Genomic features of subspecies defined by phenotypic criteria:Analyses of the mangrove species complex, Avicennia marina

Zhengzhen Wang[1], Zixiao Guo[1], Cairong Zhong[2], Haomin Lyu[1], Xinnian Li[1], Norman Duke[3], and SUHUA SHI[4]

[1]Sun Yat-Sen University
[2]Hainan Dongzhai Harbor National Nature Reserve
[3]James Cook University
[4]Sun Yat-sen University

August 3, 2020

Abstract

The designation of subspecies has often been uncertain in systematics. In addition to phenotypic divergence, designation of subspecies may need to be supplemented by population genetic analyses. In this study, we perform such a survey of the mangrove tree Avicennia marina on Indo-West Pacific coasts. This species harbors three morphological groups. We collected samples from 16 populations (577 individuals) and sequenced 94 nuclear genes. Three genetic features support the subspecies designation for the three morphological subgroups. First, the observed genetic divergence is concordant with the morphological differences, with discordance found in zones of coexistence. Second, the three groups differ in the level of genetic diversity as well as in the demographic history, suggesting a degree of ecological differentiation. Third, and most important, the divergence level varies from locus to locus across the genome. A small portion of the genome is most informative about subspecies delineation, thus hinting the uneven exchange of genes. Such locus-dependent gene flow is expected for incompletely isolated groups. This last point suggests that the reduction in gene flow can be observed at some loci, thus hinting incipient reproductive isolation. In short, the three groups of A. marina appear to have evolved far beyond the stage of structured populations, but not to the point of full species. Hence, the subspecies designation is warranted. We believe these considerations can be generalized to other taxa.

INTRODUCTION

Taxonomic rank below species has been controversial. E. Mayr (1940, 1963) defined subspecies as “a geographically defined aggregate of local populations which differ taxonomically from other subdivisions of the species.” Although critiques had challenged this subspecies classification and some taxonomists refuse to describe such groups (Wilson & Brown, 1953), the value and utility of the subspecies rank is appreciated by others (Durrant, 1955; Mayr, 1982; Phillimore & Owens, 2006). The term subspecies is now used to identify populations distinct mainly in three aspects: isolated geographic range or habitat, phylogenetically concordant phenotypic characters, and separate history (O’Brien & Mayr, 1991).

The definition of subspecies is conceptually reasonable, but the actual practice of subspecific designation is difficult and controversial. Conventional designations of subspecies mainly using phenotypic characters are often challenged by genotype-based delineation (Hawlitschek, Nagy, & Glaw, 2012; Phillimore & Owens, 2006; Torstrom, Pangle, & Swanson, 2014). Morphologically defined taxa are often found to be paraphyletic in phylogenetic analyses (Moritz, 1994; Phillimore & Owens, 2006). The discordance between morphological classification and phylogenetic assessment might be caused by biases in sampling individuals or in choosing portions of the genome for phylogenetic reconstruction. Considering that subspecies is a concept describing populations at a stage before the completion of the speciation process, methods of population genetics are particularly suited to supporting subspecies designation.

The current definition of subspecies by Mayr emphasizes that allopatric speciation is the principle mode how speciation proceeds. However, this conventional view of the Biological Species Concept that the genome evolves as a single cohesive unit has been challenged (Wu, 2001; Wu & Ting, 2004; Feder, Egan, & Nosil, 2012; Feder, Flaxman, Egan, Comeault, & Nosil, 2013; Foote, 2018; Jiggins, 2019). An increasing number of cases indicate that speciation occurs with gene flow and without geographical isolation (Brandvain, Kenney, Flagel, Coop, & Sweigart, 2014; Clarkson et al., 2014; Harr, 2006; Poelstra et al., 2014; Wang, He, Shi, & Wu, 2020).

Questions lie in what pattern of genomic divergence is typical for populations recognized as subspecies. Here we perform population genetic analyses of the mangrove tree *Avicennia marina* to assess subspecies designation. *A. marina* is the most wide-ranging mangrove species, reaching the most marginal mangrove patches of the Indo-West Pacific region (Duke, 2006; Tomlinson, 2016). The taxonomy of Indo-West Pacific (IWP) *Avicennia* had been troublesome before Duke's comprehensive revision (Duke, 1991). In that assessment, *A. marina* were divided into three varieties based on morphological variation (Duke, 1991). After that division, "varieties" or "subspecies" were used to refer to the three groups by different authors (Duke, 2006; Duke, Benzie, Goodall, & Ballment, 1998; Maguire, Peakall, Saenger, & Maguire, 2002; Maguire, Saenger, Baverstock, & Henry, 2000), but conceptually these terms describe the same phenomenon (Mallet, 2007).

Although morphological differences among the groups have been described, the genetic evidence to corroborate their subspecies status is sparse. Patterns of allozyme variation suggest that *A. marina* populations separate into only two clusters with no fixed differences among the three morphologically defined varieties (Duke et al., 1998). These findings thus provide little evidence for the subspecies division. We sought to test whether the three groups warrant the subspecies designation by collecting genotype data. Our assessment might also be useful for subspecies delineation in other taxa.

Thanks to *A. marina* 's peculiarity of inhabiting the tropical and subtropical intertidal environments and linear distribution along coastlines, its distribution range, particularly the range of each morphological group, has been amply documented (Duke, 2006, 2014). This provides a solid basis for comparing genetic and morphological variation as well as inferring population demographic history. Hence, it is advantageous to use *A. marina* to investigate the genetic variation within and between populations as they proceed to form subspecies.

The distinction between populations and subspecies is of particular significance because taxonomic assignments always affect conservation decisions as well as transplanting and breeding practices in mangrove restoration. Particularly, *A. marina* , together with other mangroves trees, are a conservation priority because these ecologically valuable species are under the threat of global climate change in combination with more direct human disturbances (Gilman, Ellison, Duke, & Field, 2008; Guo et al., 2018a).

METHODS

Morphological characters, sampling, and DNA extraction

The three morphological groups have been named *Avicennia marina var. marina, A. m. var. eucalyptifolia* , and *A. m. var. australasica* . The group *marina* is widely distributed from eastern Africa, through the Middle East, South Asia, Southeast Asia, and north to South China. It is also found in western Australia.*Eucalyptifolia* is mainly distributed in northern Australia and extends to southern Philippines, western Indonesia, and the Southwestern Pacific islands. There is a significant range overlap of the two groups in western Australia. *Australasica* is restricted to south-eastern Australia and northern New Zealand (Figure 1). *Australasica* can be morphologically distinguished from the other two groups by its fully pubescent calyx lobes and bracts (Duke, 1991, 2006). These structures are more glabrous in the other groups. The bark of *australasica*is grey fissured, with short longitudinal fissures or reticulate lines, while the bark of the other two subspecies is smooth green or chalky white with flaky patches. *Eucalyptifolia* is mainly distinguished by its lanceolate leaves (as opposed to ovate to elliptic), as well as the style in open flowers which are positioned level with upper edges of anthers (instead of the lower edges of anthers) (Duke, 1991, 2006).*Marina* may also be distinguished by its larger flowers and thicker leaves. However, these distinctions in morphological characters may be inconclusive where two putative subspecies coexist (Duke, 2006). Typical for mangrove trees, propagules of *A. marina* are bouyant on sea water and disperse over sea to nearby locations with mangorve habitats (Duke, 2006).

We sampled 16 populations, 577 individuals (16 to 100 individuals per population) from East Africa, South China, Southeast Asia, Australia to New Zealand, covering *A. marina* 's range (Table 1, Figure 1). To avoid sampling offspring from the same tree, sampled individuals were at least five meters apart. At each site, we sampled as many individuals as were available, but no more than 100. Leaves of each individual were dried, labeled, and stored for DNA extraction. DNA was extracted using the modified CTAB method (Doyle & Doyle). DNA content of each extraction was measured by NanoDrop 2000. For each population, we pooled 300ng of DNA from each individual to make one DNA mixture, ensuring that it contains the same proportion of DNA from each individual. Sixteen DNA mixtures were used in our experiments.

PCR and Illumina high-throughput sequencing

Based on about 200 DNA sequences from a library of *A. marina*expressed sequence tags (Huang et al., 2014), we developed a new set of primers anchored at exons but spanning at least one intron. The 94 pairs of primers producing amplicons 500 to 1500 bps long were used in this study. We performed polymerase chain reaction (PCR) amplification on DNA mixtures from each population using our 94 primer pairs. To reduce amplification errors, TaKaRa high-fidelity PrimerStar HS DNA polymerase was used. The 30 µL PCR mixture consists of 3 µL 10x TaqBuffer (Mg2+), 3 µL dNTPs (2mM/µL), 1.5 µL of each primer (10µM/µL), 0.5 µL HS DNA Polymerase, 3 µL DNA template (˜10ng/µL) and 19 µL deionized water. The PCR program was: 4 min at 94°C; 30 cycles of 10 s at 94°C, 30 s of annealing at the corresponding temperature (Table S1 in the online supplementary file), extension at 72°C for 2 min; followed by 8 min final extension at 72°C. Reactions were held at 16°C before PCR products were subjected to electrophoresis on 1.2% agarose gels. Target bands were excised under ultraviolet light and extracted using the Pearl DNA Gel Extraction Kit (Pearl, Guangzhou, China). Extracted DNA was examined by NanoDrop 2000 to ensure that the amount of each gene product was no less than 100ng. PCR products of the 94 loci from the same population were again pooled, using 100 ng of DNA per locus. We thus obtained 16 PCR product mixtures, each including amplicons from 94 loci.

PCR product mixtures from each population were delivered for sequencing on the Illumina Genome Analyzer and Illumina HiSeq 2000 platform at BGI (Shenzhen) following the manufacturer's instructions. 200 bp DNA libraries were constructed for these mixtures and an 8 bp index in the adapter was used to distinguish the populations. Method details used for library construction were the same as those detailed in the Supplementary materials of our previous publication (Guo et al., 2016). Raw reads produced from the Illumina Genome Analyzer platform were 90 bps in length (all populations except MC, BB, and DW; abbreviations of population names are defined in Table 1) while those from the Illumina HiSeq 2000 platform were 130 bps in length (MC, BB, and DW).

Read mapping and variant calling

The quality of short reads produced by the HiSeq2000 platform was first examined by FastQC (Andrews, 2010). Short reads were then mapped to reference sequences using MAQ 0.7.1(Li, Ruan, & Durbin, 2008). Notably, the reference sequences were obtained by sequencing DNA amplicons of all 94 loci from one *A. marina* individual using the Sanger method. We also did this for one *A. alba* individual for use as outgroup. In mapping and pileup, the mutation rate between reference and read was set to 0.002, the threshold of mismatch base quality sum was 200, and the minimum mapping quality of reads was 30. To exclude false-positive mismatches, we counted the mismatch rate for each site across the read and mismatch rate for each base quality. We trimmed the first and last 10 bases of each read and filtered bases with quality score less than 30.

By identifying variant sites using MAQ 0.7.1, we obtained nucleotide polymorphism information within each population. To avoid bias introduced by sequencing errors, we discarded sites with insufficient site coverage (<100 reads) and those with minor allele frequency less than 0.01 in each population (He et al., 2013). We obtained a list of single nucleotide polymorphisms (SNPs) per population, with allele frequencies. To reduce false SNPs introduced by homopolymers or insertions/deletions, putative variants in those regions were masked. The 16 sets of SNPs were used in the analyses below.

Genetic divergence and diversity estimation

To estimate absolute genetic divergence between populations, we computed pairwise D_{XY} following the formula derived by Nei (Nei & Li, 1979). When calculating D_{XY} , two alleles at each SNP were interpreted as two haplotypes and corresponding allele frequencies as haplotype frequencies. PairwiseD_{XY} values were summed over all SNPs and the sum was normalized by effective sequence length. For each pair of populations, the effective sequence length was defined by sites without missing data in both populations. The obtainedD_{XY} matrix was used in multidimensional scaling using the 'cmdscale' package implemented in R (Figure 2), as well as neighbor-joining tree constructed using MEGA7 (Kumar, Stecher, & Tamura, 2016). We also performed Principal Component Analysis (PCA) on the SNP frequency matrix (summarizing the frequency of each SNP in each population) using the "prcomp" function in R (Venables & Ripley, 2002) to test whether the SNP frequencies differed among populations. Finally, to assess the extent to which genetic polymorphisms were fixed,F_{ST} statistics were computed following a method for a large number of SNPs (Nei & Miller, 1990; Willing, Dreyer, & van Oosterhout, 2012).

The levels of genetic diversity within populations were measured by π and Watterson's ϑ statistics. π summarizes the average number of nucleotide differences between two sequences randomly sampled from a population (Nei, 1987), while Watterson's ϑ estimates nucleotide polymorphism based on the number of observed segregating sites (Watterson, 1977). To correct systematic errors of high-throughput sequencing, we computed ϑ values following a published algorithm (He et al., 2013).

Mantel test of D_{XY} and F_{ST}against geographic distance was performed to test the Isolation by Distance model. Geographical distances between sampling sites were approximated either by spheric distance or dispersal pathway along coasts (called coastline distance). The coastline distance is estimated according to the simulation of one-month oceanic dispersal ability using the methods described in (Van der Stocken, Carroll, Menemenlis, Simard, & Koedam, 2019), with approximate ruler of 350 km.

Geographic barriers delineating the largest genetic discontinuities between pairs of populations were identified using BARRIER 2.2 (Manni, Gue, & Heyer, 2004). By randomly selecting half of the 94 genes, we calculated one F_{ST} matrix for the 47 genes. We repeated this process 100 times and obtained 100F_{ST} matrices. Robustness of each inferred barrier was thus assessed by the 100 matrices.

Demographic history simulation

To test whether the three groups are demographically separable, we compared our real sequences against simulated sequences under eight models with different separation topologies (Simulation 1). Simulated sequences under these models were produced using the ms software (Hudson, 2002). The models are: (1) panmictic; (2) *eucalyptifolia*by itself and the other two groups together; (3) *australasica* by itself and the other two groups together; (4) *marina* by itself and the other two groups together; (5) three separate lineages with*eucalyptifolia* diverging first; (6) three lineages with*marina* diverging first; (7) three lineages with*australasica* diverging first. (8) three lineages diverging simultaneously (Figure 3b). In simulation 1, groups were divided according to morphological differences in the prior. As a control, we constructed artificial groups by pooling two populations each from one morphological type. Using these groupings, we repeated the simulations and model selection on the eight models described above (Simulation 2).

The effective population sizes of the lineages (N) and coalescent times (T) were common among all models.

Notably, to reduce the complexity of parameter setting and to speed up computation, all population size parameters were derived from a single parameter N_0randomly picked from the prior distribution. In models with more than one lineage, N_0 was assigned to any one of the lineages (using as baseline). N of other lineages were produced by multiplying N_0 by ϑ_x/ϑ_0, where ϑ_x and ϑ_0 are the observed ϑ of the current and baseline lineage respectively.

For each model, we performed 100,000 coalescent simulations using the ms program (Hudson, 2002). Each simulation contained 80 loci of 1000 base pairs. Mutation rate was set at $3.26\text{x}10^{-8}$/generation/bp, estimated from phylogenomic comparisons to closely related species with whole genomes (He et al., 2020). The sample size of each group was consistent with our real field sampling (Table 1). Demographic parameters were drawn randomly from a uniform prior distribution. Identical prior distributions of corresponding parameters were set for models within each set (Table S2 & S3).

Ten summary statistics were calculated for each simulated data set, including segregating site number (S), Watterson's estimator (ϑ), nucleotide polymorphism (π) and Tajima's D within each group, as well asD_{XY} and F_{ST} for each pair of groups. Summary statistics were calculated for each simulation independently. Euclidean distances were calculated by comparing simulated statistics with corresponding observed summary statistics. The tolerance of retaining simulated data was set to 0.05. Bayesian posterior probabilities of each model were then estimated following the Approximate Bayesian Computation (ABC) schema (Beaumont, Zhang, & Balding, 2002) using the "abc" package in R (Csilléry, François, & Blum, 2012). The "postpr" function together with "neuralnet" option in the "abc" R package was used to perform model selection.

We also built four models (v1, v2, v3, and v4) to test whether the population from Bunbury, Australia (BB, Table1) genetically belongs to the *marina* or *eucalyptifolia* group (Simulation 3, Table S4). In model v1 and v2, BB (constant effective population size of N_{bb}) and *marina* (N_{ma}) coalesced at vT_1 generations ago and then the common ancestor further coalesced with *eucalyptifolia* (effective population size N_{eu}) at vT_0 generations ago (vT_0>vT_1). Model v1 differed from v2 by presence or absence of gene flow ($m_1$and m_2) between BB and *eucalyptifolia* . Similarly, in models v3 and v4, BB (N_{bb}) coalesced with*eucalyptifolia* (N_{eu}) at vT_1generations ago. The common ancestor then coalesced with *marina*(effective population size N_{ma}) at vT_0generations ago (vT_0>vT_1) . Nine summary statistics, Watterson's estimator (ϑ) for each population and pairwise F_{ST} and D_{XY} , were used in the model selection procedure similar to the one previously described.

Detection of gene flow between subspecies

We used the statistical model implemented in TreeMix to infer patterns of splits and mixtures among populations (Pickrell & Pritchard, 2012). As revealed from the F_{ST} statistic above, some populations are genetically similar, e.g. Andaman Sea on the west of Malay Peninsula and the South China Sea (Gulf of Thailand and Hainan Island). Hence, one representative population from each region was used in this analysis. The eleven populations were related to the common ancestor through a graph of ancestral populations, which was inferred by allele frequency and a Gaussian approximation to genetic drift (Pickrell & Pritchard, 2012). Gene flow events were inferred by adding admixtures onto the Maximum Likelihood population splitting topology.

Haplotype inference and population structure mapping

The method developed by (He et al., 2019) was used to infer haplotypes. This method uses SNP linkage information in each short read pair to infer haplotypes and frequency of each haplotype in population, following an expectation-maximization algorithm (Bilmes, 1998; Dempster, Laird, & Rubin, 1977). If two adjacent SNPs were not covered by any read pair, we broke the gene into segments. In this case, the midpoint of the two adjacent SNPs is defined as the breakpoint of two consecutive segments. The accuracy of this method in inferring haplotypes has been validated by sequencing individuals using the Sanger method (He et al., 2019). We selected eight populations representing different morphological groups and different regions for inferring haplotypes: two*eucalyptifolia* (CA and DW), two *australasica* (AK and BS), and four *marina* (BB, LS, TN, and SY). Genes were split into 454 linked segments and haplotypes were inferred for each segment (Table S5). Before constructing haplotype networks, we filtered out segments with length less than 100 bps or with missing data. For each of the 231 retained segments, we computed a haplotype network using the NETWORK software (Polzin & Daneshmand, 2003).

RESULTS

Among-group genetic divergence

We obtained 76 to 87 kb of DNA sequence covering 88 to 94 genes (Table 1). By mapping short reads to reference sequences, we identified 74 to 1657 segregating sites within each population (Table 1). We calculated among-population pairwise D_{XY} values to assess genetic divergence and used the resulting distance matrix to construct a neighbor-joining tree. The D_{XY} matrix shows clear divergence between the three morphological groups (Figure 2a), with the BB population the sole exception. The largest D_{XY}values were observed between the *australasica* populations and the other two morphological groups, ranging from 7.7 to 9.9/kb (Table S6). Relatively lower divergence was observed between *eucalyptifolia*and *marina* populations, with D_{XY} values between 6.5 and 7.4/kb. By pooling populations within each morphological group, we estimated the D_{XY} to be 8.2/kb between*eucalyptifolia* and *australasica* , 6.7/kb between*marina* and *eucalyptifolia* and 9.1/kb between *marina*and *australasica* .

Genetic divergence was generally lower among populations than among morphological groups. The two *australasica* populations diverged little from each other (D_{XY} =2.2/kb). The pair of*eucalyptifolia* populations diverged more but still less than among morphological groups (D_{XY} = 5.48/kb). Within *marina* , we see two major geographical groups, one containing MC, LS, and PN (west of the Malay Peninsula) and the other TN, BK, SS, SY, WC, SB, CB, and BL (east of the Malay Peninsula, Figure S1). D_{XY} per kb ranges from 1.27 to 3.75 within the first and from 0.94 to 4.69 within the second geographical group. Between the two geographical groups, D_{XY} ranges from 4.32 to 5.69, still lower than between morphological groups. The BB population is an outlier and has diverged far from other *marina*populations (D_{XY} = 7.76-8.43/kb), to a level among morphological groups. D_{XY} provides a measurement of how far the populations diverged from each other. We also measured the extent of divergence by comparing the allele frequencies of polymorphisms within populations (Cruickshank & Hahn, 2014). Plotting principal components of the allele frequency matrix, populations of the three morphological groups generally show very different patterns, except that the DW population (*eucalyptifolia*) is close to*marina* populations and the BB population (*marina*) is again very different from all the other groups (Figure 2c).

The F_{ST} statistic quantifies these genetic differences. The 120 values of pairwise F_{ST}estimates calculated for the 16 populations are generally high, with the average value of 0.61 (first and third quartiles are 0.50 and 0.76 respectively). Populations from the South China Sea, i.e. TN, BK, SS, SY, and WC (Figure S1& S2) have relatively low pairwise divergence. The Mantel test shows a significant relationship between genetic differentiation and geographic distance. This is regardless of whether the geographic distance was estimated using the spherical or coastline method (Figure S3, see Methods for details). This correlation indicates that geographical distance contributes to, at least partly, to the high level of genetic differentiation among *A. marina* populations. However, the two geographical groups around the Malay Peninsula show genetic differentiation greater than what we would expect from the distance separating them, indicating that other factors are also important.

The BARRIER analysis reveals major barriers with >80% bootstrap support lie roughly along the Sunda shelf and between Australasia and Southeast Asia. Minor barriers are also identified between Africa and Southeast Asia, as well as between Western Australia and Northern Australia. The major barrier in the historic Sunda Land corresponds to the obvious deviation of F_{ST}values from the expectation based on distance alone (Figure S3). Geographical isolation seems to result from land barriers (e.g., the Malay Peninsula) or open ocean.

Morphological groups are demographically separable

If the morphological groups have proceeded to subspecies stage, they likely experienced different demographical histories. Both the nucleotide diversity (π) and Watterson's estimator of nucleotide polymorphism (ϑ) show different levels of within-population genetic variation. The two *eucalyptifolia* populations have the highest genetic diversity, with average ϑ (across segments) = 2.82 and 3.94/kb and average π = 3.41 and

4.06/kb (Figure 3a). In contrast, the *marina* populations are low in genetic diversity, with average ϑ ranging from 0.21 to 0.91/kb and average π ranging from 0.15 to 1.39/kb (Table1, Figure 3a). The BS population (*australasica*) has intermediate diversity, while the AK population (*australasica*) is unusually monomorphic (Table1, Figure 3a). The very low diversity in the AK population is likely due to its marginal location, similar to WC and SY.

Distinct levels of genetic diversity hint that the three morphological groups have indeed experienced different demographic events. We fitted several demographical models using approximate Bayesian computation (ABC) to test whether we can distinguish population histories. Our ABC approach shows that simulated sequences under the model with each morphological group diverging simultaneously provides the best fit to the observed data. This conclusion was validated by three repetitions and high posterior probability of this model (> 0.6, Table 2). This result indicates the three morphological groups are mostly demographically separable and diverged from each other simultaneously. In contrast, the simulations with artificial groups (Simulation 2) show no robustness in model selection.

The population BB morphologically diagnosed as *marina* shows lower genetic divergence and differentiation to *eucalyptifolia* than *marina* (Figure 2). Is it an *eucalyptifolia* mis-diagnosed as *marina* or a *marina* exchanging genes with *eucalyptifolia* ? Our ABC simulation (Simulation 3) shows that BB has descended from *marina* but experiences gene flow with *eucalyptifolia* populations (model v2, posterior probability 0.933, Table 2 and Figure 4a). This indicates that morphological groups, while significantly genetically differentiated, are genetically permeable. We also used TreeMix to capture potential gene flow events among populations (Figure 4b). We identified six such events on the population splitting graph (Table S7). Three such events occurred between morphological groups and two happened between *marina* populations. The last event occurred between BB and the outgroup species *A. alba* .

Haplotype network variation across the genome

The gold standard to define species is thought to be reproductive isolation (RI) (Feder, Egan, & Nosil, 2012; Wu, 2001; Wu & Ting, 2004). However, RI is difficult to demonstrate in practice because diverging lineages are usually distributed discontinuously and have no chance to mate. Hence, how to determine whether the morphological groups are full species or subspecies? Previous results suggest that these groups are subspecies because variation in allozyme allele frequencies among them is far less than among well-established sister species (Duke et al., 1998). Since the portion of the genome unaffected by gene flow increases as the speciation proceeds, more and more loci become informative for group delineation (Feder et al., 2012; Feder, Flaxman, Egan, Comeault, & Nosil, 2013; Nadeau et al., 2013; Wu, 2001; Wu & Ting, 2004). Subspecies are somewhere on that continuum. Therefore, we expect that only a fraction of the genome is stably differentiated among morphological groups. To test this hypothesis, we inferred haplotype networks across the 94 loci we sequenced. Using an expectation-maximization method to infer among-SNP linkage disequilibrium, we split these regions into 454 linked segments (Table S5). Segments with missing data and those less than 100bp in length were discarded and 231 segments were retained for haplotype network reconstruction, with *A. alba* as the outgroup (Figure 5).

Among these segments, 134 (58.0%) were not genetically distinguishable among groups with only one or a few haplotypes identified and all haplotypes closely related to each other and shared among the three morphological groups. The other 66 segments (28.6%) reliably distinguish *australasica* from the other two groups. Among these 66 segments, the BB population shares haplotypes with *australasica* rather than *marina* at seven loci. The third type of segments, 14 in total (6.1%), delimits *marina* from the other two groups. Five segments (2.2%) distinguish *eucalyptifolia* , but BB shares haplotypes with *eucalyptifolia* in all cases. Most importantly, 11 segments (4.8%) clearly differentiate all three morphological groups, with haplotypes split into three clusters and each morphological group containing haplotypes from a single cluster. In eight of the 11 segments, BB shares haplotypes with *eucalyptifolia* , consistent with analyses described above. Finally, one segment (0.4%) separates *marina* and *australasica,* but *eucalyptifolia* contains haplotypes from both clusters.

DISCUSSION

Substantial genetic divergence and separate demographical history

It is common for intraspecific genetic variation to be structured geographically due to isolation induced by geographical barriers. In mangroves, such barriers are usually land mass, open water, or ocean currents (Guo et al., 2016; Guo et al., 2018b; Wee et al., 2020). Such genetically structured population groups should not necessarily to be classified as subspecies, unless stable differences in morphology are also present. Conversely, morphologically recognized groups can be designated as subspecies only in the presence of clear genetic divergence. In this study we comprehensively sampled *A. marina*populations across their geographical range, assembled an extensive SNP data set, and used it to test the presence of genetic differentiation among three morphologically recognized groups. Our study finds a robust genetic split of *A. marina* into three groups, noting that this divergence was observed both in the genetic distanceD_{XY} matrix and in PCA clustering based on a SNP frequency matrix. The genetic grouping pattern is consistent with the morphological classification of the three groups, *marina* ,*eucalyptifolia* , and *australasica* .

We found clearly different levels of genetic polymorphism within each group, indicating that they might have accumulated genetic variation independently. Our approximate Bayesian computation modeling also supports the idea that the three groups are separate lineages diverging simultaneously. This approach is advantageous comparing with the previous phylogenetic analysis which indicated *australasica*diverged earlier than the split between *marina* and*eucalyptifolia* (X. Li et al., 2016).

The three groups of *A. marina* are distinct both in morphological characters and DNA polymorphism frequencies, strongly suggesting that they have evolved far beyond the stage of structured populations. Looking at the distribution pattern of genetic and morphological differentiation, the range of the three groups grade sharply (Figure 1). However, geographical barriers between *A. marina* populations are not always consistent with the distribution boundaries of morphological groups (Figure S4), which indicates the divergence among morphological groups is not caused by recent geographical isolation. However, geographical barriers as well as Isolation by Distance do seem to have contributed to genetic differentiation below subspecies. The Malay Peninsula acts as a land barrier in many sea-dispersing mangrove plants, such as *Rhizophora* (Guo et al., 2016; Wee et al., 2015)*, Ceriops* (Tan et al., 2005) *, Lumnitzera* (J. Li et al., 2016), and *Xylocarpus* (Guo et al., 2018b). Similarly to other mangrove species, propagules of *A. marina* are buoyant on sea water and disperse across long distances via currents (Steinke & Ward, 2003). The long-distance dispersal ability of *A. marina* was found to be relatively weak in both field observations and genetic surveys (Binks et al., 2019; Clarke, Kerrigan, & Westphal, 2001; Duke et al., 1998). It is thus interesting to note that the Malay Peninsula clearly contributes to among-population genetic differentiation among populations in our sample, separately from the deeper morphological group divergence.

Divergence among morphological groups was likely initially caused by a geographical barrier, although ecological factors may have also played a role. A possible explanation for the separation between *marina*and *eucalyptifolia* is the fusion of New Guinea with Australia during glacial ages when sea level was low (Duke et al., 1998). When the Torres Strait reopened, *eucalyptifolia* expanded westward until hindered by the Indonesia through flow (Gordon, 2005; Hall, 2009). The gradation from *eucalyptifolia* to *australasica* between Rockhampton and Brisbane on the east coast of Australia is probably caused by the bifurcation of the North Caledonian Jet into the North Queensland and the East Australian Currents (Ganachaud et al., 2007; Schiller et al., 2008), or by the latitudinal change in environmental conditions such as temperature. The exact mechanism may be clarified in subsequent studies.

Genomic landscape of among-subspecies divergence

The establishment of reproductive isolation is an important landmark of speciation completion (Abbott, 2017; Feder et al., 2012; Wu, 2001; Wu & Ting, 2004). Interbreeding between complete species is impeded by various forms of behavioral, ecological, or genetic incompatibilities (Abbott et al., 2013; Seehausen et al., 2014). However, reproductive isolation is usually difficult to test directly, especially in taxa naturally distributed in isolated geographical regions. Gene flow is common during the speciation process, from the initialization of speciation to its completion, as well as even between closely related full species (Brandvain et al., 2014; Clarkson et al., 2014; Harr, 2006; Poelstra et al., 2014; Wang et al., 2020). Because gene flow is expected to decrease as reproductive isolation develops, the degree of exchange in genetic material may be an indicator of the stage of differentiation.

With establishment of reproductive isolation, genetic hitchhiking grades into genomic hitchhiking at the nascent species stage (or late stage of speciation), driving previously established genomic islands of differentiation to expand and converge into a plateau, resulting in high divergence across the whole genome (Wu, 2001; Wu & Ting, 2004; Feder, Egan, & Nosil, 2012; Feder, Flaxman, Egan, Comeault, & Nosil, 2013). At the subspecies stage (or early to middle stages of speciation), gene flow is expected to be extensive, homogenizing most of the genome, with a small portion highly diverged (Wu, 2001; Wu & Ting, 2004; Feder, Egan, & Nosil, 2012; Feder, Flaxman, Egan, Comeault, & Nosil, 2013). We used TreeMix to identify gene flow events among *A. marina*morphological groups. A recent study also suggested substantial gene flow between *marina* and *eucalyptifolia* in western Australia (Binks et al., 2019). In *A. marina,* only about 5% of the genome clearly delineates the three morphological groups, not satisfying the criteria of full species designation. Even though the highly divergent genomic regions we identified may not be the ones directly determining the morphological characters used to designate them as subspecies, appearance of highly diverged islands against the background of extensive gene flow can be a feature used to recognize subspecies. Thus, our deep survey of genetic variation among populations of *A. marina* from a wide geographic range and genomic scope shows that the three *A. marina* morphological groups have proceed beyond structured populations but have not achieved full species status. We therefore argue that they should be designated as subspecies.

Our study of *A. marina* suggests that population genetic features should complement morphological differences to designate subspecies. (1) Subspecies achieve a distinct level of genetic divergence, much higher than between populations within subspecies. Genetic divergence between subspecies is strongly associated with morphological differences rather than geographical barriers. (2) Subspecies are demographically separable, as reflected by different levels of genetic diversity and demographic modeling. (3) The level of genetic divergence varies from locus to locus, hinting at continual (and uneven) exchange of genes from locus to locus. The subspecies assignments may only be clearly revealed by a small portion of the genome. These features from a population genetic prospective could be applied to assessment of subspecies in other domains of the tree of life.

The features identified above should be used collectively, together with morphological diagnostics, to resolve subspecies taxonomic issues. The BB population in this study can be used as an example. BB shows a high level of genetic divergence from every other population, comparable to the between-subspecies level. However, we do not think it should be recognized as an additional subspecies. Morphologically, *A. marina* trees on the west coast of Australia (BB in this study) appear to group with *A. m. marina* (Duke, 1991). Genetically, no loci distinguish this population exclusively from others. Instead, the BB population shows either *marina* or *eucalyptifolia*haplotypes at subspecies-informative genes. Rather than an additional subspecies, *A. marina* on the west coast of Australia, represented by the BB sample, is more likely a genetic admixture of *marina*and *eucalyptifolia.*

The utility of subspecies classification in evolutionary studies and conservation

As a rank between population and species, the subspecies is useful for predicting the evolutionary divergence levels among geographical populations (Barrowclough, 1982). Populations defined as subspecies are expected to be more highly differentiated than within groups and should have separate demographic history. The classification of subspecies should not be the end but a byproduct of investigations of genetic variation within a species if patterns that warrant designation of subspecies are found (Barrowclough, 1982).

The other important utility of subspecies is to inform conservation decisions. Mayr proposed that subspecies are of conservation importance for their potential to evolve into full species and their acquisition of unique characteristics (O'Brien & Mayr, 1991). The emphasis on species diversity in conservation policy had driven taxonomists to revise subspecies upward to species (Mallet, 2007). More recently, managers have become increasingly aware of the necessity to protect biodiversity at all levels of life. The stability of the ecosystem may be cumulatively enhanced by weak effects of individual species, analogous to the gene regulatory networks (Chen et al., 2019). As one of the most widely distributed mangrove species, *A. marina* is important for the ecological health of coastal ecosystems, especially as the global climate continues to change. Without recognition of their subspecies status, the obvious intraspecific genetic differentiation may be neglected and treated as a single conservation unit. There have been cases where cryptic species went extinct before being recognized (Yan et al., 2018). In addition, the assessment of genetic background of the subspecies will prove instructive for selecting source plants for transplanting in mangrove restoration projects. Hence, subspecies classification is meaningful for protecting biodiversity, particularly in the mangrove ecosystem.

ACKNOWLEDGEMENTS

We thank Chung-I Wu for insightful comments that improved the manuscript. This study was supported by the National Natural Science Foundation of China (grant numbers 31830005 and 31900200); the National Key Research and Development Plan (2017FY100705); the China Postdoctoral Science Foundation (grant number 2019M663212); the Science and Technology Program of Guangzhou (201707020035) and the Chang Hungta Science Foundation of Sun Yat-Sen University.

REFERENCES

Abbott, R. J. (2017). Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. *Journal of Systematics and Evolution* , *55* (4), 238–258. doi: 10.1111/jse.12267

Abbott, R. J., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., ... Zinner, D. (2013). Hybridization and speciation.*Journal of Evolutionary Biology* , *26* (2), 229–246. doi: 10.1111/j.1420-9101.2012.02599.x

Andrews, S. (2010). *FASTQC: a quality control tool for high throughput sequence data* .

Barrowclough, G. F. (1982). Geographic Variation , Predictiveness , and Subspecies. *The Auk* , *99* (3), 601–603.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* ,*162* (4), 2025–2035.

Bilmes, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *ReCALL* , *4* (510), 126. doi: 10.1080/0042098032000136147

Binks, R. M., Byrne, M., Mcmahon, K., Pitt, G., Murray, K., & Evans, R. D. (2019). Habitat discontinuities form strong barriers to gene flow among mangrove populations, despite the capacity for long-distance dispersal. *Diversity and Distributions* , *25* (2), 298–309. doi: 10.1111/ddi.12851

Brandvain, Y., Kenney, A. M., Flagel, L., Coop, G., & Sweigart, A. L. (2014). Speciation and Introgression between *Mimulus nasutus* and*Mimulus guttatus* . *Plos Genetics* , *10* (6), e1004410. doi: 10.1371/journal.pgen.1004410

Chen, Y., Shen, Y., Lin, P., Tong, D., Zhao, Y., Allesina, S., ... Wu, C. I. (2019). Gene regulatory network stabilized by pervasive weak repressions: MicroRNA functions revealed by the May-Wigner theory.*National Science Review* , *6* (6), 1176–1188. doi: 10.1093/nsr/nwz076

Clarke, P. J., Kerrigan, R. A., & Westphal, C. J. (2001). Dispersal potential and early growth in 14 tropical mangroves: Do early life history traits correlate with patterns of adult distribution?*Journal of Ecology* , *89* (4), 648–659. doi: 10.1046/j.0022-0477.2001.00584.x

Clarkson, C. S., Weetman, D., Essandoh, J., Yawson, A. E., Maslen, G., Manske, M., ... Donnelly, M. J. (2014). Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nature Communications* , *5* (May). doi: 10.1038/ncomms5248

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* , *23* (13), 3133–3157. doi: 10.1111/mec.12796

Csilléry, K., François, O., & Blum, M. G. B. (2012). Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* , *3* (3), 475–479. doi: 10.1111/j.2041-210X.2011.00179.x

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the *EM* Algorithm.*Journal of the Royal Statistical Society: Series B (Methodological)* , *39* (1), 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Duke, N. C. (1991). A systematic revision of the mangrove genus*Avicennia* (Avicenniaceae) in Australasia. *Australian Systematic Botany* , *4* (2), 299. doi: 10.1071/SB9910299

Duke, N. C. (2006). *Australia's mangroves: the authoritative guide to Australia's mangrove plants* . MER.

Duke, N. C. (2014). *'World Mangrove iD: expert information at your fingertips' Version 1.1 for Android* . MangroveWatch Pubication, Australia.

Duke, N. C., Benzie, J. A. H., Goodall, J. A., & Ballment, E. R. (1998). Genetic Structure and Evolution of Species in the Mangrove Genus*Avicennia* (Avicenniaceae) in the Indo-West Pacific.*Evolution* , *52* (6), 1612–1626.

Durrant, S. D. (1955). In Defense of the Subspecies. *Systematic Zoology* , *4* (4), 186–190.

Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in Genetics* , *28* (7), 342–350. doi: 10.1016/J.TIG.2012.03.009

Feder, J. L., Flaxman, S. M., Egan, S. P., Comeault, A. A., & Nosil, P. (2013). Geographic Mode of Speciation and Genomic Divergence.*Annual Review of Ecology, Evolution, and Systematics* ,*44* (1), 73–97. doi: 10.1146/annurev-ecolsys-110512-135825

Ganachaud, A., Kessler, W., Wijffels, S., Ridgway, K., Cai, W., Holbrook, N., ... Aung, T. (2007). *Southwest Pacific Ocean Circulation and Climate Experiment (SPICE)* . Seattle, WA.

Gilman, E. L., Ellison, J., Duke, N. C., & Field, C. (2008). Threats to mangroves from climate change and adaptation options: A review.*Aquatic Botany* , *89* (2), 237–250. doi: 10.1016/j.aquabot.2007.12.009

Gordon, A. L. (2005). Oceanography of the Indonesian seas and their throughflow. *Oceanography* , *18* (4), 14–27.

Guo, Z., Li, X., He, Z., Yang, Y., Wang, W., Zhong, C., ... Shi, S. (2018a). Extremely low genetic diversity across mangrove taxa reflects past sea level changes and hints at poor future responses. *Global Change Biology* , *24* (4). doi: 10.1111/gcb.13968

Guo, Z, Guo, W., Wu, H., Fang, X., Ng, W. L., Shi, X., ... Huang, Y. (2018b). Differing phylogeographic patterns within the Indo-West Pacific mangrove genus *Xylocarpus* (Meliaceae). *Journal of Biogeography* , *45* (3), 676–689. doi: 10.1111/jbi.13151

Guo, Z, Huang, Y., Chen, Y., Duke, N. C., Zhong, C., & Shi, S. (2016). Genetic discontinuities in a dominant mangrove *Rhizophora apiculata* (Rhizophoraceae) in the Indo-Malesian region. *Journal of Biogeography* , *43* , 1856–1868. doi: 10.1111/jbi.12770

Hall, R. (2009). Southeast Asia's changing palaeogeography. *Blumea - Biodiversity, Evolution and Biogeography of Plants* , *54* (1), 148–161. doi: 10.3767/000651909X475941

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Research* , *16* (6), 730–737. doi: 10.1101/gr.5045006

Hawlitschek, O., Nagy, Z. T., & Glaw, F. (2012). Island evolution and systematic revision of comoran snakes: Why and when subspecies still make sense. *PLoS ONE* , *7* (8). doi: 10.1371/journal.pone.0042970

He, Z., Li, X., Ling, S., Fu, Y.-X., Hungate, E., Shi, S., & Wu, C.-I. (2013). Estimating DNA polymorphism from next generation sequencing data with high error rate by dual sequencing applications. *BMC Genomics* , *14* (1), 535. doi: 10.1186/1471-2164-14-535

He, Z., Li, X., Yang, M., Wang, X., Zhong, C., Duke, N. C., ... Shi, S. (2019). Speciation with gene flow via cycles of isolation and migration : insights from multiple mangrove taxa. *National Science Review* , *6* (2), 275–288. doi: 10.1093/nsr/nwy078

He, Z., Xu, S., Zhang, Z., Guo, W., Lyu, H., Zhong, C., ... Shi, S. (2020). Convergent adaptation of the genomes of woody plants at the land-sea interface. *National Science Review* .

Huang, J., Lu, X., Zhang, W., Huang, R., Chen, S., & Zheng, Y. (2014). Transcriptome Sequencing and Analysis of Leaf Tissue of *Avicennia marina* Using the Illumina Platform. *PLoS ONE* , *9* (9), e108785. doi: 10.1371/journal.pone.0108785

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

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets.*Molecular Biology and Evolution* , *33* (7), msw054. doi: 10.1093/molbev/msw054

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* , 1851–1858. doi: 10.1101/gr.078212.108.

Li, J., Yang, Y., Chen, Q., Fang, L., He, Z., Guo, W., ... Shi, S. (2016). Pronounced genetic differentiation and recent secondary contact in the mangrove tree *Lumnitzera racemosa* revealed by population genomic analyses. *Scientific Reports* , *6* (July), 1–12. doi: 10.1038/srep29486

Li, X., Duke, N. C., Yang, Y., Huang, L., Zhu, Y., Zhang, Z., ... Shi, S. (2016). Re-evaluation of phylogenetic relationships among species of the mangrove genus *Avicennia* from Indo-West Pacific based on multilocus analyses. *PLoS ONE* , *11* (10), 1–14. doi: 10.1371/journal.pone.0164453

Maguire, T., Peakall, R., Saenger, P., & Maguire, L. (2002). Comparative analysis of genetic diversity in the mangrove species*Avicennia marina* (Forsk.) Vierh.(Avicenniaceae) detected by AFLPs and SSRs. *TAG Theoretical and Applied Genetics* , *104* (2), 388–398.

Maguire, T., Saenger, P., Baverstock, P., & Henry, R. (2000). Microsatellite analysis of genetic structure in the mangrove species*Avicennia marina* (Forsk.) Vierh.(Avicenniaceae). *Molecular Ecology* , *9* (11), 1853–1862.

Mallet, J. (2007). Subspecies, semispecies, superspecies.*Encyclopedia of Biodiversity* .

Manni, F., Gue, E., & Heyer, E. (2004). Variation : how barriers can be detected by using monmonier's algorithm. *Human Biology* ,*76* (2), 173–190.

Mayr, E. (1940). Speciation phenomena in birds. *The American Naturalist* , *74* , 249–278.

Mayr, E. (1963). *Animal Species and Evolution* . Cambridge, MA: Harvard University Press.

Mayr, E. (1982). Commentary Forum : Avian Subspecies in the 1980 ' S of What Use Are Subspecies ? *The Auk* , *99* (3), 593–595.

Moritz, C. (1994). Defining 'Evolutionarily Significant Units' for conservation. *Tree* , *9* (10), 373–375. doi: 10.1016/0169-5347(94)90057-4

Nadeau, N. J., Martin, S. H., Kozak, K. M., Salazar, C., Dasmahapatra, K. K., Davey, J. W., ... Mark, L. (2013). Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology* , *22* , 814–826. doi: 10.1111/j.1365-294X.2012.05730.x

Nei, M. (1987). *Molecular evolutionary genetics* . New York: Columbia University Press.

Nei, M., & Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* ,*76* (10), 5269–5273. doi: 10.1073/pnas.76.10.5269

Nei, M., & Miller, J. C. (1990). A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics* , *125* (4), 873–879.

O'Brien, S. J., & Mayr, E. (1991). Bureaucratic mischief: Recognizing endangered species and subspecies. *Science* , *251* (4998), 1187–1188. doi: 10.1126/science.251.4998.1187

Phillimore, A. B., & Owens, I. P. F. (2006). Are subspecies useful in evolutionary and conservation biology? *Proceedings of the Royal Society B: Biological Sciences* , *273* (1590), 1049–1053. doi: 10.1098/rspb.2005.3425

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics* , *8* (11). doi: 10.1371/journal.pgen.1002967

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Baglione, V., ... Wolf, J. B. W. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* ,*344* (6190), 1410–1414.

Polzin, T., & Daneshmand, S. V. (2003). On Steiner trees and minimum spanning trees in hypergraphs. *Operations Research Letters* ,*31* (1), 12–20. doi: 10.1016/S0167-6377(02)00185-2

Schiller, A., Oke, P. R., Brassington, G., Entel, M., Fiedler, R., Griffin, D. A., & Mansbridge, J. V. (2008). Eddy-resolving ocean circulation in the Asian–Australian region inferred from an ocean reanalysis effort. *Progress in Oceanography* , *76* (3), 334–365. doi: 10.1016/j.pocean.2008.01.003

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., ... Widmer, A. (2014). Genomics and the origin of species. *Nature Reviews Genetics* , *15* (3), 176–192. doi: 10.1038/nrg3644

Steinke, T. D., & Ward, C. J. (2003). Use of plastic drift cards as indicators of possible dispersal of propagules of the mangrove*Avicennia marina* by ocean currents. *African Journal of Marine Science* , *25* (1), 169–176. doi: 10.2989/18142320309504007

Tan, F., Huang, Y., Ge, X., Su, G., Ni, X., & Shi, S. (2005). Population genetic structure and conservation implications of*Ceriops decandra* in Malay Peninsula and North Australia.*Aquatic Botany* , *81* (2), 175–188. doi: 10.1016/j.aquabot.2004.11.004

Tomlinson, P. B. (2016). *The Botany of Mangrovess* (Second Edi). Cambridge, UK: Cambridge University Press.

Torstrom, S. M., Pangle, K. L., & Swanson, B. J. (2014). Shedding subspecies: The influence of genetics on reptile subspecies taxonomy.*Molecular Phylogenetics and Evolution* , *76* (1), 134–143. doi: 10.1016/j.ympev.2014.03.011

Van der Stocken, T., Carroll, D., Menemenlis, D., Simard, M., & Koedam, N. (2019). Global-scale dispersal and connectivity in mangroves.*Proceedings of the National Academy of Sciences of the United States of America* , *116* (3), 915–922. doi: 10.1073/pnas.1812470116

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S Springer-Verlag. *New York* .

Wang, X., He, Z., Shi, S., & Wu, C.-I. (2020). Genes and speciation – Is it time to abandon the Biological Species Concept? *National Science Review* . doi: 10.1093/nsr/nwz220

Watterson, G. A. (1977). Heterosis or Neutrality? *Genetics* ,*85* (4), 789–814.

Wee, A. K. S., Takayama, K., Chua, J. L., Asakawa, T., Meenakshisundaram, S. H., Onrizal, ... Kajita, T. (2015). Genetic differentiation and phylogeography of partially sympatric species complex *Rhizophora mucronata* Lam. and *R. stylosa* Griff. using SSR markers. *BMC Evolutionary Biology* , *15* (1), 57. doi: 10.1186/s12862-015-0331-3

Willing, E. M., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers. *PLoS ONE* , *7* (8), 1–7. doi: 10.1371/journal.pone.0042649

Wilson, E. O., & Brown, W. L. (1953). The subspecies concept and its taxonomic application. *Systematic Zoology* , *2* (3), 97–111. doi: 10.2307/2411818

Wu, C.-I. (2001). The genic view of the process of speciation.*Journal of Evolutionary Biology* , *14* (September), 851–865.

Wu, C.-I., & Ting, C. T. (2004). Genes and speciation. *Nature Reviews Genetics* , *5* (2), 114–122. doi: 10.1038/nrg1269

Yan, F., Lu, J., Zhang, B., Yuan, Z., Zhao, H., Huang, S., . . . Che, J. (2018, May 21). The Chinese giant salamander exemplifies the hidden extinction of cryptic species. *Current Biology* , Vol. 28, pp. R590–R592. doi: 10.1016/j.cub.2018.04.004

DATA ACCESSIBILITY

GenBank accession numbers of reference sequences for the genes we sequenced are KC928137-KC928228, KC954697 and KF918414-KF918415 (the detailed information could be found in the supplementary Table S2 of He et al (2019), doi: 10.1093/nsr/nwy078).

AUTHOR CONTRIBUTIONS

S. Shi and Z. Guo designed and supervised the project. S. Shi, C. Zhong, X. Li, H. Lyu and N. C. Duke collected the samples. Z. Wang and H. Lyu produced the data. Z. Wang and Z. Guo analyzed the data. Z. Guo and Z. Wang wrote the manuscript. S. Shi and N. C. Duke helped in improving the manuscript. All the authors read and approved the final manuscript.

Table 1 Sample information and population genetic statistics.

	Location	Longitude & Latitude	Site ID	N[1]	G[2]	Total reads	Depth	Total length	S[3]
1	Meed Creek, Kenya	39°58'6"E, 3deg20'33"S	MC	16	92	6870508	4670	83438	97
2	Laemson, Thailand	98°27'57"E, 9deg36'14"N	LS	35	91	10373578	5966	85999	322
3	Penang, Malaysia	100°22'5"E, 5deg31'34"N	PN	26	93	11894482	6979	88648	287
4	Thongnian, Thailand	99°48'10"E, 9deg18'6"N	TN	35	93	10605220	6100	87742	275
5	Samut Sakon, Thailand	100° 2'6"E, 13deg22'28"N	SS	19	93	12150330	6998	87532	384
6	Ban Kunsha, Thailand	100°26'33"E, 13deg30'1"N	BK	35	93	12291212	6990	87583	382
7	Sanya, China	109°41'16"E, 18deg15'33"N	SY	100	91	15241634	8087	85329	136
8	Wenchang, China	110°50'0"E, 19deg33'35"N	WC	100	93	15431782	7512	86924	118
9	Cebu, Philippines	124° 0'25"E, 10deg21'57"N	CB	26	94	11863938	6938	89399	366
10	Sabah, Malaysia	117°59'27"E, 5deg48'44"N	SB	35	93	11763230	6567	86849	89
11	Bali, Indonesia	115°14'8"E, 8deg42'59"S	BL	35	93	10450180	5837	87181	268
12	Bunbury, Australia	115°39'0"E, 33deg19'33"S	BB	40	93	6834914	3789	82804	358
13	Darwin, Australia	130°54'14"E, 12deg27'44"S	DW	40	92	6746212	4084	84700	165
14	Cairns, Australia	145°47'37"E, 16deg57'22"S	CA	35	88	11609894	6518	77737	104
15	Brisbane, Australia	153° 6'42"E, 27deg21'3"S	BS	40	93	11274220	6062	87426	759
16	Auckland, New Zealand	174°40'44"E, 36deg52'28"S	AK	22	88	11468068	5929	76119	74

Note: [1] N is the sample size, [2] G is the number of genes sequenced, [3] S is the number of segregating sites.

Table 2 Posterior probabilities of models using Approximate Bayesian Computation

		model 1	model 2	model 3	model 4	model 5	model 6	model 7	model 8
Simulation 1	replicate1	0.0007	0.3002	0.0000	0.0000	0.0001	0.0000	0.0000	0.6990
	replicate2	0.0000	0.0844	0.0000	0.0000	0.0788	0.0000	0.0000	0.8368
	replicate3	0.0927	0.1977	0.0000	0.0006	0.0804	0.0000	0.0000	0.6287
Simulation 2	replicate1	0.4001	0.1128	0.1997	0.0000	0.0000	0.0002	0.0000	0.2873
	replicate2	0.1020	0.0081	0.2922	0.0000	0.2188	0.0355	0.0010	0.3435
	replicate3	0.2007	0.2006	0.0000	0.0555	0.0994	0.0000	0.0446	0.3993
		model v1	model v1	model v2	model v2	model v3	model v3	model v4	model v4
Simulation 3		0.0515	0.0515	0.9333	0.9333	0.0118	0.0118	0.0034	0.0034

Figure 1 *Avicennia marina* distribution range and sampling locations. Ranges of the three morphological groups are shown in colors as indicated in the legend. Sampling locations are indicated by circles. Location information and population abbreviations are listed in Table 1. Leaf, flower, and fruit morphological differences are presented on the right and summarized in the imbedded table. Imbedded drawings of morphological traits were adapted from Duke (1991).

Figure 2 Genetic divergence and differentiation among*Avicennia marina* populations. (a-c): colors indicate morphological groups. (a) Multi-dimensional scaling analysis of theF_{ST} and D_{XY} matrices of 16*A. marina* populations. (b) The neighbor-joining tree on the right was constructed using the D_{XY} matrix. (c) Clustering of the *A. marina* populations using principal component analysis (PCA). PCA was performed on the SNP frequency matrix. (d) boxplots of D_{XY} values. "au," "ma," and "eu" indicate *australasica* , *marina,* and*eucalyptifolia* respectively. "maWest" and "maEast" refer to the two recognized geographical groups of *A. m. marina*populations west and east of the Malay Peninsula (see the Results section). "BB" refers to the population from Bunbury, Australia.

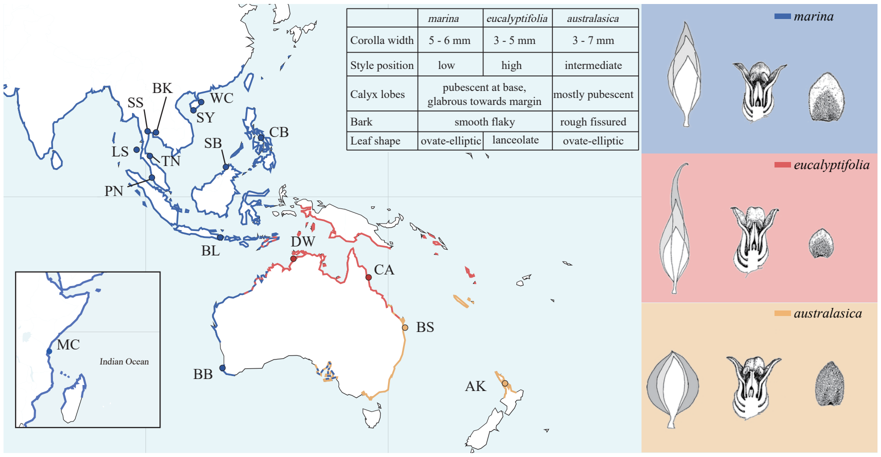
Figure 3 The subspecies evolved independently. a) Boxplots of ϑ computed for each gene in each population (upper graph) and barplots of mean ϑ and π values computed by pooling all SNPs in a population (lower graph). (b) Simulations reconstructing demographic history of*Avicennia marina* populations. Graphical presentation of the ten models of the three subspecies. N stands for effective size and T stands for time of split.

Figure 4 Gene flow between subspecies. (a) Graphical presentation of the four models to investigate the contrast between morphological and genetic characters of the maBB population in western Australia. vT_0 and vT_1 indicate divergence time points and N_{eu}, N_{bb}, and N_{ma} indicated effective population size. The constant bi-directional migration rates are denoted by m_{a} and m_{b}. (b) TreeMix to capture gene flow events on a population splitting graph. On the Maximum likelihood tree, each yellow line indicates a gene flow event between branches it links, with color indicating migration weight. Horizontal branch lengths of the tree are proportional to the amount of genetic drift that has occurred on the branch. The triangle matrix on the top-right indicates residual fit from the maximum likelihood tree. Residuals above zero imply candidate admixture events.

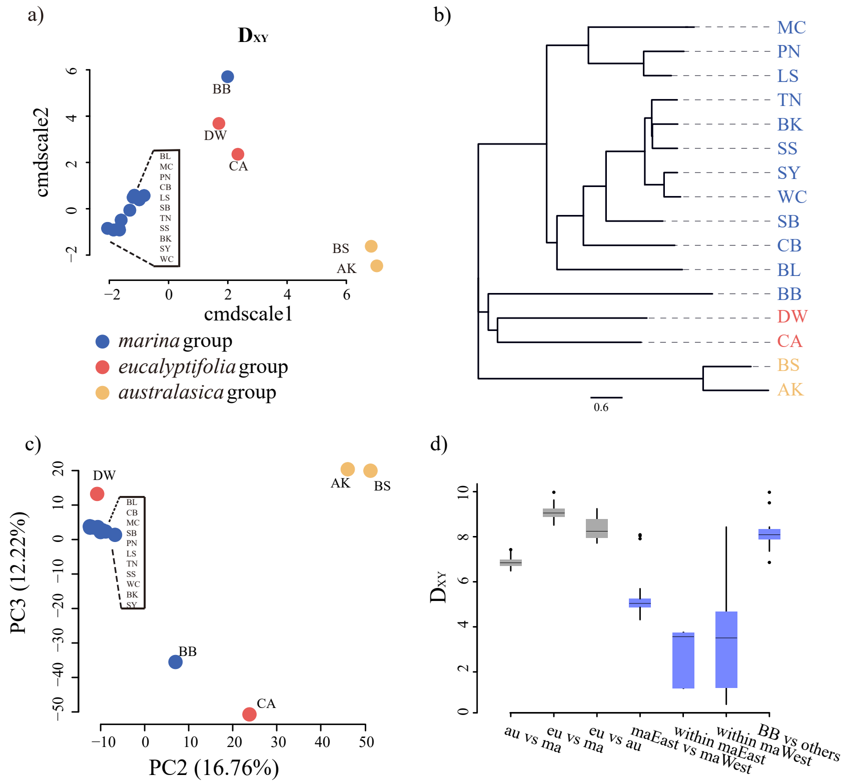
Figure 5 Networks and geographical distribution of haplotypes inferred in eight *Avicennia marina* populations. Haplotypes are indicated by different colours. Lines linking haplotypes reflect mutations, with mutations exceeding a single step marked. The geographic distribution of haplotypes is also indicated. The presented a to f cases are six typical ones to represent six types of haplotype networks. Among the 231 segments, 134, 66, 14, 11, 5, and 1 segments are classified to each type of a to f respectively.

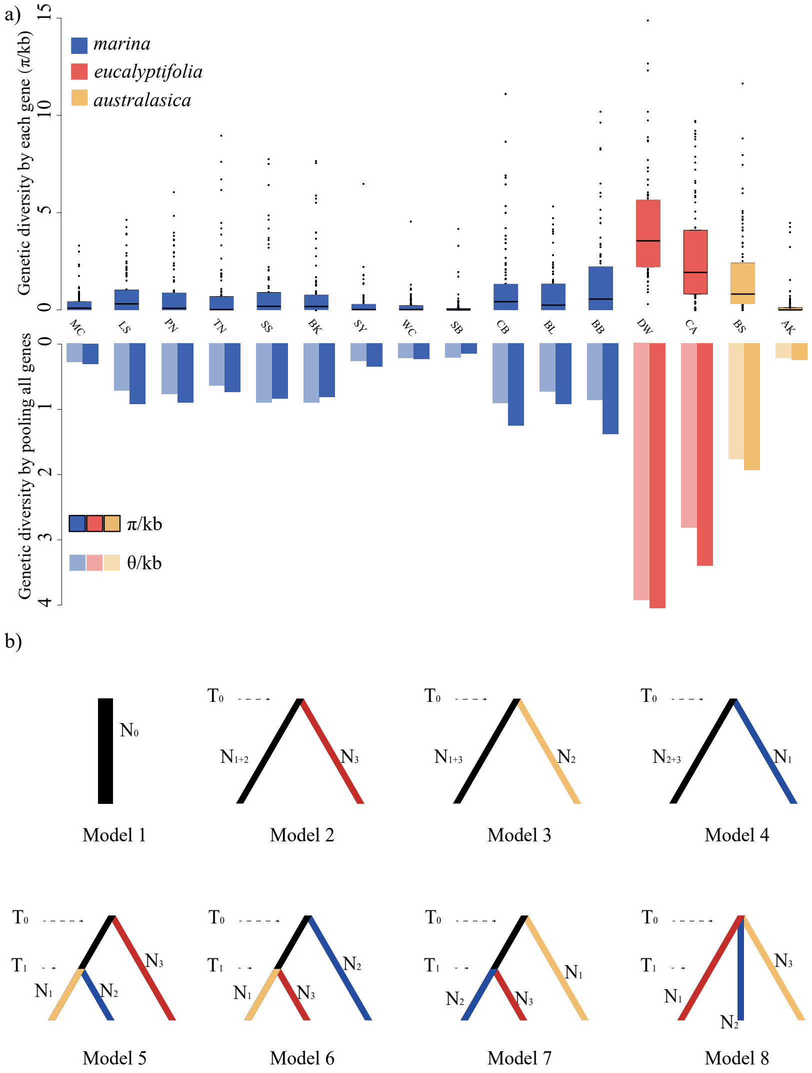
SUPPORTING INFORMATION

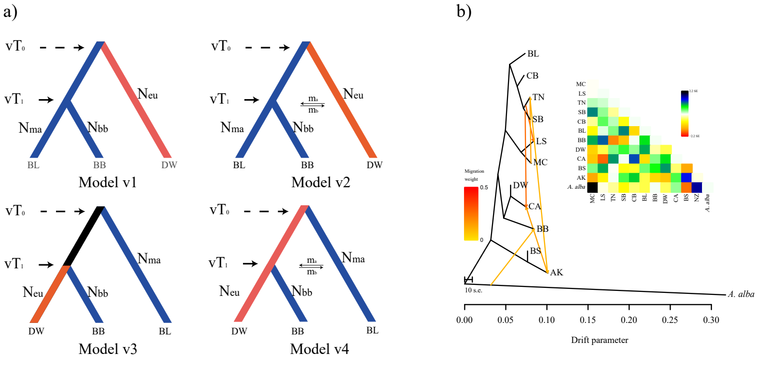
The online supplementary file contains Table S1-S7 and Figure S1-S4.

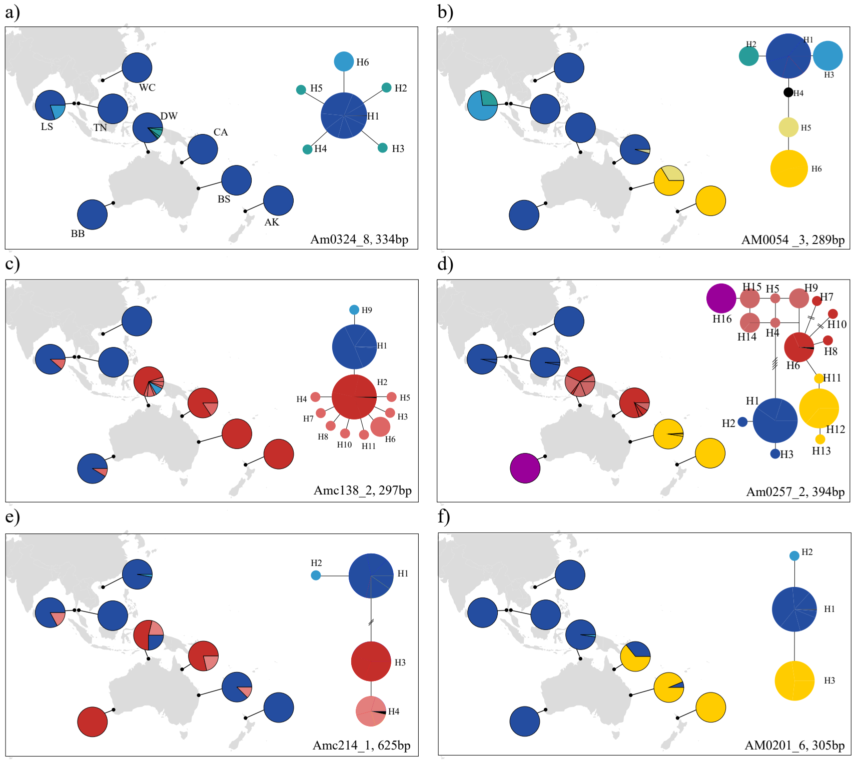


	marina	*eucalyptifolia*	*australasica*
Corolla width	5 - 6 mm	3 - 5 mm	3 - 7 mm
Style position	low	high	intermediate
Calyx lobes	pubescent at base, glabrous towards margin		mostly pubescent
Bark	smooth flaky		rough fissured
Leaf shape	ovate-elliptic	lanceolate	ovate-elliptic




a)
Genetic diversity by each gene (π/kb)
15
10
5
0
marina
eucalyptifolia
australasica
MC
LS
PN
TN
SS
BK
SY
WC
SB
CB
BL
BB
DW
CA
BS
AK
Genetic diversity by pooling all genes
0
1
2
3
4
π/kb
θ/kb
b)
N0
Model 1
T0
N1+2
N3
Model 2
N1+3
N2
Model 3
N2+3
N1
Model 4
T1
N1
N2
N3
Model 5
Model 6
Model 7
Model 8


a)
vT0
vT1
Nma
Nbb
Neu
BL
BB
DW
Model v1
m
Model v2
Model v3
Model v4
b)
BL
CB
TN
SB
LS
MC
DW
CA
BB
BS
AK
A. alba
Migration weight
0.5
0
10 s.e.
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Drift parameter


a)
WC
LS
TN
DW
CA
BS
BB
AK
H6
H5
H2
H1
H4
H3
Am0324_8, 334bp
b)
H2
H1
H3
H4
H5
H6
AM0054 _3, 289bp
c)
H9
H1
H2
H4
H5
H7
H3
H8
H10
H11
H6
Amc138_2, 297bp
d)
H16
H15
H5
H9
H7
H10
H14
H4
H8
H6
H11
H1
H2
H12
H13
H3
Am0257_2, 394bp
e)
H2
H1
H3
H4
Amc214_1, 625bp
f)
H2
H1
H3
AM0201_6, 305bp