bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Genes and the species concept - How much of the genomes can be exchanged?

Xinfeng Wang[a, *], Zixiao Guo[a, *], Shaohua Xu[a], Ming Yang[a], Qipian Chen[a], Sen Li[a], Cairong Zhong[b], Norman C. Duke[c], Ziwen He[a], Chung-I Wu[a,d,e] & Suhua Shi[a]

[a] State Key Laboratory of Biocontrol, Guangdong Key Lab of Plant Resources, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, School of Life Sciences, Sun Yat-Sen University, Guangdong, China

[b] Hainan Dongzhai Harbor National Nature Reserve Administration, Haikou, China

[c] Centre for Tropical Water and Aquatic Ecosystem Research, James Cook University, Townsville, Australia

[d] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[e] Department of Ecology and Evolution, University of Chicago, Chicago, Illinois, USA

[*] These authors contributed equally to this work.

Correspondence should be addressed to S.S. (lssssh@mail.sysu.edu.cn), C.-I.W. (ciwu@uchicago.edu) or Z.H. (heziwen@mail.sysu.edu.cn).

## Abstract

In the biological species concept, much of the genomes cannot be exchanged between species[1,2]. In the modern genic view, species are distinct as long as genes that delineate the morphological, ecological and reproductive differences remain distinct[2]. The rest (or the bulk) of the genomes should be freely interchangeable. The core of the species concept therefore demands finding out the *full potential* of introgressions between species. In a survey of two closely related mangrove species (*Rhizophora mucronata* and *R. stylosa*) on the coasts of the western Pacific and Indian oceans, we found that the genomes are well delineated in allopatry, echoing their morphological and ecological divergence. The two species are sympatric/parapatric in the Daintree River area of northeastern Australia. In sympatry, their genomes harbor 7,700 and 3,100 introgression blocks, respectively, with each block averaging about 3-4 Kb. These fine-grained and strongly-penetrant introgressions suggest that each species must have evolved many differentially-adaptive (and, hence, non-introgressable) genes that contribute to speciation. We identify 30 such genes, seven of which are about flower development, within small genomic islets with a mean size of 1.4 Kb. In sympatry, the species-specific genomic islets account for only a small fraction (< 15%) of the genomes while the rest appears interchangeable.

**Keywords:** mangroves, population genomics, speciation, introgression, species hybridization

## Introduction

The biological species concept (BSC) has been the gold standard of modern evolutionary biology[1-3]. In this concept, species are products of allopatric speciation during which geographical isolation ensures the absence of gene flow. BSC therefore makes clear assumptions about the genetics of species divergence, postulating that nearly the entire genome evolves as a cohesive unit[1]. In the lexicon of genetics, BSC postulates that the density of "speciation genes" is so high that every genomic segment would be precluded from introgressing across the species boundary. BSC is essentially a genomic concept of species[2].

This genomic concept is plausible on phenotypic considerations because even closely-related species differ by a multitude of traits. When such traits are carefully dissected, each has been found to be highly polygenic[4–9], implying extensive genetic divergence. A particularly instructive case is the spermatogenic

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

programs in Drosophila. Among sibling species, the number of genes involved in hybrid male sterility, i.e., spermatogenic failure, is in the hundreds[10–12]. The cloning of the component genes further reveals a complex web of interactions[13,14]. This level of divergence means that selection, presumably sexual in nature, drives spermatogenic programs to evolve extremely rapidly[15] with hybrid sterility being the incidental byproduct. Other "speciation traits" such as sexual isolation, genital morphology and neural development depict a genetic basis that is qualitatively similar[16–18].

Therefore, the genetics of functional divergence between species, where rigorous dissection is possible, shows a surprising degree of cohesiveness postulated by BSC. On the other hand, because BSC demands the cohesiveness across the entire genome, the empirical observations could still fall far short of the requirement by BSC. This latter consideration has led to an alternative view.

In the alternative genic view, species are define by a set of loci that govern the morphological, reproductive, behavioral and ecological characters. These "speciation genes" may, collectively, account for no more than a fraction of the genome[10,12–14,19]. This fraction should be fitness-reducing upon introgression, whereas the rest of the genome can be freely exchanged without a fitness consequence. In short, the diverging genomes comprise both introgressable and non-introgressable DNA segments. These non-introgressable segments are often referred to as "genomic islands" which are, in theory, more divergent than the rest of the genome[2,4,5,12,20–24]. Based on this conjecture, genomic islands have been identified either by the relative level of divergence or by the absolute divergence[4,20,22–29].

Assuming that the current evidence for "speciation with gene flow"[30,31] is convincing, (but see ref. [32,33]), we ask what it takes to corroborate the genic concept of species. It requires answering the following questions: i) the proportion of the genome that is non-introgressable, ii) the number and size distribution of such non-introgressable segments and, in particular, iii) the genic content within these segments. Ideally, the genome of a species can be entirely replaced by that of a closely related species, except for the "speciation genes" themselves as shown in Fig. 1A. The dynamics of such replacement has been modeled in the "Recurrent Selection and Backcross (RSB)" theory[34] but none of these questions has been answered by the current empirical approaches which often do not have the resolution (see Fig. 1A legends).

In this study, we propose a different solution by comparing the same pair of species in allopatry vs. in sympatry. To realize the full introgression potentials, the sympatric taxa also have to meet a number of criteria including the stage of speciation, the timing of secondary contact and the duration of sympatry. Such opportunities may not be common. These criteria will be reviewed in Discussion after the results are presented.

## Results

We study two closely related mangrove species – *Rhizhophora mucronata* vs. *R. stylosa*[35,36]. Mangroves are woody plants that have colonized the intertidal zones of the tropical coasts[35–37]. Because of the narrow band of suitable habitats along the coasts (or near the river mouths), the global distributions of mangroves are essentially one dimensional, making them ideal for biogeographical studies. In particular, the genomes of *R. mucronata* and *R. stylosa* have been published[36] and their speciation history has been analyzed[35]. Built on these previous analyses, this study surveys the allopatric and sympatric populations in their full ranges of distributions.

*Rhizophora mucronata* has a wide distribution in the Indo-Western Pacific (IWP), particularly to the west of the Strait of Malacca and all the way to East Africa. In contrast, *R. stylosa* differs by its extension

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

eastward from the Strait of Malacca to the western Pacific Islands (Fig. 1B). The two species have been reported to overlap in scattered locales along a number of western Pacific coastlines. However, in our own field trips, their relative abundance is often skewed in favor of one species and the co-occurrence has been rarely found. The sole exception in our collection is in the Daintree River (DR) area of northeastern Australia, where both species are quite abundant (Fig. 1B). It is the evidence from this site of sympatry that is instructive about the genic makeup of species.

***Genomic diversity within Rhizophora mucronata and R. stylosa, respectively***

For genomic studies, 21 *R. stylosa* individuals from five locations (labeled s1-s5) and 31 *R. mucronata* individuals from seven locations (named m1-m7) were analyzed (Fig. 1B). Note that m1 and s1 designate the sympatric DR samples. All the samples are sequenced for the whole genome on the Illumina Hiseq 2000 platform, yielding a mean depth of 15X (ranging from 11X to 22X) (Supplementary Table S2 and Table S3). The short reads of each individual are mapped to the reference genome of *R. apiculata*[36], with a genomic coverage of 81% (79% - 82%) (Table S1). The level of genetic diversity shows two patterns. Low genetic diversity is found in all allopatric populations (average $\boldsymbol{\theta}_\pi$ at 0.44 and 0.40 per Kb for *R. mucronata* and *R. stylosa*, respectively) and the level is much higher in the sympatric DR populations ($\boldsymbol{\theta}_\pi$ = 1.05/Kb and 1.22/Kb, respectively). The Watterson's estimates ($\boldsymbol{\theta}_w$) are similar (Table S1) (see Materials and Methods).

***Divergence between the two species in allopatry***

We first constructed a Maximum Likelihood (ML) tree using RAXML on the sequences of the 31 *R. mucronata* and 21 *R. stylosa* individuals from the 11 populations[38]. The ML tree bifurcates with a clear delineation between species across all allopatric populations. However, the m1 and s1 (i.e. DR) samples show strong signs of admixture as they are "in the middle" of the bifurcated tree (Fig. 1B). When the DR samples are removed, the phylogeny shows clear delineation (Fig. 1C). Those two trees are robust when rebuilt using the ML method in IQTREE[39] or the Neighbour-Joining (NJ) method in MEGA7 (Supplementary Fig. S1 and S2)[40]. The monophyletic delineation of *R. mucronata* and *R. stylosa* in allopatry is also supported by the principle component analysis (PCA) (Supplementary Fig. S3)[41].

In total, 1.2 million variable sites are detected across all populations of the two species. We first partition these sites by excluding the DR samples (see Materials and Methods). Each site is then represented by an $F_{ST}$ value with $F_{ST} = 0$ indicating no differentiation between the two species in allopatry and $F_{ST} = 1$ indicating complete differentiation. Figure 1D shows the U-shaped distribution whereby the abundance of sites at the far right reveals the extensive differentiation between species. Such a U-shape distribution is typical of species of some divergence with little gene flow[3].

Morphologically, the two species are distinguished by style length[37,42], as pictured in Fig. 2A. The morphological differences between *R. stylosa* and *R. mucronata* across populations are shown in Fig. 2C. *R. mucronata* is readily distinguished by its short style, in the range of 0.9-1.6mm (Fig. 2A). In contrast, the style of *R. stylosa* is long, 2.4-5.3 mm (Fig. 2A) with no overlap between the two species (Fig. 2C)[37,42]. While the style length varies from locale to locale in both species, this trait is a species-diagnostic one across locales. The two species also show different habitat preferences with *R. mucronata* usually found upriver while *R. stylosa* is close to the river mouth (Figs. 2D). Additional diagnostic morphological characters, which are less stable, are listed in the Supplement (Table S4).

***Characterizations of R. stylosa and R. mucronata in sympatry (the DR samples)***

For the sympatric DR samples, which appear admixed in their DNA sequences, the morphological characters remain distinct. The style length of each sample is concordant with that of the allopatric

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

populations of the same species (Figs. 2B). In fact, the two species in all sympatric populations can be clearly delineated by this character (see Fig. 2C). In the DR area, these two species are parapatric-sympatric with distributions up- or down-river and extensive overlaps in the middle (Figs. 2D). This difference in habitat preference is seen in all locales[42,43].

Corroborating the phylogenetic positions of the DR samples in Fig. 1B, we use the Bayesian clustering analysis, ADMIXTURE[44]. The analysis identifies two genetic components that make up the genomes of the DR samples (Fig. 3A and Supplementary Fig. S4). PCA results also indicate significant admixture in m1 and s1 individuals (Supplementary Fig. S3). Furthermore, because species divergence is monophyletic in all allopatric comparisons, incomplete lineage sorting as the cause of the observed admixture in the DR samples is rejected. In short, we interpret the high $F_{ST}$ sites as manifesting the divergence after speciation (Fig. 1D) with subsequent admixture in the DR area. Additional tests of introgression (LD analysis, *D* stataistic and the modified $f_d$ statistic) are presented in the Supplement (Tables S5-S6 and Figs. S5-S6).

***Extensive introgressions in sympatry***

For the two species in sympatry in the DR area, we ask the following questions: 1) How many introgressed segments can be found in each species? 2) Is the introgression symmetric between the two species? 3) How fine-grained are the introgressed segments (i.e., many small segments or a few large ones)? A few large blocks are expected after recent hybridizations but many fine-grained blocks may result from old introgressions that have been eroded by recombination. If true, introgression (and hence speciation itself) might be a prolonged process. 4) How many genomic segments are non-introgressable and what are their genic contents? Question 4 will be the subject of the next section.

To quantify the introgressions between *R. stylosa* and *R. mucronata* within the DR area, we use the divergent sites of the non-DR samples. There are 228,778 sites with $F_{ST} > 0.8$ between the two species, now referred to as d-sites (d for divergence). Note that the bulk of d-sites (163,089 sites) are fully divergent with $F_{ST} = 1.0$ outside the DR area (Fig. 1D). Furthermore, a fraction of the d-sites are introgression sites in the DR samples (referred to as i-sites). An i-site is where the introgressed allele (or i-allele) is found in >= n of the 10 genomes. (Note that both m1 and s1 samples have five diploid individuals, or 10 genomes; see Materials and Methods). We further impose a condition that the reciprocal introgression of the i-allele can happen at most once (<= 1) in the samples. This second condition is less crucial since high-frequency introgressions in both directions are rare.

It is necessary to set n close to the maximum of 10 for strongly penetrant introgressions. Fig. 3B shows the level of introgressions in the two directions. We set n = 8 for the m1 samples where the i-allele is usually found >= 8 times (the orange bars in Fig. 3B). Hence, the results with n= 2 and n = 8 would not be very different. Furthermore, to avoid the confounding presence of remnant ancient polymorphisms, we require introgressions at an i-site to be strongly asymmetric: >= n one way (say, from *R. stylosa* to *R. mucronata*) and <= 1 in the reciprocal direction (Supplementary Fig. S7). For the *R. stylosa* (s1) samples, the occurrence of the i-allele is rather even between 2 and 10 (the green bars in Fig. 3B). The asymmetry is probably due to the geography of the DR area, which is at the fringe of the *R. mucronata* distribution. Consequently, gene flow from *R. mucronata* into *R. stylosa* may be more limited here, resulting in the lower frequency of introgressions in the s1 samples. In this regard, setting n = 8 would miss many introgressions in *R. stylosa* leading to a much lower introgression rate than in *R. mucronata*. Nevertheless, the final estimations appear robust even when n is set as low as 2 (see below). Simulations of these scenarios are presented in the Supplement.

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Obviously, introgressions do not happen site-by-site, but appear as long segments of DNA consisting of consecutive i-sites. We shall label these segments "introgression blocks" (or i-blocks). Fig. 4A shows a segment of the genome that comprises a string of d-sites, some of which are i-sites as defined above. These d-sites and i-sites are embedded in a background of low-$F_{ST}$ or invariant sites ($F_{ST}$ <= 0.8). This figure shows 3 i-blocks, each consisting of one, two or three i-sites. The length of each block is defined by the distance between the two breakpoints flanking the block. Unless specified, we remove the singleton i-blocks that harbor only a single i-site when presenting the length distribution of i-blocks.

The analysis of i-blocks is summarized in Table 1. We shall focus on the results with n = 8 but the results of n = 2 and n = 10 are given for comparison. In the DR area, samples of *R. mucronata* (m1) harbor far more introgressions than those of *R. stylosa* (s1). The bottom of Table 1 at n=8 shows that 15.8 or 23.4% of the *R. mucronata* genomes are introgressions from *R. stylosa*, the two values depending on whether singleton i-blocks are counted. In the opposite direction, 7.8 – 11.3% of the *R. stylosa* genomes are introgressions. The introgressions of Table 1 can be visualized in Figs. 4-5. The salient observation is the highly fine-grained nature of the introgressions. In *R. mucronata*, the introgressions are distributed over 7,714 i-blocks with an average length of 3.40 Kb. In *R. stylosa*, there are 3,070 i-blocks with an average size of 4.21 Kb. During the evolution, there should be numerous recombination events that break the introgressions into thousands of tiny i-blocks.

It should be noted that Table 1 and Figs. 4-5 present the extreme cases of introgressions that rise to very high frequencies in a non-native genomic background. Because introgressions happen in both directions, beneath these highly penetrant i-blocks are many more introgressions that do not meet the stringent criteria (see the next section for more details).

The distributions of i-blocks are shown at the large genomic scale in Fig. 4B, at the scaffold scale in Fig. 4C and as individual sites in Fig. 5A-5C. Note that only d-sites and i-sites are portrayed in these figures, which convey the visual impression of the fine-grained nature of the i-blocks. (As shown in Fig. 1D, the d- and i-sites are the 228,778 sites with $F_{ST}$ > 0.8; the rest are invariant and lowly divergent sites.) Specifically, the i-blocks are dispersed across the whole genome (Fig. 4B and Supplementary Fig. S9). Indeed, 93 (in s1 genomes) and 96 (in m1 genomes) of the top 100 scaffolds harbor the switching between i- and d-blocks (Table 1). Figure 4C shows that the switching between i- and d-blocks can occur in a few to tens of Kbs. At the site level, i-blocks and d-blocks can switch within a small distance (Fig. 5A-5C). An i-block (or d-block) may harbor only one i-site (or d-site), referred to as singleton block (Table 1 and Supplementary Table S7). Singleton blocks, not uncommon but less reliable, are not used in the tally.

The extensive fine-grained introgressions convey two messages. First, hybridizations may happen continually over a long span of time. Each hybridization event would initially bring in whole-chromosome introgressions that are subsequently broken down by recombination. Small DNA fragments may have been introgressed in this piece-meal manner continually. Second, loci of differential adaptation between species may be very common such that introgressions tend to be small, and thus free of the introgressed alleles that are deleterious in the genetic background of another species[45]. In the next section, we will direct the attention toward non-introgressions, which are blocks of native alleles flanked by introgressed DNA segments.

***Very fine-grained interspersion between "introgressable" and "non-introgressable" blocks***

Some DNA segments may not be introgressable due to the presence of genes of adaptive differences. Such loci, by definition, contribute to reproductive isolation or ecological speciation[2,46] and have sometimes

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

been referred to "speciation genes"[12,13,19,47–49]. The number, size and direction of introgressions are therefore functions of a number of parameters: 1) the rate of hybridization; 2) the strength of selection against the speciation genes when introgressed; 3) the number and location of speciation loci; 4) the rate of recombination that free neutral genes from the linkage to speciation genes; and 5) the length of time since the time of initial hybridization.

To probe the influences of these parameters, we carry out computer simulations based on the Recurrent Selection and Backcross (RSB) model (see Luo et al., 2002[34] and the Materials and Methods). The RSB model is proposed for identifying genes of complex traits[34]. In its execution, one dilutes the genome of breed A (say, the bull dog) with that of breed B (e.g., the border collie) but retains all the desired phenotypic traits of the former. This is done by continually selecting for the traits of breed A while backcrossing the culled products to breed B. The scheme is almost identical with the process of "speciation with gene flow" in their model structure. They differ only in the parameter values; for example, the length of time in speciation is far larger and the gene flow is much smaller, and often bidirectional as well. The differences entail separate simulations for speciation with gene flow.

One particular scenario of speciation with introgression is simulated in Fig. 5D-5E, where two speciation loci, at position 51 and 71, are assumed (see legends). In this demonstration, the introgression occurs in one direction only. At generation 1000, extensive admixture is evident but, as the process continues to generation 10,000, the genome is almost entirely replaced by the introgressed alleles (shown in blue). Importantly, the two speciation loci (shown in pink) resist replacement and bring with them traces of nearby native segments. In this scenario, the selection against introgression is strong (s = -0.05) and the recombination rate for a 100Kb simulated sequence is high (r = 1).

Additional scenarios are presented in Fig. S11A, which shows that a lower recombination rate (r = 0.1) would increase the size of the non-introgressed DNA segments, because the neutral genes near positions 51 and 71 are selected against along with the speciation loci. Figures S11B and S11C show that a reduced selection intensity (s= -0.01) or a 10-fold higher introgression rate would give rise to extensive introgressions. Interestingly, partial introgressions are detected even at positions 51 and 71, where selection acts against the invading alleles. The simulations suggest that, given the right parameter values, the pattern of introgression would follow exactly the prediction based solely on selection, whereby only the alleles of the speciation loci cannot be introgressed. The rest of the genome, even right next to the speciation loci, is freely shared between species.

In the previous section, we define introgressions as the invading DNA segments that rise to 0.8 – 1.0 in frequency (i-blocks), conditional on the reciprocal direction being 0 – 0.1 in frequency. In this section, we focus on the fraction of sites in the genome where the introgression frequency is <= 0.1 in either direction, referred to as j-site (j-site is used as the antonym of i-site). Due to the various patterns of introgressions (Fig. 5D-5E, Fig. S7 and Fig. S11), a large grey area of partial introgression exists in frequencies between 0.1 and 0.8. Many such DNA segments are introgressable but have not risen to a high frequency whereas other segments may harbor non-introgressable loci but reach an appreciable frequency transiently. Therefore, while we will use the stringent low cutoff of 0.1 for identifying putative "speciation genes", we use the much less strong cutoff of <=0.2 for evaluating the size of the non-introgessable genomes. Fig. S7 shows a total of 228,778 sites, among which 31,564 have the i-allele at <=0.2 in both directions. Counting the sites alone, we non-conservatively estimate the fraction of non-introgressable genomes to be < 15% (i.e., 31.6/228.8 = 13.8% < 15%) as stated in the abstract. At present, showing that > 85% of the genomes should be introgressable is the best that we can do.

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

With the j-allele define above, a j-block is defined as a DNA segment containing >= 2 j-sites. Of particular interest within j-blocks are the coding genes that have at least one j-site (Table 2 and Supplementary Table S10). By these stringent criteria, there are only 159 j-blocks which together account for < 0.1% of the genome (Table 2). While only 30 genes containing j-sites are found in these j-blocks (Table 2 and Supplementary Table S10), it is remarkable that 7 of the 30 genes function in flower development and/or gamete production as shown in Table 2 (see the WEGO gene ontology in Supplementary Fig. S10, where a larger set of genes is presented under less stringent criteria). Two of the seven genes regulate flowering period, which is later and shorter in *R. stylosa* than in *R. mucronata*[50]. Mutants of *RA_08689* (encoding MRG family protein) and *RA_19120* (known as *SPF1*) exhibit a late-flowering phenotype in Arabidopsis[51,52]. *RA_11619* and *RA_19120* are involved in female gametophyte development [53,54]. *RA_08689*, *RA_10417*, *RA_13641*, *RA_19120* and *RA_20369* all play a role in pollen germination, pollen tube growth and cotyledon development[55–57]. In particular, *RA_08699* (known as *LFR*) is required for all stages of pollen development[58,59] and the null allele of *LFR* is male-sterile in *A. thaliana*[59]. Since all seven genes contain highly differentiated amino acids and non-introgressable sites (j-sites) (Table 2, Supplementary Table S10 and Fig. S12), their involvement in the speciation between *R. mucornata* and *R. stylosa* seems plausible.

**Discussion**

The species of *R. mucronata* and *R. stylosa* in the DR area are unusual, or possibly unique, among sympatric species reported in the literature[37,60,61] as explained below. These features may be the primary reason that they are fairly close to corroborating the genic view of species, whereby a small fraction of the genomes delineate species.

The DR populations stand out even among comparisons between these two species in other locales, where the sympatric species remain clearly delineated in their genomic sequences. For example, the m2/s2 collections from Singapore both show the expected phylogenetic relationship of their species designation (Fig. 1B). This expected pattern is consistently found in other locales of sympatry: in Brandan, Indonesia[60], in Panay Island, Philippines, in Kosrae, Micronesia, in Yap, Micronesia and in North Sulawesi, Indonesia[61]. The two species in sympatry outside of the DR area occasionally show a slight tendency of being "on the fringe" of their phylogenetic cluster, thus suggesting low-level introgressions. Nevertheless, the extensive fine-grained introgression observed in the DR samples has not been reported before. Importantly, Yan et al. (2016)[61] did notice samples from northeastern Australia (from Trinity Inlet and Daintree River, Queensland) to be different without further clarification.

The near absence of prior reports of fine-grained introgression is understandable as several conditions have to be met for this phenomenon to be realized. The first condition is that the two diverging populations have to be in the right stage when they first come into contact. This stage roughly corresponds to Stage III defined in Wu (2001)[2] whereby speciation is nearly complete but gene flow is still possible. Had the secondary contact happened before this stage (in Stage I or II), the process of speciation could be arrested or even reversed. On the other hand, if the contact starts too late, there would be too little gene flow to give rise to the extensive introgressions observed in the DR area.

Among the locales of sympatry reported for *R. mucronata* and *R. stylosa*, northern Australia has been suggested to be where the two species came into contact in their incipient stage of speciation[37]. In this view, the two diverging taxa moved eastward crossing of the southern Indian Ocean to Australia[37]. They then dispersed north from Australia before spreading east- and westward. By this time, the two species may be

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

too divergent to experience gene flow, thus explaining their clean phylogenetic relationship in sympatry in these other locales.

The second condition may be even more difficult to satisfy – that the two species need to remain in contact for a long period of time after establishing the secondary contact[62]. As discussed, numerous recombination events accumulated over a long period of time are necessary to achieve the fine-grained introgression. Continual gene flow also prevents further build-ups of functional divergence that would lead to the complete cessation of gene flow[2].

The third condition is ecological. Two sympatric species without niche separation would face the problem of competitive exclusion[63], making long-term coexistence unlikely. *R. mucronata* and *R. stylosa* had evolved a degree of niche separation that results in limited overlaps in habit preference (Fig. 2D). Given the necessary confluence of all these conditions, *R. mucronata* and *R. stylosa* in the DR area may be truly exceptional.

In conclusion, non-introgressable DNA segments, or genomic islands, are often portrayed to be large segments of the genome. Instead of a few large "genomic islands"[4,5,20,22,23,29], we observe in the DR samples a large number of tiny islets. In a previous study, the genomic island surrounding the speciation gene, Odysseus, is indeed found to be < 2 Kb[64] and, hence, a veritable islet.

Small introgressions are obviously conducive for the identification of genes driving the adaptive divergence (or speciation genes) as only a few candidate genes are involved (see Tables 2 and Supplementary Table 10). Finally, the contrast between large islands and small islets is important for understanding the process of speciation. The simple model presented above is intended to probe the relative importance of various parameters such as the introgression rate, selection intensity and length of hybridization. Realistic predictions will require careful measurements of all relevant parameters. In particular, measuring the selection intensity against "speciation genes" may reveal how selection drives speciation at the molecular level[2,12,45,48,65–68].

## Acknowledgments

We thank Wei Lun Ng for the photo of *R. mucronata* style in Fig. 2A. This study was supported by the National Natural Science Foundation of China (91731301, 31600182 and 31830005); the National Key Research and Development Plan (2017FY100705); the 985 Project (33000-18841204) and the Fundamental Research Funds for the Central Universities (17lgpy99).

## References

1. Mayr, E. *Animal Species and Evolution*. (Cambridge, MA: Harvard University Press, 1963).
2. Wu C.-I The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
3. Futuyma, D. J. *Evolution*. (Sinauer Associates Inc, 2005).
4. Toews, D. P. L. *et al.* Plumage Genes and Little Else Distinguish the Genomes of Hybridizing Warblers. *Curr. Biol.* **26**, 2313–2318 (2016).
5. Poelstra, J. W. *et al.* The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**, 1410–1414 (2014).
6. Delmore, K. E., Toews, D. P. L., Germain, R. R., Owens, G. L. & Irwin, D. E. The Genetics of Seasonal Migration and Plumage Color. *Curr. Biol.* **26**, 2167–2173 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

7. Van Belleghem, S. M. *et al.* Complex modular architecture around a simple toolkit of wing pattern genes. *Nat. Ecol. Evol.* **1**, (2017).
8. Bay, R. A. *et al.* Genetic Coupling of Female Mate Choice with Polygenic Ecological Divergence Facilitates Stickleback Speciation. *Curr. Biol.* **27**, 3344-3349.e4 (2017).
9. Tavares, H. *et al.* Selection and gene flow shape genomic islands that control floral guides. *Proc. Natl. Acad. Sci.* **115**, 11006–11011 (2018).
10. Wu C, Palopoli MF Genetics of Postmating Reproductive Isolation in Animals. *Annu. Rev. Genet.* **28**, 283–308 (1994).
11. Sawamura, K., Davis, A. W. & Wu, C.-I Genetic analysis of speciation by means of introgression into Drosophila melanogaster. *Proc. Natl. Acad. Sci.* **97**, 2652–2655 (2002).
12. Wu, C.-I & Ting, C. T. Genes and speciation. *Nat. Rev. Genet.* **5**, 114–122 (2004).
13. Ting, C. T., Tsaur, S. C., Wu, M. L. & Wu, C.-I A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* **282**, 1501–1504 (1998).
14. Sun, S., Ting, C. T. & Wu, C.-I The normal function of a speciation gene, Odysseus, and its hybrid sterility effect. *Science* **305**, 81–83 (2004).
15. Coulthart, M. B. & Singh, R. S. High level of divergence of male-reproductive-tract proteins, between Drosophila melanogaster and its sibling species, D. simulans. *Mol. Biol. Evol.* **5**, 182–191 (1988).
16. Hollocher, H., Ting, C. T., Wu, M. L. & Wu, C.-I Incipient speciation by sexual isolation in Drosophila melanogaster: Extensive genetic divergence without reinforcement. *Genetics* **147**, 1191–1201 (1997).
17. Tao, Y., Zeng, Z., Li, J., Hartl, D. L. & Laurie, C. C. Genetic Dissection of Hybrid Incompatibilities Between Drosophila simulans and D. mauritiana. II. Mapping Hybrid Male Sterility Loci on the Third Chromosome. *Genetics* **164**, 1399–1418 (2003).
18. Shi, L., Ji, W. & Su, B. Transgenic rhesus monkeys carrying the human MCPH1 gene copies show human-like neoteny of brain development. *Natl. Sci. Rev.* **6**, in press (2019).
19. Mallet, J. What does Drosophila genetics tell us about speciation? *Trends Ecol. Evol.* **21**, 386–393 (2006).
20. Ellegren, H. *et al.* The genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**, 756–760 (2012).
21. Osada, N. & Wu, C.-I Inferring the Mode of Speciation From Genomic Data. *Genetics* **169**, 259–264 (2005).
22. Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. Genomic islands of speciation in Anopheles gambiae. *PLoS Biol.* **3**, 1572–1578 (2005).
23. Malinsky, M. *et al.* Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350**, 1493–1498 (2015).
24. Harr, B. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**, 730–737 (2006).
25. Renaut, S. *et al.* Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat. Commun.* **4**, (2013).
26. Clarkson, C. S. *et al.* Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. *Nat. Commun.* **5**, (2014).
27. Wang, J., Street, N. R., Scofield, D. G. & Ingvarsson, P. K. Variation in Linked Selection and Recombination Drive Genomic Divergence during Allopatric Speciation of European and American Aspens. *Mol. Biol. Evol.* **33**, 1754–1767 (2016).
28. Ma, T. *et al.* Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. *Proc. Natl. Acad. Sci.* **115**, E236–E243 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. Carneiro, M. *et al.* The genomic architecture of population divergence between subspecies of the European Rabbit. *PLoS Genet.* **10**, (2014).
30. Brandvain, Y., Kenney, A. M., Flagel, L., Coop, G. & Sweigart, A. L. Speciation and Introgression between Mimulus nasutus and Mimulus guttatus. *PLoS Genet.* **10**, (2014).
31. Schumer, M. *et al.* Natural selection interacts with recombination to shape the evolution of hybrid genomes. *Science* **360**, 656–660 (2018).
32. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157 (2014).
33. Yang, M., He, Z., Shi, S. & Wu, C.-I Can genomic data alone tell us whether speciation happened with gene flow? *Mol. Ecol.* **26**, 2845–2849 (2017).
34. Luo, Z. W., Wu, C.-I & Kearsey, M. J. Precision and high-resolution mapping of quantitative trait loci by use of recurrent selection, backcross or intercross schemes. *Genetics* **161**, 915–929 (2002).
35. He, Z. *et al.* Speciation with gene flow via cycles of isolation and migration: Insights from multiple mangrove taxa. *Natl. Sci. Rev.* **6**, 275–288 (2019).
36. Xu, S. *et al.* The origin, diversification and adaptation of a major mangrove clade (Rhizophoreae) revealed by whole-genome sequencing. *Natl. Sci. Rev.* **4**, 721–734 (2017).
37. Duke, N. C., Lo, E. & Sun, M. Global distribution and genetic discontinuities of mangroves - Emerging patterns in the evolution of Rhizophora. *Trees - Struct. Funct.* **16**, 65–79 (2002).
38. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
39. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
40. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **33**, 1870–4 (2016).
41. Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).
42. Duke, N. C. Indo-West Pacific stilt mangroves: Rhizophora apiculata, R. mucronata, R.stylosa, R. x annamalai, R. x lamarckii. In: Elevitch CR (ed), Traditional trees of Pacific Islands: their culture, environment, and use: 641–660.Permanent Agriculture Resources, Hol. in (2006).
43. Duke, N. C., Ball, M. C. & Ellison, J. C. Factors Influencing Biodiversity and Distributional Gradients in Mangroves. *Glob. Ecol. Biogeogr. Lett.* **7**, 27 (2006).
44. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–64 (2009).
45. Fang, S. *et al.* Incompatibility and competitive exclusion of genomic segments between sibling Drosophila species. *PLoS Genet.* **8**, (2012).
46. Schluter, D. Evidence for ecological speciation and its alternative. *Science* **323**, 737–741 (2009).
47. Nosil, P. & Schluter, D. The genes underlying the process of speciation. *Trends Ecol. Evol.* **26**, 160–167 (2011).
48. Presgraves, D. C. The molecular evolutionary basis of species formation. *Nat. Rev. Genet.* **11**, 175–180 (2010).
49. Wittbrodt, J. *et al.* Novel putative receptor tyrosine kinase encoded by the melanoma-inducing Tu locus in Xiphophorus. *Nature* **341**, 415–421 (1989).
50. Duke, N. C. *'World Mangrove iD: expert information at your fingertips' App Store Version 1.2,July 2017. MangroveWatch Publication, Australia - e-book.* (2017).
51. Bu, Z. *et al.* Regulation of Arabidopsis Flowering by the Histone Mark Readers MRG1/2 via Interaction with CONSTANS to Modulate FT Expression. *PLoS Genet.* **10**, (2014).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

52. Kong, X., Luo, X., Qu, G. P., Liu, P. & Jin, J. B. Arabidopsis SUMO protease ASP1 positively regulates flowering time partially through regulating FLC stability. *J. Integr. Plant Biol.* **59**, 15–29 (2017).
53. Yu, H.-J. Analysis of the Female Gametophyte Transcriptome of Arabidopsis by Comparative Expression Profiling. *PLANT Physiol.* (2005). doi:10.1104/pp.105.067314
54. Liu, L. *et al.* Two SUMO Proteases SUMO PROTEASE RELATED TO FERTILITY1 and 2 Are Required for Fertility in Arabidopsis. *Plant Physiol.* **175**, 1703–1719 (2017).
55. Ikeda, Y., Banno, H., Niu, Q. W., Howell, S. H. & Chua, N. H. The ENHANCER of SHOOT REGENERATION 2 gene in Arabidopsis regulates CUP-SHAPED COTYLEDON 1 at the transcriptional level and controls cotyledon development. *Plant Cell Physiol.* **47**, 1443–1456 (2006).
56. Wang, Y. *et al.* Transcriptome Analyses Show Changes in Gene Expression to Accompany Pollen Germination and Tube Growth in Arabidopsis. *Plant Physiol.* **148**, 1201–1211 (2008).
57. Cartagena, J. A. *et al.* The Arabidopsis SDG4 contributes to the regulation of pollen tube growth by methylation of histone H3 lysines 4 and 36 in mature pollen. *Dev. Biol.* **315**, 355–368 (2008).
58. Wang, Z. *et al.* LFR, which encodes a novel nuclear-localized Armadillo-repeat protein, affects multiple developmental processes in the aerial organs in Arabidopsis. *Plant Mol. Biol.* **69**, 121–131 (2009).
59. Wang, X. T., Yuan, C., Yuan, T. T. & Cui, S. J. The arabidopsis LFR gene is required for the formation of anther cell layers and normal expression of key regulatory genes. *Mol. Plant* **5**, 993–1000 (2012).
60. Wee, A. K. S. *et al.* Genetic differentiation and phylogeography of partially sympatric species complex Rhizophora mucronata Lam. and R. stylosa Griff. using SSR markers Phylogenetics and phylogeography. *BMC Evol. Biol.* **15**, 1–13 (2015).
61. Yan, Y.-B., Duke, N. C. & Sun, M. Comparative Analysis of the Pattern of Population Genetic Diversity in Three Indo-West Pacific Rhizophora Mangrove Species. *Front. Plant Sci.* **7**, 1–17 (2016).
62. Coyne, J. A. & Orr, H. A. Appendix: A Catalogue and Critique of Species Concepts. in *Speciation* 447–472 (2004).
63. Hardin, G. The competitive exclusion principle. *Science* **131**, 1292–1297 (1960).
64. Ting, C.-T., Tsaur, S.-C. & Wu, C.-I The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. *Proc. Natl. Acad. Sci.* **97**, 5313–5316 (2002).
65. Payseur, B. A. & Rieseberg, L. H. A genomic perspective on hybridization and speciation. *Mol. Ecol.* **25**, 2337–60 (2016).
66. Seehausen, O. *et al.* Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192 (2014).
67. Soria-Carrasco, V. *et al.* Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**, 738–742 (2014).
68. Terai, Y. *et al.* Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biol.* **4**, 2244–2251 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Materials and Methods

### *Sampling and genome re-sequencing*

To make the samples of *R. mucronata* and *R. stylosa* more representative, we collected individuals both in allopatry and sympatry in the Indo-West Pacific region (Fig. 1). We re-sequenced 31 *R. mucronata* individuals from seven populations and 21 *R. stylosa* individuals from four populations (Fig. 1 and Supplementary Tables S1-S3). To tell apart the two species by morphology, we observed the style length and shape in the bud and took photos (Fig. 2). Fresh leaves were sampled from individual trees and dried with silica gel. DNA isolation was done following the CTAB method[69]. Short-read libraries were sequenced using the Illumina Hiseq 2000 platform with insert size of 350bp and constructed following the TruSeq DNA Sample Preparation Guide. We obtained high quality sequencing data for each individual genome, with coverage in the 12 to 22X range (Supplementary Tables S2-S3).

### *SNP calling and genetic diversity detection*

We used the Genome Analysis Toolkit (GATK)[70] to call variants. Filtered reads from all 52 individuals were mapped to the *R. apiculata* reference genome using the Burrows-Wheeler Aligner (BWA) [71]. Our reference is the *de novo* genome sequence of *R. apiculata*[36]. SAMtools were used to import, sort, and pair bam files and remove duplications. To obtain high quality variants, only SNPs called by both GATK and SAMtools/bcftools were retained[72]. To remove low-quality variants , we eliminated all loci that had base quality (Q) or mapping quality (q) smaller than 20. We additionally applied the following filters: 1) at least two reads had to support the minor allele to call a heterozygote; 2) homozygous SNPs were only retained if read depth was at least 2. After filtering, we selected these high quality sites for further analyses, with multi-allelic (>=3) sites, insertions, and deletions excluded. To estimate genetic diversity in each population, we calculated $\theta_w$ (Watterson's $\theta_w$) and $\theta_\pi$ (Nei and Li's $\theta_\pi$)[73,74] within each population (Supplementary Table S1). To estimate genomic divergence between *R. mucronata* and *R. stylosa* populations, we calculated the genetic differentiation coefficient ($F_{ST}$) (Fig. 1D)[74,75].

### *Detecting gene flow*

We applied Patterson's *D* statistic and a modified $f_d$ statistic to quantify gene flow[76,77]. A positive *D* or $f_d$ value is an indicator of introgression (Supplementary Fig. S8 and Table S6). The basic model has three ingroups ($P_1$, $P_2$, and $P_3$) and the outgroup (O) in the genealogical relationship (((P$_1$,P$_2$),P$_3$),O). In our analysis, $P_1$ and $P_2$ are different populations from the same species *R. mucronata* (or *R. stylosa*), while $P_3$ corresponds to the other species. The outgroup is the reference *R. apiculata*[36]. Positive *D* values imply that $P_2$ and $P_3$ have more shared alleles than $P_1$ and $P_3$ (see Supplement Table S6 and Fig. S6). A software package (plink-1.07) was used to estimate linkage disequilibrium (LD), represented by the $r^2$ statistic within each population or group (Supplementary Fig. S5)[78]. LD decay was used to test for the presence of admixture events. We also calculated LD decay in sympatric populations in Singapore (s2 and m2) and allopatric *R. mucronata* and *R. stylosa* populations (Supplementary Fig. S5) as controls.

### *Genomic scan for introgressed and non-introgressable blocks*

We used four predefined taxa: m1 (*R. mucronata* population in Daintree River), s1 (*R. stylosa* population in Daintree River), $M_{allo}$ (allopatric *R. mucronata* populations m2-m7), and $S_{allo}$ (allopatric *R. stylosa* populations s2-s4). To get a more informative data set, we filtered out sites with too many missing genotypes in each taxon or low divergence ($F_{ST}$ <= 0.8) between $M_{allo}$ and $S_{allo}$. We retained 228,778 SNPs ($F_{ST}$ > 0.8, which we call divergent sites or d-sites between $M_{allo}$ and $S_{allo}$). 163,089 of the d-sites are fixed ($F_{ST}$ = 1.0 and $D_{xy}$ = 1.0) between $M_{allo}$ and $S_{allo}$. There are four possible states of each d-site: homozygous *R.*

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

*mucronata* variant (M type or MM), homozygous *R. stylosa* variant (S type or SS), heterozygote (MS), or missing data.

We then looked for introgressed sites (i-sites) and non-introgressable sites (j-sites) among all the d-sites across m1 and s1 genomes. We have five diploid individuals (10 genomes) from the m1 and s1 populations. We have define allele classes as follows. **Introgressed allele** (**i-allele)**: an *R. stylosa* variant in m1 populations or an *R. mucronata* variant in s1 populations. **i-site:** an i-site in m1 or in s1 genomes is defined as >= 8 occurrences of i-allele out of the 10 genomes (Fig. 3B and Supplementary Fig. S7). **j-site:** a d-site with <=1 occurrences of i-allele in both m1 and s1 populations (Fig. 3B and Supplementary Fig. S7). **i-block:** A genomic block in one species is considered to be introgressed from the other species if one or more i-sites continuously (without disruption by other d-sites) are present (Fig. 4A). The length of an i-block is determined by the midpoint between the flanking (d-sites, i-sites) intervals (as shown in Fig. 4A). **j-block:** a genomic block with one or more j-sites continuously. We define the boundaries the same as for i-blocks.

***Simulations of genomic sequences under hybridization, selection, and recombination***

To probe the influences of hybridization, selection, and recombination on genomic sequences, we carry out computer simulations based on the Recurrent Selection and Backcross (RSB) model (see Luo et al., 2002). We set high and low levels for each parameter. Population size was set at 1000. The length of simulated sequences is 100 kb (for convenience, 1 kb is the basic unit that cannot be separated by recombination). The original allele in the sequence and an i-allele from the other species are differentially labeled. Hence, at the beginning of the simulations, the sequences of all individuals are in original alleles states (100 x). After several generations of hybridization, selection, and recombination, the sequences become shuffled (Fig. 5D-5E and Supplementary Fig. S11).

We first set a low hybridization rate (or introgression, 1/1000 per generation) and recombination (10E-6 per generation between adjacent base pairs). For every generation, 999 individuals are picked from the original population and one is from the other population (or species). The recombination probability (r) for a 100 kb sequence is about 0.1. Since population size is 1000, there will be an average of 100 individuals with recombination in each generation. Two negatively selected loci (#51 and #71) are defined in the simulated sequences. If one or both sites harbor an i-allele, the relative fitness of this sequence is 0.95 (Supplementary Fig. S11A) or 0.99 (Supplementary Fig. S11B). We also examined a high introgression rate regime (10/1000). In this case, four loci (#41, #51, #71 and #76) were set as negatively selected (relative fitness = 0.95 for an i-allele) (Supplementary Fig. S11C). Finally, we simulated genomic sequences under a high recombination rate (10E-5, r = 1.0 for a 100Kb simulated sequence per generation) and a low introgression rate (1/1000 per generation). Two loci (#51 and #71) were negatively selected (relative fitness = 0.95 for an i-allele) (Fig. 5D-5E and Supplementary Fig. S11D-S11F).

69. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
70. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).
71. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr. arXiv* (2013). doi:arXiv:1303.3997 [q-bio.GN]
72. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
73. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 5269–73 (1979).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

74. Nei, M. *Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences of the United States of America* (1973).
75. Wright, S. The genetic structure of populations. *Ann. Eugenetics* **16**, 97–159 (1951).
76. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
77. Green, R. E. *et al.* A draft sequence of the neandertal genome. *Science* **328**, 710–722 (2010).
78. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–75 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1 Summary of high-penetrance introgressed i-blocks between sympatric species

| | >=2 occurrences of i-allele | | >=8 occurrences of i-allele | | =10 occurrences of i-allele | |
|---|---|---|---|---|---|---|
| | **m1** | **s1** | **m1** | **s1** | **m1** | **s1** |
| **No. of i-blocks** | 7,654 | 5,786 | 7,714 | 3,070 | 7,046 | 1,689 |
| **(No. scaffolds with i-blocks)** | (96) | (97) | (96) | (93) | (96) | (88) |
| **Length of i-block** | 12bp–332.5 Kb | 13bp-413.9 Kb | 12bp–332.5 Kb | 13bp-141.1 Kb | 12bp–332.5 Kb | 18bp-141.1 Kb |
| **(mean)** | (3,503 bp) | (3,436 bp) | (3,394 bp) | (4,207 bp) | (2,794 bp) | (4,600 bp) |
| **No. of i-sites in a block** | 2 - 197 bp | 2 - 110 bp | 2 – 140 bp | 2 – 110 bp | 2 - 84 bp | 2 - 85 bp |
| **(total number)** | (30,711 bp) | (26,514 bp) | ( 30,892 bp) | (15,136 bp) | (25,974 bp) | (8,399 bp) |
| **Total length of i-blocks** | 26,812 Mb | 19,881 Mb | 26.181 Mb | 12.917 Mb | 19.686 Mb | 7.769 Mb |
| **(% of the genome)** | (16.19%) | (12.01%) | (15.81%) | (7.80%) | (11.89%) | (4.69%) |
| **Total length of i-blocks[1]** | 40.335 Mb | 28.733 Mb | 38.688 Mb | 18.749 Mb | 32,137 Mb | 11,919 Mb |
| **(% of the genome)** | (24.36%) | (17.35%) | (23.36%) | (11.32%) | (19.41%) | (7.20%) |

Note that the species origin of introgressed alleles (i-alleles) is first defined in the allopatric populations. Hence, i-alleles in the DR area can be identified even when they are bi-directional. All i-alleles in this table are uni-directional with, for example, >=8, in one direction, while the reciprocal direction has <= 1 i-allele. An introgressed block (i-block), unless explicitly stated, should be large enough to have >= 2 introgressed sites (i-sites).

The 100 scaffolds collectively account for 72.617% (165.6 Mb) of the whole genome (228.1 Mb).

[1] These include the singleton i-blocks.

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. High-confidence non-introgressable j-blocks for the identification of genes involved in speciation

| | |
|---|---|
| No. of j-blocks (No. scaffolds with j-blocks) | 159 (59) |
| Length of j-blocks - Range (mean) | 13 bp – 29.1 Kb (1,279 bp) |
| No. of j-sites in a block – Range (total j-sites) | 2 - 7 bp (368 bp) |
| Total length of j-blocks (% of the genome) | 203,286 bp (0.089%) |
| No. of genes with j-sites | 30 |
| No. of genes of flower development with j-sites | 7 (see below) |

| **Gene name** | **L(aa)** | **Site** [1] | **Function in *Arabidopsis thaliana*** |
|---|---|---|---|
| ***RA_08689*** **(*AT4G37280*)** | 259 | 2 | MRG family protein. Regulating flowering through elevating the expression of flowering genes *FLC* and *FT* (FLOWERING LOCUS C and T). The mutant shows a late-flowering phenotype. |
| ***RA_08699*** **(*AT3G22990*)** | 452 | 1 | Armadillo-repeat containing protein. Other names: *LFR*. Required for all stages of pollen development. The expression is particularly strong in the tapetal cells and pollen grains. The null allele is male-sterile. |
| ***RA_10417*** **(*AT4G32440*)** | 461 | 3 | Plant Tudor-like RNA-binding protein. Participating in pollen germination and tube growth. |
| ***RA_11619*** **(*AT5G17200*)** | 300 | 2 | Pectin lyase-like superfamily protein. Involved in carbohydrate metabolic process. Participating in early stage of female gametophyte development. |
| ***RA_13641*** **(*AT3G56600*)** | 551 | 1 | Phosphatidylinositol 4-kinase gamma-like protein. Expressed during: pollen and flower development stages. Participating in pollen germination and tube growth. |
| ***RA_19120*** **(*AT1G09730*)** | 1053 | 4 | SUMO protease. Other names: *ASP1*, *SPF1*. Positively regulating flowering time. Along with *SPF2*, its activity is required for fertility as double mutants have defects in gametogenesis and embroygenesis. |
| ***RA_20369*** **(*AT5G45950*)** | 367 | 1 | GDSL-motif esterase/acyltransferase/lipase. Expressed during flower, leave and plant embryo development stages. Participating in pollen germination and tube growth. |

A j-block, unless explicitly stated, should have >= 2 non-introgressable sites (j-sites).

L(aa): amino acid sequence length of the gene.

[1] Site: No. of highly differentiated amino acids between *R. mucronata* and *R. stylosa* are given (see Supplementary Fig. S12).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

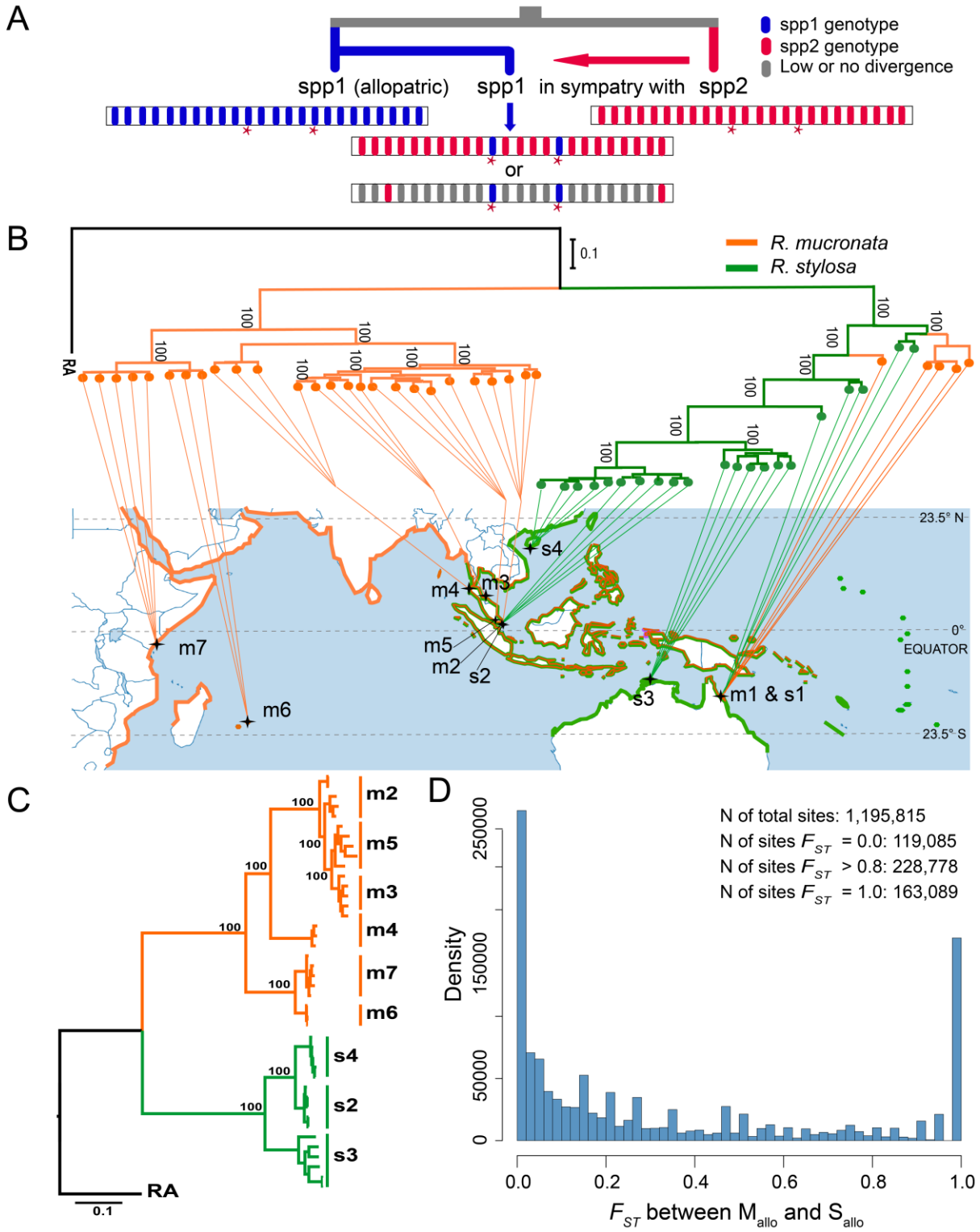


Fig. 1. The phylogeny and biogeography of *R. mucronata* and *R. stylosa*. (A) Hypothetical scenario where a population of species 1 is sympatric with species 2. The one-way introgression (red arrow) may eventually result in the complete replacement of the genomic segments of species 1 (blue) by those of species 2 (red). Massive introgression is predicted by the genic view of species. Note that species with a low level of genomic divergence (gray) cannot provide the conclusive resolution. (B) The maximum-likelihood (ML) tree, generated by RAxML with 100 bootstraps, is superimposed on the biogeography. The numbers on the nodes indicate the supporting values and the short bar near the root indicates the unit of genetic distances. *R. mucronata* is colored in orange while *R. stylosa* is in green. (C) The same phylogeny excluding the sympatric m1 and s1 samples from the Daintree area. We denote the allopatric populations as $M_{allo}$ (m2-m7) and $S_{allo}$ (s2-s3). (D) The spectrum of the $F_{ST}$ statistic between the $M_{allo}$ and $S_{allo}$ samples.

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

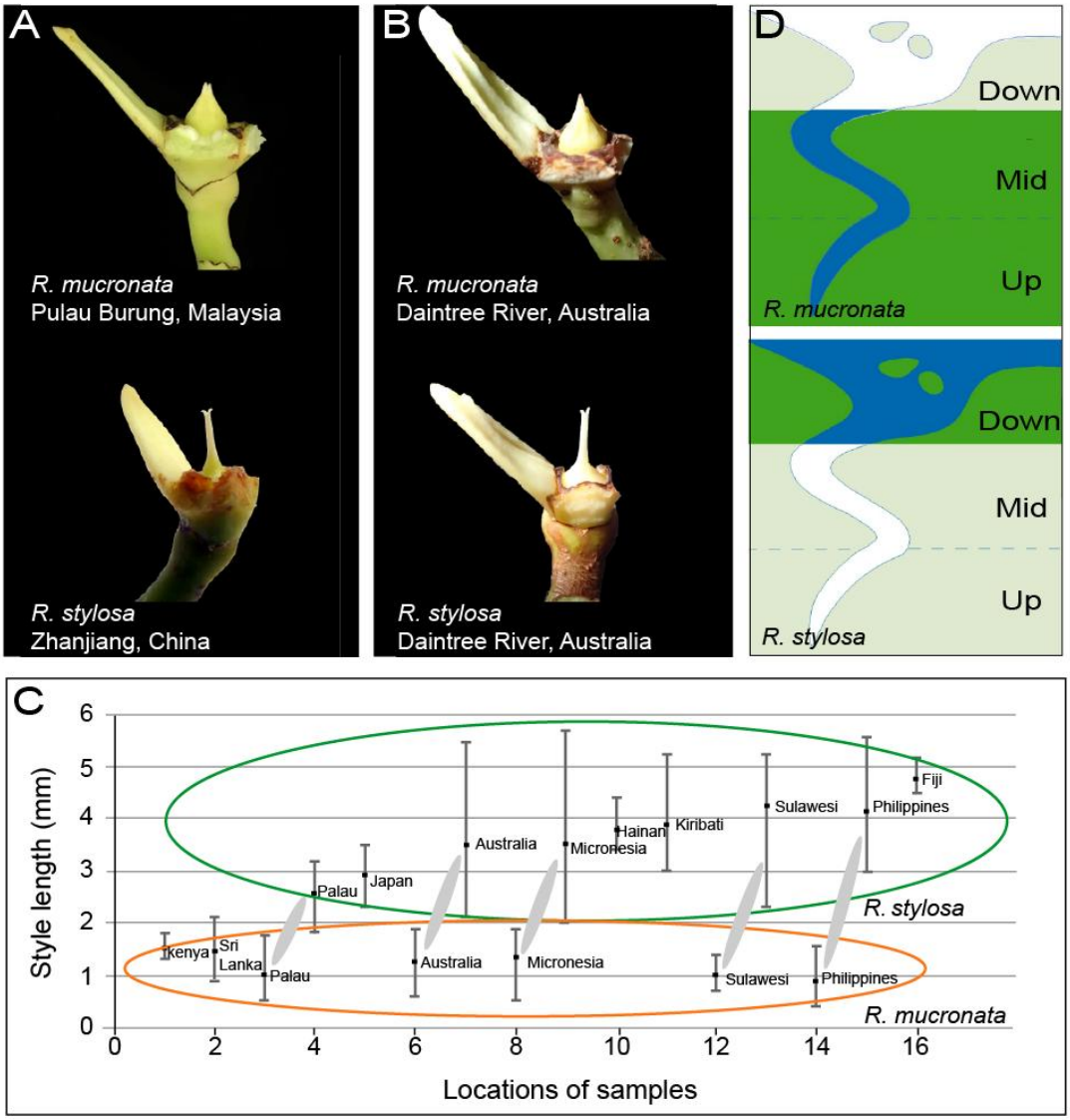


Fig. 2. Key diagnostic characters between *R. mucronata* and *R. stylosa*. (A) The styles of *R. mucronata* (Pulau Burung, Malaysia, 101 °50'14.6"E, 2 °29'33.7'N) and *R. stylosa* (Zhanjiang, China, 109 °45'46.50"E, 21 °34'7.32"N) in allopatric samples. (B) The styles of *R. mucronata* and *R. stylosa* in sympatry in the Daintree River area, Australia. (C) Statistics of the style length of *R. stylosa* (green oval) and *R. mucronata* (orange oval) from different sites throughout the Indo-West Pacific region. Sites where the two species grow in sympatry are linked by a lightgray mark. (D) The habitat preferences of *R. mucronata* and *R. stylosa* in a typical estuary (reproduced from mangrove ID;[50]).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

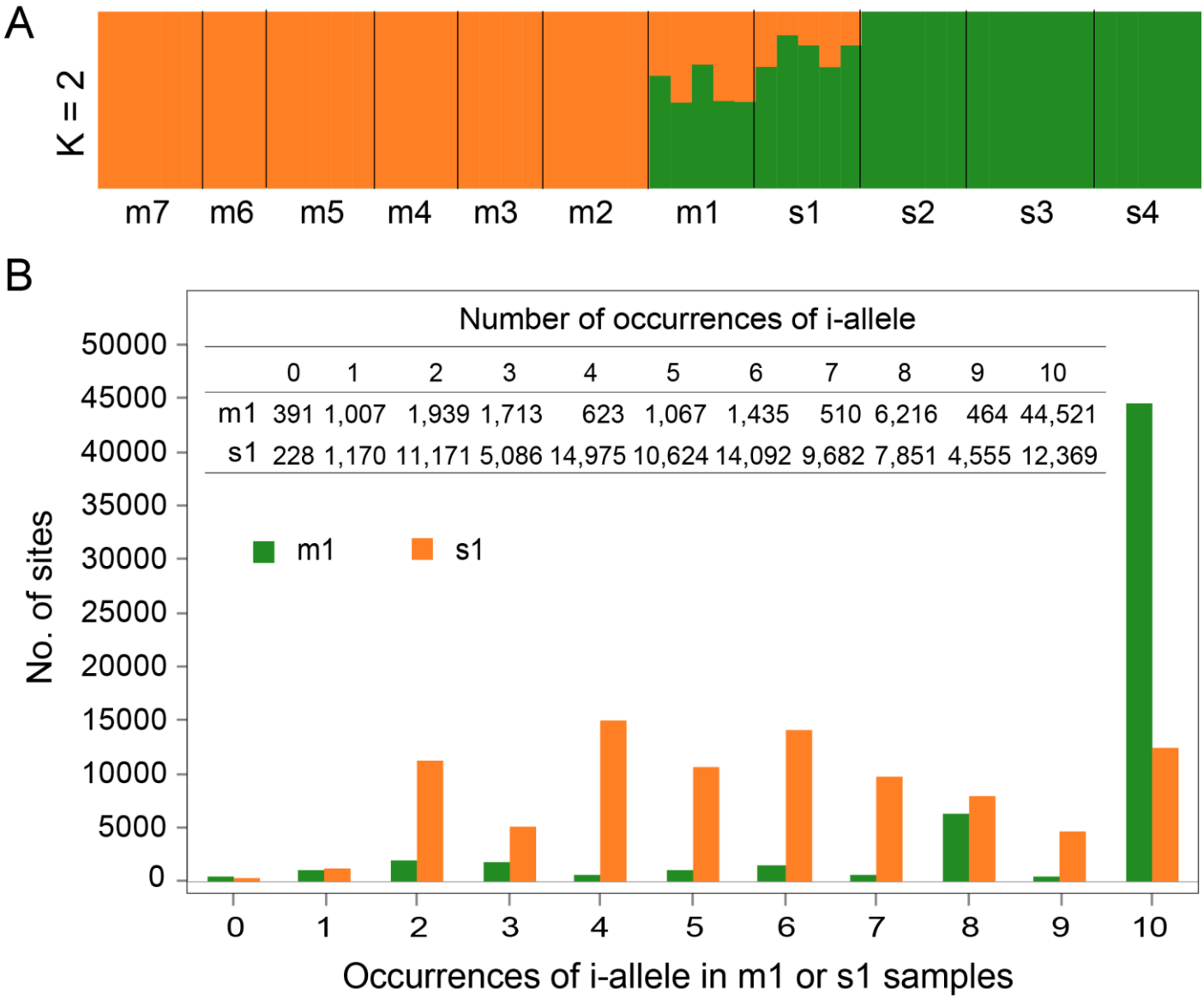


Fig. 3. Admixture of genomes between *R. mucronata* and *R. stylosa*. (A) Genetic clustering of all 52 individuals of the two species is done by ADMIXTURE (K =2). Orange or green color denotes, respectively, the components of *R. mucronata* (m) and *R. stylosa* (s). Each box represents a population with m1 and s1 indicating the sympatric populations in Daintree river, Australia. (B) The site distributions in m1 (orange) and s1 (green) samples, classified by the occurrence of the i-allele (introgressed allele), which ranges from 0 to 10 (i.e., five diploid individuals, or 10 genomes). The actual numbers of sites are shown in the inset table (see also Supplementary Fig. S7).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

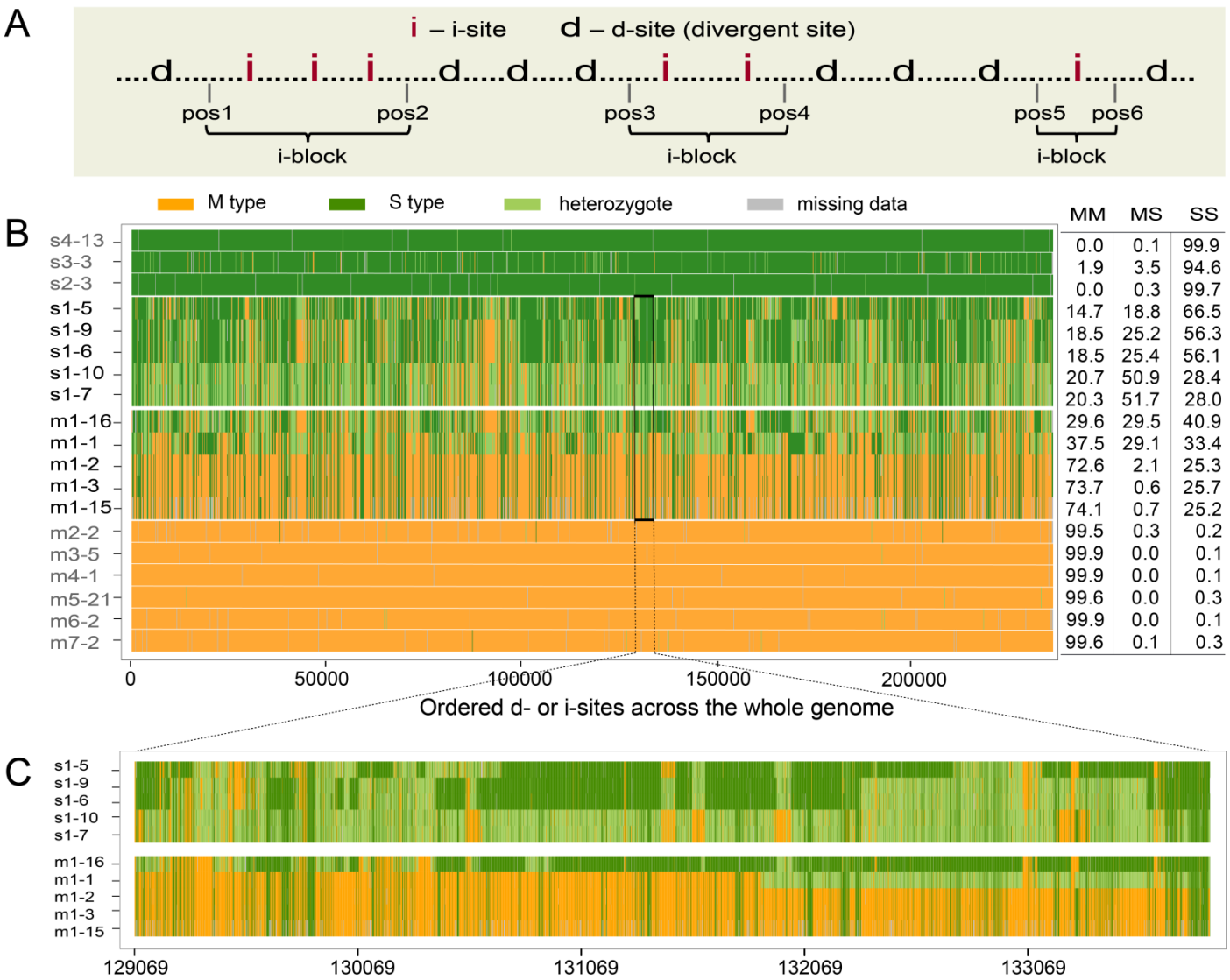


Fig. 4. The genomic compositions of introgressions and non-introgressions. (A) The schematic diagram for defining the i-block, which harbors consecutive i-sites without being interrupted by d-sites. The length of an i-block is determined by the midpoints of the flanking (d, i) intervals. The lengths of the 3 i-blocks are shown. (B) The genome-wide landscape of i-blocks in m1 and s1 samples. The 10 individuals from the sympatric s1 and m1 populations and one individual each from the allopatric populations are shown (see Fig. S9 for the full display). In each ideogram, all d- and i-sites are displayed consecutively. Each site is color coded for the MM (orange), MS (light green) and SS (green) type (M for the *R. mucronata* variant and S for the *R. stylosa* variant). The percentage (%) for the MM, MS and SS sites are summarized to the right of the display. Note the extensive intermingling of the genomes only in the sympatric samples. (C) A fine scale view of the i-blocks in one scaffold (scaffold 864).

bioRxiv preprint doi: https://doi.org/10.1101/622019; this version posted May 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

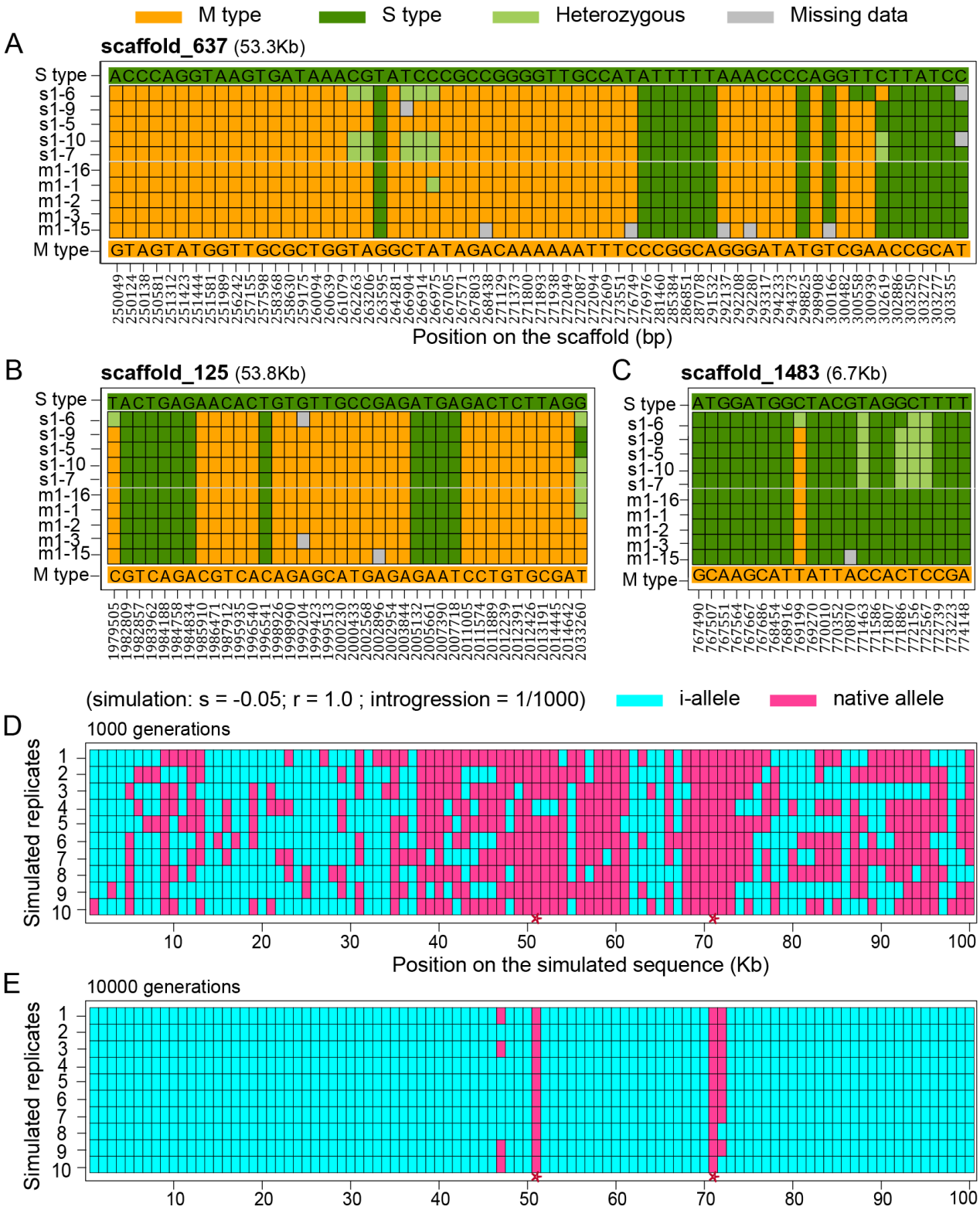


Fig. 5. (A-C) Examples of i-blocks in m1 and s1 samples at the site level. Only the d- and i-sites are displayed and the total length is given next to the scaffold name. Each site is colored coded for the MM, MS and SS types as in Fig. 4. Note the very fine-scale delineation of the blocks. (D-E) Simulated introgressions in haploid 100 Kb genomes. This example is done under strong selection (s = -0.05), high recombination (r = 1.0 for per 100Kb per generation) and low introgression (1/1000 per generation). Two time points are given (see Materials and Methods and Fig. S11 for details). Two speciation genes at 51 and 71 Kb are marked by red stars at the bottom. Sites of introgression and non-introgression are marked blue and pink, respectively. Note that very fine delineations of blocks are possible under the simulated conditions.