Manuscript type: Article

Speciation in *Howea* palms occurred in sympatry, was preceded by ancestral admixture, and was associated with edaphic and phenological adaptation

Owen G. Osborne[a,b], Adam Ciezarek[a], Trevor Wilson[c], Darren Crayn[d], Ian Hutton[e], William J. Baker[f], Colin G.N. Turnbull[g], Vincent Savolainen[a,f,1]

[a]Department of Life Sciences, Silwood Park Campus, Imperial College London, Ascot SL5 7PY, UK; [b]current address: Molecular Ecology and Fisheries Genetics Laboratory, School of Natural Sciences, Bangor University, Bangor UK; [c]Royal Botanic Gardens Sydney, Mrs Macquaries Road, Sydney NSW 2000, Australia; [d]James Cook University, Australian Tropical Herbarium, Sir Robert Norman Building, McGregor Road, Smithfield, QLD 4878, Australia; [e]Lord Howe Island Museum, NSW 2898, Lord Howe Island, Australia; [f]Royal Botanic Gardens Kew, Richmond, Surrey TW9 3AB, UK; [g]Department of Life Sciences, South Kensington Campus, Imperial College London, London SW7 2AZ, UK

[1]Corresponding Author: Vincent Savolainen; Tel +442075942374; Email v.savolainen@imperial.ac.uk

Abstract

Howea palms are viewed as one of the most clear-cut cases of speciation in sympatry. The sister species *H. belmoreana* and *H. forsteriana* are endemic to the oceanic Lord Howe Island, Australia, where they have overlapping distributions and are reproductively isolated mainly by flowering time differences. However, the potential role of introgression from Australian mainland relatives had not previously been investigated, a process that has recently put other examples of sympatric speciation into question. Furthermore, the drivers of flowering time-based reproductive isolation remain unclear. We sequenced an RNA-seq dataset that comprehensively sampled *Howea* and their closest mainland relatives (*Linospadix*, *Laccospadix*), and collected detailed soil chemistry data on Lord Howe Island to evaluate whether secondary gene flow had taken place and to examine the role of soil preference in speciation. *D*-statistics analyses strongly support a scenario whereby ancestral *Howea* hybridised frequently with its mainland relatives, but this only occurred prior to speciation. Expression analysis, population genetic and phylogenetic tests of selection, identified several flowering time genes with evidence of adaptive divergence between the *Howea* species. We found expression plasticity in flowering time genes in response to soil chemistry as well as adaptive expression and sequence divergence in genes pleiotropically linked to soil adaptation and flowering time. Ancestral hybridisation may have provided the genetic diversity that promoted their subsequent adaptive divergence and speciation, a process that may be common for rapid ecological speciation.

Introduction

The geographic context of speciation has been a controversial topic in evolutionary biology. Theoretical models suggest that speciation can occur in sympatry with initial gene flow (Dieckmann and Doebeli 1999; Kondrashov and Kondrashov 1999; Doebeli et al. 2005; Bolnick and Fitzpatrick 2007), however it is likely to require far stronger divergent selection than speciation with spatial separation, and convincing examples in nature have been rare. It is unclear whether this reflects its genuine rarity, or the difficulty of convincing demonstration. Coyne and Orr (2004) set out four criteria for identifying cases of sympatric speciation: species are 1) currently sympatric, 2) sister taxa and 3) reproductively isolated and that 4) a period of allopatry during divergence is highly unlikely. The fourth criterion is the most difficult to demonstrate, particularly in species with broad continental distributions. Therefore, tests of sympatric speciation have attempted to reduce the possibility of an allopatric phase by focusing on species pairs restricted to small and isolated habitats such as islands. The assumption here is that the habitat island is too small for geographical isolation to have occurred within it, and too distant from other suitable habitats for speciation to have occurred during geographic isolation between the current habitat and a second one. This approach has thus far been used to infer sympatric speciation in plants and finches on remote oceanic islands (Ryan et al. 2007; Papadopulos et al. 2013) and cichlid fishes inhabiting crater lakes (Schliewen et al. 1994; Barluenga et al. 2006; Malinsky et al. 2015).

One of the best-known examples of speciation in sympatry in plants are the only two extant species of *Howea* palms (Savolainen et al. 2006). *Howea belmoreana* and *H. forsteriana* are restricted to the small (ca. 15 km^2) and remote (600 km from the nearest other landmass) Lord Howe Island (LHI; Australia), where they overlap in distribution across two soil types; although the former species is restricted to volcanic soil, the latter is present on both volcanic and calcareous soil. The evidence of sympatric speciation is further upheld by a sister relationship between the species in molecular phylogenetic trees (Savolainen et al. 2006; Baker et al. 2011), as well as prezygotic reproductive isolation by differences in flowering phenology (Savolainen et al. 2006; Hipperson et al. 2016) and some evidence of postzygotic isolation in the form of reduced hybrid fitness (Hipperson et al. 2016). They are both diploid, so polyploid speciation is excluded (Savolainen et al. 2006). However, a rigorous demonstration of a sister-relationship is required to provide evidence for sympatric speciation, since secondary contact (i.e. introgression) by more distantly related species can be concealed in phylogenetic trees constructed by a small number of genetic markers, as has been recently shown for seven crater lake cichlid radiations (Malinsky et al. 2015; C.H. Martin et al. 2015; Kautt et al. 2016; Meier et al. 2017; Poelstra et al. 2018). Relative to the closest extant outgroup species to *Howea*, the monotypic *Laccospadix australasicus* and species of *Linospadix*, the monophyly of *Howea* is supported by phylogenetic reconstruction using only two nuclear markers. In comparison to the cichlid examples above, two markers provide insufficient evidence to detect introgression. Furthermore, the existing phylogenetic tree includes only three of the seven species of *Linospadix*: *L. albertisianus, L. minor* and *L. palmerianus* (Savolainen et al. 2006). *Linospadix monostachyos,* which was not included in this phylogenetic analysis, inhabits the most proximal part of Australian mainland to LHI, making it the most likely known candidate for gene flow with *Howea*. A phylogenomic analysis to explicitly test secondary introgression into *Howea* from mainland relatives is therefore needed.

Beyond the prevalence of speciation in sympatry, many questions remain regarding its genomic underpinning and the conditions that may promote it. For example, the role of ancestral gene flow in providing genetic diversity on which selection can act may be more important than previously appreciated (Meier et al. 2017), and the mechanisms by which reproductive isolation evolves following initial local adaptation remain opaque in most species. These factors are largely unknown in *Howea*, but since the species have overlapping, yet distinct soil preferences, it is possible that soil characteristics have in part been responsible for the divergence of species. One hypothesis is that soil acted as a driver of speciation in *Howea* via selection on genes pleiotropically affecting both flowering time and soil preference, thereby producing reproductive isolation as a by-product of soil adaptation (Dunning et al. 2016). Another hypothesis is that a switch in soil preference induced a plastic response in flowering time, allowing initial divergence. These differences could have later been canalised (Pfennig et al. 2010) or bolstered by a reinforcement-like mechanism (Kirkpatrick 2001). Several genes have been identified with evidence of divergent selection between the *Howea* species (Dunning et al. 2016), although a lack of corresponding outgroup data has prevented our ability to pinpoint in which species changes have occurred. Furthermore, the characteristics of the soil have not been analysed beyond broad categories (calcareous versus volcanic) and pH measurements (Savolainen et al. 2006; Papadopulos et al. 2013), lacking variation in soil components (e.g. macronutrients, micronutrients and water availability), which may have been crucial to local adaptation and the evolution of flowering time differences.

To further develop our understanding of speciation in sympatry, and test the implicated mechanisms driving this process in *Howea*, we combined a detailed chemical analysis of soil on LHI with an RNA-seq dataset (transcriptomes) of three tissue types derived from 54 individuals, which included all species of *Linospadix* and *Laccospadix* as well as both *Howea* species. Specifically, we aimed to (i) determine whether external secondary gene flow from mainland relatives caused speciation in *Howea*; (ii) determine whether gene flow between the ancestor of *Howea* and its mainland relatives preceded colonisation of LHI; (iii) characterise the soil types that the two *Howea* species inhabit, and correlate this to gene expression and sequence divergence in the species; and (iv) identify which genes have undergone adaptive evolution in each of the two *Howea* species, and evaluate whether these include loci that could have driven the evolution of reproductive isolation.

Materials and methods

Tissue sampling and RNA sequencing

For *H. belmoreana* and *H. forsteriana*, we took the data from Dunning et al. (2016), that is, 19 and 17 individuals from each species, respectively. For outgroup species, we sampled between one and six individuals from the wild and at the Royal Botanic Gardens, Sydney (*Linospadix* albertisianus: 1, *L.* apetiolatus: 2, *L. microcaryus*: 1, *L. minor*: 3, *L. monostachyos*: 6, *L. palmerianus*: 2, *Laccospadix australasicus*: 4; Table S1). Leaf, inflorescence and root tissue was taken for each individual for RNA-sequencing. Tissue was cut into <5mm^2 sections and stored in RNAlater (Sigma) at -20°C. All tissue samples were sent to the BGI Tech solutions (Hong Kong) for RNA extraction, library construction and sequencing. Paired-end 100 base pairs (bp) libraries were multiplexed and sequenced on an Illumina HiSeq 4000 sequencer. These data were supplemented with publicly-available RNA-seq data for other palms (Arecaceae). All data are available from SRA accession no (PRJNA528594).

Bioinformatic processing

Illumina primer and adaptor sequences were removed and initial quality control was completed by removing reads with an average PHRED-scaled quality score < 20 (both completed by BGI Tech Solutions). Reads were then corrected for each individual using Rcorrector v.1.0.2 (Song and Florea 2015), followed by trimming in Trimmomatic v.0.33 (Bolger et al. 2014) with the following settings LEADING:5, TRAILING:5, SLIDINGWINDOW:4:5, MINLEN:75 (following recommendations by MacManes (2014)).

To improve the reference transcriptome of Dunning et al. (2016), we used a comprehensive multi-assembler, multi-K-mer approach as follows. For the individual with the highest number of corrected, trimmed read pairs for each of the two *Howea* species, we separately assembled de novo transcriptomes using eight transcriptome-specific assemblers: BinPacker v.1.0 (Liu et al. 2016), Bridger v. 2014-12-01 (Chang et al. 2015), IDBA-tran v.1.1.0 (Peng et al. 2013), Oases v.0.2.08 (Schulz et al. 2012), Shannon v.0.0.2 (Kannan et al. 2016), SOAPdenovo-Trans v.1.0.4 (Xie et al. 2014), TransABySS v.1.5.5 (Robertson et al. 2010) and Trinity v.2.4.0 (Grabherr et al. 2013). For Trinity, Bridger and BinPacker, we used three K-mer lengths, representing the maximum, minimum and default settings of 19, 25 and 33. For IDBA-tran, Oases, Shannon, SOAPdenovo-Trans and TransABySS, we used six K-mer lengths, 21,31,41,51,61 and 71. To determine fragment length distributions, necessary for Bridger, Oases and BinPacker, we mapped all reads to their 25 K-mer Trinity assembly using BWA-MEM v.0.7.8 (Li and Durbin 2009) with default settings. Input sizes were determined from the resulting mappings using the *CollectInsertSizeMetrics* function of Picard Tools v.2.6.0 (available at http://broadinstitute.github.io/picard/). This first-pass assembly produced 34 independent assemblies for each species (three for Trinity, Bridger and BinPacker, six for Velvet-Oases, transABYSS, SOAPdenovo and Shannon, and one for IDBA-tran, which builds on each K-mer length assembly iteratively to produce one final assembly).

For each of the 68 de novo assemblies, we then used TransDecoder v.3.0.1 (Haas et al. 2013) to identify open reading frames (ORF) using the *single_best_ORF* option. Contigs lacking an ORF over 100 amino acids in length were discarded, since these are likely to represent assembly artefacts. The retained coding sequences (CDS) for all 34 assemblies per species were then combined and clustered using CD-HIT-EST v.4.6.1 (Fu et al. 2012) with a local sequence identity threshold of 0.99, and coverage length settings of aL= 0.005, aS = 1 as recommended by Cerveau and Jackson (2016). Only the longest sequence from each cluster was retained to remove redundancy. To reduce the chance of assembly errors in our final assembly, we removed sequences that were only recovered by a single assembler or in less than four individual (i.e. assembler and K-mer length-specific) assemblies, following Cerveau and Jackson (2016). This resulted in a single non-redundant reference dataset for each *Howea* species. These were then matched between species using reciprocal best BLAST. Each assembly was blast searched against the other with BLASTN (Camacho et al. 2009) and those that were reciprocally each other's top hit were retained as reciprocal best BLAST pairs (RBB-pairs). RBB-pairs were aligned using MAFFT v.7.245 (Katoh et al. 2002) using automatic selection of the appropriate alignment strategy. Consensus sequences for each pairwise alignment were produced using an in-house python script (available at https://github.com/ogosborne/fasta_alignment_filters/), in which all non-matching nucleotides were coded as ambiguity characters. To control for insertions and deletions within coding regions, we reran TransDecoder v.3.0.1 (Haas et al. 2013) on the consensus sequences as above. The resulting consensus sequences were used as a mapping reference. To determine locus-to-transcript relationships, we used a mapping-based sequence clustering pipeline. Mapping of all reads from *Howea, Linospadix* and *Laccospadix* to the reference was first conducted using STAR v.2.5.3a (Dobin et al. 2013), with each tissue from each individual mapped separately, allowing unlimited matches per read. The mappings were then used to cluster transcripts with Corset v.1.06 (Davidson and Oshlack 2014) using a distance threshold of 0.3 and considering each species-tissue type combination as a distinct experimental grouping. For each resulting Corset cluster, only the transcript with the longest ORF was retained. The resulting sequences formed our final transcriptome assembly, and these were carried forward into downstream analyses. These were matched to the previous *Howea* transcriptome assembly (Dunning et al. 2016) using reciprocal best BLASTn (Camacho et al. 2009), and both assemblies were assessed using BUSCO v.2.0 (Simao et al. 2015).

To identify sequence variation, STAR was run on all data, including the publicly available Arecaceae data from the NCBI Short Sequence Archive (SRA), mapping all reads from each individual together and allowing only unique read mappings. Variants were then called and filtered using the samtools-bcftools v.1.3.1 pipeline (Li 2011). First samtools *mpileup* function was used to create a pileup file for each individual separately, considering only reads with a PHRED scaled mapping quality over 20. Pileups were then used to call SNPs using bcftools *call* function with the multiallelic caller model, keeping sites that are ambiguous in the reference, and outputting gVCF homozygous reference blocks. The resulting SNPs were filtered using bcftools *filter* function excluding SNPs within three bases of an indel, with a genotype quality < 20 or with a base quality < 20. Homozygous calls were required to be supported by three reads and heterozygous calls were required to be supported by two reads for each allele. FASTA formatted sequences were produced from the resulting VCF files and reference sequences using vcf2fas, where heterozygous bases were coded as IUPAC ambiguity codes (Bruno Nevado, available from https://github.com/brunonevado/vcf2fas). Resulting sequences were then concatenated to produce an aligned FASTA file for each gene.

Prior to phylogenetic analysis, alignments were filtered using inhouse python scripts (available at https://github.com/ogosborne/fasta_alignment_filters/). To produce input data for gene tree-based analyses (which used *Howea*, *Linospadix* and *Laccospadix*, as well as the closest outgroup, *Areca catechu*, in order to root the trees), we first removed sequences where over 2% of bases were heterozygous, as these may represent erroneous mapping of paralogues to the same reference. We then removed alignment columns with over 90% missing data and sequences with over 50% missing data. Following these filters, alignments were retained if they contained four or more sequences and were over 100 bases in length. We refer to this as the ‘individual sequence dataset’. To produce input data for species tree-based analysis (using all Arecaceae species), we first removed highly heterozygous sequences as above, before producing a consensus sequence for each species where any intraspecific variants were coded as missing data. The resulting sequences were concatenated into an alignment for each gene that was then filtered for missing data as above, we refer to this as the ‘species-consensus sequence dataset’.

Functional annotation

All genes in the final reference transcriptome were annotated by BLAST. Model species proteomes (primary isoforms only) and accompanying Gene Ontology (GO) terms were downloaded from Phytozome v. (Goodstein et al. 2012) (Available from https://phytozome.jgi.doe.gov, downloaded 11/08/18). Amino acid sequences of each gene were searched against four model species proteomes, *Arabidopsis thaliana*, *Brachypodium distachyon*, *Oryza sativa* and *Zea mays,* using BLASTP v2.2.25 (Camacho et al. 2009) with an e-value cut-off of 0.0001, and only the top hit was retained. Genes were annotated with the GO terms of their homologues from each of the reference proteomes. Redundancy was removed and ancestor terms were added. GO term enrichment amongst genes of interest was then tested using the topGO v.2.26.0 package in R (Alexa et al. 2006) using the *weight* algorithm.

To specifically identify genes with the potential to drive reproductive isolation by pleiotropically linking soil adaptation and flowering time in *Howea*, we identified all genes with known involvement in flowering time and LHI-relevant soil characteristics. To computationally identify flowering-time related genes, we took two approaches. Firstly, we identified all genes for which the *A. thaliana* homologue was in the FLOR-ID flowering time gene database (Bouché et al. 2016). Secondly, we identified all genes annotated with the GO terms GO:0009909: “regulation of flower development” and GO:0010228: “vegetative to reproductive phase transition of meristem”. To computationally identify genes potentially involved in soil adaptation in *Howea*, we took three approaches. Firstly, we identified all genes that had significantly differential expression with regard to soil chemistry in *H. forsteriana* (see “Analysis of gene expression” above) that we considered to have direct evidence of a link with soil chemistry in *Howea*. Secondly, since water content differed significantly between the soil types of LHI (see results), we identified all genes whose homologues were in the DroughtDB drought gene database (Alter et al. 2015). Thirdly, we identified all genes annotated with the following GO terms, which were relevant to our findings from the soil chemistry analysis: GO:0006970: “response to osmotic stress”, GO:0009414: “response to water deprivation”, GO:0042221: “response to chemical”, GO:0036377: “arbuscular mycorrhizal association”, GO:0031667: “response to nutrient levels”, and GO:0006811: “ion transport”. All genes which were annotated as potentially involved in both flowering and soil related functions were then individually assessed with an extensive literature search. We only considered genes to be potential pleiotropic for soil adaptation and reproductive isolation when they had (i) published evidence of a mutant flowering time phenotype and (ii) either differential expression according to soil in *Howea* or published evidence of a mutant phenotype relevant to the differences we found between LHI soil types.

Phylogenetic inference

Firstly, we used the species-consensus dataset to produce a dated species tree. We inferred a maximum likelihood gene tree with this data using RAxML v.8.2.9 (Stamatakis 2014) with 200 bootstraps using the GTRGAMMA model of evolution. The branch lengths were then re-estimated on 100 bootstraps of the data using the GTRGAMMA model in RAxML with the topology fixed to that of the best maximum likelihood tree. These bootstrapped trees were then rooted with the clade formed by *Daemonorops jenkinsiana* and *Mauritia flexuosa* as the outgroup, as in previous studies (Baker et al. 2011; Couvreur et al. 2011; Faurby et al. 2016). Divergence times were estimated from these trees using the penalised likelihood method implemented in r8s v. 1.80 (Sanderson 2003). We used three fossil calibrations, taken from Faurby et al. (2016): minimum ages of 65 mya for the most recent common ancestor (MRCA) of *Daemonorops* and *Mauritia*; 54.8 mya for the MRCA of *Cocos* and *Elaeis*; and 85.8 mya for the MRCA of *Phoenix* and *Borassus*. We also set the root age to 100 mya, the crown age of Arecaceae found in previous work (Couvreur et al. 2011). For each tree, we identified the optimal rate-smoothing parameter using cross-validation in r8s (with the following settings: *method* = pl, *penalty* = add, *algorithm* = tn, *cvstart* = -8, *cvinc* = 0.5, *cvnum* = 32). The optimal rate-smoothing parameter for each tree was then used to estimate divergence times and the solutions were checked with the *checkGradient* function. Mean node age estimates from the 100 bootstrap replicates were taken as point estimates and standard deviations were used to compute 95% confidence intervals. Because topological discordance among gene trees can affect branch length estimation, we produced a second species tree for which we attempted to limit its influence by removing highly discordant trees from the analysis. For each gene, the best maximum likelihood gene tree was compared to the best maximum likelihood species tree (see above) in a Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa 1999) implemented in RAxML (Stamatakis 2014). Genes for which the species tree (estimated with all genes) had a significantly worse likelihood than the gene tree estimated with only the gene in question (*P* < 0.05) were then removed from the analysis, and RAxML and r8s were rerun on this filtered dataset as above.

Secondly, we used the individual sequence dataset to infer a multi-species coalescence-based tree, because gene tree discordance can obscure phylogenetic inference in closely related species. For each gene in the individual sequence dataset, we inferred a maximum likelihood gene tree with RAxML v.8.2.9 (Stamatakis 2014) with 200 bootstraps using the GTRGAMMA model of evolution. SH-like branch supports were also calculated using RAxML (Anisimova et al. 2011). To examine gene tree discordance using DensiTree plots, all trees that contained every individual were filtered to remove trees with fewer than 10 nodes with under 80% SH-like support. These were then rooted using *Areca catechu* as the outgroup and made ultrametric using the *root* and *chronos* functions in the APE package (Paradis et al. 2004) in R v.3.3.1 (R Core Development Team 2008). These trees were then visualised in DensiTree v.2.2 (Bouckaert 2010). To infer the species phylogeny while explicitly accounting for gene-tree discordance we produced a coalescent-based tree. We collapsed low support nodes (SH-like support < 80) as recommended by Zhang et al. (2018) and retained all genes with over 50% of nodes un-collapsed. These were then input into ASTRAL v.5.5.6 (Mirarab and Warnow 2015) for phylogenetic reconstruction with enforced intraspecific monophyly, which inferred the topology and calculated concordance factor and posterior probability based branch-support.

Detection of introgression

Firstly, to detect introgression we used a multidimensional scaling (MDS) approach with the individual sequence dataset. SNPs for all genes were filtered to remove singletons and those with over 90% missing data across all individuals. To produce a set of unlinked SNPs, these were then further filtered to keep only the SNP with the least missing data per gene. MDS analysis was then carried out using the *mds-plot* function in PLINK v1.9 with two dimensions. If one *Howea* species were the result of hybridisation with one of the outgroups we would expect it to cluster more closely with outgroups than the other *Howea* species in this analysis.

Secondly, we used a *D*-statistic approach with the species-consensus dataset with the aim of differentiating between three scenarios: (i) sympatric speciation in *Howea* following allopatric separation from their sister taxa with no subsequent gene flow between *Howea* and outgroups; (ii) sympatric speciation in *Howea* following ancestral gene flow with their sister taxa; and (iii) speciation in *Howea* being driven by introgression from outgroups (Fig. 1). We calculated Patterson's *D* statistic (Green et al. 2010) for each four taxon subtree of species with the topology (((*H. belmoreana*, *H. forsteriana*),P_3), outgroup) where each species from *Linospadix* and *Laccospadix* was used as the third 'population', P_3, separately. Counts of two discordant site patterns *ABBA* (((A,B),B),A) and *BABA* (((B,A),B),A) were compared using Patterson's *D* statistic, with a value of *D* significantly different from zero implying introgression between one of the *Howea* species and P_3, i.e. supporting scenario (iii) above. To test for more complex introgression scenarios, including introgression between the ancestor of *Howea* and the outgroups, we used five-taxon D_{FOIL} statistics (Pease and Hahn 2015) for each five-taxon subtree with the topology: (((*H. belmoreana*, *H. forsteriana*),(P_3,P_4)),outgroup) in which the split of P_3 and P_4 predates the split of *Howea*. Four D_{FOIL} statistics were calculated: D_{FO}, D_{IL}, D_{FI} and D_{OL} (Pease and Hahn 2015). The combination of positive, negative and zero results of these four statistics can be used to infer ancestral introgression. Specifically, test results in which D_{FO} and D_{IL} were either both positive or both negative, while D_{FI} and D_{OL} were both zero, imply introgression between P_3 or P_4 and the ancestor of *Howea* (scenario ii) (Pease and Hahn 2015). Zero values for Patterson's *D* are also expected under scenario (ii). Zero values for all *D*-statistics would support scenario (i) (Fig. 1). *Areca catechu* was used as the outgroup in four and five-taxon tests because it was the closest relative to *Howea*, *Linospadix* and *Laccospadix* in our dataset. Site patterns were counted using the *fasta2dfoil.py* script from dfoil (Pease and Hahn 2015). To estimate confidence intervals and *P*-values for estimates of the *D*-statistics, 1,000 bootstrap replicates were used, in which genes were sampled with replacement to the total number of genes, and *D* was re-estimated. *P*-values were calculated for each of these tests and corrected for multiple testing using Bonferroni correction. Any test with a corrected $P < 0.05$ would be considered to show evidence of introgression.

Soil analysis

To analyse the soil characteristics in which *Howea* grows, we collected soil samples from 34 sites from which the *Howea* individuals with sequence data in this study were sourced. Soil samples were sent to the Diagnostic and Analytical Services Environmental Laboratory (Wollongbar, NSW, Australia) for analysis. This included three analyses: (i) 20 acid extractable elements (Al, As, B, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Mo, Na, Ni, P, Pb, S, Se and Zn) were quantified using inductively coupled plasma atomic emission spectroscopy (ICP-AES); (ii) four Diethylenetriamine pentaacetate extractable micronutrients (Cu, Fe, Mn and Zn) were quantified using ICP-AES, in a protocol which represent a closer approximation of the phytoavailability of these elements; (iii) soil electrical conductance, a metric that correlates with multiple soil properties related to plant health (Peverill et al. 1999), was measured. For the majority of sites, water availability (27 sites) and pH (33 sites) was also measured. A total of 50ml of soil was collected and weighed; samples were then dried in an oven for 48 hours at 80°C. They were then re-weighed and the percent water content was calculated. All samples for water content analysis were collected during a two-week period following at least two weeks without rainfall from the 2nd to 15th of April 2018. Soil pH was measured using Inoculo soil pH test kits (EnviroEquip Pty). To investigate overall soil variation, we used a principal component analysis (PCA) approach. Missing values for pH and water content were first converted to their respective median values. PCA was then performed using the *prcomp* function in R with scaling to account for variable scales for different components of soil variation. Each component of soil variation was also compared separately. Three comparisons were applied: (i) calcareous versus volcanic soil; (ii) *H. belmoreana* presence versus absence; and (iii) *H. forsteriana* presence versus absence. These were compared for each soil type using Mann-Whitney U tests and *P*-values were corrected for multiple testing using the false discovery rate (Benjamini and Hochberg 1995).

Population structure

To estimate the extent of isolation-by-distance within each *Howea* species, and to determine whether there was any evidence of population structure according to soil type within *H. forsteriana*, we calculated the pairwise coefficient of relatedness for all individuals in PLINK using the set of unlinked SNPs used for our MDS analysis. To detect isolation-by-distance, pairwise matrices of log geographic distance and coefficients of relatedness were used to conduct Mantel tests using *mantel.randtest* in the R package adegenet (v. 2.1.1; Jombart and Ahmed 2011); this was done for each species separately. To test whether there was more divergence between soil types for *H. forsteriana*, we compared coefficients of relatedness between all pairs of individuals from different soils types with all pairs from the same soil type using a t-test.

Tests of selection

We calculated several population genetic statistics to look for evidence of selection, using the PopGenome package in R (Pfeifer et al. 2014) and the individual sequence dataset. Differentiation (as measured by F_{ST}; Weir and Cockerham 1984) and net divergence (d_{XY}; Nei, 1987) were calculated for each contig between the two *Howea* species, and Tajima's D (Tajima 1989) and average pairwise diversity (π) were calculated within each species. Genes with F_{ST} of 1 (indicating complete fixation), d_{XY} over the 95th percentile, or Tajima's D below the 95th percentile were considered genes of interest for downstream analyses.

To identify genes potentially evolving under positive selection in the two *Howea* species, we also used a phylogenetic d_N/d_S-based approach with the species-consensus dataset, which was trimmed to include only CDS sequences inferred by TransDecoder (Haas et al. 2013). We then implemented various codon filters. Sequences with premature stop codons were removed, as well as the final stop codon of each alignment. To ensure that the amount of selection could be compared between the two *Howea* species, we retained only codons that contained no missing data in either species. Following these filters, alignments containing at least 33 codons, at least one SNP between the two *Howea* species, and at least three species in the alignment were taken forward for tests of selection. We used the branch-site test of positive selection (Zhang et al. 2005) implemented in the codeml program in PAML v.4.8 (Yang 2007). The branch-site models allow selection to vary across both sites and branches of the phylogeny. For each *Howea* species, we implemented two models. The alternative model allows d_N/d_S to vary above 1 on some sites on the branch being tested (the foreground branch, designated as the tips leading to each *Howea* species separately) whereas other branches (background branches) only vary between 0 and 1. The alternative model is compared to a null model where d_N/d_S is fixed at 1 for these sites on the foreground branch. It is compared in a Likelihood Ratio Tests (LRT), which approximates a chi-squared distribution with one degree of freedom. Phylogenetic uncertainty can affect the results of the branch-site test (Pie 2006), so we took two approaches to ensure our results were robust to it. First, we reran the significant branch-site tests using a species tree estimated with only genes that were not significantly discordant with the overall species topology (see "Phylogenetic inference" section above). Second, we reran the significant branch-site tests using the best maximum likelihood gene tree for the gene tested, rather than the species tree.

Analysis of gene expression

Read counts for each tissue type and each gene for all individuals in *Howea*, *Linospadix* and *Laccospadix* were produced by CORSET. These were used for phylogenetic analysis of gene expression. Counts were converted to reads per million to correct for differences in total numbers of reads per individual. Normalised read data were then used for LRT for branch-specific expression shift tests implemented in EVE (Rohlfs and Nielsen 2015). This approach models the evolution of gene expression as an Ornstein–Uhlenbeck process, where the parameter θ_i^{a} represents the optimal expression level for gene *i* in lineage *a*, and $\theta_i^{non\text{-}a}$ represents the optimal expression level for gene *i* in all other lineages. The test compares a null model where $\theta_i^{a} = \theta_i^{non\text{-}a}$ with an alternative model where $\theta_i^{a} \neq \theta_i^{non\text{-}a}$. The models are compared with an LRT to identify significant expression shifts in the focal lineage *a*. Six LRTs were implemented, testing for significant expression shifts in the branches leading to *H. belmoreana* and *H. forsteriana* separately in each of the three tissue types. LRT statistics were used to calculate *P*-values using chi-squared tests with one degree of freedom, and the *P*-values were corrected for multiple testing using FDR. Genes that showed significant expression shifts ($P < 0.05$) following multiple test correction were considered genes of interest in downstream analyses.

Finally, we also investigated whether gene expression within each species was related to variation in soil chemistry. The soil chemistry dataset is highly multidimensional, so we used the first principal component of the soil PCA (above), which separates the two soil types, as an explanatory variable in an analysis of differential expression. We used DESeq2 (Love et al. 2014) to test for differential expression across PC1 of soil chemistry within each *Howea* species and tissue type combination separately. Following Dunning and colleagues (2016), who published the transcriptome data for the two *Howea* species, we also included sampling date (categorised into three collecting trips) as a confounding variable in the model. DESeq2 *P*-values were corrected for multiple testing using FDR.

Results

Transcriptome

For each of the 19 *Linospadix* and *Laccospadix* individuals (Table S1), between 26,173,535 and 45,115,994 paired-end fragments were sequenced using RNA-seq. These were supplemented with previously published RNA-seq data for the two *Howea* species (36 individuals) and data from ten other Arecaceae species. Following read correction and trimming, all newly sequenced individuals had between 25,990,695 and 44,902,175 reads remaining (Table S2). Utilising a multi-assembler and multi-Kmer pipeline using the individual of each *Howea* species with the most reads, we produced a final transcriptome assembly containing 26,972 genes. BUSCO analysis found that the reference transcriptome had 88.7% completeness, a substantial improvement on the 77.5% completeness of the previous assembly by Dunning et al. (2016), demonstrating the utility of our approach. Read-mapping to the transcriptome assembly resulted in between 65% and 84% of reads being uniquely mapped for *Howea* individuals, between 58% and 81% in *Linospadix* and *Laccospadix*, and between 19% and 74% for other palms (Table S3).

Ancestral hybrid swarm followed by sympatric speciation in *Howea*

The two *Howea* species were supported as sister species with a 100% of bootstrap support for every node in our RAxML tree using all concatenated transcripts (Fig. S1). The coalescent-based ASTRAL tree was topologically identical, and posterior probability support for all nodes was high (>0.99). Unlike previous analyses that resolved *Laccospadix* as sister to *Howea* (Savolainen et al. 2006), we found a sister relationship between *Laccospadix* and *Linospadix*. The short branch length between the common ancestors of *Howea-Linospadix-Laccospadix* and *Linospadix-Laccospadix* and high level of ancestral introgression (see below) likely explains the difference between studies. We estimated the divergence time of the two *Howea* species as 3.3 million years ago (Fig. 2a). The second dated tree we produced, in which genes that were phylogenetically incongruent with the species topology were removed (Fig. S2), had a divergence time for the two *Howea* species of 4.4 million years ago. These dates are older than previous estimates (Savolainen et al. 2006) but are still well within the age of LHI, which was formed between 6.4 and 6.9 million years ago (McDougall et al. 1981).

There was no evidence of hybridisation between extant *Howea* and *Linospadix* or *Laccospadix*. Firstly, tree topologies were highly consistent between loci (Fig. 2a) with 95% of gene trees supporting the monophlyly of *Howea* (Fig 2a-b; i.e. 95% quartet support). Quartet support for the other interspecific nodes between species in *Linospadix, Laccospadix* and *Howea* ranged from 38% to 87%, demonstrating a relatively high level of gene tree-species tree discordance within the dataset. Patterson's *D* Statistics were not significantly different from zero when any *Laccospadix* or *Linospadix* species were tested as a potential introgressant (Fig. 3b, Table S4). However, all five-taxon *D*-statistic tests showed evidence for introgression between *Linospadix* or *Laccospadix* and the *ancestor* of *Howea* (scenario (ii) in Fig. 1; Fig. 3a; Table S5). All comparisons in which *Laccospadix* and one of the *Linospadix* species were P_3 and P_4 showed evidence for *Laccospadix* as the introgressing taxon, suggesting that admixture between *Laccospadix* and ancestral *Howea* was either stronger or more recent than it was between *Linospadix* and ancestral *Howea* (Pease and Hahn 2015). There was also evidence for admixture between some *Linospadix* species and ancestral *Howea* when two *Linospadix* species were assigned as P_3 and P_4. The only species with no evidence of admixture was *L. albertisianus*, which is restricted to New Guinea, whereas all other species tested are found on the Australian mainland. Defining the exact patterns of introgression is not possible, but they indicate a complex history in which admixture has occurred independently between ancestral *Howea* and several lineages within *Linospadix-Laccospadix* independently (one interpretation which is consistent with the results is shown in Fig. 3c).

The MDS analysis of all sequence polymorphism data was consistent with the results above. It resolved both *Howea* species as distinct clusters, and outgroup species were equidistant from the two *Howea* species (Fig. 4). If one species was a product of hybridisation between ancestral *Howea* and an outgroup (as shown in scenario (iii) in Fig. 1), it would be expected to cluster more closely to the outgroup than the non-hybrid species. Instead, the reported equidistance in the MDS supports lack of hybridisation during or after speciation in *Howea*. Furthermore, the fact that *Laccospadix* is closer to *Howea* than is its sister taxon *Linospadix*, is in line with the higher level of admixture between *Laccospadix* and ancestral *Howea* (*D*-statistic results above).

Our Mantel tests revealed no evidence for isolation by distance in either *Howea* species (*H. belmoreana*: P = 0.440; *H. forsteriana*: P = 0.588; Fig. S3), indicating that geographically-based isolation within LHI is unlikely.

Taken together, our results strongly support a scenario whereby ancestral *Howea* was part of a hybrid swarm on mainland Australia, but neither hybridisation with outgroups or allopatric isolation within LHI drove speciation following the colonisation of LHI by the ancestral *Howea*.

Soil chemistry drives expression shifts in flowering time genes in *Howea*

Detailed soil chemical analysis (Table S6) revealed substantial differences between the two soil types on LHI, and between specific sites that each species inhabits. The PCA analysis showed that most volcanic sites were clustered, with two outliers (Fig. 5b). In contrast to this, calcareous soils were more diffusely distributed. Notably, both volcanic outliers were sites only occupied by *H. forsteriana* and it is possible these represented intermediate soil types

(Fig. 5a). All sites inhabited by *H. belmoreana* were clustered in the PCA (Fig. 5c). Conversely, *H. forsteriana*-inhabited sites were widely distributed across both of the first two principal components (Fig. 5d).

When individual constituents of soil variation were compared, 21 constituents were significantly different between volcanic and calcareous soils, 15 were significantly different between *H. belmoreana* present versus absent sites, and three were significantly different between *H. forsteriana* present versus absent sites (Fig. S4). Calcareous soil was characterised by significantly higher concentrations of arsenic, boron, calcium, cadmium, sodium, phosphorus and sulphur, high pH and lower water content. Volcanic soils were characterised by significantly higher concentrations of aluminium, cobalt, chromium, copper, iron, potassium, manganese, nickel and zinc, higher water content and neutral pH (Fig. S4). Overall, our soil analysis emphasises that, whereas *H. belmoreana* is an edaphic specialist, *H. forsteriana* is a generalist able to grow on a far broader range of soil types.

When the first principal component of soil chemistry variation was used as an explanatory variable in a differential expression analysis (for each *Howea* species and tissue type separately), very few genes were differentially expressed according to soil variation in *H. belmoreana* (inflorescence: four genes, leaf: two genes, root: seven genes). In *H. forsteriana,* while again there was minimal soil-related differential expression in inflorescences and roots (inflorescence: 15 genes, root: 22 genes), there was a very high level of differential expression in leaves (1,118 genes, Table S7). This included 18 genes known to be involved in flowering time differences in model plants, potentially indicating a link between soil chemistry and flowering time divergence in *Howea*. We found that pairs of *H. forsteriana* individuals from the same soil type were no more closely related than pairs of individuals from different soil types (t-test: $P = 0.78$), indicating that there were no ‘ecotypes’ within *H. forsteriana* that would explain gene expression differences, especially considering that such a large number of genes were differentially expressed.

Adaptive evolution of protein sequence and gene expression is species and tissue specific in *Howea*

We used phylogenetic approaches to search for genes with amino acid substitutions under positive selection (Yang 2007) as well as genes with a significant shifts in expression level (Rohlfs and Nielsen 2015) in the branches leading to each *Howea* species. Genes that have undergone significant expression shifts were unevenly distributed across tissue types and species. In total, 1,736 genes have undergone an expression shift in at least one tissue in at least one species. In inflorescence and root tissue, there were significantly more in *H. forsteriana* (inflorescence: 560 in *H. forsteriana* versus 100 in *H. belmoreana*, Fishers Exact Test *P* < 0.001; root: 131 in *H. forsteriana* versus 20 in *H. belmoreana*, Fishers Exact Test *P* < 0.001), whereas in leaf tissue there were significantly more in *H. belmoreana* (892 in *H. belmoreana* versus 227 in *H. forsteriana*, Fishers Exact Test *P* < 0.001). The tests of positive selection on amino acid sequence revealed that 104 genes likely evolved under positive selection in *H. belmoreana* while 132 were under positive selection in *H. forsteriana*; although this difference in numbers was not significant (*P* = 0.077; Fisher's exact test). Of the 1,972 genes that had any evidence of adaptive evolution from these tests, only 9% were found in more than one of these sets (Fig. S5).

Genes under adaptive evolution are enriched for edaphic and phenology-related functions

There was significant enrichment of 103 GO terms amongst our genes of interest (e.g. genes showing an expression shift or significant sequence-based evidence of positive selection, see Methods), ranging from two GO terms amongst genes which underwent a significant expression shift in the roots of *H. belmoreana* to 27 GO terms amongst genes with any evidence of adaptive evolution (Table S8). Several of these genes were relevant to the speciation scenario of *Howea.* For example, genes with either evidence of an expression shift or with a signature of positive selection in *H. forsteriana* were significantly enriched for the GO term “negative regulation of flower development”, indicating that the differing flowering times of the two species has evolved adaptively (Table S8, GO term assignment and evidence of selection listed in Table S7). Several GO terms likely to be involved in soil preference differences between the two species are also enriched amongst candidate genes (Table S8). The term “response to cadmium ion” is over-represented among genes showing evidence for positive selection in *H. forsteriana*, “response to water deprivation“ is over-represented amongst genes under positive selection in *H. belmoreana*, “cellular calcium ion homeostasis” is over-represented amongst genes which have undergone an expression shift in leaf tissue in *H. forsteriana*, and “cellular response to phosphate starvation is over-represented amongst genes that have undergone a significant expression shift in the inflorescence of *H. forsteriana*. Cadmium, phosphorus, calcium and water content all vary significantly between the soils of the two species. Several less specific GO terms relevant to soil chemistry differences between calcareous and volcanic soil such as “transition metal ion transport”, “regulation of ion transport” and “divalent metal ion transport” were also enriched amongst genes of interest. Furthermore, several terms involved in biotic interactions known to differ between the soil types and species were over-represented. This included several defence related GO terms: “defence response”, “defence response signalling pathway”, “defence response to bacterium”, “response to bacterium” and “regulation of defence response”. Osborne et al. (2018) showed that multiple plant pathogens, both fungal and bacterial, are differentially abundant between the two soil types and even between the two species on the same (volcanic) soil type. Our results here indicate that pathogens may act as selection pressure on the species. Finally, we found that the term “response to karrikin” was significantly over-represented in genes that had experienced an expression shift in *H. forsteriana*. Karrikin response genes are involved in the initiation of arbuscular mycorrhizal fungi symbiosis in rice (Gutjahr et al. 2015). This may be important in *Howea* too, since the two species have divergent soil-specific interactions with arbuscular mycorrhizal fungi, which likely affect their relative fitness on calcareous versus volcanic soils (Osborne et al. 2018).

We also identified 122 genes within our dataset (Tables S7 and S9) that have the potential to pleiotropically link soil adaptation and flowering time in *Howea*. This is based on their mutant phenotypes in model plant species (Table S9; references listed in table). Nine of these showed evidence of adaptive evolution in one or both of the species (Table S10). Two of these nine were amongst the six candidate ‘speciation genes’ identified by Dunning et al. (2016), and another one, *DCL1*, was annotated to the same *Arabidopsis* homologue as in Dunning et al., although we did not return each other as best reciprocal match in our BLAST searches. Two of the six candidates for positive selection in Dunning et al. ($d_N/d_S > 1$), however, did not show evidence of positive selection in our branch-site test, although because this test is highly conservative, this may not necessarily surprising (Gharib and Robinson-Rechavi 2013). In total, combining results from this study with that of Dunning et al (2016) identified 13 candidate ‘speciation genes’ in *Howea* (Table S10).

Discussion

The role of admixture in *Howea*

We found no evidence for gene flow from outgroups into *Howea* following initial colonisation, supporting a model of sympatric speciation. Traditionally, the main difficulty of demonstrating sympatric speciation has been to show that there has been no potential for geographic separation within their habitat such that reproductive isolation could have evolved in allopatry. For this reason, while many potential examples of sympatric speciation may exist on continental landmasses (Sorenson et al. 2003; Hadid et al. 2013; Osborne et al. 2013; Hadid et al. 2014), the most convincing case studies have been found in tiny habitat islands such as crater lakes and oceanic islands (Schliewen et al. 1994; Savolainen et al. 2006;

Papadopulos et al. 2011; Malinsky et al. 2015). Recent evidence of secondary gene flow into several crater lakes hosting cichlid radiations that were previously thought to have evolved in complete sympatry has cast doubt on these examples (Martin et al. 2015; Poelstra et al. 2018). All four cichlid radiations in Cameroon were shown to involve secondary gene flow from external riverine populations, with some even being more closely related to the river populations than other species within their lakes (Martin et al. 2015). Furthermore, in one of these lakes, a genomic region containing several olfactory genes involved in mate choice in cichlids was introgressed prior to the first speciation event (Poelstra et al. 2018). This indicates a mechanism by which secondary gene flow may have played a causative role in their speciation (Poelstra et al. 2018). While some authors considered secondary gene flow to be unlikely to have been causative in some other cichlid radiations (Malinsky et al. 2015), this cannot currently be ruled out. Given that our results make a role for secondary gene flow highly unlikely, *Howea* appears to be one of the strongest examples of sympatric speciation in nature.

Nevertheless, admixture with outgroups may be important in the evolutionary history of *Howea*. Admixture can play a key role in generating genetic diversity on which selection can act (Seehausen 2004; Seehausen 2015; Arnold and Kunte 2017) and has been shown to precede adaptive divergence in multiple taxa. For example in Lake Victoria cichlids, ancestral admixture produced exceptional variation in opsin genes known to be involved in speciation and adaptation, thereby facilitating adaptive radiation in the lake (Meier et al. 2017). In Darwin’s finches, admixture between two species increased the standing genetic and evolutionary responsiveness to fluctuating environmental conditions (Grant and Grant 2014). In light of these and other examples (Pardo-Diaz et al. 2012; Stankowski and Streisfeld 2015), our finding that ancestral *Howea* may have been part of a mainland hybrid swarm opens the possibility that ancestrally-introgressed variation could have been important in their divergence following the colonisation of LHI. This may be critical given the likelihood of a genetic bottleneck upon colonisation of LHI. Unfortunately, identifying introgressed regions requires longer genomic windows than we can derive from transcriptomic data and so it is outside the scope of this study (Martin et al. 2015), although our ongoing sequencing of the *Howea* genome will provide an opportunity to test this hypothesis in the future.

We have interpreted the *D*-statistic results in terms of introgression; however, non-zero results could also be obtained because of ancestral population structure (Pease and Hahn 2015). For example, consider three ancestral populations *A, B* and *C*, in which gene flow between *B* and *C* is stronger than between *A* and *C*. Then, consider that the three populations later diverge into three species as per the phylogenetic tree ((A,B),C), with no gene flow following speciation. The *D*-statistic may then imply, wrongly, introgression between species *B* and species *C*. However, in our case, a population structure-only scenario to explain all our results would be highly complex. The relatively ancient age of the *Howea* species splits makes it implausible that all D_{FOIL} results are explained by ancestral population structure alone, which has been subsequently preserved to the present day. Furthermore, since all Patterson's *D*-statistics were zero, there was no ancestral population structure by which one nascent *Howea* species experienced more gene flow with outgroups than the other. This bolsters the case that ancestral *Howea* was one homogenous population upon colonisation of LHI.

Ball's Pyramid is an unlikely source of geographic isolation

Ball's Pyramid is an inhospitable sea stack 30 km off LHI. It is known that Ball's Pyramid, as well as LHI, had a greater terrestrial extent in the past due to lower sea levels at some time in the Pleistocene (Papadopulos et al. 2011; Woodroffe et al. 2006; Linklater et al. 2018). This raised the possibility of allopatric speciation in *Howea* between populations isolated on LHI and Ball's Pyramid. Papadopulos and colleagues (2011) investigated this possibility and concluded that allopatric divergence of *Howea* was unlikely. The argument was that the distance that would have separated populations of *Howea* on LHI and Ball's Pyramid would not be greater than the current length of LHI, on which populations are not geographically structured in this wind-pollinated genus. Furthermore, a recent demographic modelling analysis supports a model in which gene flow following speciation was high and reduced towards the present over models which include an allopatric period (Papadopulos et al., 2019a). Our new analyses strengthen the case further. While sea level has been lower (and thus the terrestrial extent of both islands has been larger) in the past, our dates indicates that *Howea* split in the Pliocene, when sea level was actually about 22 m higher than today (Dwyer and Chandler 2009; Miller et al. 2012). Erosion of Ball's pyramid and LHI to their current state, where Ball's pyramid is a shear sea stack completely unsuitable for *Howea* colonisation, occurred rapidly following their formation and substantially earlier than the *Howea* split (6-7 mya; Linklater et al 2018). In any case, during periods of lowest sea level, the distance between the islands was around 4 km, whereas we found no evidence of isolation-by-distance within either species with a maximum distance of 5.8 km (Fig. S3). While this does not necessarily mean that pollen travels over these distances, taken together with the evidence from demographic modelling and the likely unsuitability of Ball's Pyramid for *Howea* colonisation at the time of speciation, all available evidence indicate that allopatric isolation between Ball's pyramid and LHI is unlikely to have been responsible for speciation in *Howea*.

The ecological circumstances of speciation in *Howea*

The main ecological difference between the *Howea* species today is their soil preferences. Our analyses highlight the fact that while *H. belmoreana* is a specialist on volcanic soil, *H. forsteriana* is a soil generalist. However, the observation that *H. forsteriana* has not displaced *H. belmoreana* indicates that this generalism comes at a cost. Common garden experiments have provided evidence for this, showing that *H. belmoreana* has a higher survival rate than *H. forsteriana* on volcanic soil (Hipperson et al. 2016). One explanation for this difference is that *H. forsteriana* is less able to form arbuscular mycorrhizal associations on volcanic soil than either *H. belmoreana* on volcanic soil or *H. forsteriana* on calcareous soil (Osborne et al. 2018).

Several soil characteristics significantly differed depending on broad soil type categories (volcanic versus calcareous) and the presence or absence of each species. The differences, which are likely to exert contrasting stresses on plants, comprised changes in essential primary (P, K) and secondary (S, Ca) macronutrients, micronutrients (Al, B, Co, Cu, Fe, Na, Mg, Mn, Ni, Zi) and toxins (As, Cd, Cr). This highlights the fact that a switch in soil is a multidimensional environmental change, and is likely to affect multiple genes, potentially increasing the barriers to gene flow more than simpler environmental switches (Nosil et al. 2017; Riesch et al. 2017).

Given their current divergent soil preferences in the two *Howea* species, it is likely that soil adaptation played a role in their divergence. Since there was admixture between ancestral *Howea* and *L. minor* following the split of *L. minor* and *L.* albertisianus, we can approximately date the colonisation of LHI to between 4.96 and 3.27 mya (the speciation times of *L. minor* and *L. albertisianus*, and *H. belmoreana* and *H. forsteriana*, respectively; this is in line with our dating calculations with r8s). The current calcareous formations on LHI were deposited in the last 350,000 years (Brooke et al. 2003), substantially later than this. However, we had hypothesised that it was more likely that the ancestral *Howea* first colonised volcanic soils, and subsequently moved to calcareous deposits due to the fact that the volcanic soils are older and more similar to mainland soils (Savolainen et al. 2006; Papadopulos et al., 2019b). Since the erosion rate of calcarenite can be around 2.35 mm/year (Balaguer et al. 2019), calcareous deposits from c.a. 3-5 mya would unlikely have survived today. Of course, it is also possible that the ancestral *Howea* colonised calcareous soils first, before colonising volcanic ones. While we do not know the detailed edaphic composition on the island at the time of colonisation and speciation, we can get insight into the ecological history of *Howea* from the proportion and identity of genes that have undergone adaptive divergence in the species.

As would be expected given the relatively old divergence of *Howea*, there are a multitude of sequence and gene expression differences between the two species. Previous research in *Howea* (Dunning et al. 2016) could not polarise the trajectory of change in these loci because of the absence of outgroup data. Under our default scenario in which ancestral *Howea* was a volcanic specialist and speciation was precipitated by an invasion of calcareous soil by the ancestor of *H. forsteriana*, it may be expected that more adaptive evolution should be found in this latter species. This was indeed the case for expression shifts in inflorescence and root, as well as in positive selection on coding sequences (Fig. S5). Furthermore, these genes in *H. forsteriana* were significantly enriched for several soil adaptation-related functions (Table S8). In contrast, significantly more expression shifts were found in the leaves of *H. belmoreana*, and while the reason for this is unclear, it should be noted that the leaf morphology of *H. belmoreana* is unusual relative to both *H. forsteriana* and outgroup species in *Linospadix* and *Laccospadix*, featuring recurved leaves with ascending leaflets (Savolainen et al. 2006) (compare Fig. 2c and 2d).

We note that there are, of course, caveats to consider in gene expression results. Since our samples are wild trees, tissues were collected in different locations and at different times, and information such as plant age and health cannot be known. The environment affects gene expression, and therefore our samples from the wild cannot be standardised as they would be if derived from greenhouse experiments. Nevertheless, we have controlled for sampling date in our models, and we found that the several significant shifts are consistent with our hypotheses.

While our overall results are consistent with a scenario of ancestral volcanic-specialism, alternatives may still be possible. What we can say conclusively, is that the evolution of the two species involved adaptation to the abiotic (Hipperson et al. 2016) and biotic (Osborne et al. 2018) soil variation on the island, so we then investigated a link between soil adaptation and the main component of reproductive isolation, flowering time.

The evolution of reproductive isolation in *Howea*

We identified genes that could have linked ecological adaptation to soil and reproductive isolation via flowering time via two distinct mechanisms: plasticity and pleiotropy. Notably, cadmium (Cd) and zinc (Zi) both differ between the soil types, with cadmium being significantly higher in calcareous soil, and zinc being significantly higher in volcanic soil. Substrate concentrations of these two elements have both been experimentally shown to alter flowering time in *Arabidopsis* (Wang et al. 2012; Przedpelska-Wasowicz and Wasowicz 2013). Therefore, migration between the soil types has the potential to cause a plastic shift in flowering time, a mechanism which may be common in plants (Levin 2009). Such a shift could have instantaneously reduced gene flow between the two nascent *Howea* species in the early stages of their evolution. This scenario is supported in *Howea* given that we found multiple known ‘flowering time genes’, which were differentially expressed according to soil chemistry in *H. forsteriana*. While mean flowering time is not significantly different between the two soil types in *H. forsteriana*, it flowers protoandrously (male flowers are produced earlier) on calcareous soil, whereas on volcanic soil male and female flowering have been found to be synchronous in at least one population (Savolainen et al. 2006), showing that soil chemistry may affect aspects of flowering phenology. Furthermore, there are many examples of a loss of plasticity during or following speciation (Aubret and Shine 2009; Pfennig et al. 2010; Palmer 2014; Levis et al. 2018), so soil-specific flowering time differences could have been more pronounced in the past. Flowering time plasticity would be expected to drive speciation most strongly when pollen dispersal is high but seed dispersal is low, since pollen-mediated gene flow should be directly affected by flowering time whereas seed-mediated gene flow should not. This is likely the situation in *Howea*, which is wind-pollinated but has large and immobile seeds (Savolainen et al. 2006). Therefore soil-mediated flowering time plasticity is a plausible mechanism of speciation in *Howea*. With the combined results of this study and Dunning et al (2016), we also identified 13 genes showing evidence of adaptive expression or sequence divergence in *Howea,* with functions that could pleiotropically link soil adaptation and flowering time (Table S10). While we tentatively consider these genes to be candidate ‘speciation genes’, their potential function was inferred from sequence similarity to model plant genes of known function, which does not guarantee that the orthologous palm genes have the same function. Given the long generation times of palm trees, functional assays using palm mutants are not practical. However, assays involving knockout mutants in model systems for candidate orthologous genes, followed by phenotype rescue using palm genes would provide clearer insight into the function of key genes. This work is ongoing by Savolainen and Turnbull and should further elucidate the genomic basis for speciation in *Howea*.

Author contributions

VS conceived the study with contributions from WJB, CGNT and OGO; OGO, TW, DC and IH collected samples; OGO conducted analyses with help from AC; OGO and VS wrote the manuscript with comments from all authors.

Acknowledgments

We thank Robert Kooyman, Liloa Dunn, Luke Dunning, Helen Hipperson and Rishi De-Kayne for help with field collection, Igor Lysenko for help producing the GIS map, Lucy Smith for some illustrations, Roger Butlin for comments, and NERC, The Leverhulme Trust and the ERC for funding.

References

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22:1600–1607.

Alter S, Bader KC, Spannagl M, Wang Y, Bauer E, Schön CC, Mayer KFX. 2015. DroughtDB: An expert-curated compilation of plant drought stress genes and their homologs in nine species. *Database* 2015:1–7.

Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst. Biol*. 60:685–699.

Arnold ML, Kunte K. 2017. Adaptive Genetic Exchange: A Tangled History of Admixture and Evolutionary Innovation. *Trends Ecol. Evol*. 32:601–611.

Aubret F, Shine R. 2009. Genetic Assimilation and the Postcolonization Erosion of Phenotypic Plasticity in Island Tiger Snakes. *Curr. Biol*. 19:1932–1936.

Baker WJ, Norup M V., Clarkson JJ, Couvreur TLP, Dowe JL, Lewis CE, Pintaud JC, Savolainen V, Wilmot T, Chase MW. 2011. Phylogenetic relationships among arecoid palms (Arecaceae: Arecoideae). *Ann. Bot*. 108:1417–1432.

Balaguer P, Pons GX, Mir-Gual M. 2019. The Rocky Coasts of Balearic Islands: Dynamic Processes, Sediments and Management. In: Morales JA, editor. The Spanish Coastal Systems. Springer International. p. 116–141.

Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A. 2006. Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* 439:719–723.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* 57:289–300.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120.

Bolnick DI, Fitzpatrick BM. 2007. Sympatric Speciation: Models and Empirical Evidence. *Annu. Rev. Ecol. Evol. Syst*. 38:459–487.

Bouché F, Lobet G, Tocquin P, Périlleux C. 2016. FLOR-ID: An interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res*. 44:1167–1171.

Bouckaert RR. 2010. DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics* 26:1372–1373.

Brooke BP, Woodroffe CD, Murray-Wallace C V., Heijnis H, Jones BG. 2003. Quaternary calcarenite stratigraphy on Lord Howe Island, southwestern Pacific Ocean and the record of coastal carbonate deposition. *Quat. Sci. Rev.* 22:859–880.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10:421.

Cerveau N, Jackson DJ. 2016. Combining independent de novo assemblies optimizes the coding transcriptome for nonconventional model eukaryotic organisms. *BMC Bioinformatics* 17:525.

Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X. 2015. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol.* 16:30.

Couvreur TLP, Forest F, Baker WJ. 2011. Origin and global diversification patterns of tropical rain forests: Inferences from a complete genus-level phylogeny of palms. *BMC Biol.* 9:44.

Coyne JA, Orr HA. 2004. Speciation. Sunderland (MA): Sinauer Associates.

Davidson NM, Oshlack A. 2014. Corset: Enabling differential gene expression analysis for de novo assembled transcriptomes. *Genome Biol*. 15:410.

Dieckmann U, Doebeli MO. 1999. On the origin of species by sympatric speciation. *Nature* 400:354–357.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21.

Doebeli M, Dieckmann U, Metz JA, Tautz D. 2005. What we have also learned: adaptive speciation is theoretically plausible. *Evolution* 59:691–699.

Dunning LT, Hipperson H, Baker WJ, Butlin RK, Devaux C, Hutton I, Igea J, Papadopulos AST, Quan X, Smadja CM, et al. 2016. Ecological speciation in sympatric palms: 1. Gene expression, selection and pleiotropy. *J. Evol. Biol*. 29:1472–1487.

Dwyer GS, Chandler MA. 2009. Mid-Pliocene sea level and continental ice volume based on coupled benthic Mg/Ca palaeotemperatures and oxygen isotopes. *Phil. Trans. R. Soc. A* 367:157-168.

Faurby S, Eiserhardt WL, Baker WJ, Svenning JC. 2016. An all-evidence species-level supertree for the palms (Arecaceae). *Mol. Phylogenet. Evol.* 100:57–69.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.

Gharib WH, Robinson-Rechavi M. 2013. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol. Biol. Evol*. 30:1675–1686.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* 40:1178–1186.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. 2013. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol*. 29:644–652.

Grant PR, Grant BR. 2014. Synergism of Natural Selection and Introgression in the Origin of a New Species. *Am. Nat.* 183:671–681.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al. 2010. A Draft Sequence of the Neandertal Genome. *Science* 328:710–722.

Gutjahr C, Gobbato E, Choi J, Riemann M, Johnston MG, Summers W, Carbonnel S, Mansfield C, Yang S, Nadal M, et al. 2015. Rice perception of symbiotic arbuscular mycorrhizal fungi requires the karrikin receptor complex. *Science* 350: 1521-1524.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc*. 8:1494–1512.

Hadid Y, Pavlicek T, Beiles A, Ianovici R, Raz S, Nevo E. 2014. Sympatric incipient speciation of spiny mice *Acomys* at “Evolution Canyon,” Israel. *Proc. Natl. Acad. Sci.* 111:1043–1048.

Hadid Y, Tzur S, Pavlicek T, Sumbera R, Skliba J, Lovy M, Fragman-Sapir O, Beiles A, Arieli R, Raz S, et al. 2013. Possible incipient sympatric ecological speciation in blind mole rats (*Spalax*). *Proc. Natl. Acad. Sci*. 110:2587–2592.

Hipperson H, Dunning LT, Baker WJ, Butlin RK, Hutton I, Papadopulos AST, Smadja CM, Wilson TC, Devaux C, Savolainen V. 2016. Ecological speciation in sympatric palms: 2. Pre- and post-zygotic isolation. *J. Evol. Biol*. 29:2143–2156.

Jombart, T, Ahmed I. 2011. Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27:3070–3071.

Kannan S, Hui J, Mazooji K. 2016. Shannon: An Information-Optimal de Novo RNA-Seq Assembler. 2016. *bioRxiv* https://www.biorxiv.org/content/10.1101/039230v1

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30:3059–3066.

Kautt AF, Machado-Schiaffino G, Torres-Dowdall J, Meyer A. 2016. Incipient sympatric speciation in Midas cichlid fish from the youngest and one of the smallest crater lakes in Nicaragua due to differential use of the benthic and limnetic habitats? *Ecol. Evol*. 6:5342–5357.

Kirkpatrick M. 2001. Reinforcement during ecological speciation. *Proc. R. Soc. B Biol. Sci*. 268:1259–1263.

Kondrashov AS, Kondrashov FA. 1999. Interactions among quantitative traits in the course of sympatric speciation. *Nature* 400:351–354.

Levin DA. 2009. Flowering-time plasticity facilitates niche shifts in adjacent populations. *New Phytol.* 183:661–666.

Levis NA, Isdaner AJ, Pfennig DW. 2018. Morphological novelty emerges from pre-existing phenotypic plasticity. *Nat. Ecol. Evol.* 2:1289–1297.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.

Linklater M, Hamylton S, Brooke B, Nichol S, Jordan A, Woodroffe C. 2018. Development of a seamless, high-resolution bathymetric model to compare Reef morphology around the subtropical island shelves of Lord Howe Island and Balls Pyramid, southwest Pacific Ocean. *Geosciences* 8:11.

Liu J, Li G, Chang Z, Yu T, Liu B, Mcmullen R. 2016. BinPacker: Packing-Based De Novo Transcriptome Assembly from RNA-seq Data. *PLoS Comput. Biol*. 12:1004772.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15:550.

Macmanes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. *Front. Genet.* 5:13.

Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. 2015. Genomic islands of speciation separate Cichlid ectomorphs in an East African crater lake. *Science* 350:1493–1498.

Martin CH, Cutler JS, Friel JP, Touokong CD, Coop G, Wainwright PC. 2015. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution* 69:1406–1422.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol*. 32:244–257.

McDougall I, Embleton BJJ, Stone DBI. 1981. Origin and evolution of Lord Howe Island, Southwest Pacific Ocean. *J. Geol. Soc. Aust.* 28:155–176.

Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat. Commun.* 8:1–11.

Miller KG, Wright JD, Browning J V, Kulpecz A, Kominz M, Naish TR, Cramer BS, Rosenthal Y, Peltier WR, Sosdian S. 2012. High tide of the warm Pliocene: Implications of global sea level for Antarctic deglaciation. *Geology* 40:407–410.

Mirarab S, Warnow T. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31:44–52.

Nei M. 1987. Molecular Evolutionary Genetics. New York: Columbia University Press

Nosil P, Feder JL, Flaxman SM, Gompert Z. 2017. Tipping points in the dynamics of speciation. *Nat. Ecol. Evol*. 1:1–8.

Osborne OG, Batstone TE, Hiscock SJ, Filatov DA. 2013. Rapid speciation with gene flow following the formation of Mt. Etna. *Genome Biol. Evol*. 5:1704–1715.

Osborne OG, De-Kayne R, Bidartondo MI, Hutton I, Baker WJ, Turnbull CGN, Savolainen V. 2018. Arbuscular mycorrhizal fungi promote coexistence and niche divergence of sympatric palm species on a remote oceanic island. *New Phytol*. 217:1254–1266.

Palmer AR. 2014. Symmetry Breaking and the Evolution of Development. *Science* 306:828–833.

Papadopulos AST, Baker WJ, Crayn D, Butlin RK, Kynast RG, Hutton I, Savolainen V. 2011. Speciation with gene flow on Lord Howe Island. *Proc. Natl. Acad. Sci.* 108:13188–13193.

Papadopulos AST, Price Z, Devaux C, Hipperson H, Smadja CM, Hutton I, Baker WJ, Butlin RK, Savolainen V. 2013. A comparative analysis of the mechanisms underlying speciation on Lord Howe Island. *J. Evol. Biol.* 26:733–745.

Papadopulos AST, Igea J, Smith TP, Hutton I, Baker WJ, Butlin RK, Savolainen V. 2019a. Ecological speciation in sympatric palms: 4. Demographic analyses support speciation of *Howea* in the face of high gene flow. *Evolution* in press

Papadopulos AST, Igea J, Dunning LT, Osborne OG, Quan X, Pellicer J, Turnbull C, Hutton I, Baker WJ, Butlin RK, Savolainen V. 2019b. Ecological speciation in sympatric palms: 3. Genetic map reveals genomic islands underlying species divergence in *Howea*. *Evolution* in pressParadis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Pardo-Diaz C, Baxter SW, Joron M, McMillan WO, Jiggins CD, Merot C, Salazar C, Figueiredo-Ready W. 2012. Adaptive Introgression across Species Boundaries in Heliconius Butterflies. *PLoS Genet*. 8:1002752.

Pease JB, Hahn MW. 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Syst. Biol.* 64:651–662.

Peng Y, Leung HCM, Yiu SM, Lv MJ, Zhu XG, Chin FYL. 2013. IDBA-tran: A more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics* 29:326–334.

Peverill KI, Sparrow LA, Reuter DJ. 1999. Soil Analysis: An Interpretation Manual. Clayton, Australia: Csiro Publishing

Pie, MR. The Influence of Phylogenetic Uncertainty on the Detection of Positive Darwinian Selection. 2006. *Mol. Biol. Evol.* 23, 2274-2278.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: An efficient swiss army knife for population genomic analyses in R. *Mol. Biol. Evol*. 31:1929–1936.

Pfennig DW, Wund MA, Snell-Rood EC, Cruickshank T, Schlichting CD, Moczek AP. 2010. Phenotypic plasticity's impacts on diversification and speciation. *Trends Ecol. Evol*. 25:459–467.

Poelstra JW, Richards EJ, Martin CH. 2018. Speciation in sympatry with ongoing secondary gene flow and a potential olfactory trigger in a radiation of Cameroon cichlids. *Mol. Ecol.* 27:4270-4288

Przedpelska-Wasowicz E, Wasowicz P. 2013. Does zinc concentration in the substrate influence the onset of flowering in *Arabidopsis arenosa* (Brassicaceae)? *Plant Growth Regul.* 69:87–97.

R Core Development Team. 2008. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing

Riesch R, Muschick M, Lindtke D, Villoutreix R, Comeault AA, Farkas TE, Lucek K, Hellen E, Soria-Carrasco V, Dennis SR, et al. 2017. Transitions between phases of genomic differentiation during stick-insect speciation. *Nat. Ecol. Evol.* 1:0082.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, et al. 2010. De novo assembly and analysis of RNA-seq data. *Nat. Methods* 7:909–912.

Rohlfs R V., Nielsen R. 2015. Phylogenetic ANOVA: The expression variance and evolution model for quantitative trait evolution. *Syst. Biol*. 64:695–708.

Ryan PG, Bloomer P, Moloney CL, Grant TJ, Delport W. 2007. Ecological Speciation in South Atlantic Finches. *Science* 315:1420–1422.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.

Savolainen V, Anstett M-C, Lexer C, Hutton I, Clarkson JJ, Norup M V, Powell MP, Springate D, Salamin N, Baker WJ. 2006. Sympatric speciation in palms on an oceanic island. *Nature* 441:210–213.

Schliewen UK, Tautz D, Pääbo S. 1994. Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature* 368:629–632.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28:1086–1092.

Seehausen O. 2004. Hybridization and adaptive radiation. *Trends Ecol. Evol.* 19:198–207.

Seehausen O. 2015. Process and pattern in cichlid radiations - inferences for understanding unusually high rates of evolutionary diversification. *New Phytol.* 207:304–312.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol*. 16:1114-1116.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

Song L, Florea L. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience* 4:48.

Sorenson MD, Sefc KM, Payne RB. 2003. Speciation by host switch in brood parasitic indigobirds. *Nature* 424:928–931.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Stankowski S, Streisfeld MA. 2015. Introgressive hybridization facilitates adaptive divergence in a recent radiation of monkeyflowers. *Proc. R. Soc. B Biol. Sci.* 282.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Wang B, Jin SH, Hu HQ, Sun YG, Wang YW, Han P, Hou BK. 2012. UGT87A2, an *Arabidopsis* glycosyltransferase, regulates flowering time via FLOWERING LOCUS C. *New Phytol*. 194:666–675.

Wang WY, Xu J, Liu XJ, Yu Y, Ge Q. 2012. Cadmium induces early flowering in *Arabidopsis. Biol. Plant*. 56:117–120.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38:1358–1370.

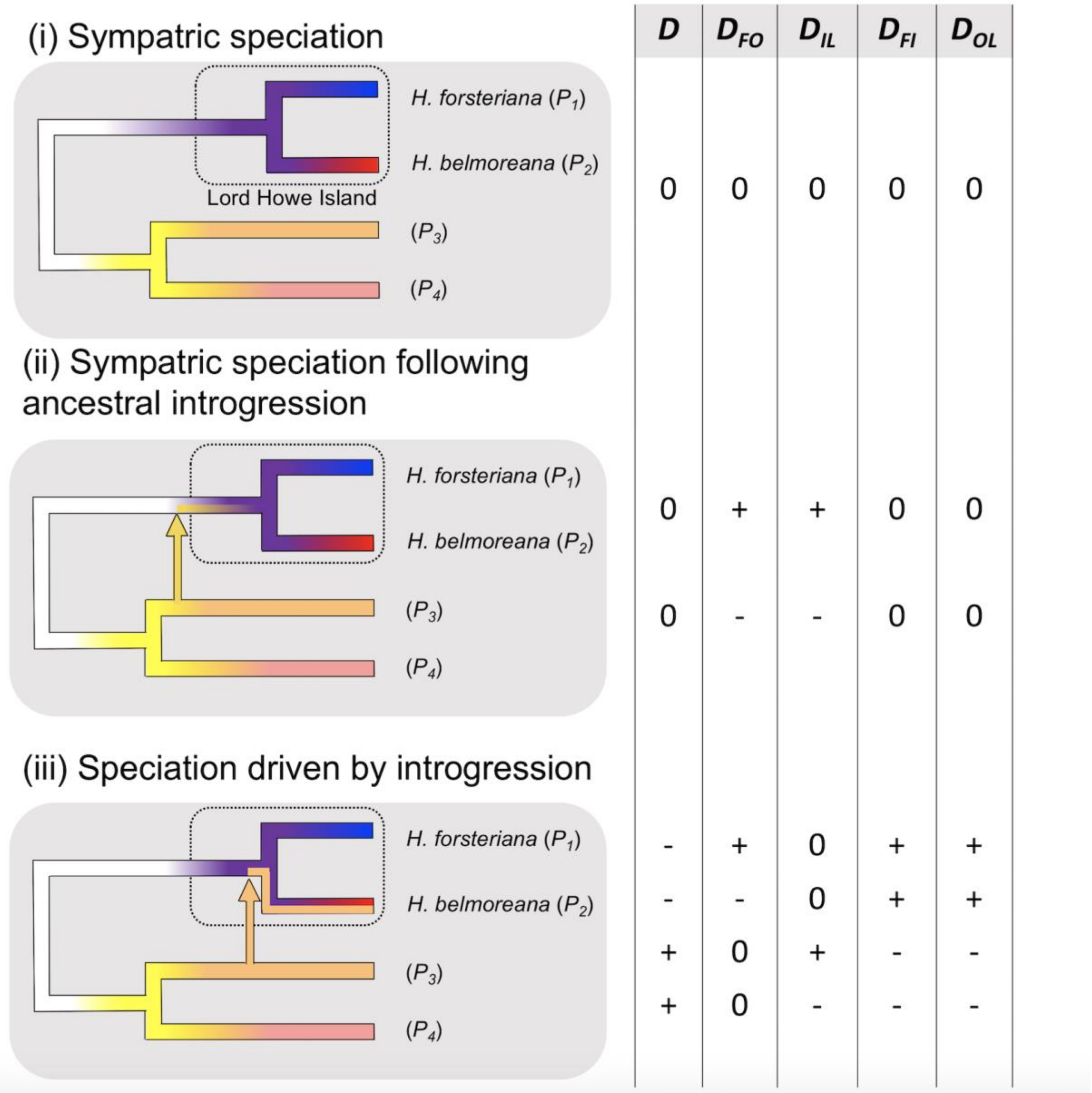
Woodroffe CD, Kennedy DM, Brooke BP, Dickson ME. 2006. Geomorphological evolution of Lord Howe Island and carbonate production at the latitudinal limit to reef growth. *J Coast Res* 22:188–201.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, et al. 2014. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30:1660–1666.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. 24:1586–1591.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:15–30.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22:2472–2479.



D	D_{FO}	D_{IL}	D_{FI}	D_{OL}
0	0	0	0	0
0	+	+	0	0
0	-	-	0	0
-	+	0	+	+
-	-	0	+	+
+	0	+	-	-
+	0	-	-	-

Figure 1. Three possible scenarios of introgression during the evolution of *Howea*, with expected *D*-statistic results for each scenario: (i) ancestral *Howea* speciates allopatrically from the ancestor of *Linospadix* and *Laccospadix*, and later speciate sympatrically on LHI. In this case, all *D* statistics should not be significantly different from zero; (ii) sympatric speciation in *Howea* follows ancestral introgression between *Howea* and *Linospadix* or *Laccospadix* lineages. In this case, *D* should be zero, D_{FO} and D_{IL} should either both be positive (in the case of P_3 introgression) or both negative (in the case of P_4 introgression) whereas D_{FI} and D_{OL} should both be zero; (iii) *Howea* speciation is a direct result of introgression. In this case, a wide range of combinations of D_{FOIL} statistics are possible (see Pease and Hahn, 2015, for details) but *D* will always be significantly positive or negative. The arrows on the phylogenies represent introgression events, colour change represents allele frequency changes over time

and the dotted boxes represents LHI. The table on the right shows the expected sign for Patterson's *D* (Green et al. 2010) and D_{FOIL} statistics (D_{FO}, D_{IL}, D_{FI} and D_{OL}) for each scenario, with multiple possible combinations supporting some scenarios (Following Pease and Hahn 2015). While not represented in this figure, non-zero *D*-statistics can also result from ancestral population structure (see discussion).

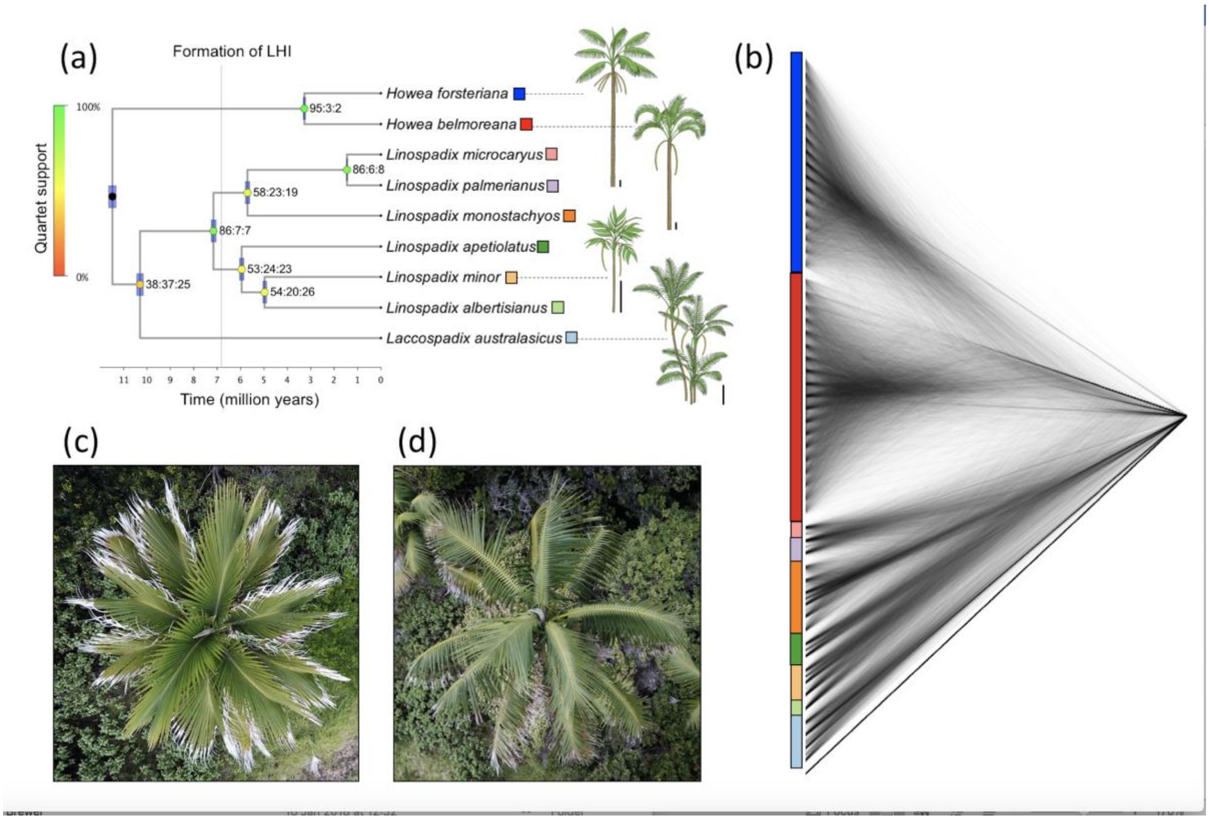


Figure 2. Phylogeny and morphology of *Howea* and its closest relatives. (a) A dated species tree of *Howea*, *Linospadix* and *Laccospadix*. Node labels show the percentage of gene trees supporting the dominant topology followed by the other two possible (unrooted) topologies. Nodes are coloured by proportion of gene trees supporting the dominant topology and blue node bars show 95% confidence intervals of node ages. Illustrations of selected species are drawn to the right of the phylogenetic tree, with a 0.5 m scale bars to show approximate relative heights of the species. A DensiTree plot (b) shows the level of concordance between gene trees. Following filtering for low confidence nodes, each unique gene tree topology is transparently plotted such that gene tree discordance is apparent. Coloured bars denote species and are coloured according to the boxes beside the species names on panel (a). Photographs of the crowns of *H. belmoreana* (c) and *H. forsteriana* (d) show differences in leaf morphology.

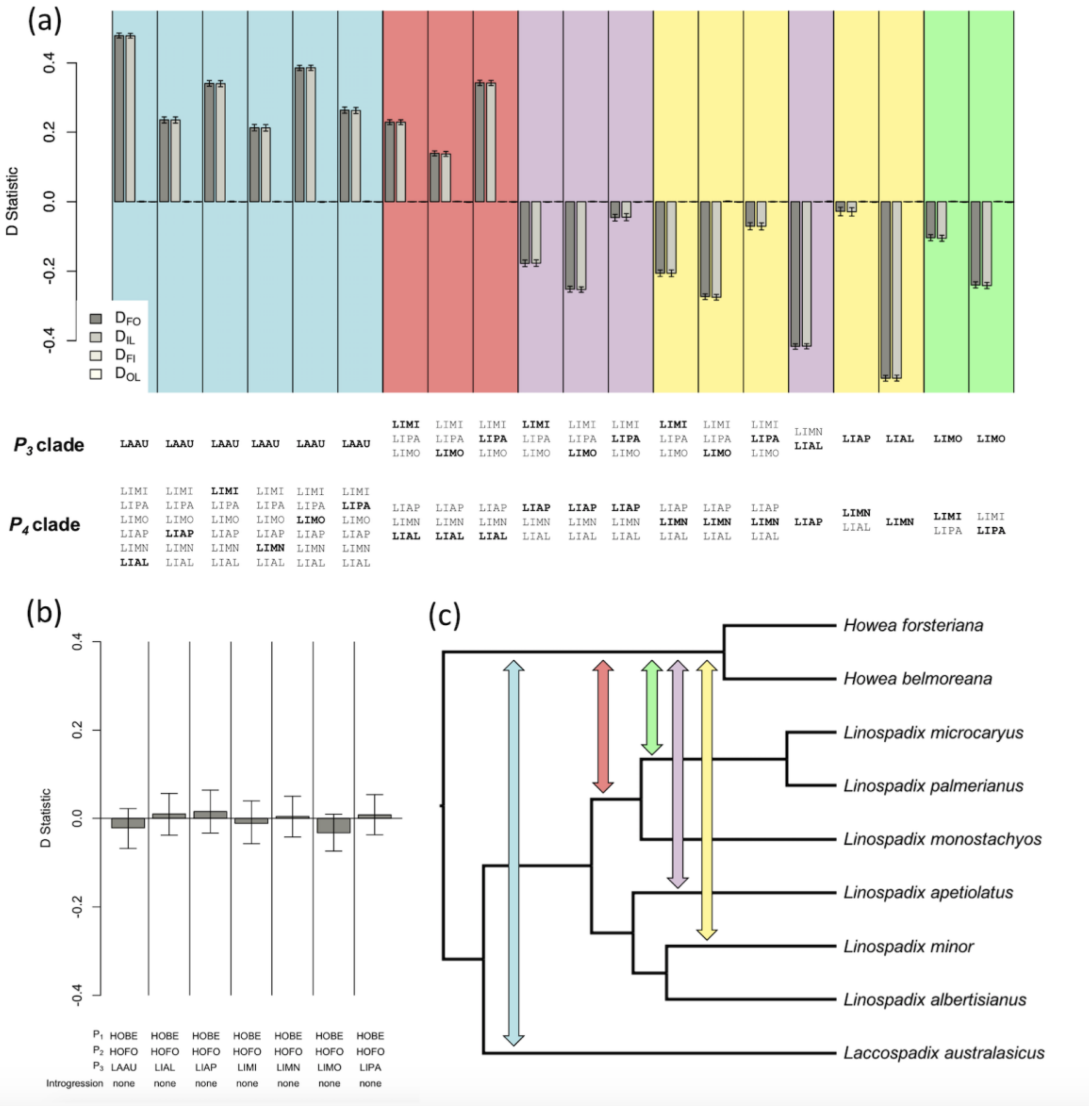


Figure 3. Introgression between taxa. Values and 95% confidence intervals are shown for all four D_{FOIL} statistics (a), and Patterson's *D* statistic (b) for each testable subtree. The identity of species in clades represented by P_3 and P_4 are shown below each test, with the actual species tested highlighted in bold. Tests unanimously support introgression between *Linospadix* or *Laccospadix* species and ancestral *Howea* (a), but not between extant species of *Howea* and any of the outgroups (b). Species abbreviations are as follows: HOBE: *Howea belmoreana*, HOFO: *H. forsteriana*, LAAU: *Laccospadix australasicus*, LIAL: *Linospadix albertisianus*, LIAP:*L. apetiolatus*, LIMI: *L. microcaryus*, LIMN: *L. minor*, LIMO: *L. monostachyos*, LIPA: *L. palmerianus*. One interpretation of the results which minimises the number of introgression events is shown in panel (c), with introgression events shown on the tree as coloured arrows corresponding to the tests that support them in panel (a). The horizontal positions of introgression arrows are arbitrary and do not reflect the

timing of introgression. While we have interpreted the results in terms of introgression in panel (c), ancestral population structure can also lead to non-zero *D*-statistics.

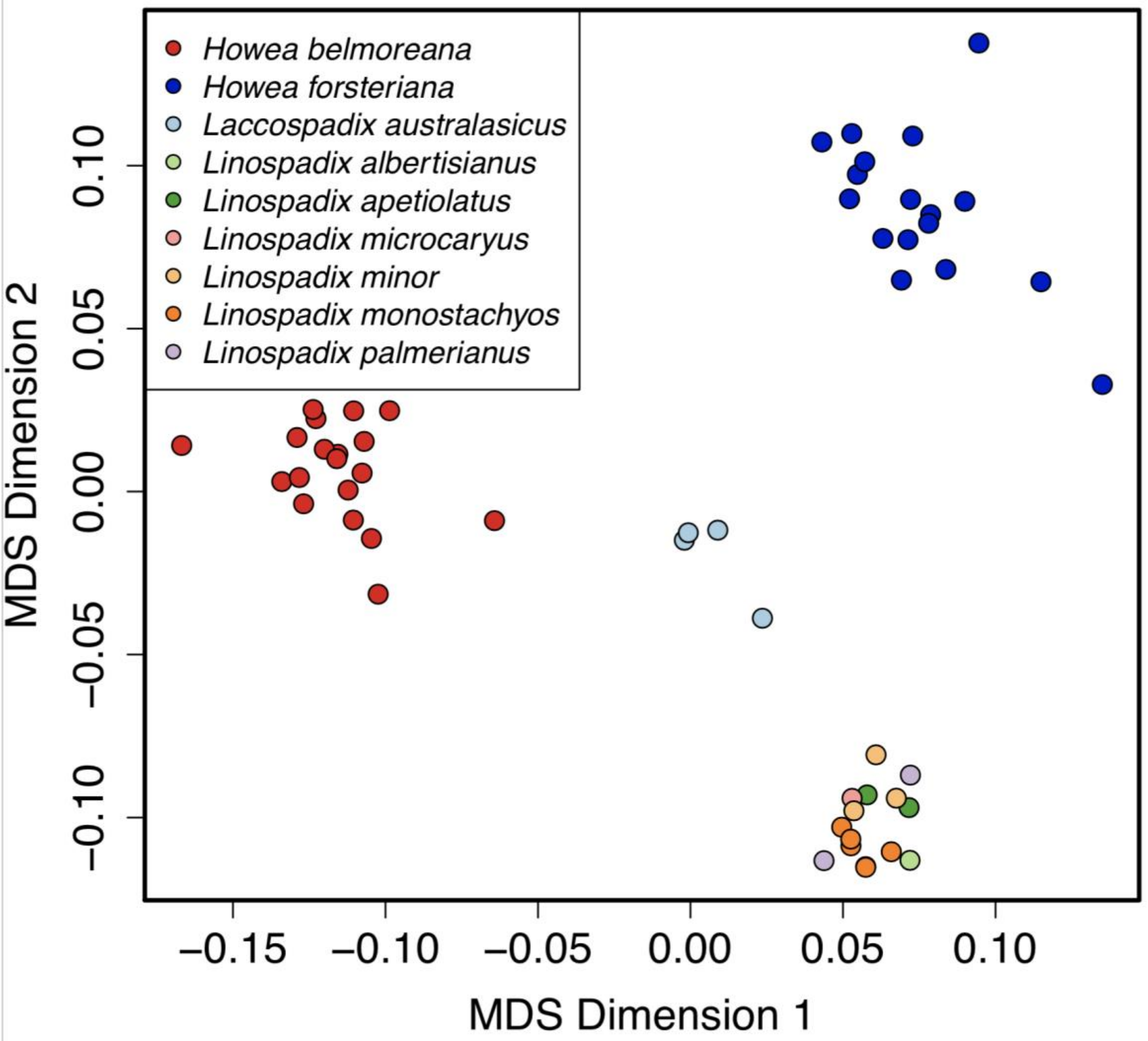


Figure 4. Genetic clustering amongst *Howea* and their outgroups revealed by multidimensional scaling of all single nucleotide polymorphism data. The two *Howea* species are equidistant from outgroup species, however *Laccospadix* is closer to *Howea* than its sister genus *Linospadix* (see text).

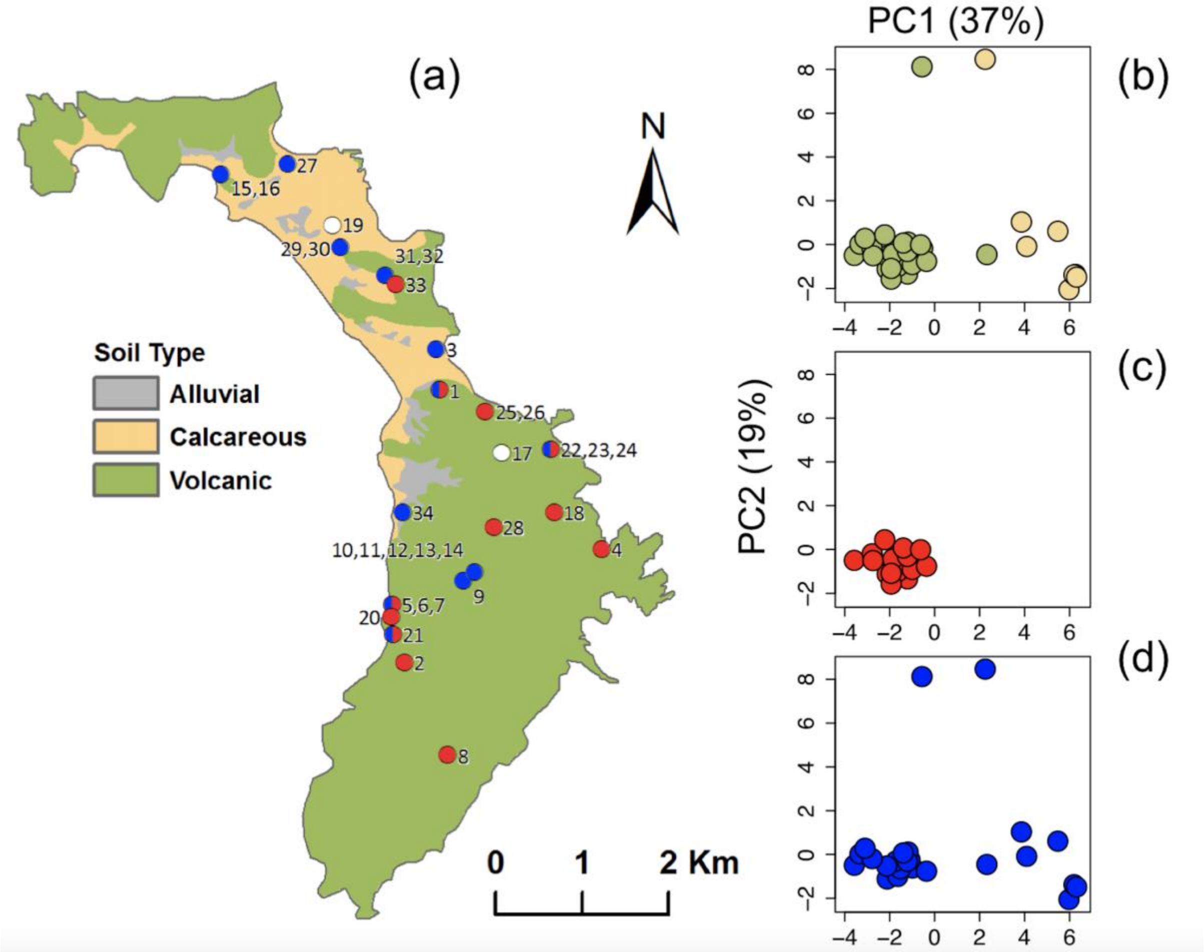


Figure 5. Soil characteristics of the two *Howea* species habitats on LHI. (a) A map of the island showing broad soil classifications and sampling locations. Sampling sites with only *H. forsteriana* are shown in blue, with only *H. belmoreana* are shown in red and those with both species are shown half blue and half red. (b-d) The first two principle components (PCs) of a PCA of normalised soil metrics for water content, pH, concentrations of 20 acid extractable elements and four DTPA-extractable micronutrients. Plot (b) is coloured by soil type (volcanic: green, calcareous: yellow) and numbers on the plot correspond to those in (a). (c) shows all sites with *H. belmoreana* present (red) and (d) shows all sites with *H. forsteriana* present (blue).