggest that Hawaii-Island *glaberrima-newellii* may have derived from a hybrid swarm-like population (Fig 2). Harsh environments such as newly formed volcanic islands may promote colonization by hybrid progenies [108] with transgressive traits that are not seen in either parent [109,110]. A denser sampling of *M. polymorpha* on the islands of Maui Nui may distinguish whether the Hawaii Island *glaberrima-newellii* complex has a deeper ancestry (*i.e.* hybridization of multiple progenitors) or a recent ancestry (*i.e.* deriving from a single ancestral background).

Genome-wide F_{ST} levels suggested that gene flow has largely homogenized the genomes of *newellii* and *glaberrima*, but we discovered regions resistant to gene flow based on elevated absolute and relative divergence, and these genomic islands of divergence were enriched for evidence of positive selection in both varieties. Given the contrasting ecological niches of *newellii* and *glaberrima* [68,74], it seems likely that these islands of divergence are associated with ecological divergence between the varieties [111,112]. Our results further suggested that selection may have been stronger or more frequent in the island-endemic *newellii*, consistent with isolation by adaptation to Hawaii Island's harsh riparian environment. These islands of divergence were enriched for genes with functions related to transcription factor activity, suggesting that divergence in the gene expression landscape underlies the emergence of this riparian specialist. Indeed, many "speciation genes" are known to have functions relating to DNA binding and

transcriptional regulation [113,114], and global gene expression patterns of hybrids from reproductively isolated species are often mis-regulated [115]. Future work is needed to understand the changes in genetic architecture that coincide with incipient speciation of *newellii*.

Materials and Methods

Sequencing data generated from this study were deposited under NCBI project ID PRJNA534153. Codes used in the analysis can be found from github (https://github.com/cjy8709/Hawaii_Metrosideros_polymorpha) and data generated from this study can be found in dryad.

Sample genome sequencing. Samples analyzed within this study can be found in S1 Table. DNA was extracted from leaf buds and 2×100bp paired-end libraries were constructed using Nextera library kit. Libraries were sequenced on an Illumina HiSeq 2500 system.

Sequencing reads were aligned to the reference *Metrosideros polymorpha* var. *glaberrima* genome sequence from Izuno et al. [101]. FASTQ reads were aligned using the program bwa-mem version 0.7.16a-r1181 [116], and duplicate reads from PCR during the library preparation step were removed using picard version 2.9.0 (<http://broadinstitute.github.io/picard/>). Alignment statistics were reported using GATK version 3.8–0 (<https://software.broadinstitute.org/gatk/>) and samtools version 1.9.

SNP calling and genotyping. Genotype calls were executed using GATK. Alignment BAM files were used by GATK HaplotypeCaller engine with the option ‘-ERC GVCF’ to output variants as the genomic variant call format (gVCF). The gVCFs of each sample were merged together to conduct a multi-sample joint genotyping procedure using the GATK GenotypeGVCFs engine. Variants were divided into SNPs and INDELs using GATK SelectVariants engine, and variant filtration were conducted using GATK VariantFiltration engine with GATK bestpractice hard filter guidelines. All SNPs located within 5bps of an INDEL variant and any SNPs for which less than 80% of the samples were genotyped were removed.

Population relationship analysis. Initially, genetic relationships between and within the individuals were examined using genotype likelihoods under a complete probabilistic framework using ANGSD version 0.929 [83] and ngsTools [82]. For all analyses we required a minimum base and mapping quality score of 30, while requiring genotypes from a minimum of 75% of individuals per SNP site, a minimum population read coverage of one-third the average population read depth, and a maximum population read coverage of three times the average population read depth. Because no adequate outgroup species genome sequence is available for *M. polymorpha* we implemented a folded site frequency spectrum. To minimize the effect of linkage on the population relation inferences we randomly pruned the polymorphic sites. With a sliding window of size 10,000 bp, a polymorphic site was randomly chosen, while requiring a minimum distance of 5,000 bps between adjacent SNP sites.

Ancestry proportions (K) were estimated with NGSadmix [85] using genotype likelihoods calculated from ANGSD. K was estimated from 2 to 7, and for each K the analysis was repeated 100 times to choose the run with the highest log-likelihood. Genotype posterior probabilities (GPP) were calculated with ANGSD for use in principal components analysis (PCA) and phylogeny reconstruction; GPP were translated to genetic distances between individuals using ngsCovar [82] for PCA and NGSdist [117] for genetic distances. Using the genetic distances, FastME ver. 2.1.5 [118] was used to reconstruct a neighbor-joining tree, which was visualized using iTOL version 3.4.3 (<http://itol.embl.de/>). Bootstrap resampling of the GPP was conducted using NGSdist generating a 1,000-bootstrap resampled dataset. Each resampled data set was used to estimate genetic distances, which were then used to estimate the bootstrap confidence level for the phylogenetic tree.

Genotype calls from Hawaii Island individuals plus the outgroups, *M. rugosa* and *M. tremuloides*, from Oahu were used to generate a neighbor-joining tree (S2 Fig). Genetic distance between two individuals (X and Y) was estimated using the Kronecker delta function-based equation [119]:

$$d(X,Y) = \frac{1}{L} \sum_{i=1}^L \left[1 - \frac{1}{2} \max(\delta_{a_i c_i} + \delta_{b_i d_i}, \delta_{a_i d_i} + \delta_{b_i c_i}) \right]$$

where L is the number of sites, a_i and b_i are the two allele copies in sample X, c_i and d_i are the two allele copies in sample Y, and δ_{jk} is the Kronecker delta function (equals one if allele j is identical to allele k and 0 otherwise). From the genetic distance matrix, FastME was used to build the neighbor-joining tree.

Tests for admixture. The local genomic topological relationship was examined using the topological weighting procedure implemented in twisst [86] (<https://github.com/simonhmartin/twiss>). Initially, local phylogenetic trees were estimated in sliding windows with a size of 100 polymorphic sites using RAxML with the GTRCAT substitution model. The script `raxml_sliding_windows.py` from the `genomics_general` package by Simon Martin (https://github.com/simonhmartin/genomics_general/tree/master/phylo) was used. We then used the ‘complete’ option of twisst to calculate the exact weighting of each local window by considering all subtrees.

Treemix version 1.13 [87] was used to fit migration edges on a bifurcating tree. We thinned the polymorphism data by randomly sampling a SNP every 1 kbp using plink version 2.0 to minimize the effects of LD between SNPs. The four-population test [91] was conducted using the fourpop program from the Treemix package.

Demography modeling with dadi. Demography models were tested using the diffusion approximation method of dadi [92]. A visual representation of the 27 demographic models we tested can be found in S6 Fig and S8 Fig. Because we lacked an appropriate outgroup, folded 2-dimensional site frequency spectra (2D-SFS) were analyzed. 2D-SFS were generated using two different approaches: 1) using ANGSD to calculate the likelihood of the sample allele frequency followed by realSFS from the ANGSD package to calculate the expected number of sites within a given allele frequency based on that likelihood, and 2) using the genotype calls to calculate the allele frequency. With ANGSD the random polymorphic sites that were chosen in the “Population relationship

analysis” were used to calculate the site allele frequency. With the genotype call dataset we randomly sampled a SNP every 10 kbp using plink and used this thinned variant call dataset. The script easySFS.py (<https://github.com/isaacovercast/easySFS>) was then used to generate the site frequency spectrum and choose the best sample size to project down and maximize the number of segregating sites to be analyzed.

Optimization of the model parameters was performed through 4 rounds of randomly perturbing the parameter values using the Nelder-Mead method. In the first round, random starting parameters were three-fold perturbed for a total of 10 replicates with a maximum of 3 iterations. Each optimized parameter was then used to simulate the 2D-SFS, and a multinomial approach was used to compare and estimate the log-likelihood of the observed 2D-SFS. Best-scoring likelihood was used in the second round for 20 replicates with parameters perturbed two fold and maximum iterations of 5. The best likelihood model of the previous round was used in a third round for 30 replicates with parameters perturbed two-fold and maximum iterations of 10. In the final round the best likelihood model of the third round was used for 40 replicates, with parameters perturbed one-fold, and maximum iterations of 15. Parameters from the round with the highest likelihood were selected to represent each demographic model. Akaike Information Criteria (AIC) values were used to compare demography models. The three best-fitting models were further analyzed with increased replicates and maximum iteration values. Specifically, with increasing rounds the replicates were 20, 40, 50, and 70; with maximum iterations of 10, 10, 25, and 50; and 3, 2, 2, and 1 fold parameter perturbations. The demography model with the lowest AIC was chosen as the best-fitting model.

Because ANGSD-based and genotype call-based site frequency spectra gave largely concordant results, all further analysis used only the ANGSD-based site frequency spectrum. Confidence of the best-fitting demography model parameter values was obtained through bootstrapping the site frequency spectrum and re-estimating the demography parameter values using the bootstrap data. The bootstrap site frequency spectrum was obtained using the realSFS program and randomly selected polymorphic sites by chromosome.

Divergence time between G_{HI} and N was a main parameter of interest, and the unscaled divergence times, T , were converted into years using the following equation:

$$Divergence\ Time\ (Year) = 2 \times N_{anc} \times T \times g$$

where N_{anc} is the ancestral population size and g is the generation time in years. N_{anc} is unknown but can be inferred from the effective mutation rate θ ($4N_{anc}\mu L$, where μ is the mutation rate and L is the total length of the sequenced region that was examined to analyze the SNPs), which was calculated by *dadi* in the optimal demographic model. Here, the previous equation can be rewritten as:

$$Divergence\ Time\ (Year) = 2 \times \frac{\theta}{4\mu L} \times T \times g$$

For L , ANGSD had analyzed 280,279,511 positions to detect polymorphic sites in 2,261,660 positions. But because we pruned the sites to 28,742 positions L was calculated as $280,279,511 \times (28,742/2,261,660)$. Because μ and g for *M. polymorpha* are unknown, various biologically plausible values were used to estimate the divergence time.

Population genetic analysis. The gVCFs from the SNP-calling step were used again for the GATK GenotypeGVCFs engine but to call genotypes for all sites including the

nonvariant positions (option ‘--includeNonVariantSites’). This was done in order to correctly infer the number of variable and invariant sites for downstream calculation of window-based population genetics statistics. The `genomics_general` package was used to calculate θ , D_{xy} , and F_{ST} in 10-kbp sliding windows and 5-kbp increments. Within each window sites were required to have a minimum quality score of 30 and a minimum depth of $5\times$. Only windows that had a minimum of 3 kbp of sites that passed the quality filter were further analyzed. For the calculation of sliding windows of f_4 statistics, we used the script from Richards and Martin [106] and analyzed 10-kbp windows that had a minimum of 50 polymorphic sites.

Using the genotype calls, `plink` was used to calculate linkage disequilibrium (LD). LD was measured in squared correlations (r^2) between polymorphic sites that were less than 3 kbp apart.

Evidence of selective sweeps was detected using haplotype-based or LD-based methods [97–99]. These approaches were taken because with no appropriate outgroup sequence it was not possible to polarize the variants and infer the high-frequency alleles, which are necessary for many allele frequency-based methods of detecting selective sweeps. The focal populations G_{H1} and N were phased and imputed using the program `beagle` ver. 5.0 [120]. Sweeps were detected using the haplotype homozygosity (H_{12}), haplotype length (H), and LD-based tests of selection (ω_{max}).

For the calculation of ω_{max} statistics, the total number of grids varied across scaffolds [121]. Grids were set so that ω_{max} would be calculated every 5,000 bp, and for each grid the minimum and maximum window size was set at 5,000 bp and 100,000 bp respectively. The H_{12} statistic was calculated by setting the analysis window size to 100

SNPs and sliding the analysis window by 10 SNPs. The H statistics were estimated for every SNP position. For any given window, statistics were averaged across all haplotype-based tests of selection to represent the selective sweep value for that window.

Gene ontology enrichment. Coding sequences of each gene model were assigned a computationally predicted function and gene ontology using the eggnoG pipeline [122]. An ontology could be assigned to 12,448 genes out of the total 37,956 gene models. We required an ontology to have more than 1 gene group member for further consideration. Gene ontology enrichment was tested through a hypergeometric test.

Multiple testing corrections. All statistical tests with a p-value (Mann Whitney U test, Pearson's correlation, f_4 test, jack-knife bootstrap p-value, and hypergeometric test of gene ontology enrichment) were pooled together and corrected for multiple hypotheses testing using the Benjamini and Hochberg [123] correction method.

Acknowledgements

We thank the Hawaii Department of Forestry and Wildlife for permission to collect leaf samples from state forests and J. Johansen for assistance with sample collection. We are grateful also to T. Sakishima, A. Veillet, and the Core Genetics Facility at the University of Hawaii Hilo for assistance with DNA isolation. We are also grateful to S. Ferrand at New York University Abu Dhabi (NYUAD) for assistance with the library preparation, and M. Gros-Balthazard at NYUAD for support with the genomic data. We thank E. Richards at University of North Carolina Chapel Hill for help with the

analysis. We also thank the NYUAD Center for Genomics and Systems Biology for sequencing support and the New York University IT High Performance Computing for supplying the computational resources, services, and staff expertise.

References

1. Mayr E. Animal species and evolution. Harvard University Press; 1963.
2. Mallet J, Meyer A, Nosil P, Feder JL. Space, sympatry and speciation. *J Evol Biol.* 2009;22: 2332–2341.
3. Turelli M, Barton NH, Coyne JA. Theory and speciation. *Trends Ecol Evol.* 2001;16: 330–343.
4. Bolnick DI, Fitzpatrick BM. Sympatric Speciation: Models and Empirical Evidence. *Annu Rev Ecol Evol Syst.* 2007;38: 459–487.
5. Nosil P. Speciation with gene flow could be common. *Mol Ecol.* 2008;17: 2103–2106.
6. Fitzpatrick BM, Fordyce JA, Gavrilets S. What, if anything, is sympatric speciation? *J Evol Biol.* 2008;21: 1452–1459.
7. Felsenstein J. Skepticism Towards Santa Rosalia, or Why are There so Few Kinds of Animals? *Evolution.* 1981;35: 124–138.
8. Kondrashov AS, Kondrashov FA. Interactions among quantitative traits in the course of sympatric speciation. *Nature.* 1999;400: 351–354.
9. Dieckmann U, Doebeli M. On the origin of species by sympatric speciation. *Nature.* 1999;400: 354–357.
10. Foote AD. Sympatric Speciation in the Genomic Era. *Trends Ecol Evol.* 2018;33: 85–95.
11. Alcaide M, Scordato ESC, Price TD, Irwin DE. Genomic divergence in a ring species complex. *Nature.* 2014;511: 83–85.
12. Martin CH, Cutler JS, Friel JP, Denning Touokong C, Coop G, Wainwright PC. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution.* 2015;69: 1406–1422.

- 602 13. Malinsky M, Svoldal H, Tyers AM, Miska EA, Genner MJ, Turner GF, et al. Whole-
603 genome sequences of Malawi cichlids reveal multiple radiations
604 interconnected by gene flow. *Nat Ecol Evol.* 2018;2: 1940–1955.
- 605 14. Richards E, Servedio M, Martin C. Searching for sympatric speciation in the
606 genomic era. *bioRxiv.* 2019; 367623.
- 607 15. Kirkpatrick M, Ravigné V. Speciation by Natural and Sexual Selection: Models
608 and Experiments. *Am Nat.* 2002;159: S22–S35.
- 609 16. Via S. Divergence hitchhiking and the spread of genomic isolation during
610 ecological speciation-with-gene-flow. *Philos Trans R Soc B Biol Sci.* 2012;367:
611 451–460.
- 612 17. Rundle HD, Nosil P. Ecological speciation. *Ecol Lett.* 2005;8: 336–352.
- 613 18. Schluter D. *The Ecology of Adaptive Radiation.* Oxford University Press; 2000.
- 614 19. Coyne J, Orr H. *Speciation.* Sutherland; 2004.
- 615 20. Nosil P, Vines TH, Funk DJ. Perspective: Reproductive isolation caused by natural
616 selection against immigrants from divergent habitats. *Evolution.* 2005;59:
617 705–719.
- 618 21. Lowry DB, Modliszewski JL, Wright KM, Wu CA, Willis JH. The strength and
619 genetic basis of reproductive isolating barriers in flowering plants. *Philos*
620 *Trans R Soc B Biol Sci.* 2008;363: 3009–3021.
- 621 22. Lowry DB, Rockwood RC, Willis JH. Ecological reproductive isolation of coast
622 and inland races of *Mimulus guttatus*. *Evolution.* 2008;62: 2196–2214.
- 623 23. Melo MC, Grealy A, Brittain B, Walter GM, Ortiz-Barrientos D. Strong extrinsic
624 reproductive isolation between parapatric populations of an Australian
625 groundsel. *New Phytol.* 2014;203: 323–334.
- 626 24. Westberg E, Poppendieck H-H, Kadereit JW. Ecological differentiation and
627 reproductive isolation of two closely related sympatric species of *Oenanthe*
628 (*Apiaceae*). *Biol J Linn Soc.* 2010;101: 526–535.
- 629 25. Baack E, Melo MC, Rieseberg LH, Ortiz-Barrientos D. The origins of reproductive
630 isolation in plants. *New Phytol.* 2015;207: 968–984.
- 631 26. Servedio MR, Noor MAF. The Role of Reinforcement in Speciation: Theory and
632 Data. *Annu Rev Ecol Evol Syst.* 2003;34: 339–364.
- 633 27. Via S. Natural selection in action during speciation. *Proc Natl Acad Sci.* 2009;106:
634 9939–9946.

- 635 28. Matute DR, Butler IA, Turissini DA, Coyne JA. A Test of the Snowball Theory for
636 the Rate of Evolution of Hybrid Incompatibilities. *Science*. 2010;329: 1518–
637 1521.
- 638 29. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al.
639 Genomics and the origin of species. *Nat Rev Genet*. 2014;15: 176–192.
- 640 30. Feder JL, Egan SP, Nosil P. The genomics of speciation-with-gene-flow. *Trends*
641 *Genet*. 2012;28: 342–350.
- 642 31. Nosil P, Harmon LJ, Seehausen O. Ecological explanations for (incomplete)
643 speciation. *Trends Ecol Evol*. 2009;24: 145–156.
- 644 32. Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in *Anopheles*
645 *gambiae*. *PLoS Biol*. 2005;3: e285.
- 646 33. Harr B. Genomic islands of differentiation between house mouse subspecies.
647 *Genome Res*. 2006;16: 730–737.
- 648 34. Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, et al. Genomic
649 islands of speciation separate cichlid ecomorphs in an East African crater lake.
650 *Science*. 2015;350: 1493–1498.
- 651 35. Han F, Lamichhaney S, Grant BR, Grant PR, Andersson L, Webster MT. Gene flow,
652 ancient polymorphism, and ecological adaptation shape the genomic
653 landscape of divergence among Darwin’s finches. *Genome Res*. 2017;27:
654 1004–1015.
- 655 36. Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, et al.
656 Genomic islands of divergence in hybridizing *Heliconius* butterflies identified
657 by large-scale targeted sequencing. *Philos Trans R Soc Lond B Biol Sci*.
658 2012;367: 343–353.
- 659 37. Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, et al. Genomic islands
660 of divergence are not affected by geography of speciation in sunflowers. *Nat*
661 *Commun*. 2013;4: 1827.
- 662 38. Carneiro M, Albert FW, Afonso S, Pereira RJ, Burbano H, Campos R, et al. The
663 genomic architecture of population divergence between subspecies of the
664 European rabbit. *PLoS Genet*. 2014;10: e1003519.
- 665 39. Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Müller I, et al. The genomic
666 landscape underlying phenotypic integrity in the face of gene flow in crows.
667 *Science*. 2014;344: 1410–1414.

- 668 40. Marques DA, Lucek K, Meier JI, Mwaiko S, Wagner CE, Excoffier L, et al. Genomics
669 of Rapid Incipient Speciation in Sympatric Threespine Stickleback. *PLoS*
670 *Genet.* 2016;12: e1005887.
- 671 41. Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, Farkas TE, et al.
672 Transitions between phases of genomic differentiation during stick-insect
673 speciation. *Nat Ecol Evol.* 2017;1: 82.
- 674 42. Wolf JBW, Ellegren H. Making sense of genomic islands of differentiation in light
675 of speciation. *Nat Rev Genet.* 2017;18: 87–100.
- 676 43. Burri R, Nater A, Kawakami T, Mugal CF, Olason PI, Smeds L, et al. Linked
677 selection and recombination rate variation drive the evolution of the genomic
678 landscape of differentiation across the speciation continuum of *Ficedula*
679 flycatchers. *Genome Res.* 2015; gr.196485.115.
- 680 44. Delmore KE, Hübner S, Kane NC, Schuster R, Andrew RL, Câmara F, et al.
681 Genomic analysis of a migratory divide reveals candidate genes for migration
682 and implicates selective sweeps in generating islands of differentiation. *Mol*
683 *Ecol.* 2015;24: 1873–1888.
- 684 45. Irwin DE, Alcaide M, Delmore KE, Irwin JH, Owens GL. Recurrent selection
685 explains parallel evolution of genomic regions of high relative but low
686 absolute differentiation in a ring species. *Mol Ecol.* 2016;25: 4488–4507.
- 687 46. Cruickshank TE, Hahn MW. Reanalysis suggests that genomic islands of
688 speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol.*
689 2014;23: 3133–3157.
- 690 47. Nachman MW, Payseur BA. Recombination rate variation and speciation:
691 theoretical predictions and empirical results from rabbits and mice. *Philos*
692 *Trans R Soc Lond B Biol Sci.* 2012;367: 409–421.
- 693 48. Lexer C, Widmer A. The genic view of plant speciation: recent progress and
694 emerging questions. *Philos Trans R Soc B Biol Sci.* 2008;363: 3023–3036.
- 695 49. Ma T, Wang K, Hu Q, Xi Z, Wan D, Wang Q, et al. Ancient polymorphisms and
696 divergence hitchhiking contribute to genomic islands of divergence within a
697 poplar species complex. *Proc Natl Acad Sci.* 2018;115: E236–E243.
- 698 50. Percy DM, Garver AM, Wagner WL, James HF, Cunningham CW, Miller SE, et al.
699 Progressive island colonization and ancient origin of Hawaiian *Metrosideros*
700 (*Myrtaceae*). *Proc Biol Sci. The Royal Society;* 2008;275: 1479–90.
- 701 51. Dawson J, Stemmermann L. *Metrosideros* (Gaud). In: Wagner W, Herbst D,
702 Sohmer S, editors. *Manual of the flowering plants of Hawai'i.* Honolulu,
703 Hawai'i: Univ. Hawai'i Press; 1990. pp. 964–970.

- 704 52. Sur GL, Keating R, Snow N, Stacy EA. Leaf Micromorphology Aids Taxonomic
705 Delineation within the Hypervariable Genus *Metrosideros* (Myrtaceae) on
706 O'ahu. *Pac Sci.* 2018;72: 345–361.
- 707 53. Mueller-Dombois D. Forest dynamics in Hawaii. *Trends Ecol Evol.* 1987;2: 216–
708 220.
- 709 54. Drake DR. Germination Requirements of *Metrosideros polymorpha*, the
710 Dominant Tree of Hawaiian Lava Flows and Rain Forests. *Biotropica.* 1993;25:
711 461–467.
- 712 55. Morrison KR, Stacy EA. Intraspecific divergence and evolution of a life-history
713 trade-off along a successional gradient in Hawaii's *Metrosideros polymorpha*. *J*
714 *Evol Biol.* 2014;27: 1192–1204.
- 715 56. Stacy EA, Johansen JB, Sakishima T, Price DK. Genetic analysis of an ephemeral
716 intraspecific hybrid zone in the hypervariable tree, *Metrosideros polymorpha*,
717 on Hawai'i Island. *Heredity.* 2016;117: 173–183.
- 718 57. Cordell S, Goldstein G, Mueller-Dombois D, Webb D, Vitousek PM. Physiological
719 and morphological variation in *Metrosideros polymorpha*, a dominant
720 Hawaiian tree species, along an altitudinal gradient: the role of phenotypic
721 plasticity. *Oecologia.* 1998;113: 188–196.
- 722 58. Corn CA, Hiesey WM. Altitudinal variation in hawaiian metrosideros. *Am J Bot.*
723 1973;60: 991–1002.
- 724 59. Kitayama K, Pattison R, Cordell S, Webb D, Mueller-Dombois D. Ecological and
725 Genetic Implications of Foliar Polymorphism in *Metrosideros polymorpha*
726 Gaud. (Myrtaceae) in a Habitat Matrix on Mauna Loa, Hawaii. *Ann Bot.*
727 1997;80: 491–497.
- 728 60. Stemmermann L. Ecological studies of Hawaiian *Metrosideros* in a successional
729 context. *Pac Sci.* 1983;37: 361–373.
- 730 61. Vitousek PM, Turner DR, Kitayama K. Foliar Nutrients During Long-Term Soil
731 Development in Hawaiian Montane Rain Forest. *Ecology.* 1995;76: 712–720.
- 732 62. Stacy EA, Paritosh B, Johnson MA, Price DK. Incipient ecological speciation
733 between successional varieties of a dominant tree involves intrinsic
734 postzygotic isolating barriers. *Ecol Evol. John Wiley & Sons, Ltd;* 2017;7:
735 2501–2512.
- 736 63. Rhoades A. The evolution of reproductive barriers within an endemic Hawaiian
737 tree species (*Metrosideros polymorpha*) across environmental extremes.
738 University of Hawai'i at Hilo. 2012.

- 739 64. Hart PJ. Tree growth and age in an ancient Hawaiian wet forest: vegetation
740 dynamics at two spatial scales. *J Trop Ecol.* 2010;26: 1–11.
- 741 65. Carpenter FL. Plant-Pollinator Interactions in Hawaii: Pollination Energetics of
742 *Metrosideros Collina* (Myrtaceae). *Ecology.* 1976;57: 1125–1144.
- 743 66. Corn C. Variation in Hawaiian *Metrosideros*. University of Hawai'i at Manoa,
744 Honolulu. 1979.
- 745 67. Petit RJ, Hampe A. Some Evolutionary Consequences of Being a Tree. *Annu Rev*
746 *Ecol Evol Syst.* 2006;37: 187–214.
- 747 68. Stacy EA, Johansen JB, Sakishima T, Price DK, Pillon Y. Incipient radiation within
748 the dominant Hawaiian tree *Metrosideros polymorpha*. *Heredity.* 2014;113:
749 334–342.
- 750 69. Crawford DJ. Progenitor-derivative species pairs and plant speciation. *Taxon.*
751 2010;59: 1413–1423.
- 752 70. Gavrillets S. Perspective: models of speciation: what have we learned in 40 years?
753 *Evolution.* 2003;57: 2197–2215.
- 754 71. Barraclough TG, Vogler AP. Detecting the Geographical Pattern of Speciation
755 from Species-Level Phylogenies. *Am Nat.* 2000;155: 419–434.
- 756 72. Butlin Roger K, Galindo Juan, Grahame John W. Sympatric, parapatric or
757 allopatric: the most important way to classify speciation? *Philos Trans R Soc B*
758 *Biol Sci.* 2008;363: 2997–3007.
- 759 73. Fitzpatrick BM, Fordyce JA, Gavrillets S. Pattern, process and geographic modes
760 of speciation. *J Evol Biol.* 2009;22: 2342–2347.
- 761 74. Ekar J, Price D, Johnson M, Stacy E. Varieties of the highly dispersible and
762 hypervariable tree, *Metrosideros polymorpha*, differ in response to
763 mechanical stress and light across a sharp ecotone. *Am J Bot.* 2019;In press.
- 764 75. Motley T, Carr G. Artificial hybridization in the Hawaiian endemic genus
765 *Labordia* (Loganiaceae). *Am J Bot.* 1998;85: 654.
- 766 76. Kocher TD. Adaptive evolution and explosive speciation: the cichlid fish model.
767 *Nat Rev Genet.* 2004;5: 288–298.
- 768 77. Huber SK, León LFD, Hendry AP, Bermingham E, Podos J. Reproductive isolation
769 of sympatric morphs in a population of Darwin's finches. *Proc R Soc B Biol Sci.*
770 2007;274: 1709–1714.
- 771 78. Knope ML, Morden CW, Funk VA, Fukami T. Area and the rapid radiation of
772 Hawaiian *Bidens* (Asteraceae). *J Biogeogr.* 2012;39: 1206–1216.

- 773 79. Clague D. The growth and subsidence of the Hawaiian-Emperor volcanic chain.
774 The origin and evolution of Pacific island biotas, New Guinea to eastern
775 Polynesia: patterns and processes. Amsterdam, The Netherlands: SPB
776 Academic Publishing; 1996. pp. 35–50.
- 777 80. Price JP, Elliott-Fisk D. Topographic History of the Maui Nui Complex, Hawai'i,
778 and Its Implications for Biogeography. Pacific Science. University of Hawai'i
779 Press; 2004. pp. 27–45. Available:
780 <http://scholarspace.manoa.hawaii.edu/handle/10125/2709>
- 781 81. Izuno A, Hatakeyama M, Nishiyama T, Tamaki I, Shimizu-Inatsugi R, Sasaki R, et
782 al. Genome sequencing of *Metrosideros polymorpha* (Myrtaceae), a dominant
783 species in various habitats in the Hawaiian Islands with remarkable
784 phenotypic variations. J Plant Res. 2016;129: 727–736.
- 785 82. Fumagalli M, Vieira FG, Linderth T, Nielsen R. ngsTools: methods for population
786 genetics analyses from next-generation sequencing data. Bioinformatics.
787 2014;30: 1486–7.
- 788 83. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation
789 Sequencing Data. BMC Bioinformatics. 2014;15: 356.
- 790 84. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in
791 unrelated individuals. Genome Res. 2009;19: 1655–64.
- 792 85. Skotte L, Korneliussen TS, Albrechtsen A. Estimating Individual Admixture
793 Proportions from Next Generation Sequencing Data. Genetics. 2013;195: 693–
794 702.
- 795 86. Martin SH, Van Belleghem SM. Exploring Evolutionary Relationships Across the
796 Genome Using Topology Weighting. Genetics. 2017;206: 429–438.
- 797 87. Pickrell JK, Pritchard JK. Inference of Population Splits and Mixtures from
798 Genome-Wide Allele Frequency Data. PLoS Genet. 2012;8: e1002967.
- 799 88. Richards EJ, Poelstra JW, Martin CH. Don't throw out the sympatric speciation
800 with the crater lake water: fine-scale investigation of introgression provides
801 equivocal support for causal role of secondary gene flow in one of the clearest
802 examples of sympatric speciation. Evol Lett. 2018;2: 524–540.
- 803 89. Wang H, Vieira FG, Crawford JE, Chu C, Nielsen R. Asian wild rice is a hybrid
804 swarm with extensive gene flow and feralization from domesticated rice.
805 Genome Res. 2017;27: 1029–1038.
- 806 90. Jeffery NW, DiBacco C, Wringe BF, Stanley RRE, Hamilton LC, Ravindran PN, et al.
807 Genomic evidence of hybridization between two independent invasions of

808 European green crab (*Carcinus maenas*) in the Northwest Atlantic. *Heredity*.
809 2017;119: 154–165.

810 91. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian
811 population history. *Nature*. 2009;461: 489–94.

812 92. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the
813 joint demographic history of multiple populations from multidimensional SNP
814 frequency data. *PLoS Genet*. 2009;5: e1000695.

815 93. Portik DM, Leaché AD, Rivera D, Barej MF, Burger M, Hirschfeld M, et al.
816 Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest
817 frog using demographic model selection. *Mol Ecol*. 2017;26: 5245–5263.

818 94. Charles KL, Bell RC, Blackburn DC, Burger M, Fujita MK, Gvoždík V, et al. Sky, sea,
819 and forest islands: Diversification in the African leaf-folding frog *Afraxalus*
820 *paradorsalis* (Anura: Hyperoliidae) of the Lower Guineo-Congolian rain forest.
821 *J Biogeogr*. 2018;45: 1781–1794.

822 95. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG,
823 et al. The Rate and Molecular Spectrum of Spontaneous Mutations in
824 *Arabidopsis thaliana*. *Science*. 2010;327: 92–94.

825 96. Xie Zhengqing, Wang Long, Wang Lirong, Wang Zhiqiang, Lu Zhenhua, Tian
826 Dacheng, et al. Mutation rate analysis via parent–progeny sequencing of the
827 perennial peach. I. A low rate in woody perennials and a higher mutagenicity
828 in hybrids. *Proc R Soc B Biol Sci*. 2016;283: 20161016.

829 97. Garud NR, Messer PW, Buzbas EO, Petrov DA. Recent selective sweeps in North
830 American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS*
831 *Genet*. 2015;11: e1005004.

832 98. Schlamp F, van der Made J, Stambler R, Chesebrough L, Boyko AR, Messer PW.
833 Evaluating the performance of selection scans to detect selective sweeps in
834 domestic dogs. *Mol Ecol*. 2016;25: 342–356.

835 99. Kim Y, Nielsen R. Linkage disequilibrium as a signature of selective sweeps.
836 *Genetics*. 2004;167: 1513–24.

837 100. Nosil P, Funk DJ, Ortiz-Barrientos D. Divergent selection and heterogeneous
838 genomic divergence. *Mol Ecol*. 2009;18: 375–402.

839 101. Izuno A, Kitayama K, Onoda Y, Tsujii Y, Hatakeyama M, Nagano AJ, et al. The
840 population genomic signature of environmental association and gene flow in
841 an ecologically divergent tree species *Metrosideros polymorpha* (Myrtaceae).
842 *Mol Ecol*. 2017;26: 1515–1532.

- 843 102. Marques DA, Meier JI, Seehausen O. A Combinatorial View on Speciation and
844 Adaptive Radiation. *Trends Ecol Evol.* 2019;34: 531–544.
- 845 103. Heliconius Genome Consortium. Butterfly genome reveals promiscuous
846 exchange of mimicry adaptations among species. *Nature.* 2012;487: 94–98.
- 847 104. Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-
848 Barrio A, et al. Evolution of Darwin’s finches and their beaks revealed by
849 genome sequencing. *Nature.* 2015;518: 371–375.
- 850 105. Miao B, Wang Z, Li Y. Genomic Analysis Reveals Hypoxia Adaptation in the
851 Tibetan Mastiff by Introgression of the Grey Wolf from the Tibetan Plateau.
852 *Mol Biol Evol.* 2016;34: msw274.
- 853 106. Richards EJ, Martin CH. Adaptive introgression from distant Caribbean islands
854 contributed to the diversification of a microendemic adaptive radiation of
855 trophic specialist pupfishes. *PLOS Genet.* 2017;13: e1006919.
- 856 107. Enard D, Petrov DA. Evidence that RNA Viruses Drove Adaptive Introgression
857 between Neanderthals and Modern Humans. *Cell.* 2018;175: 360-371.e13.
- 858 108. Seehausen O. Hybridization and adaptive radiation. *Trends Ecol Evol.*
859 2004;19: 198–207.
- 860 109. Gross BL, Rieseberg LH. The Ecological Genetics of Homoploid Hybrid
861 Speciation. *J Hered.* 2005;96: 241–252.
- 862 110. Yakimowski SB, Rieseberg LH. The role of homoploid hybridization in
863 evolution: A century of studies synthesizing genetics and ecology. *Am J Bot.*
864 2014;101: 1247–1258.
- 865 111. Pease JB, Haak DC, Hahn MW, Moyle LC. Phylogenomics Reveals Three Sources
866 of Adaptive Variation during a Rapid Radiation. *PLOS Biol.* 2016;14:
867 e1002379.
- 868 112. Sicard A, Kappel C, Josephs EB, Lee YW, Marona C, Stinchcombe JR, et al.
869 Divergent sorting of a balanced ancestral polymorphism underlies the
870 establishment of gene-flow barriers in *Capsella*. *Nat Commun.* 2015;6: 7960.
- 871 113. Presgraves DC. The molecular evolutionary basis of species formation. *Nat*
872 *Rev Genet.* 2010;11: 175–180.
- 873 114. Maheshwari S, Barbash DA. The genetics of hybrid incompatibilities. *Annu Rev*
874 *Genet.* 2011;45: 331–355.
- 875 115. Mack KL, Nachman MW. Gene Regulation and Speciation. *Trends Genet.*
876 2017;33: 68–80.

- 877 116. Li H. Aligning sequence reads, clone sequences and assembly contigs with
878 BWA-MEM. arXiv. 2013; 1303.3997v2.
- 879 117. Vieira FG, Lassalle F, Korneliussen TS, Fumagalli M. Improving the estimation
880 of genetic distances from Next-Generation Sequencing data. Biol J Linn Soc.
881 2016;117: 139–149.
- 882 118. Lefort V, Desper R, Gascuel O. FastME 2.0: A Comprehensive, Accurate, and
883 Fast Distance-Based Phylogeny Inference Program. Mol Biol Evol. 2015;32:
884 2798–2800.
- 885 119. Freedman AH, Gronau I, Schweizer RM, Ortega-Del Vecchyo D, Han E, Silva
886 PM, et al. Genome sequencing highlights the dynamic early history of dogs.
887 PLoS Genet. 2014;10: e1004016.
- 888 120. Browning BL, Browning SR. Genotype Imputation with Millions of Reference
889 Samples. Am J Hum Genet. 2016;98: 116–126.
- 890 121. Pavlidis P, Jensen JD, Stephan W. Searching for Footprints of Positive Selection
891 in Whole-Genome SNP Data From Nonequilibrium Populations. Genetics.
892 2010;185: 907–922.
- 893 122. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C,
894 et al. Fast Genome-Wide Functional Annotation through Orthology
895 Assignment by eggNOG-Mapper. Mol Biol Evol. 2017;34: 2115–2122.
- 896 123. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and
897 Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57:
898 289–300.

899

900 **Figure Legends**

901

902 **Figure 1. Genetic relationship between glaberrima and newellii, two varieties of *M.***

903 ***polymorpha*.** G_{H1}, G_{H2}, and N are the genetic populations identified through this study.

904 (A) Geographic locations of the 40 sampled trees of glaberrima and newellii. Inset shows
905 the leaf morphology of glaberrima and newellii. (B) ADMIXTURE plot of K=2, 3, 4, and
906 5 for the 40 trees. Individuals from G_{H1}, G_{H2}, and N are indicated with ^, * and #,

respectively. (C) PCA plot of the 40 trees. (D) Neighbor-joining tree of the 40 trees.

Nodes with greater than 90% bootstrap support are indicated with black circles.

Figure 2. Relationships among the three focal populations of *M. polymorpha* on

Hawaii Island. (A) Topology weight for a three-population topology under three

different outgroups. Dotted box represents the topological relationship obtained from

genome-wide data. (B) Treemix graph of the three focal populations and three outgroups

assuming 4 migration edges. (C) Results of f_4 tests of admixture in the $m=4$ treemix

graph.

Figure 3. Estimation of demographic parameters for newellii and G_{HI} (the

glaberrima closely related and purported progenitor to newellii) on Hawaii Island.

(A) dadi parameter estimates for the best-fitting model. Effective population sizes of

current populations are scaled by the effective population size of the ancestral population

(N_{anc}) that is arbitrarily set to 1. Time is reported in unscaled units of T where $T \times 2N_{anc} = \tau$

(divergence time in generations). Migration rates are reported in units of m_{12} where

$m_{12}/2N_{anc} = M_{12}$ (proportion of individuals in each generation that are new migrants from

population 2 to population 1). Bootstrap parameter estimates are shown in square

brackets. nuG_{HI} : effective population size of G_{HI} ; nuN : effective population size of N ;

m_{12} : migration rate from N to G_{HI} ; m_{21} : migration rate from G_{HI} to N ; and T : unscaled

divergence time between G_{HI} and N . (B) Divergence time (years) between G_{HI} and N

under various mutation rates and generation times. Bootstrap confidence intervals are

shown in shaded areas.

Figure 4. Patterns of genome-wide divergence between G_{HI} and N. (A) Genome-wide zF_{ST} values calculated from 10-kbp windows. Scaffolds were ordered from longest to shortest. (B) Absolute measures of divergence (D_{xy}) between genomic islands of divergence ($zF_{ST} > 4$) and the genomic background ($zF_{ST} < 4$). (C) Selective sweep test statistics for genomic islands of divergence and the genomic background. Significant differences are indicated with * < 0.05 , ** < 0.01 , and *** < 0.001 .

Supporting information

S1 Fig. ADMIXTURE plot of $K=2$ to 7 for the 40 *M. polymorpha* individuals.

S2 Fig. Phylogenetic relationship of *M. polymorpha*, *M. rugosa*, and *M. tremuloides*.

S3 Fig. Treemix graph of the Hawaii-Island *M. polymorpha* and its outgroup from $m=0$ to $m=5$.

S4 Fig. Plot of squared correlation coefficient (r^2) between polymorphic sites. Decay of linkage disequilibrium was plotted by averaging r^2 in 100-bp bins.

S5 Fig. Genome-wide F_{ST} calculated in 10-kbp windows between G_{HI} and N.

S6 Fig. Visual representation of the 20 demographic models.

S7 Fig. dadi model fit for the best-fitting demographic model. Above row shows the observed and model fit folded site frequency spectrum. Below shows the map and histogram of the residuals.

S8 Fig. Visual representation of the 7 vicariance demographic models.

S9 Fig. Levels of polymorphism (θ_π) between islands of divergence ($zF_{ST} > 4$) and the genomic background ($zF_{ST} < 4$) between G_{HI} and N.

S10 Fig. Plot of the proportion of genic sequences to f_4 statistics of the topology $[G_{HI}, N; G_M, T]$ in 10-kbp windows.

S1 Table. Information on the sequenced and analyzed individuals of this study.

S2 Table. dadi parameter estimates for 20 different models using genotype likelihood based 2D-SFS for the G_{HI} and N populations. nu1, nu2: population 1 and 2 effective population size under models without a population size change; nu1a, nu2a: population 1 and 2 effective population size before a population size change; nu1b, nu2b: population 1 and 2 effective population size after a population size change; m12, m21: migration rate. For symmetric migration models only m12 is shown; m12a, m21a: migration rate of first epoch; m12b, m21b; migration rate of second epoch; T1, T2, T3: divergence times for different epoch periods of the model.

S3 Table. dadi parameter estimates for 20 different models using genotype call-

based 2D-SFS for G_{HI} and N populations. nu1,nu2: population 1 and 2 effective

population size under models without a population size change; nu1a, nu2a: population 1

and 2 effective population size before a population size change; nu1b, nu2b: population 1

and 2 effective population size after a population size change; m12, m21: migration rate.

For symmetric migration models only m12 is shown; m12a, m21a: migration rate of first

epoch; m12b, m21b; migration rate of second epoch; T1,T2, T3: divergence times for

different epoch periods of the model.

S4 Table. dadi parameter estimate for top 3 best-fitting models after further

optimization steps for G_{HI} and N populations. nu1,nu2: population 1 and 2 effective

population size; m12, m21: migration rate. For symmetric migration models only m12 is

shown; T1,T2: divergence times for different epoch periods of the model.

S5 Table. dadi parameter estimates for the 7 vicariance models for G_{HI} and N

populations. nu1,nu2: population 1 and 2 effective population size; m12, m21: migration

rate. For symmetric migration models only m12 is shown; T1,T2: divergence times for

different epoch periods of the model. s: proportion of ancestral population that lead to

population 1, and population 2 was established with 1-s.

S6 Table. dadi parameter estimates for the top 3 best-fitting vicariance models after

further optimization steps for G_{HI} and N populations. nu1,nu2: population 1 and 2

effective population size; m12, m21: migration rate. For symmetric migration models

only m12 is shown; T1,T2: divergence times for different epoch periods of the model. s: proportion of ancestral population that lead to population 1, and population 2 was established with 1-s.

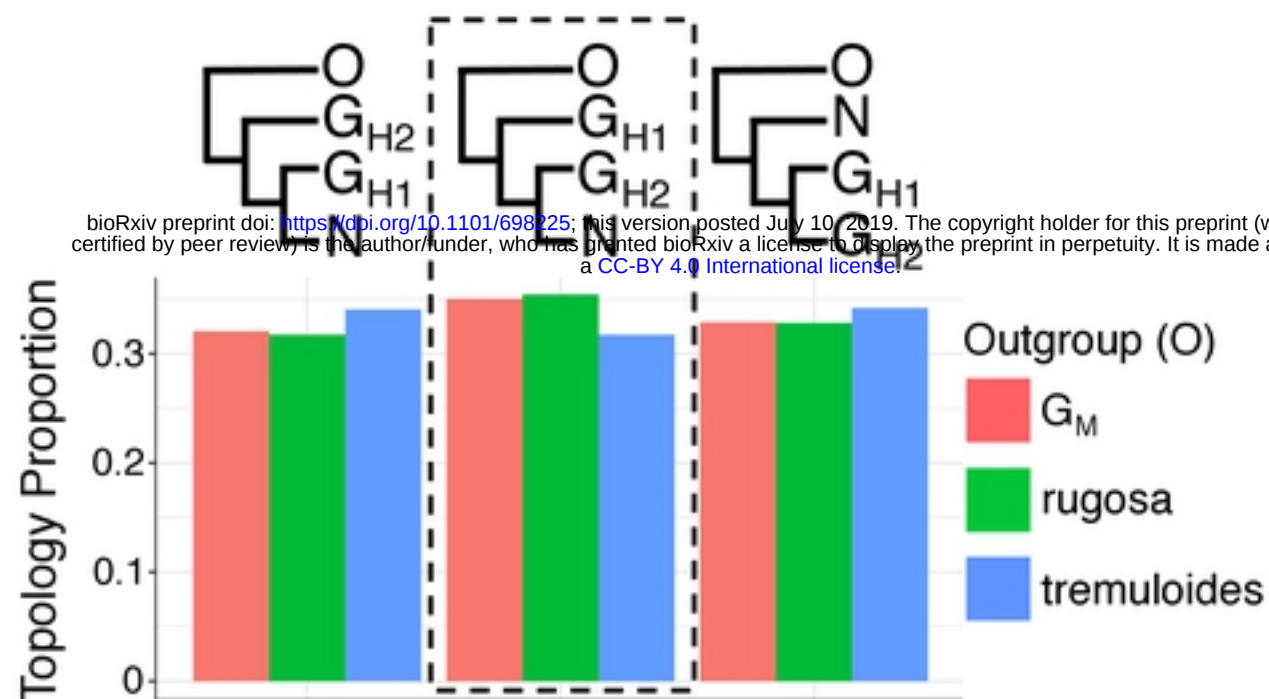
S7 Table. dadi parameter estimates for 20 different models using genotype call-based 2D-SFS for G_{H2} and N populations. nu1,nu2: population 1 and 2 effective population size under models without a population size change; nu1a, nu2a: population 1 and 2 effective population size before a population size change; nu1b, nu2b: population 1 and 2 effective population size after a population size change; m12, m21: migration rate. For symmetric migration models only m12 is shown; m12a, m21a: migration rate of first epoch; m12b, m21b; migration rate of second epoch; T1,T2, T3: divergence times for different epoch periods of the model.

S8 Table. dadi parameter estimate for top 3 best-fitting models after further optimization steps for G_{H2} and N populations. nu1,nu2: population 1 and 2 effective population size; m12, m21: migration rate. For symmetric migration models only m12 is shown; T1,T2: divergence times for different epoch periods of the model.

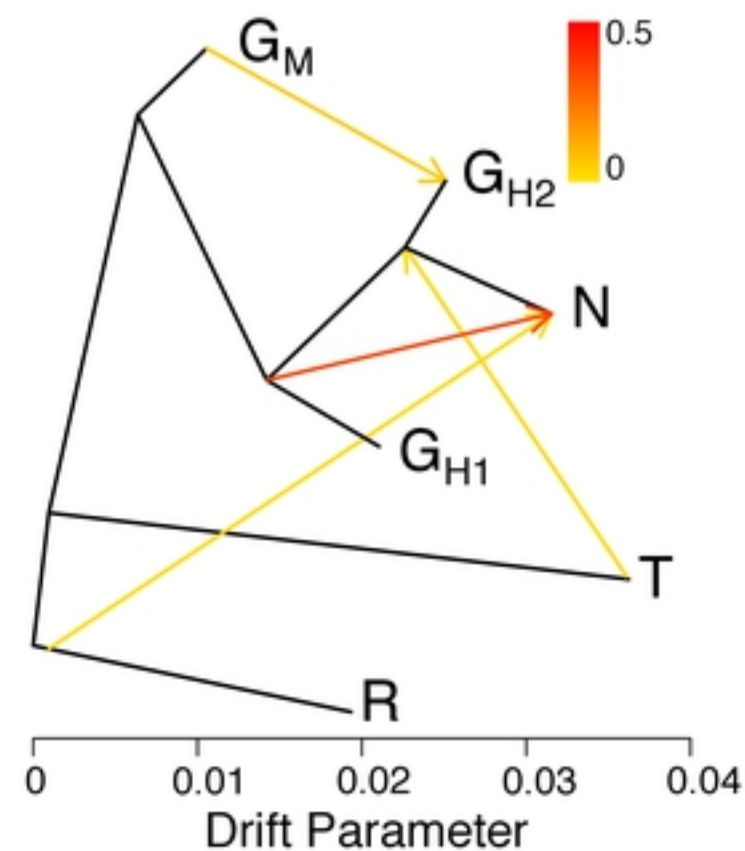
S9 Table. Median D_{xy} within and outside the genomic islands of divergence found in G_{H1} and N for various population comparisons.

S10 Table. List of gene names overlapping genomic islands of divergence.

A



B



C

Population Topology	f_4 statistics	S.E.	Z-score	P-value
G2,N;G1,R	-1.41E-04	1.02E-04	-1.38	1.68E-01
G2,N;G1,T	-5.10E-04	1.08E-04	-4.73	2.25E-06
G2,N;G1,Gm	-5.44E-04	8.18E-05	-6.65	2.93E-11
G2,G1;R,Gm	-2.00E-04	1.14E-04	-1.75	8.01E-02
G1,N;R,Gm	-2.03E-04	1.04E-04	-1.96	5.00E-02
G2,N;R,Gm	-4.03E-04	9.56E-05	-4.22	2.44E-05
G2,G1;T,Gm	7.05E-04	1.19E-04	5.91	3.42E-09
G1,N;T,Gm	-7.40E-04	1.08E-04	-6.83	8.49E-12
G2,N;T,Gm	-3.48E-05	1.01E-04	-0.34	7.34E-01
G2,G1;R,T	-9.05E-04	1.26E-04	-7.19	6.48E-13
G1,N;R,T	5.36E-04	1.17E-04	4.59	4.43E-06
G2,N;R,T	-3.68E-04	1.08E-04	-3.41	6.50E-04
G2,Gm;R,T	-8.34E-04	1.63E-04	-5.13	2.90E-07
G1,Gm;R,T	7.05E-05	1.51E-04	0.47	6.38E-01
N,Gm;R,T	-4.66E-04	1.57E-04	-2.96	3.08E-03

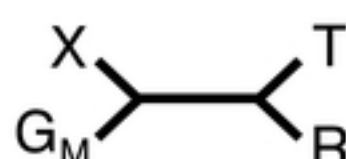
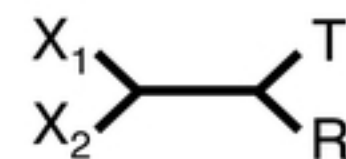
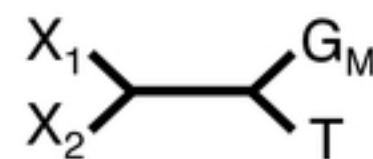
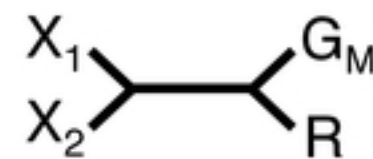
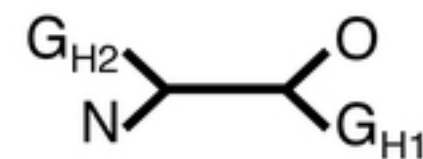
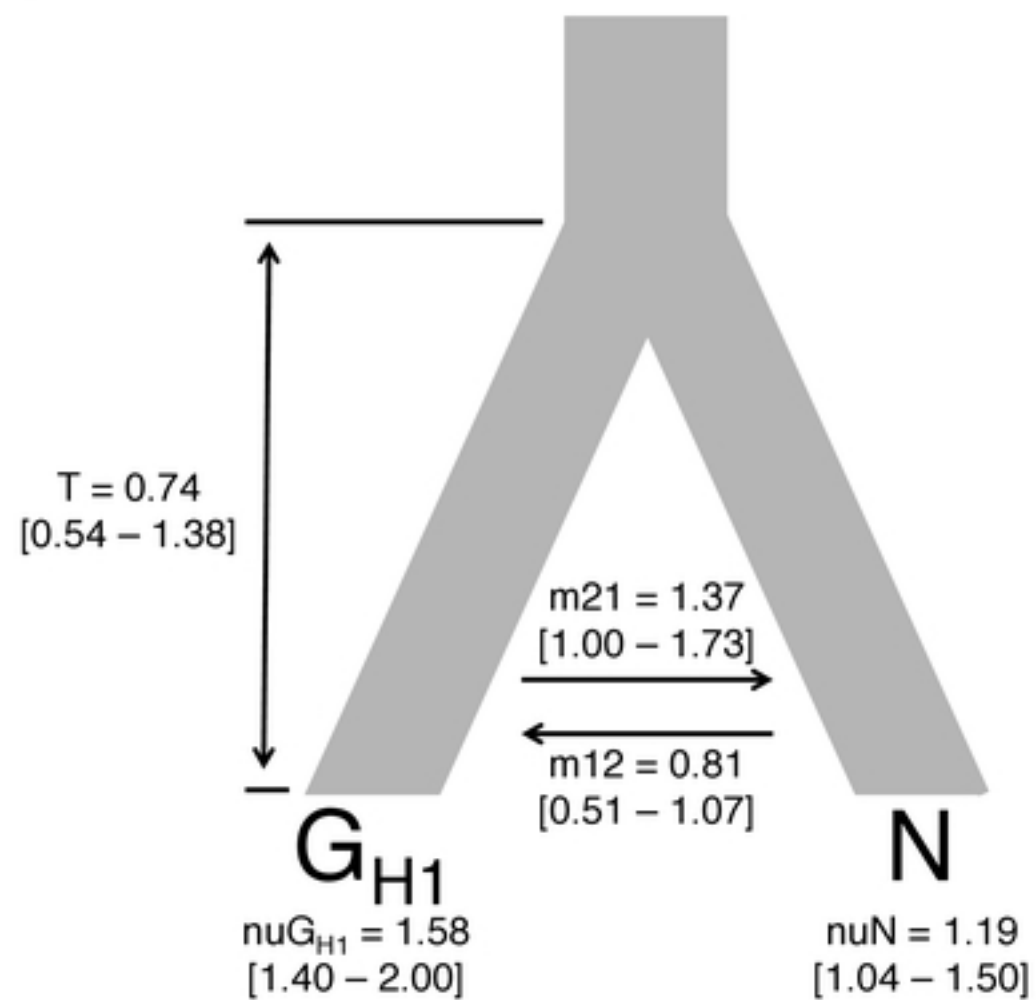


Figure2

A



B

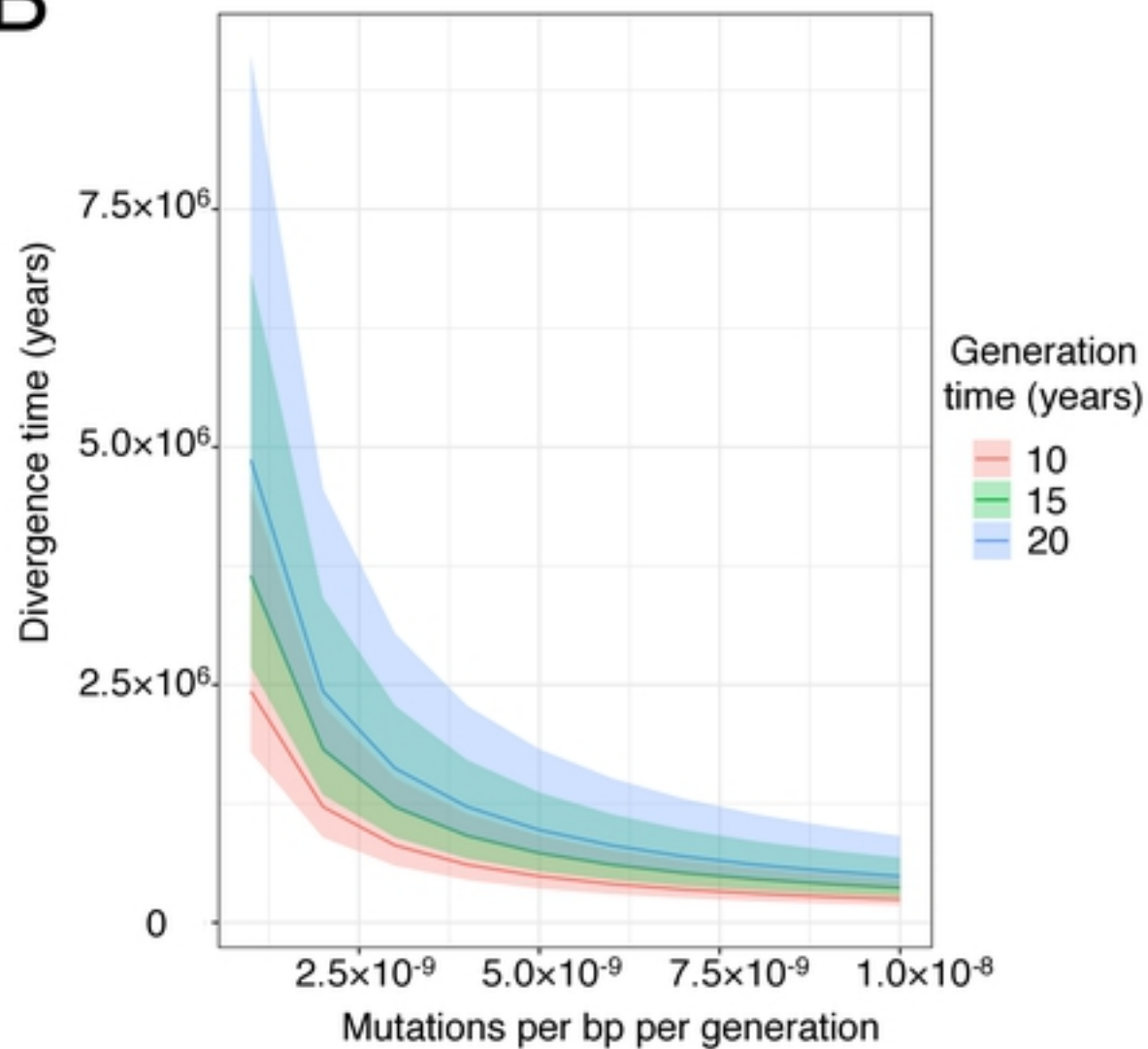
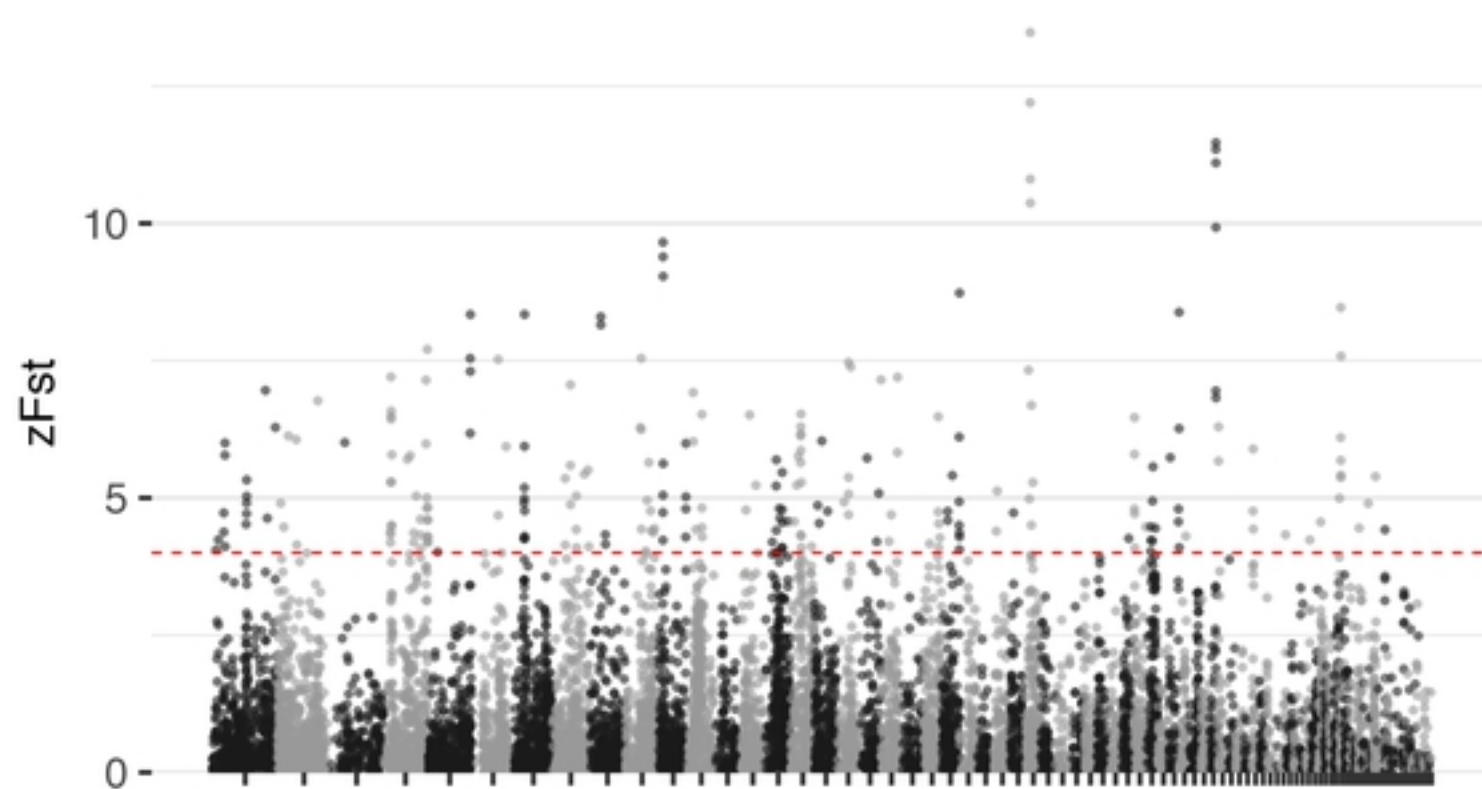
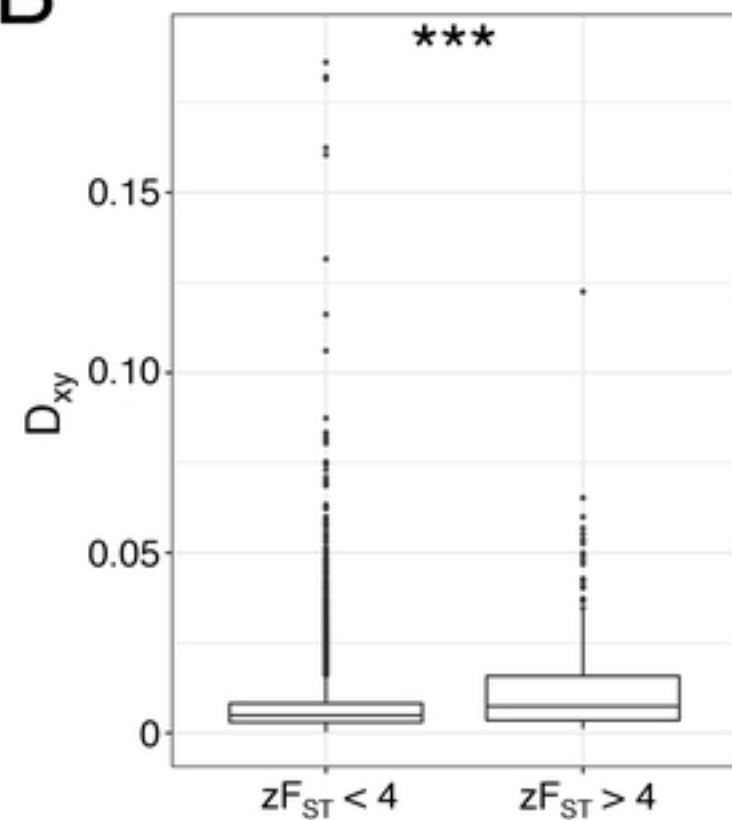


Figure3

A



B



C

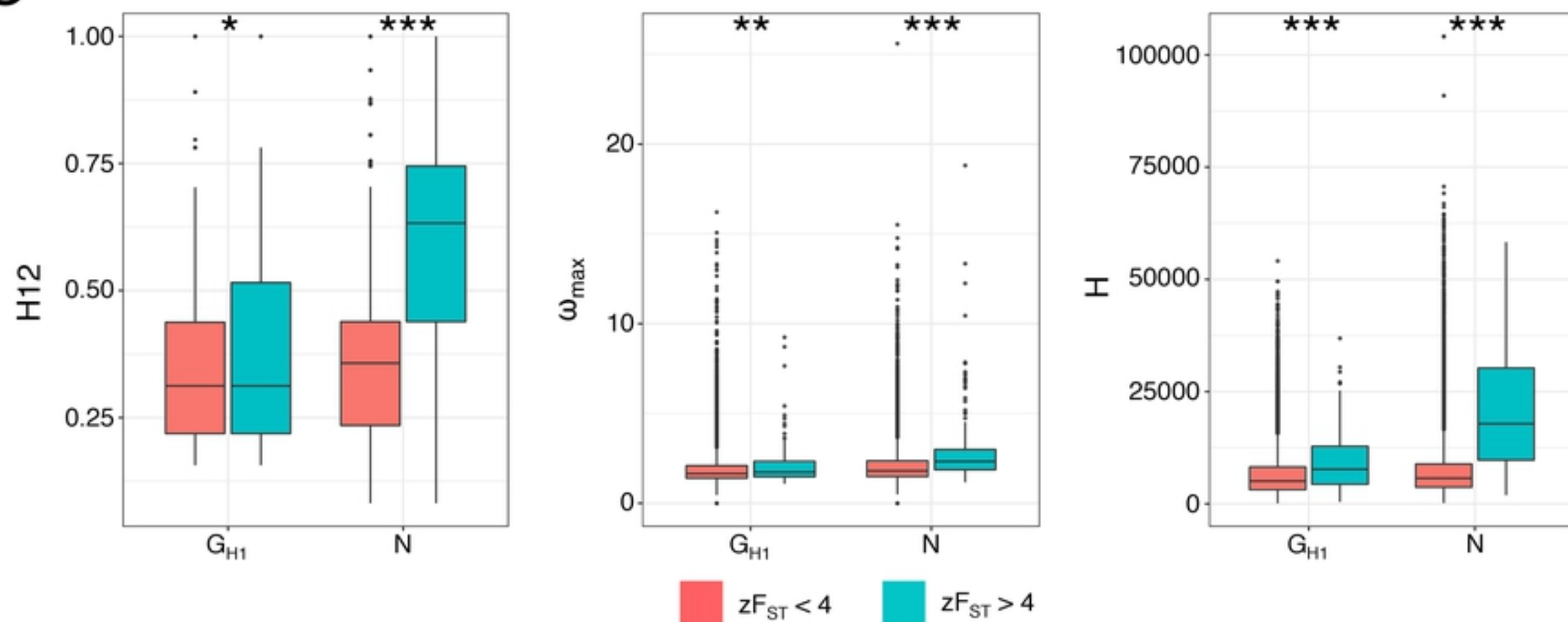


Figure4