BRIEF COMMUNICATION

Ecological speciation in sympatric palms: 3. Genetic map reveals genomic islands underlying species divergence in *Howea*

Abstract

Although it is now widely accepted that speciation can occur in the face of continuous gene flow, with little or no spatial separation, the mechanisms and genomic architectures that permit such divergence are still debated. Here, we examined speciation in the face of gene flow in the *Howea* palms of Lord Howe Island, Australia. We built a genetic map using a novel method applicable to long-lived tree species, combining it with double digest restriction-site associated DNA sequencing of multiple individuals. Based upon various metrics, we detected 46 highly differentiated regions throughout the genome, four of which contained genes with functions that are particularly relevant to the speciation scenario for *Howea*, specifically salt and drought tolerance.

Introduction

We investigated the genomic basis of speciation in *Howea* palms, which is a genus of only two species endemic to a minute oceanic island, Lord Howe Island (LHI), in the Tasman sea. LHI sits 600 km off mainland Australia and is less than 16 km^2, meaning that for any pair of endemic sister species that have diverged within the lifetime of the island (6.9 my), an allopatric phase in their divergence is unlikely (Savolainen *et al.* 2006; Papadopulos *et al.* 2011). Hence, *Howea* is a solid example of speciation in sympatry (Savolainen *et al.* 2006; Coyne 2011; Papadopulos *et al.* 2011, 2019). Furthermore, it has been hypothesised that the two *Howea* species diverged in sympatry as a result of ecological speciation facilitated by soil adaptation and a shift in flowering phenology (Fig. 1; Savolainen *et al.* 2006; Babik *et al.* 2009; Papadopulos *et al.* 2011, 2013b, 2014; Hipperson *et al.* 2016). *Howea* is widespread on LHI, although *H. belmoreana* is restricted to the older volcanic rocks whereas *H. forsteriana* is found predominantly on Pleistocene calcareous deposits (calcarenite) around the coast (Savolainen *et al.* 2006; Woodroffe *et al.* 2006; Papadopulos *et al.* 2013). Marked flowering time differences between the species indicate that prezygotic isolation is now strong and current levels of gene flow are low (Savolainen *et al.* 2006; Babik *et al.* 2009; Dunning *et al.* 2016; Hipperson *et al.* 2016; Papadopulos *et al.* 2019). Indirect evidence of post-zygotic isolation due to selection against juvenile hybrids supports the hypothesis that divergent selection has influenced the speciation process (Hipperson *et al.* 2016). Given that the distributions of *Howea* palms overlap extensively and that *Howea* is wind pollinated, speciation is likely to have occurred in the face of gene flow, which has reduced quickly as divergence progressed (Savolainen *et al.* 2006; Babik *et al.* 2009; Papadopulos *et al.* 2011, 2013, 2014, 2019). Here, we built a genetic map using a novel method applicable to long-lived tree species, combining it with double digest restriction-site associated DNA sequencing of multiple individuals, and we then examined the landscape of genomic differentiation that has arisen during and after speciation in *Howea* palms.

Material and Methods

DNA EXTRACTION

For linkage mapping, a single, wild *H. belmoreana* tree was selected on LHI, and leaf tissue was collected and preserved in silica gel. Ninety-four immature seeds were collected from this tree, dissected, and the endosperm tissue was removed and preserved in RNAlater (Sigma-Aldritch). Genomic DNA was then extracted using CTAB (Doyle & Doyle 1987) and purified using a caesium chloride gradient and dialysis. DNA samples were further cleaned and concentrated using DNeasy Mini spin columns (Qiagen). For the genome scan, leaf tissue was collected and preserved in silica gel from 42 *H. belmoreana* and 54 *H. forsteriana* individuals sampled at Far Flats, a plot on LHI where both species co-occur (Papadopulos *et al.* 2019). For shotgun sequencing, a single, wild collected *H. forsteriana* individual was used. Genomic DNA was extracted from these 97 individuals using DNeasy Plant Mini kits (Qiagen).

GENOTYPING AND LINKAGE MAP

Double digest RAD-sequencing (ddRAD) was performed following Papadopulos *et al.* (2019). For the map, we genotyped a mother tree and 94 of its seeds. During the formation of the female megagametophyte a single cell undergoes meiosis and programmed cell death eliminates three of the four descendent haploid spores (Fig. S1). Three sequential mitotic nuclear divisions take place in the remaining megaspore to produce eight nuclei. Cellular division produces seven cells that make up the embryo sac, one of which – the central cell - contains two polar nuclei (Sundaresan & Alandete-Saez 2010). When fertilised, this homo-diploid cell develops into the triploid endosperm containing a single copy of the paternal genome and two identical copies of the maternal genome. Identification of which maternal allele is inherited by the offspring at any given heterozygous position in the mother was achieved by ddRAD sequencing of the maternal and endosperm tissue. The 2:1 ratio of maternal to paternal alleles is maintained in the relative read depth of alleles at each locus in the endosperm, allowing the maternally inherited allele at each locus to be determined in each seed sample (Fig. S1). The raw sequencing data were processed and individuals genotyped using components of the *STACKS* (Catchen *et al.* 2011) pipeline, *perl* and *R* scripts (R Development

Core Team 2019). The 'process-radtags' component of *STACKS* was used to de-multiplex the barcoded samples in each library, remove tags of low quality, with ambiguous barcodes or missing base calls, and truncate each sequence to 95 bp. The paired ends of each read were then merged into a single contiguous sequence to minimise the inclusion of paralogous sequences in the same RAD loci in subsequent steps. The genotyping process was composed of four main steps: (i) construction of a reliable, high coverage catalogue of heterozygous sites in the maternal tree; (ii) genotyping of endosperm tissue at these sites; (iii) addition of loci/haplotypes present in trees from the LHI site to the maternal catalogue; and (iv) genotyping of the wild trees. First (i), to remove highly similar clusters of *STACKS* and error prone loci from the maternal dataset the *STACKS* pipeline was run using at least five exactly matching reads to create a stack, allowing one mismatch between stacks to create a locus, allowing up to 200 stacks to form a single locus, disabling the deleveraging algorithm and disabling haplotype calling from secondary reads. Reads assigned to loci composed of more than two haplotypes were then removed from the dataset. The remaining reads were then processed using the denovo_map.pl perl wrapper for *STACKS* to generate a catalogue of loci and haplotypes present in the mother using at least 50 exactly matching reads to create a stack and allowing three mismatches between stacks to create a locus. (ii) Reads for each endosperm tissue were assembled into loci using *USTACKS* (minimum depth to create a stack = 10, mismatches = 3) and these were then mapped to the maternal catalogue using *SSTACKS*. For all heterozygous loci in the mother, the two haplotypes in the endosperms were randomly assigned as an A or B allele and the read depths of haplotypes were extracted for each seed using custom *perl* scripts. To determine the maternally inherited allele (A or B) in each seed, the relative read depth of the A allele at each locus (read depth of A/read depth of A + B) was analysed using the *kmeans* clustering algorithm in R with 4 predefined clusters (corresponding to triploid genotypes of AAA = 1.00, AAB = 0.66, BBA = 0.33 and BBB = 0.00). (iii) To expand the catalogue to encompass haplotypes present in both *Howea* species, the Far Flats samples were assembled into loci using *USTACKS* (-m20, -M3) and these stacks were merged into the existing catalogue allowing 3 mismatches between loci in different individuals. (iv) To genotype the Far Flats individuals, loci were assembled with lower coverage in *USTACKS* (-m5, -M3) and these stacks were mapped to the catalogue loci. For the genome scan analyses, haplotypes of these individuals were extracted for loci included in the linkage map. Genotypic data for the *H. belmoreana* seeds were initially processed using the *R/qtl* (Broman *et al.* 2003) package. Four individuals were excluded due to high levels of missing data. After exclusion of these samples, loci with more than 22 missing genotypes out of 90 progeny were also removed from further analysis. The remaining 3,772 loci were phased and assembled into linkage groups using the *formlinkagegroups* function with a minimum logarithm of odds threshold of 7.0 and maximum recombination fraction of 0.25. The loci within each assembled linkage group were then ordered in *JoinMap* v4.1 (Kyazma) using the regression mapping algorithm (three rounds), and inter-marker distances were calculated in centimorgans (cM) using the Kosambi mapping function. The mean coverage of mapped loci was 1,117 reads in the mother (s.d. +/- 1,145) and 176 (s.d. +/- 89) in the endosperm. Unequivocal homozygote genotypes accounted for 51% of endosperm allele calls, 32% of calls were derived from proportional differences between alleles and 17% were treated as missing data.

IDENTIFYING GENOMIC ISLANDS

Differentiation (i.e., F_{ST}; Weir & Cockerham 1984) between *H. belmoreana* and *H. forsteriana* was calculated at each ddRAD locus using the *diveRsity* package in *R* (Keenan *et al.* 2013). Divergence (d_{XY}) was also calculated for each locus using equation 10.20 of Nei (1987). Genome-wide distributions of F_{ST} and d_{XY} were generated using a local Gaussian kernel smoothing technique within each chromosome (Hohenlohe *et al.* (2010). Kernel smoothing was performed using the *ksmooth* function in *R* with a bandwidth value of 2 cM, defined as the standard deviation of the kernel. The bandwidth of 2 cM was chosen because it is similar to the average distance between the markers (1.6 cM). To identify genomic islands with both high F_{ST} and d_{XY}, *fastsimcoal2* was used to generate a null distribution of expected F_{ST} and d_{XY} values (i.e. without selection) that incorporated a demographic scenario (Papadopulos *et al.* 2019) and the position of markers in the genetic map. Under the best fitting *fastsimcoal2* model (a model with initial strong gene flow followed by a reduced gene flow, model 5 in Papadopulos *et al.* 2019; see Table S1 for parameters), we simulated the same number of 190 bp DNA fragments as contained in the genetic map with the positions preserved by separating simulated loci by the same recombination distances as in the map (i.e., recombination rate between loci varied across the genome). Within fragments, recombination was fixed at the genome wide average (6.85 x 10^{-9} base^{-1} generation^{-1}). Each chromosome was simulated 1,000 times separately. Kernel smoothed F_{ST} values were calculated for each simulation using the methods applied to the observed data above. These data were used to calculate P-values at each centimorgan, and outlier F_{ST} islands were identified at alpha = 0.05. Outlier d_{XY} regions were identified as those positions with d_{XY} values in the 90th percentile of the observed data. Observed rather than simulated data was used as the random assignment of mutation in the simulation lead to very broad confidence intervals for d_{XY}. To define the full extent of the high F_{ST} + d_{XY} islands – these islands were joined or extended only if the position next to an F_{ST} + d_{XY} outlier had a high (but not significant) probability of being an F_{ST} outlier (assessed using Hidden Markov Models, HMM) and also coincided with a region of high d_{XY}. To do this, F_{ST} P-values were converted into z-scores using *qnorm* and three hidden states were fitted to detect regions of the genome with low, intermediate and high probabilities of belonging to an outlier region. For each state, a Gaussian distribution of the z-scores was assumed. Means and standard deviations for each hidden state, as well as the transition matrix defining probabilities of transferring from one state to another, were all estimated from the data. Direct transitions from low to high states were not permitted. Parameters were estimated using the Baum-Welch algorithm and the probable sequence of hidden states was determined from the data and parameter estimates using the Verbiti algorithm. The results of the HMM procedure were only used to define the size of the regions identified at $P < 0.05$, rather than locate the position of islands. An island was only extended when an adjacent position (i) was assigned the high F_{ST} state by the HMM and (ii) was an outlier d_{XY} position.

ESTIMATION OF RECOMBINATION RATE

To estimate recombination rates in genomic islands versus the rest of the genome, we assembled a draft genome of *Howea*. We estimated the genome size of *H. forsteriana* and *H. belmoreana* following the one-step flow cytometry procedure described by Doležel *et al.* (2007). Then, a shotgun genome assembly was performed for *H. forsteriana*. A total of 432.98 Gigabases (Gb) of cleaned, paired-end, Illumina reads (49-150 bp reads, insert sizes = 170 bp, 250 bp, 800 bp, 2 kilobases (kb), 5 kb, 10 kb and 20 kb) were assembled into genomic contigs using *SOAPdenovo* (Luo *et al.* 2012). *SSPACE* (Boetzer *et al.* 2011) was then used to extend and scaffold contigs. Summary statistics for the shotgun assembly are provided in the supplementary material (Table S2). *BUSCO* (Simão *et al.* 2015) analysis was performed in genome mode using the Embryophyta BUSCOs (Benchmarking Universal Single-Copy Orthologs) to assess genome completeness. Consensus sequences of the ddRAD markers included in the genetic map were mapped to genomic scaffolds using *BLASTn* (Camacho *et al.* 2009), retaining only the best hits. As suggested by Tang *et al.* (2015), scaffolds (n = 3,980 and total length = 0.42 Gb) were ordered based on the average map location of the ddRAD markers for each scaffold. The physical length of each chromosome was calculated using the proportion of the total length of scaffolds (0.42 Gb) that mapped to that chromosome. As only 13.3% of the genome is covered by our scaffolds, we then estimated the length of that chromosome as the corresponding proportion of the total genome size of *H. forsteriana* (estimated here as 3.15 Gb). Finally, we calculated the recombination rate as the genetic distance from the map divided by the estimated physical length of a given genomic region as above (chromosomes and genomic islands) in 10 cM sliding windows.

GENE CONTENT IN GENOMIC ISLANDS

To assign transcripts from the *Howea* reference transcriptome (Dunning *et al.* 2016) to genomic scaffolds, *BLASTn* was used with *max_target_seqs*=1 and an E-value cut-off of 1 x 10^{-20}. Only the highest scoring match for each transcript was retained. Using transcriptome data from Dunning *et* *al.* (2016), the proportions of transcripts showing evidence of differential expression or signatures of selection within and outside speciation islands were compared using Fisher's Exact Tests. This was done using highly differentiated genes ($F_{ST} > 0.8$), genes with evidence for positive selection ($d_N/d_S > 1$), and differentially expressed genes in any tissue (Dunning *et al.* 2016). The transcripts from Dunning *et al.* (2016) were mapped to genomic scaffolds using *BLAT* with default settings (Kent 2002). Alignments were then filtered to include only the best hit for each transcript and alignments covering 80% of the transcript. Filtered *BLAT* alignments were then converted to *AUGUSTUS* hints (Stanke *et al.* 2006). *AUGUSTUS* was used to predict genes in the genomic scaffolds, using the transcript-derived hints and annotation training files from *Zea mays* using the following settings: no UTR prediction, no in-frame stop codons, and gene prediction on both strands. The resulting predicted amino acid sequences were *BLASTp*-searched against the *Arabidopsis thaliana* proteome (Araport11_genes.201606.pep downloaded on 31/01/17) and only the best scoring hit from each predicted amino acid sequence was retained. Gene ontology (GO) enrichment for genomic islands was compared to all scaffolded transcripts; this was performed for both transcriptome-based (Dunning *et al.* 2016) and the above *AUGUSTUS*-based genome annotations. To test for enrichment of GO terms among genes within particular genomic islands we used the R package *TopGO* (Alexa *et al.* 2006) using the "elim" algorithm and Fisher's Exact tests to assess significance. Preliminary assessment of gene functions in genomic islands was made from The Arabidopsis Information Resource (TAIR) descriptions of gene functions, GO terms and associated references. Further published records of functional assessments were acquired from the TAIR known phenotypes database (https://www.arabidopsis.org), the drought stress genes database (http://pgsb.helmholtz-muenchen.de/droughtdb/drought_db.html) and the flowering interactive database (http://www.phytosystems.ulg.ac.be/florid/). Finally, systematic web searches were performed using gene names with and without the terms "stress" and "flowering", given the speciation scenario for *Howea* (Fig. 1).

Results and discussion

GENE FLOW AND GENOMIC DIFFERENTIATION

The linkage map contains 3,772 ddRAD loci ordered onto 16 linkage groups corresponding to the 16 pairs of chromosomes in *Howea* (Savolainen *et al.* 2006) and spanning 2,399 cM (0.70 cM/Mb or 1.42 Mb/cM; Fig. 2 and S2, Table S3). Across the map, we observed a positive correlation between F_{ST} and d_{XY} ($p < 0.0001$, $r^2 = 0.18$; Figs. S3 and S4), which, in these relatively recently diverged species, may be an indication that gene flow has played a role in shaping genomic differentiation. To characterise the genomic landscape, F_{ST} and d_{XY} were calculated for 1,498 high-quality ddRAD loci in the map, which were present in both species (genome wide, F_{ST} = 0.46, d_{XY} = 7.6 x 10^{-3}). Genetic differentiation during sympatric speciation should be substantially greater for loci that have been subject to divergent selection than for loci in neutral regions (Wu 2001). In the course of speciation with gene flow, genomic regions in proximity with those barrier loci that are the target of selection may experience reduced effective gene flow. Meanwhile, the rest of the genome would still be subject to the homogenising effects of genetic exchange (Nosil *et al.* 2008; Via & West 2008; Soria-Carrasco *et al.* 2014). This may lead to a pattern of elevated differentiation (F_{ST}) and divergence (d_{XY}) in regions containing barrier loci compared to the rest of the genome (Hohenlohe *et al.* 2010; Ellegren *et al.* 2012; Nadeau *et al.* 2012; Martin *et al.* 2013; Renaut *et al.* 2013; Poelstra *et al.* 2014). These patterns of heterogeneous genomic differentiation have been used in attempts to identify regions of the genome that harbour barrier loci responsible for adaptation and speciation (Ellegren *et al.* 2012; Nadeau *et al.* 2012; Martin *et al.* 2013; Poelstra *et al.* 2014). However, there are several complicating factors that may mean that these regions do not act as barriers to gene flow (see sections below for discussion of these; Noor & Bennett 2010; Turner & Hahn 2010; Cruickshank & Hahn 2014; Ravinet *et al.* 2017). Here, high-F_{ST} islands (mean F_{ST} = 0.87, range = 0.64 – 0.99, mean d_{XY} = 9.9 x 10^{-3}, range = 5.0 x 10^{-3} – 1.7 x 10^{-2}) were numerous (38 islands), relatively small (mean size = 1.7 cM, range = 1 – 5 cM) and accounted for 3.3% of the genome (total length = 80 cM; Table S4). In contrast, we detected only eight islands with both higher F_{ST} and d_{XY} (high-F_{ST}+d_{XY}) than the rest of the genome (mean F_{ST} = 0.88, range = 0.58 – 0.98, mean d_{XY} = 1.5 x 10^{-2}, range = 1.2 x 10^{-2} – 2.0 x 10$^{-2;}$ Welch's t-test, P < 0.0001), which were located on seven pairs of chromosomes. These high-F_{ST}+d_{XY} genomic islands were, on average, marginally larger (mean size = 2.38 cM, range = 1 – 4 cM; Mann-Whitney U test, P = 0.05) than other high-F_{ST} islands, and represented only 0.8% of the genome (19 cM, Table S5). A permutation test showed that high-F_{ST}+d_{XY} islands were not the result of high-F_{ST} and high-d_{XY} positions co-occurring by chance (P < 0.0001). These high-F_{ST}+d_{XY} islands are more likely to have been involved in speciation in the face of gene flow than islands with high-F_{ST} but no elevation in d_{XY} (Hohenlohe *et al.* 2010; Ellegren *et al.* 2012; Nadeau *et al.* 2012; Martin *et al.* 2013; Renaut *et al.* 2013; Poelstra *et al.* 2014). In both species, nucleotide diversity (π) was significantly lower in high-F_{ST} islands than the genome average (Table S6) as has been observed in other plants (Chapman *et al.* 2016), but was only lower in high-F_{ST}+d_{XY} islands for *H. forsteriana*. In seven out of eight of high-F_{ST}+d_{XY} islands, π was substantially lower in *H. forsteriana* than in *H. belmoreana*, a possible indication of a selective sweep having taken place in this species. The generally small size of high-F_{ST}+d_{XY} islands indicates that these islands did not expand gradually over time, as would be expected under divergence hitchhiking theory when gene flow is ongoing (Via 2009; Feder *et al.* 2012; Rafajlović *et al.* 2016). Our results contrast with those in a comparable analysis of another case of sympatric speciation, i.e., the cichlids of lake Massoko in Tanzania (Malinsky *et al.* 2015). In a whole genome analysis of these fish, and measuring F_{ST} and d_{XY} as here, 55 high-F_{ST}+d_{XY} islands were identified. This is substantially more than in the palms here and from a much smaller genome. Similar numbers of islands were found in flycatchers and many more in other systems (Ellegren *et al.* 2012; Renaut *et al.* 2013; Soria-Carrasco *et al.* 2014). This is likely to be, in part, due to the resolution of our map as we have probably missed finer scale islands (< 1cM). However, it is noteworthy that in the Massoko cichlids, 27 islands formed clusters extending 5 - 45 cM across five linkage groups, which are larger than the islands detected in *Howea*, suggesting the resolution we use may be sufficient to detect a substantial proportion of the larger islands in our system.

An alternative explanation for high-F_{ST}+d_{XY} islands has been proposed recently (Guerrero & Hahn 2017). Guerrero and Hahn showed that these regions can be the result of balanced polymorphisms in the ancestral population that have been ‘sieved’ by the speciation process when different alleles are fixed in each descendent population. Because ancestral balanced polymorphisms have had longer to accumulate divergence than those that only diverged following speciation, this process is expected to have the most pronounced effect early in speciation ($t < 2N_{\mathrm{e}}$; were t = divergence time and N_e = effective population size). If sieved polymorphisms are responsible for these islands, then d_{XY} in these regions should substantially exceed the expected level of d_{XY} based on the time since speciation, which can be calculated as $\mathrm{E}(d_{\mathrm{XY}}) = 2\mu t + \Theta_{\mathrm{ANC}}$ (where μ = the neutral mutation rate and Θ_{ANC} = the ancestral level of diversity). Using respective estimates of t and μ of 266,136 and 1.3×10^{-8} from (Papadopulos *et al.* 2019) and assuming $\Theta_{\mathrm{ANC}} = 0$ (because of the bottleneck caused by long-distance colonisation of LHI), we arrive at an expected d_{XY} of 6.9×10^{-3}. This differs from the level of d_{XY} estimated from our map by only 0.0007, which may constitute the contribution of ancestral polymorphism to our estimate. Alternatively, this small discrepancy may arise if our estimate of t is wrong by approximately 100,000 years (within the 95% CI of t) or if our d_{XY} estimate is derived only from the subset of the data used for the demographic inference that was also included in the map. These estimates are substantially lower than our mean observed d_{XY} for high-F_{ST}+d_{XY} islands (1.5×10^{-2}), but similar to that of the high-F_{ST} only islands. These data may point to a role for sieved balanced polymorphisms in the origin of our high-F_{ST}+d_{XY} islands, and that sympatric speciation may have been reliant on existing genetic variation in the ancestral population. However, we cannot rule out the possibility that some of these high-F_{ST}+d_{XY} regions may contain sieved polymorphisms that did not play a direct role in the speciation process.

HIGH-F_{ST} ISLANDS ARE IN REGIONS OF LOW RECOMBINATION

Whole genome shotgun sequencing for *H. forsteriana* produced 432.98 Gb of Illumina reads (126x coverage), which assembled into a total length of 3.15 Gb (contig N50 = 3783, scaffold N50 =

37,986; Table S2), similar to the genome size estimated from flow cytometry: For *H. forsteriana*, 1C = 3.50 ± 0.01 pg (3,423 ± 9.78 Mb); for *H. belmoreana*, 1C = 3.08 ± 0.02 pg (3,012.24 ± 19.56 Mb). *BUSCO* (Simão *et al.* 2015) analysis of the assembled genome found that 73.2% of BUSCOs were complete.

In four of eight high-F_{ST}+d_{XY} islands, the genetic distance (cM) per Mb was lower, i.e. recombination rate was lower, than the average rate for the rest of the chromosome where the island was located, but no different from a random draw (Table S3 and S5, sign test, P = 0.5). As a whole, high-F_{ST}+d_{XY} islands did not have significantly lower estimated recombination rates than the rest of the genome (Fig. 3, Welch's t-test, P = 0.39), but high-F_{ST} differentiation islands did (P < 0.0001). In line with the findings in *Howea,* a recent analysis of sympatric populations of divergent stickleback ecotypes showed that signatures of adaptation were considerably more frequent in regions of low recombination when compared to the same ecotype in sympatry or parallel and divergent ecotypes in allopatry (Samuk *et al.* 2017). It has been proposed that limited marker resolution can result in a reduced ability to detect islands outside regions of low recombination (Lowry *et al.* 2017). Given the resolution of our data this is a possibility. However, not all of the high-F_{ST} only islands detected fall within regions of low recombination and our null model explicitly accounts for the recombination distance between markers, suggesting this is unlikely to be the case. Furthermore, Samuk *et al.* (2017) compared whole genome and 'genotyping by sequencing' and found that the pattern of ecotype-associated divergence correlated with recombination rate was consistent between datasets, and therefore was not an artefact of low marker density. In addition, our high-F_{ST}+d_{XY} islands are not associated with low recombination, indicating that their detection was not an artefact of limited marker density.

The association of high-F_{ST} only islands with low recombination could be the result of linked selection (either background selection or hitchhiking) (Burri *et al.* 2015). If genomic diversity was largely shaped by linked selection, we would expect a positive correlation between π and recombination rate; in fact, the opposite is observed across the whole genome (Fig. 4) as well as within both high-F_{ST} and high-F_{ST}+d_{XY} islands (Fig S5). Also, d_{XY} was negatively correlated with recombination rate (Fig. 4, P < 0.0001) - a pattern that was also observed in sympatric stickleback ecotypes and interpreted as a joint effect of gene flow and divergent selection (Samuk *et al.* 2017). These findings are consistent with a limited role for linked selection in the evolution of heterogeneous differentiation in *Howea*. Instead, high-F_{ST} only islands are more likely to have arisen as a product of selection after speciation.

STRESS AND FLOWERING TIME GENES ARE PRESENT IN SPECIATION ISLANDS

We detected 37 genes in high-F_{ST}+d_{XY} islands, 19 of which could be annotated by comparison to the *Arabidopsis* Araport11 protein sequences. An additional 233 genes were located in high-F_{ST} only islands, of which we annotated 120. In total 5,309 genes were assigned to the genetic map, of which 3,020 were annotated, including 2,844 with GO terms. We then examined whether these islands were enriched for any GO terms. There was an excess of genes involved in responses to abiotic stimuli and catabolism of organic compounds in our annotated 120 (Table S7). We also evaluated whether genes that were found to be potentially involved in *Howea* speciation by Dunning *et al.* (2016) were present in our islands. Out of 2,250 such candidate genes that were included on our genetic map, 19 were found in high-F_{ST}+d_{XY} islands (Table S8), although this did not represent a higher proportion than expected by chance (Fisher's exact test, $P = 0.863$). Note that 4,598 candidate genes from Dunning *et al.* (2016) were not found on the map, and these could still have been important during speciation. Furthermore, it may be that the transcriptome-derived candidate genes (Dunning *et al.* 2016) are regulated by genes within our high-F_{ST}+d_{XY} islands. Alternatively, it is possible that some candidates were not involved in speciation, but diverged subsequently.

Finally, we performed a systematic review of the known functions of the 19 annotated genes in high-F_{ST}+d_{XY} islands. We could ascribe 13 of these genes with functions relevant to the speciation scenario, that is, environmental stresses (including those stemming from soil preferences) and flowering time (Fig. 1 and Table S9). Three genes were linked to salt stress, four to drought stress, two to alterations in flowering time, three to osmotic stress, two to cold and three to light stresses (references in Table S9). High-F_{ST}+d_{XY} islands No. 3.1, 5.1, 12.1 and 15.1 contained multiple genes with relevant functions (Table S9), and it is noteworthy that both islands 3.1 and 15.1 contained genes with $F_{\mathrm{ST}} > 0.9$ (Dunning *et al.* 2016) as well as a combination of genes involved in both environmental responses and flowering time control. These are good candidate genes for adaptation and speciation as the habitat that *H. forsteriana* occupies is characterised by low soil moisture and increased salt, light and wind exposure (Papadopulos *et al.* 2019).

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.

Babik, W., Butlin, R.K., Baker, W.J., Papadopulos, A.S.T., Boulesteix, M., Anstett, M.C., *et al.* (2009). How sympatric is speciation in the Howea palms of Lord Howe Island? *Mol. Ecol.* 18: 3629–3638.

Boetzer, M., Henkel, C. V, Jansen, H.J., Butler, D. & Pirovano, W. (2011). Scaffolding preassembled contigs using SSPACE. *Bioinformatics* 27: 578–579.

Broman, K.W., Wu, H., Sen, Å. & Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19: 889–890.

Burri, R., Nater, A., Kawakami, T., Mugal, C.F., Olason, P.I., Smeds, L., *et al.* (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. *Genome Res.* 25: 1656–1665.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., *et al.* (2009). BLAST plus: architecture and applications. *BMC Bioinformatics* 10: 1.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J.H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3 Genes, Genomes, Genet.* 1: 171–182.

Chapman, M.A., Hiscock, S.J. & Filatov, D.A. (2016). The genomic bases of morphological divergence and reproductive isolation driven by ecological speciation in Senecio (Asteraceae). *J. Evol. Biol.* 29: 98–113.

Coyne, J.A. (2011). Speciation in a small space. *Proc. Natl. Acad. Sci.* 108: 12975–12976.

Cruickshank, T.E. & Hahn, M.W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* 23: 3133–3157.

Doležel, J., Greilhuber, J. & Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* 2: 2233–44.

Doyle, J.J. & Doyle, J.L. (1987). A rapid DNA isolation procedure for small amounts of fresh leaf tissue. *Phytochem. Bull.* 19: 11–15.

Dunning, L.T., H. Hipperson, Baker, W.J., Butlin, R.K., Devaux, C., Hutton, I., *et al.* (2016). Ecological speciation in sympatric palms: 1. Gene expression, selection and pleiotropy. *J. Evol. Biol.* 29: 1472–1487.

Ellegren, H., Smeds, L., Burri, R., Olason, P.I., Backstrom, N., Kawakami, T., *et al.* (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature* 491: 756–760. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

Feder, J.L., Egan, S.P. & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends Genet.* 28: 342–350. Elsevier Trends Journals.

Guerrero, R.F. & Hahn, M.W. (2017). Speciation as a sieve for ancestral polymorphism. *Mol. Ecol.* 1–7.

Hipperson, H., Dunning, L.T., Baker, W.J., Butlin, R.K., Hutton, I., Papadopulos, A.S.T.T., *et al.* (2016). Ecological speciation in sympatric palms: 2. Pre- and post-zygotic isolation. *J. Evol. Biol.* 29: 2143–2156.

Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A. & Cresko, W.A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* 6: e1000862. Public Library of Science.

Keenan, K., McGinnity, P., Cross, T.F., Crozier, W.W. & Prodöhl, P.A. (2013). diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods Ecol. Evol.* 4: 782–788.

Kent, W.J. (2002). BLAT — The BLAST -Like Alignment Tool. *Genome Res.* 12: 656–664.

Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos, K.E., Reed, L.K., Antolin, M.F., *et al.* (2017). Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* 17: 142–152.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1: 18.

Malinsky, M., Challis, R.J., Tyers, A.M., Schiffels, S., Terai, Y., Ngatunga, B.P., *et al.* (2015). Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science (80-.).* 350: 1493–1498.

Martin, S.H., Dasmahapatra, K.K., Nadeau, N.J., Salazar, C., Walters, J.R., Simpson, F., *et al.* (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* 23: 1817–1828.

Nadeau, N.J., Whibley, A., Jones, R.T., Davey, J.W., Dasmahapatra, K.K., Baxter, S.W., *et al.* (2012). Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* 367: 343–353.

Nei, M. (1987). *Molecular evolutionary genetics*. Columbia university press.

Noor, M.A.F. & Bennett, S.M. (2010). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity (Edinb).* 103: 439–444.

Nosil, P., Egan, S.P. & Funk, D.J. (2008). Heterogeneous genomic differentiation between walking-stick ecotypes: 'isolation-by-adaptation' and multiple roles for divergent selection. *Evolution (N. Y).* 62: 316–336.

Papadopulos, A.S.T., Baker, W.J., Crayn, D., Butlin, R.K., Kynast, R.G., Hutton, I., *et al.* (2011). Speciation with gene flow on Lord Howe Island. *Proc. Natl. Acad. Sci. U. S. A.* 108: 13188–13193.

Papadopulos, A.S.T., Igea, J., Smith, T.P., Osborne, O., Dunning, L., Turnbull, C., *et al.* (2019). Ecological speciation in sympatric palms: 4. Demographic analyses support that Howea did speciate in the face of high gene flowe. *Evolution* (*in revision)*.

Papadopulos, A.S.T., Kaye, M., Devaux, C., Hipperson, H., Lighten, J., Dunning, L.T., *et al.* (2014). Evaluation of genetic isolation within an island flora reveals unusually widespread local adaptation and supports sympatric speciation. *Philos. Trans. R. Soc. B Biol. Sci.* 369: 20130342.

Papadopulos, A.S.T., Price, Z., Devaux, C., Hipperson, H., Smadja, C.M., Hutton, I., *et al.* (2013). A comparative analysis of the mechanisms underlying speciation on Lord Howe Island. *J. Evol. Biol.* 26: 733–745.

Poelstra, J.W., Vijay, N., Bossu, C.M., Lantz, H., Ryll, B., Müller, I., *et al.* (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science (80-.).* 344: 1410–1414.

R_Development_Core_Team. (2019). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rafajlović, M., Emanuelsson, A., Johannesson, K., Butlin, R.K. & Mehlig, B. (2016). A universal mechanism generating clusters of differentiated loci during divergence-with-migration. *Evolution (N. Y).* 70: 1609–1621.

Ravinet, M., Faria, R., Butlin, R.K., Galindo, J., Bierne, N., Rafajlović, M., *et al.* (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *J. Evol. Biol.* 30: 1450–1477.

Renaut, S., Grassa, C.J., Yeaman, S., Moyers, B.T., Lai, Z., Kane, N.C., *et al.* (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat. Commun.* 4: 1827.

Samuk, K., Owens, G.L., Delmore, K.E., Miller, S.E., Rennison, D.J. & Schluter, D. (2017). Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Mol. Ecol.* 26: 4378–4390.

Savolainen, V., Anstett, M.-C., Lexer, C., Hutton, I., Clarkson, J.J., Norup, M. V, *et al.* (2006). Sympatric speciation in palms on an oceanic island. *Nature* 441: 210–213.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.

Soria-Carrasco, V., Gompert, Z., Comeault, A.A., Farkas, T.E., Parchman, T.L., Johnston, J.S., *et al.* (2014). Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. *Science (80-.).* 344: 738–742.

Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7: 62.

Sundaresan, V. & Alandete-Saez, M. (2010). Pattern formation in miniature: the female gametophyte of flowering plants. *Development* 137: 179–189.

Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J.C., *et al.* (2015). ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* 1–15.

Turner, T.L. & Hahn, M.W. (2010). Genomic islands of speciation or genomic islands and speciation? *Mol. Ecol.* 19: 848–850.

Via, S. (2009). Natural selection in action during speciation. *Proc. Natl. Acad. Sci.* 106: 9939–9946.

Via, S. & West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol. Ecol.* 17: 4334–4345. Blackwell Publishing Ltd.

Weir, B.S. & Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution (N. Y).* 38: 1358–1370. Society for the Study of Evolution.

Woodroffe, C.D., Kennedy, D.M., Brooke, B.P. & Dickson, M.E. (2006). Geomorphological evolution of Lord Howe Island and carbonate production at the latitudinal limit to reef growth. *J. Coast. Res.* 22: 188–201. Coastal Education and Research Foundation.

Wu, C.I. (2001). The genic view of the process of speciation. *J. Evol. Biol.* 14: 851–865.

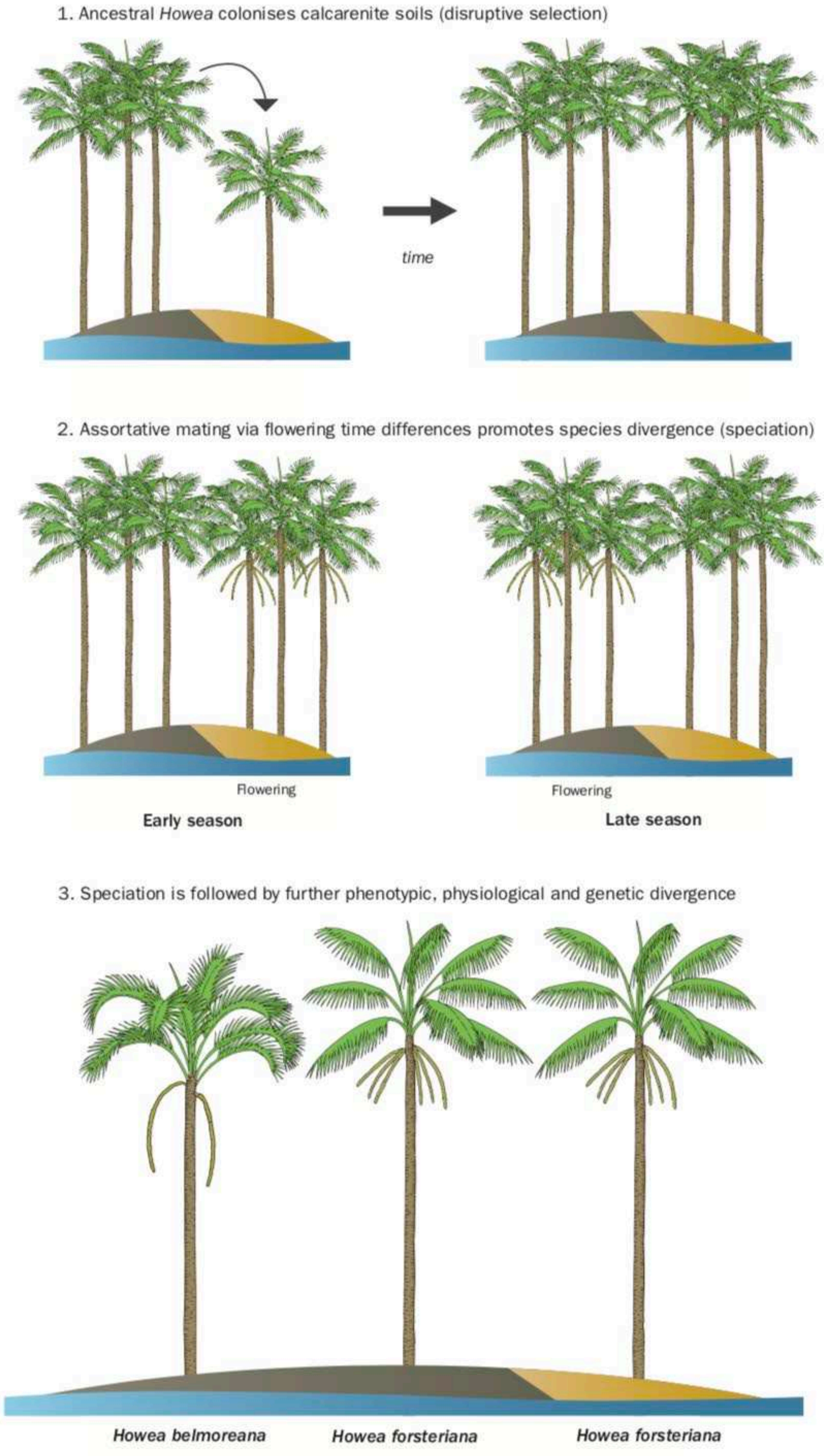
1. Ancestral *Howea* colonises calcarenite soils (disruptive selection)
time
2. Assortative mating via flowering time differences promotes species divergence (speciation)
Flowering
Early season
Flowering
Late season
3. Speciation is followed by further phenotypic, physiological and genetic divergence
Howea belmoreana
Howea forsteriana
Howea forsteriana

Fig. 1 Hypothesised speciation scenario for *Howea*. (1) Lord Howe Island is composed of two main soil types, volcanic (the initial soil type; dark brown) and calcareous soils (subsequent calcarenite deposits; light brown). An ancestral *Howea* colonised Pleistocene calcarenite deposits from volcanic soils, resulting in disruptive selection via adaptation to environmental stresses (e.g., stemming from soil preferences) and triggering flowering time differences. (2) Assortative mating via displacement of flowering phenology promoted reproductive isolation. (3) Further divergence arose after speciation. Today, the curly palm, *H. belmoreana*, grows on volcanic soils, has erect leaflets, a single spike per inflorescence, and flowers late in the season. The kentia palm, *H. forsteriana*, has colonised both calcareous and volcanic soils, has pendulous leaflets, multiple spikes per inflorescence, and it flowers earlier in the season; it is also one of the world's most commonly traded houseplants.

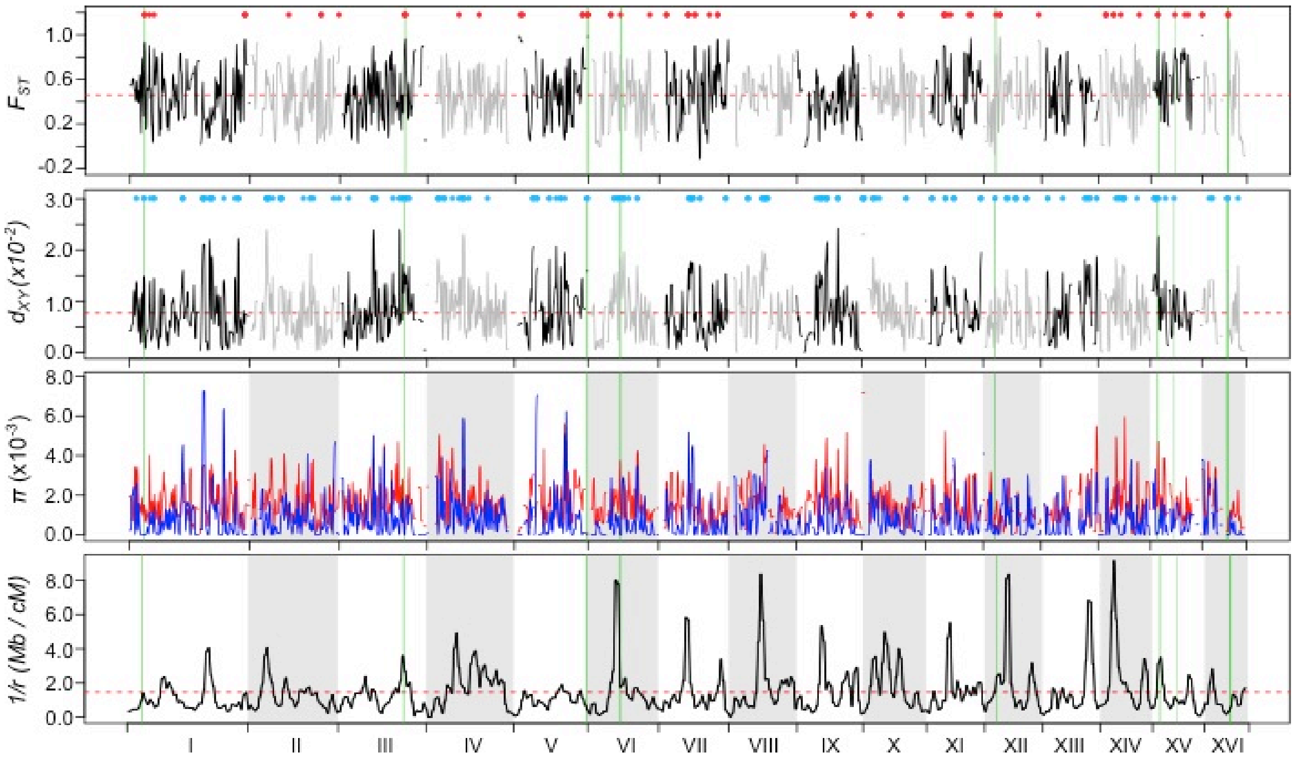

Fig. 2 Genomic divergence between *H. belmoreana* and *H. forsteriana*. The x-axis denotes the genomic order of ddRAD markers on each chromosome. Chromosomes are ordered by length in cM. In the F_{ST} and d_{XY} plots (top two panels) kernel smoothed values of F_{ST} and d_{XY} on each chromosome are shown in alternately black and grey for adjacent chromosomes, coloured dots denote the positions of outlier positions (red = F_{ST} and blue = d_{XY}), dashed red lines denote the genome average. Vertical green bars signify the positions of the 15 high-F_{ST}+d_{XY} islands more likely involved in sympatric speciation, and red dots are the high-F_{ST} only islands more likely to have occurred post-speciation (see text). Genetic diversity (π) for *H. belmoreana* and *H. forsteriana* are shown by red and blue lines, respectively. The lower panel shows the recombination rate in 10cM windows. High-F_{ST} and high-F_{ST}+d_{XY} islands appear to be associated with regions of low recombination (i.e., low cM per Mb; high 1/r), but this association is not statistically significant. Note that genetic diversity is not lower in speciation islands despite low recombination.

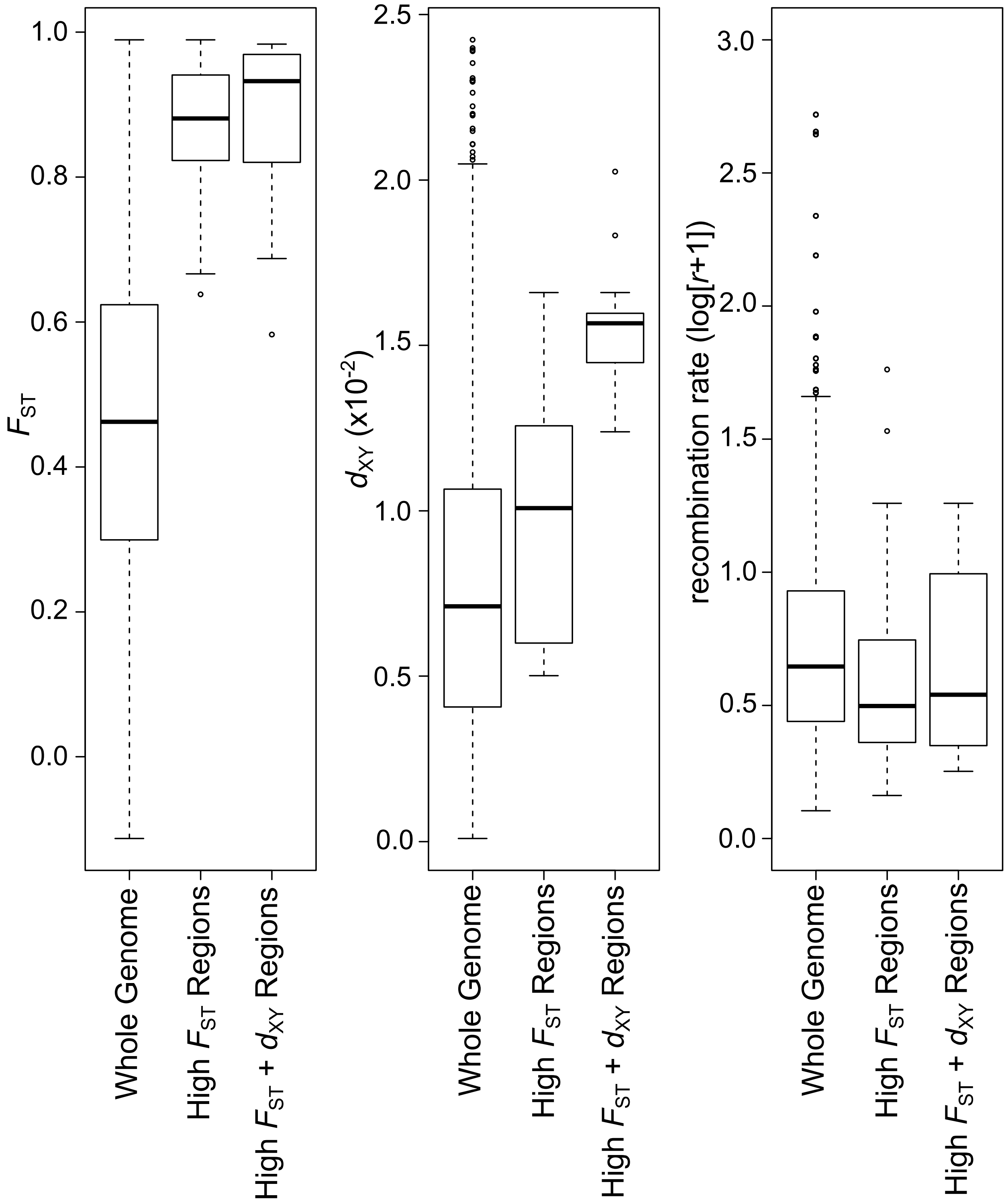

Fig. 3 Comparison of divergence metrics. Boxplots depict the median (bold line), interquartile range (box), and 1.5 times the interquartile range (whiskers).

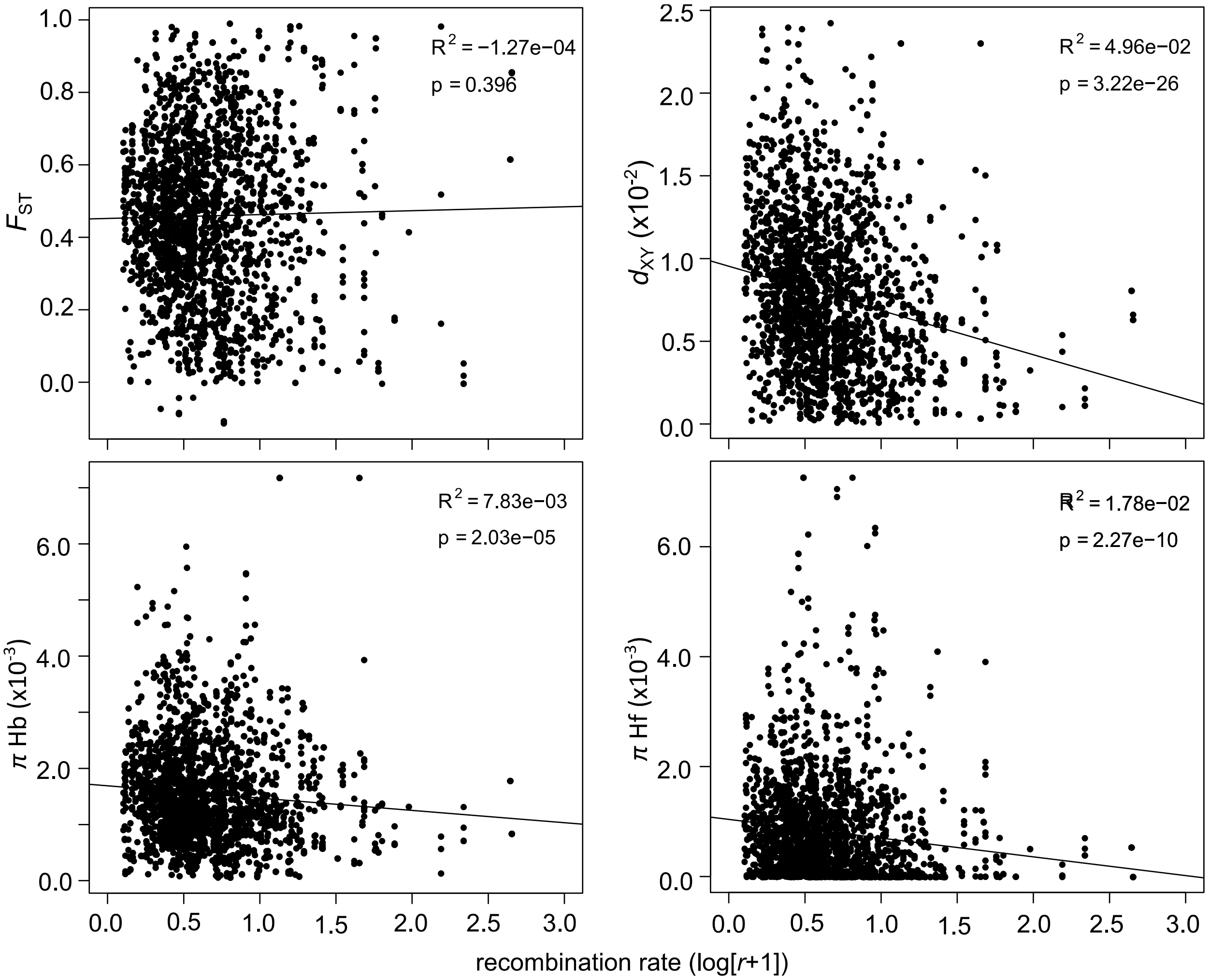

Fig. 4 Genome wide relationships of recombination rate with population genetic metrics indicate no role for linked selection in shaping differentiation in *Howea*. Recombination rate was not correlated with F_{ST} but was negatively correlated with d_{XY} and π in both species.