Genomic insights into rapid speciation within the world's largest tree genus

Yee Wen Low (lowyeewen@icloud.com)
Singapore Botanic Gardens https://orcid.org/0000-0002-0939-9068

Sitaram Rajaraman
Nanyang Technological University

Crystal Tomlin
University at Buffalo

Joffre Ali Ahmad
Forestry Department

Wisnu Ardi
Kebun Raya Bogor

Kate Armstrong
New York Botanical Garden

Parusuraman Athen
Singapore Botanic Gardens

Ahmad Berhaman
Universiti Malaysia Sabah

Ruth Bone
Royal Botanic Gardens, Kew

Martin Cheek
Royal Botanic Gardens, Kew

Nicholas Rui Wen Cho
Nanyang Technological University

Le Min Choo
Singapore Botanic Gardens

Ian Cowie
Department of Environment, Parks and Water Security

Darren Crayn
James Cook University https://orcid.org/0000-0001-6614-4216

Steve Fleck
University at Buffalo

Andrew Ford
CSIRO

Paul Forster

Brisbane Botanic Gardens

Deden Girmansyah
Herbarium Bogoriense

David Goyder
Royal Botanic Gardens, Kew https://orcid.org/0000-0002-3449-7313

Bruce Gray
James Cook University

Charlie Heatubun
BALITBANGDA Papua Barat

Ali Ibrahim
Singapore Botanic Gardens

Bazilah Ibrahim
Singapore Botanic Gardens

Himesh Jayasinghe
University of Colombo

Muhammad Ariffin Kalat
Forestry Department

Hashendra Kathriarachchi
University of Colombo

Endang Kintamani
Herbarium Bogoriense

Sin Lan Koh
Singapore Botanic Gardens

Joseph Tuck Kwong Lai
National Parks Board

Serena Mei Lynn Lee
Singapore Botanic Gardens

Paul Kiam Fee Leong
Singapore Botanic Gardens

Weihao Lim
Singapore Botanic Gardens

Shawn Kaihekulani Yamauchi Lum
Nanyang Technological University

Ridha Mahyuni
Herbarium Bogoriense

William McDonald
Brisbane Botanic Gardens

Faizah Metali

Environmental and Life Sciences Programme, Faculty of Science, Universiti Brunei Darussalam, Gadong, Brunei Darussalam

Wendy Mustaqim
Universitas Samudra https://orcid.org/0000-0001-9902-830X

Akiyo Naiki
University of the Ryukyus

Kang Min Ngo
Nanyang Technological University

Matti Niissalo
Singapore Botanic Gardens

Subhani Ranasinghe
Department of National Botanic Gardens

Remi Repin
Sabah Parks

Himmah Rustiami
Herbarium Bogoriense

Victor Simbiak
Universitas Papua

Rahayu Sukri
Universiti Brunei Darussalam

Siti Sunarti
Herbarium Bogoriense

Liam Trethowan
Royal Botanic Gardens, Kew

Anna Trias-Blasi
Royal Botanic Gardens, Kew

Thais Vasconcelos
University of Arkansas

Jimmy Wanma
Universitas Papua

Pudji Widodo
Universitas Jenderal Soedirman

Douglas Siril Wijesundara
National Institute of Fundamental Studies

Stuart Worboys
James Cook University https://orcid.org/0000-0001-6706-4509

Jing Wei Yap
Universiti Tun Hussein Onn Malaysia

Kien Thai Yong

Universiti Malaya
Gillian S.W. Khew
Singapore Botanic Gardens
Jarkko Salojärvi
Nanyang Technological University https://orcid.org/0000-0002-4096-6278
Todd Michael
The Salk Institute https://orcid.org/0000-0001-6272-2875
David Middleton
Singapore Botanic Gardens
David Burslem
University of Aberdeen https://orcid.org/0000-0001-6033-0990
Charlotte Lindqvist
University at Buffalo (SUNY)
Eve Lucas
Royal Botanic Gardens, Kew
Victor Albert
University at Buffalo https://orcid.org/0000-0002-0262-826X

Article

Keywords: Species radiation, Syzygium, plant genomics

Posted Date: October 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-969304/v1

Version of Record: A version of this preprint was published at Nature Communications on September 12th, 2022. See the published version at https://doi.org/10.1038/s41467-022-32637-x.

Genomic insights into rapid speciation within the world's largest tree genus

Yee Wen Low[1,2,3,*,†], Sitaram Rajaraman[4,5,*], Crystal M. Tomlin[6,*], Joffre Ali Ahmad[7], Wisnu H. Ardi[8], Kate Armstrong[9], Parusuraman Athen[1], Ahmad Berhaman[10], Ruth E. Bone[2], Martin Cheek[2], Nicholas R.W. Cho[4], Le Min Choo[1], Ian D. Cowie[11], Darren Crayn[12], Steven Fleck[6], Andrew J. Ford[13], Paul I. Forster[14], Deden Girmansyah[15], David J. Goyder[2], Bruce Gray[12], Charlie D. Heatubun[2,16,17], Ali Ibrahim[1], Bazilah Ibrahim[1], Himesh D. Jayasinghe[18,19], Muhammad Ariffin Kalat[7], Hashendra S. Kathriarachchi[18], Endang Kintamani[15], Sin Lan Koh[1], Joseph T.K. Lai[20], Serena M.L. Lee[1], Paul K.F. Leong[1], Wei Hao Lim[1], Shawn K.Y. Lum[21], Ridha Mahyuni[15], William J.F. McDonald[14], Faizah Metali[22], Wendy A. Mustaqim[23], Akiyo Naiki[24], Kang Min Ngo[21], Matti Niissalo[1], Subhani Ranasinghe[25], Rimi Repin[26], Himmah Rustiami[15], Victor I. Simbiak[17], Rahayu S. Sukri[22], Siti Sunarti[15], Liam A. Trethowan[2], Anna Trias-Blasi[2], Thais N.C. Vasconcelos[2,27], Jimmy F. Wanma[17], Pudji Widodo[28], Douglas Siril A. Wijesundara[19], Stuart Worboys[12], Jing Wei Yap[29], Kien Thai Yong[30], Gillian S.W. Khew[1], Jarkko Salojärvi[4,5], Todd P. Michael[31], David J. Middleton[1], David F.R.P. Burslem[3], Charlotte Lindqvist[4,6,†], Eve J. Lucas[2,†], Victor A. Albert[4,6,†]

[1]Singapore Botanic Gardens, National Parks Board, Singapore.
[2]Royal Botanic Gardens, Kew, London, UK.
[3]School of Biological Sciences, University of Aberdeen, Aberdeen, UK.
[4]School of Biological Sciences, Nanyang Technological University, Singapore.
[5]Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland.
[6]Department of Biological Sciences, University at Buffalo, New York, USA.
[7]Brunei National Herbarium, Forestry Department, Ministry of Primary Resources and Tourism, Brunei Darussalam.
[8]Bogor Botanical Garden, Bogor, Indonesia.
[9]New York Botanical Garden, Bronx, New York, USA.
[10]Faculty of Tropical Forestry, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia.
[11]Northern Territory Herbarium, Department of Environment, Parks and Water Security, Darwin, Northern Territory, Australia.
[12]Australian Tropical Herbarium, James Cook University, Cairns, Queensland, Australia.
[13]CSIRO, Land and Water, Tropical Forest Research Centre, Atherton, Queensland, Australia.
[14]Queensland Herbarium, Department of Environment and Science, Brisbane Botanic Gardens, Brisbane, Queensland, Australia.
[15]Herbarium Bogoriense, Cibinong, Indonesia.
[16]BALITBANGDA Papua Barat, Manokwari, Papua Barat, Indonesia.
[17]Universitas Papua, Manokwari, Papua Barat, Indonesia.
[18]Department of Plant Sciences, Faculty of Science, University of Colombo, Sri Lanka.

[19]National Institute of Fundamental Studies, Kandy, Sri Lanka.
[20]Pulau Ubin, Conservation, National Parks Board, Singapore.
[21]Asian School of the Environment, Nanyang Technological University, Singapore.
[22]Universiti Brunei Darussalam, Brunei Darussalam.
[23]Program Studi Biologi, Fakultas Teknik, Universitas Samudra, Langsa, Aceh, Indonesia.
[24]Tropical Biosphere Research Center, University of the Ryukyus, Okinawa, Japan.
[25]National Herbarium, Department of National Botanic Gardens, Peradeniya, Sri Lanka.
[26]Sabah Parks, Kota Kinabalu, Sabah, Malaysia.
[27]Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas, USA.
[28]Faculty of Biology, Universitas Jenderal Soedirman, Puwokerto, Indonesia.
[29]Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Panchor, Johor, Malaysia.
[30]Institute of Biological Sciences, Faculty of Science, Universiti Malaya, Kuala Lumpur, Malaysia.
[31]The Plant Molecular and Cellular Biology Laboratory, Salk Institute for Biological Studies, La Jolla, California, USA.

*These authors contributed equally to this work
[†]Corresponding authors

Author contributions

YWL, VAA, CL, EJL, DFRPB and DJM conceived the study. TPM generated the Oxford Nanopore Technology sequence for and assembled the reference genome. YWL, SR, NRWC, SF, CMT, CL and VAA analysed the genomic data. YWL assembled the morphological data with contributions from HDJ on Sri Lankan taxa; YWL, CMT and VAA performed the character evolutionary analyses. YWL and EJL assembled the biogeographic data; YWL, CMT and VAA carried out the biogeographic analyses. YWL, SR, CMT, CL and VAA wrote the first draft of the paper. EJL, DFRPB, JS, and DJM provided comments on the first draft. All other authors participated in fieldwork and/or provided plant tissue samples or other materials for this work. All authors approved the final paper version.

Acknowledgements

YWL was supported by a postgraduate scholarship research grant from the Ministry of National Development, Singapore awarded through the National Parks Board, Singapore (NParks). Principal research funding from NParks and the School of Biological Sciences (SBS), Nanyang Technological University (NTU), Singapore, is gratefully acknowledged. We thank Peter Preiser, Associate Vice President for Biomedical and Life Sciences, for facilitating NTU support. We are also grateful to Kenneth Er, CEO of NParks, for facilitating research funding through that organisation. VAA and CL were funded by SBS,

NTU for a one-year research leave. JS acknowledges funding from an NTU start-up grant and the Academy of Finland (decision 318288). VAA and CL also acknowledge support from the United States National Science Foundation (grants 2030871 and 1854550, respectively). Fieldwork conducted by YWL was supported by an Indonesian Government RISTEK research permit (Application ID: 1517217008) and an Access License from the Sabah State government [JKM/MBS.1000-2/2JLD.7(84)]. TNCV is grateful to the Assemblée de la Province Nord and Assemblée de la Province Sud (New Caledonia) for facilitating relevant collection permits. AN was partly supported by the Research Project Promotion Grant (Strategic Research Grant No. 17SP01302) from the University of the Ryukyus, and partly by the Environment Research and Technology Development Fund (JPMEERF20204003) from the Environmental Restoration and Conservation Agency of Japan. Administrative support provided by Mui Hwang Khoo-Woon and Peter Ang at the molecular laboratory of the Singapore Botanic Gardens (SBG) are gratefully acknowledged. Rosie Woods and Imalka Kahandawala (DNA and Tissue Bank, Royal Botanic Gardens, Kew) facilitated with additional DNA samples. Daniel Thomas (SBG) and Yan Yu (Sichuan University) provided valuable comments for biogeographical analyses. NovogeneAIT in Singapore is acknowledged for personalised sequencing service. Loh Xiang Yun kindly prepared the line drawing used in this study. This work is dedicated to all field technical assistants who have assisted and supported this study in many ways.

Abstract

Species radiations have long fascinated biologists, but the contribution of adaptation to observed diversity and speciation is still an open question. Here, we explore this question using the clove genus, *Syzygium*, the world's largest genus of tree species comprising approximately 1200 species. We dissect *Syzygium* diversity through shotgun sequencing of 182 distinct species and 58 additional as-yet unidentified taxa, and assess their genetic diversity against a chromosome-level reference genome of the sea apple, *Syzygium grande*. We show that *Syzygium grande* shares a whole genome duplication (WGD) event with other Myrtales. Genomic analyses confirm that *Syzygium* originated in Sahul (Australia-New Guinea), and later diversified eastward to the Hawaiian Islands and westward in multiple independent migration events. The migrations were associated with bursts of speciation events, visible by poorly resolved branches on phylogenies and networks, some of which were likely confounded by incomplete lineage sorting. Clinal genomic variation in some sublineages follows phylogenetic progression, which coupled with sympatric occurrences of distantly related species suggests that both geographic and ecological speciation have been important in the diversification of *Syzygium*. Together, these results point to a mixture of both neutral and adaptive drivers having contributed to the radiation of the genus.

Introduction

Species radiations - wherein perplexing amounts of diversity appear to have formed extremely rapidly - have figured prominently as models in the history of evolutionary theory[1]. Alfred Russell Wallace's independent formulation of the theory of natural selection was motivated by his observations on the immense species diversity of Malesian archipelago in the tropical Far East[2,3], consisting of thousands of islands and including New Guinea and Borneo, the second and third largest islands in the world. Still today the archipelago serves as an ideal observation ground for studying speciation and its processes, as it harbours one of the greatest plant diversities in the world[4,5]. Island radiations, while often holding immense morphological and ecological variation, are often poorly resolved phylogenetically, leading to the impression that evolutionary change can be a swift process that does not require substantial underlying genetic change[6]. Here, we investigate speciation patterns and their potential drivers in the most species-rich tree genus worldwide, *Syzygium*[7].

Syzygium is a member of the myrtle family (Myrtaceae), which includes 1,193 species recognised worldwide[8]. The genus is restricted to tropical and subtropical regions of the Old World, where it is distributed from Africa through to India, across Southeast Asia and extending to Hawaii in the Pacific Ocean, with the centre of species diversity in Indomalesia[8]. The type species of *Syzygium* is *S. caryophyllatum*, a poorly-known, small to medium-sized tree endemic to southern India and Sri Lanka[9]. The best-known species in the genus is the clove tree, *Syzygium aromaticum*, from which flower buds are gathered, dried and used as a spice, a preservative and in pharmacology[10]. In addition, *Syzygium aqueum, S. cumini, S. jambos, S. malaccense* and *S. samarangense* are widely cultivated in the tropics for their large edible fruits[11]. *Syzygium samarangense* is cultivated commercially in Southeast Asia, where it is marketed as the wax apple, java apple, rose apple or samarang rose apple. Apart from being used as cooking ingredients or cultivated for fruits, *Syzygium* species with dense and bushy crowns, such as *S. antisepticum*, *S. australe*, *S. leuhmanii*, *S. myrtifolium* and *S. zeylanicum*, are used in the horticulture industry in Australia, Indonesia, Malaysia and Singapore for hedges, natural fences, natural sound barriers and privacy screens[12].

Syzygium species are generally medium-sized to large, characteristically sub-canopy trees that are sometimes emergent, while some also form shrubs, small forest understorey treelets, swamp and mangrove forest trees, and rheophytic vegetation[13]. As is true of many tropical trees, *Syzygium* flowers are visited by a large diversity of insects and vertebrates, and their fruits are typically eaten by a variety of flying and arboreal vertebrates and even terrestrial bird, mammal and reptile (e.g., tortoises) browsers. *Syzygium* species also occur as dominant mid-level canopy trees, affecting the ecosystems of plants, animals, and fungi in lower forest layers[13]. Many species co-occur; for example, there exist ca. 50 taxa on a single 52-ha. ecological plot in the Lambir Hills National Park (Sarawak, East Malaysia, Borneo[14]), where they display fine-scale differentiation in habitat occupancy and stature[4]. The genus is notorious as one of the most difficult to identify due to the paucity of clear, diagnostic morphological characters for distinguishing species[13,15,16]; morphological variation in the genus can appear as

continua of traits rather than collections of discrete units. Given the immense number of species assigned to *Syzygium*, it contributes disproportionately to the diversity of Southeast Asian tropical forests. Local tree species richness across forests in Southeast Asia is largely driven by a small number of highly species-rich genera, of which *Syzygium* one of the most important. However, geographical variation in genus richness is much less pronounced[4], and therefore understanding diversification within *Syzygium* helps explain large-scale patterns of diversity.

Phylogenetic studies of *Syzygium* have so far involved only a few PCR-amplified plastid and nuclear marker genes[17,18]. An infrageneric classification proposed in 2010 was based on three plastid loci[19], and resolved some major clades in the genus while showing that several previously segregated genera were better placed within a broad monophyletic concept as the single genus *Syzygium*. However, interrelationships within the bulk of the genus, species of *Syzygium* subg. *Syzygium*, were left largely unresolved. Here, we sequenced whole genomes to vastly increase the available data in an attempt to more fully resolve phylogenetic relationships among *Syzygium* species. We used Oxford Nanopore Technology (ONT) long-read sequencing[20] to assemble and annotate a chromosome-scale reference genome for the sea apple, *Syzygium grande*[11]. We examined the palaeopolyploid history of *Syzygium* to assess whether whole genome duplications may have played a role in speciation though sub- or neo-functionalisation events, eventually fixed by natural selection or drift processes during species transitions[21]. We further carried out whole-genome sequencing of 292 *Syzygium* individuals and outgroups to address evolutionary relationships among the species. Both Illumina short-read assemblies as well as mapping of the read data to the *S. grande* genome were carried out for phylogenomic investigations of possible rapid diversification in the group.

Results and Discussion

Assembly and annotation of the reference and resequenced *Syzygium* genomes

A chromosome-level assembly of *Syzygium grande* (Fig. 1A) was carried out using wtdbg2[22] to generate a 405,179,882 bp genome in 174 contigs (N50 of 39,560,356 bp) from more than 60 Gb of ONT long reads, and the assembly was subsequently polished with 30 Gb Illumina short reads. Finally, scaffolding into pseudo-chromosomes was carried out using Dovetail HiC technology[23] to generate 11 pseudo-chromosomes (Fig. 1B) (Supplementary Data S1).

Following the assembly, repeat masking (Supplementary Information, Table S1) and gene prediction was carried out using evidence from *Syzygium grande* RNA-seq data and protein sequences from *Arabidopsis thaliana* and *Populus trichocarpa*. Altogether 39,903 gene models (Supplementary Data S2) were predicted with 86.6% of benchmarked universal single-copy (BUSCO 3.0.2[24]) genes being present. In addition to the reference assembly, 30 Gb of Illumina HiSeqX sequencing data was generated for each of 289 *Syzygium* individuals and three outgroup taxa (two *Metrosideros* and one *Eugenia* species, both Myrtaceae; Supplementary Information, Table S2) and assembled *de novo*

using the MaSuRCA assembler[25] (Supplementary Data S3; Supplementary Information, Table S3 and Fig. S1). The average single-copy completeness across this set of genomes was 89.23% (Supplementary Table S3), indicating that the draft assemblies were of acceptable quality.

Genome structure of *Syzygium* and its phylogenetic context among angiosperms

We used our chromosome-level assembly of *Syzygium grande* to re-evaluate the polyploid history of its family, Myrtaceae, and order, Myrtales. Myrtales are a diverse rosid lineage comprising approximately 13,000 species across 380 genera and 9 families[26]. All rosids share the *gamma* triplication event that occurred in the core eudicot common ancestor[27-29]. Sequencing of the *Eucalyptus grandis* (Myrtaceae) genome revealed an additional whole genome duplication (WGD) in its lineage[30], and later analyses of the *Punica granatum* (pomegranate) genome in the related family Lythraceae suggested that this polyploidy event may have been shared[31,32], occurring near the base of the order. Further work on the *Psidium guajava* (guava) genome came to a similar conclusion[33]. However, the broad, transcriptome-based 1KP project suggested that the Lythraceae and Myrtaceae WGDs might be independent events. Indeed, seven independent, lineage-specific WGDs were predicted by 1KP (their Supplementary Figure 8) to characterise a larger lineage containing *Larrea, Tribulus* (both Zygophyllaceae), Combretaceae, Onagraceae, Melastomataceae, Lythraceae and Myrtaceae[34].

Syntenic alignments of the *Syzygium grande* genome against itself revealed at least one whole genome multiplication event since the *gamma* palaeohexaploidy (Supplementary Information, Fig. S2), and alignment against the *Vitis vinifera* genome confirmed the single lineage-specific WGD (Supplementary Information, Fig. S3). A more detailed study against both *Eucalyptus grandis* and *Punica granatum* revealed 1:1 syntenic relationships (Supplementary Information, Figs. S4 and S5), strongly suggesting a shared polyploid history. We investigated this further by extracting internally syntenic gene pairs in *Eucalyptus grandis, Punica granatum, Vitis vinifera* and *Populus trichocarpa*. When rate-corrected against the *gamma* hexaploidy event[35], an ancient pan-Myrtales WGD was supported, approaching *gamma* in age (Fig. 1D). Furthermore, subgenome-wise fractionation patterns were extremely similar in *Syzygium grande*, *Eucalyptus grandis*, and *Punica granatum*, supporting the hypothesis that a single polyploidy event underlies all Myrtales (Fig. 1E). Furthermore, alignment of *Populus trichocarpa* against given *Syzygium grande* chromosomes showed the expected 2:1 syntenic pattern indicative of an independent Salicaceae-specific WGD in the rosid order Malpighiales (Supplementary Information, Fig. S6). Based on phylogenetic relationships recently solidified for Myrtales families[36], we conclude that earlier genome-based determinations of shared polyploid status within Myrtales are correct in indicating one basal WGD, and that the transcriptome-based 1KP study erroneously inflated the number of WGDs within the clade (Fig. 1C).

Since some polyploid events such as the *gamma* triplication[27-29] and the pan-angiosperm WGD[37] co-occur with major flowering plant radiations[21] (here, the core eudicots and all angiosperms, respectively), a single polyploid event shared by all Myrtales might hold implications for early diversification in the order. However, it is well-

known that some large angiosperm diversifications, such as Gentianales (an even larger lineage than Myrtales at >20,000 species across 1,121 genera and 5 families[26]), are not marked by ancestral WGDs, leaving polyploidy as a causal mechanism for diversification rather inconclusive, or at the very least an incomplete explanation.

Phylogeny and population genomics

Species-level interrelationships within *Syzygium* have not yet been investigated in depth. To obtain a whole genome-level phylogeny we used the *Syzygium grande* genome assembly as a reference for mapping variants from three outgroup taxa and 289 independently sequenced *Syzygium* accessions representing at least 182 distinct species, 49 repeated species samples, and 58 additional as-yet unidentified taxa. SNP calling yielded 1,867,173 variants across all 292 samples, from which we determined genome-wide phylogenetic relationships using RAxML[38]. Since the SNPs could be identified only from the relatively conserved parts of the genome, we also collected predicted universal single copy genes from BUSCO analyses (Supplementary Data S4) of our Illumina-based draft assemblies to generate a complementary phylogenetic estimate using ASTRAL[39], a coalescence-based approach that incorporates individual gene trees into species tree estimation.

Phylogenetic analysis using genome-wide SNPs (Supplementary Data S5) resulted in a phylogeny that was robust and well-resolved with the outgroup-based rooting, namely *Metrosideros excelsa*, *M. nervulosa* (tribe Metrosidereae) and *Eugenia reinwardtiana* (tribe Myrteae). Five major clades resolved in the phylogeny were all well-supported, with most branches receiving 100% bootstrap support at the nodes, indicating strong internal consistency within the dataset. These five major clades represent the previously characterised *Syzygium* subg. *Syzygium*, *S.* subg. *Acmena*, *S.* subg. *Perikion*, *S.* subg. *Sequestratum* and a subgenus yet to be named (matching the circumscription of the genus *Aphanomyrtus*) that includes *S.* cf. *attenuatum*, *S. rugosum* and an unidentified species from Sulawesi labelled here as “SULAWESI2”. It is noteworthy that internal branch lengths are heterogeneous in length, indicating that the clades are differentially divergent either in time, diversification rate, population size, or all of these factors[40]. The largest clade in the phylogeny, having both the most recognised species and the most representative individuals in our current sample, is the *Syzygium* subg. *Syzygium* clade. Relationships within this clade are largely well resolved and supported, as are interrelationships among the five subgenera. Despite strong support, it is important to note that such an analysis generates a phylogeny that represents a genome-wide average, rather than taking into account the independent inheritance of different loci across the genome characteristic of incomplete lineage sorting or adaptive processes.

To obtain an independent view of the *Syzygium* species tree, we used our BUSCO single-copy gene sets (Supplementary Data S4) to compare the gene trees derived from independent nuclear loci in a coalescence species tree approach. We analysed two different BUSCO gene sets that differed in their completeness among accessions: the set of 229 genes containing representatives from all sequenced individuals, and a second set with 1227 genes present in ~95% of accessions. The phylogenies obtained from both single copy genes and genome-wide SNPs (Supplementary Information, Fig. S7)

concordantly displayed the five major, well-supported clades representing the five subgenera of *Syzygium*, including also their relative branching order from the outgroup root, albeit with some minor disagreement of taxon placement within clades. These corroborating results inferred from two different approaches indicate strong and consistent phylogenetic signal within our genomes. Furthermore, *Syzygium* interrelationships based on plastid mappings (Supplementary Data S6), derived by mapping the Illumina sequence reads for each accession onto the *Syzygium grande* plastid genome, yielded partly incongruent results that may be traceable to ancient hybridisation and plastome capture, or to incomplete lineage sorting (ILS) (Supplementary Information, Fig. S8).

Despite strong overall signal supporting a multifurcating evolutionary history, the many extremely short coalescent branch lengths generated by the ASTRAL approach suggest that ILS[40] may have been a confounding biological factor at various points during the *Syzygium* radiation. These branch lengths, which are interpretable in terms of time in generations (g) divided by effective population size (N_e)[40], provide evidence that many *Syzygium* clades either radiated extremely rapidly, or that their ancestral population sizes were comparatively large, or both. Such g and N_e conditions are known to promote gene-tree/species-tree discordance through ILS[40]. We sought to investigate signatures of ILS in the data further using NeighborNet, a distance-based method based on neighbour-joining that generates phylogenetic networks[41]. The character incongruence that is manifested as extra edges in these networks beyond a perfectly bifurcating tree have been interpreted both in terms of interspecies admixture and/or incomplete lineage sorting phenomena[42,43].

NeighborNet analysis of our genome-wide SNP data for *Syzygium* subg. *Syzygium* including a single outgroup species, *S. rugosum*, showed that while many of the evolutionary relationships among taxa were strongly tree-like, at least one major clade (which we informally term here the "*Syzygium grande* group") likely involved a burst of lineage splits (Fig. 3D), as evidenced by the predominantly neutrally evolving SNPs which illustrate a highly webbed network of splits at its base. While the stem lineage of the *Syzygium grande* group was strongly supported in the BUSCO and SNP trees, it is noteworthy that in the SNP analysis many parallel edges nonetheless appear along it, suggesting internal incongruence among SNPs, possibly reflecting differential inheritance with ILS (see Suh et al., 2015[43]). A further, larger lineage including the *Syzygium grande* group and its outgroups was similarly well supported in the BUSCO and SNP trees; however, its own stem lineage contained even more parallel edges and potentially even more severe ILS (Fig. 3D). To rule out admixture generating these results, we formally tested for gene flow within the *Syzygium grande* group using Patterson's f_3 statistic[44] (Supplementary Information, Table S4). Significant negative Z-scores revealed no evidence for admixture, instead marking only source-target taxon pairs that represented joint population membership (i.e., demonstrating shared drift only, via extremely close genetic interrelationships) (Supplementary Information, Table S6).

Next, we used the same SNP data with the ADMIXTURE software[45] to search for genomic partitioning among the clades and accessions that might be attributable to

admixture (introgression) or differential blockwise inheritance through extremely narrow species splits (ILS). ADMIXTURE assumes *K* ancestral population clusters on the data; it is not decisive regarding mechanisms underlying any *K*-cluster mixtures within individuals analysed. The approach was developed for population-level data wherein mixed *K*-clusters are most likely attributable to admixture rather than ILS through lineage splits (e.g., speciation events). However, results at the interspecific level are often interpreted uncritically as actually indicative of cross-lineage admixture[46] (for an example, see Zhang et al., 2020[47]). Indeed, the *K* components from ADMIXTURE simply represent subsets of inherited SNP variation that could reflect any underlying mixtures, of which ILS can be one mechanistic basis (Supplementary Information, Fig. S9). We do not provide any formal tests of these alternatives here, however, ILS does seem a likely underlying causal factor for some of the *K* mixtures given both the short coalescence branch lengths on the ASTRAL species tree and the reticulation of the NeighborNet.

Our ADMIXTURE analysis cross-validation scores supported *K*=14 as the best representation of ancestral population structure (Supplementary Information, Figs. S10 and S11). At this *K*, the *Syzygium grande* group is almost entirely assigned to one *K* in the ADMIXTURE results (orange). However, at other *K* values (e.g., *K*=10,11,12; Supplementary Information, Fig. S9), *K* mixtures within this group are apparent. It is worth noting that the fold level of cross validation affects the preferred number of components, since the optimum results in a case where for each component there is at least one representative in the test set. The outgroups to the *Syzygium grande* group also contain the orange-coloured cluster at *K*=14, but they additionally include mixtures with other ancestral populations (Supplementary Information, Fig. S10). These *K*-cluster mixtures appear to be consistent with the multiple edges underlying this larger lineage in the NeighborNet analysis. In other words, they are likely indicative of differential inheritance of genomic regions and their SNPs through ILS. To rule out admixture generating these results, we formally tested for gene flow within the *Syzygium grande* group using Patterson's f_3 statistic[44] (Supplementary Information, Table S4). Significant negative Z-scores revealed no evidence for admixture, instead marking only source-target taxon pairs that represented joint population membership (i.e., demonstrating shared drift only, via extremely close genetic interrelationships) (Supplementary Information, Table S6).

With ILS the more likely explanation for these results, we used local principal component analysis (PCA)[48] to examine whether patterns of SNP-based relatedness differed instead by location along chromosomes. Clear distinctions in the sample projections on PCA components along a scaffold would indicate that different genomic blocks have different evolutionary histories, of which introgression, ILS, local selection, or even drift are suspect source mechanisms. Local PCA takes window-wise PCA projections of SNP variation and arrays differences among them on a multidimensional scaling (MDS) plot; three distinct "corners"- are then selected from the MDS plot, and the corner-wise variation is pooled for final analyses[48]. We analysed both whole-chromosomal variation as well as repeat-masked data, the latter to ensure that distinct patterns obtained were not solely related to ambiguous mappings due to different transposable element families. The *Syzygium grande* group characteristically appears as a tight cluster across different corners on the 11 chromosomes (Supplementary

Information, Fig. S12). However, in some of these collections of windows, the group is unresolved from its closest outgroups and from the rest of *Syzygium* subg. *Syzygium*; in other corners, these outgroups are poorly distinguished from the remainder of the subgenus, while the *S. grande* group stays distinct. We infer that these results support the hypothesis of underlying ILS – i.e., regional block-wise genomic distinction vs. indistinction of these taxa, as reflected by the many-paralleled edges of their corresponding stem lineages in the NeighborNet result (Fig. 3).

We further studied the SNP data genome-wide using standard PCA[49]. Projections to main principal components illustrated clear clines that mostly correspond to lineages on the BUSCO and SNP trees (Fig. 3B and Supplementary Information, Fig. S13). Several filtrations of data (Supplementary Information, Table S5), including analyses of homozygous sites only (as well as checks for coverage that suggested no apparent biases) yielded similar results and therefore increased confidence that the clinal patterns were not artefactual (Supplementary Information, Fig. S14). A simple explanation for these linear gradations is that allelic variation in *Syzygium* became fixed in consecutive speciation events, along an ongoing cladogenetic process. A PCA analysis of *Syzygium* subg. *Syzygium* (Supplementary Information, Fig. S15) recapitulates this pattern at a smaller scale, and a PCA of the *S. grande* group alone (Fig. 3C) highlights that its two main clades partly overlap in PCA, consistent with short coalescence branch lengths on the BUSCO tree. In other words, the clinal patterns may reflect a neutral process akin to isolation by distance[50,51] (IBD; see, e.g., Seeholzer et al., 2018[52]), for example, comprising serial founder events in an island-hopping model of geographic speciation (but see Barton, 1996[53]). Similar clinal variation among Big Island (Island of Hawai'i) accessions of a closely related Myrtaceae species, *Metrosideros polymorpha* (see Choi et al., 2021, their Figure 3C[54]), might also reflect simple IBD processes in an extremely young and rapidly expanding/dissecting volcanic environment.

Allopatric speciation does not necessarily require adaptive differences, only the null model of reproductive isolation and genetic drift[55]. The possibility of entirely neutral phenotypic clines forming in a model of progressive cladogenesis, such as we hypothesise here for diagnosable *Syzygium* species, may attest to IBD, reflect environmental gradients that accompany spatial population expansion, or even involve admixture between previously isolated populations or clades[52]. However, many *Syzygium* species are sympatrically distributed, which may suggest that ecological speciation[56-58] (and therefore adaptive differences, such as flowering allochrony and other gene flow barriers) could also be operative. Even weak selection via such local adaptation can significantly speed up an entirely drift-based geographic speciation process[55]. For example, the phylogenetic results presented here show that *Syzygium* species sympatric in the Bukit Timah and Danum Valley forest plots are broadly distributed across the phylogenetic tree, but there are also clear clusters of species within some subclades (Fig. 2 and Supplementary Information, Supplementary Note 1 and Table S2). The former phylogenetic scattering suggests that considerable reproductive isolation and lineage diversification evolved prior to any migration to sympatric niches that may have occurred. However, the local phylogenetic clustering seems consistent with *in situ* ecological speciation following simple allopatric lineage splits.

Biogeography and character evolution

Next, we sought to evaluate potential correlates of *Syzygium* phylogenetic structure with geographic distribution patterns and morphological traits (Supplementary Data S7). In a review on the origins and assembly of Malesian rainforests[59], *Syzygium* was highlighted as a key genus for understanding the floristic evolution of the region. Formal biogeographic analyses using the BioGeoBEARS[60] and RASP (Reconstruct Ancestral State in Phylogenies)[61] software each demonstrate, despite limited taxon sampling of outgroups, that the genus *Syzygium* is of Sahul origin, i.e., centred on Australia and New Guinea (Supplementary Information, Figs. S16 and S17). This finding is consistent with previous work on *Syzygium* and Myrtaceae as a whole, which similarly finds Sahul as the ancestral area[62]. We also generated a dated ultrametric (BUSCO) tree to provide split and crown group times for subclades and species diversifications. We used as a calibration point the minimum and maximum ages of a fossil assignable to *Syzygium* subg. *Acmena* (20.9-22.1 Mya)[63]. The crown group of the entire genus *Syzygium* is dated at 51.2 Mya, and the crown groups of subgenera *Sequestratum*, *Perikion*, *Acmena*, the unnamed clade, and *Syzygium* date to 34.1, 24.1, 15.7, 7.0 and 9.3 Mya, respectively (Supplementary Information, Fig. S10). As such, *Syzygium* itself dates to before the Sunda-Sahul convergence which occurred ~25 Mya[64], with most subgenera diversifying after the convergence.

Repeated invasions both westward and northward from Sahul that correspond with species diversifications are clearly apparent. For example, parallel migrations into Sunda occurred at least 12 times (Fig. 2), sometimes corresponding with large radiations, but only within *Syzygium* subg. *Syzygium* (Supplementary Information, Figs. S16 and S17). The earliest migration to Sunda was by 17.1 Mya, the crown group age for *Syzygium* subg. *Sequestratum* (Supplementary Information, Fig. S10). The unnamed clade migrated to Sunda by 7.0 Mya, and *Syzygium* subg. *Perikion* had entered Sunda (Peninsular Malaysia) and later migrated to Sri Lanka by 3.0 Mya. Within *Syzygium* subg. *Acmena*, Sunda had been accessed by 777 Kya. *Syzygium* subg. *Syzygium* is resolved as having a Sahul origin, with a crown group age of 9.3 Mya; according to Hall's 2013 land/sea level reconstructions[65], its entry into Sunda may have involved considerable island hopping from Australia. As many as seven invasions of Sunda occurred, at least three of which (according to our sampling) resulted in hyperdiverse subclades. The earliest Sunda migrations within the type subgenus involved the hyperdiverse *Syzygium pustulatum* group, with a minimum crown group age of 2.8 Mya, and the large *S. creaghii* group, which has a similar minimum crown age of 2.5 Mya. These lineages entered Sunda following the New Guinea uplift, which began about 5 Mya[66]. The extremely diverse *Syzygium grande* group migrated much later from Sahul into Sunda by 466 Kya; it subsequently radiated broadly and very recently into the North Pacific (by 38.8 Kya), the Indian subcontinent (by 58.3 Kya), and from there on to Africa (by 17.3 Kya). The *Syzygium pustulatum* and *S. creaghii* groups, which are marked by fan-like radiations in the NeighborNet analysis, unlike the *S. grande* group, do not show considerable character incongruence suggestive of ILS at its base. The *Syzygium pustulatum* group and smaller and late-migrating *S. jambos* group (the latter having entered Sunda by 233 Kya) also represent rapid diversifications into Sunda with significant tree-like structure at their stem-

lineage bases in the NeighborNet analysis. The last 1 My in Southeast Asian biogeography was marked by cyclical sea level changes that repeatedly divided and rejoined Sunda-Sahul vegetation[67,68], and the minimum invasion dates for the *Syzygium grande* and *S. jambos* groups correspond with periods when sea levels were lower than today[69] and therefore lowland rainforest vegetation more continuous. To summarise, parallel dispersals from Sahul into Sunda and beyond sometimes correlate with what appear to be rapid radiations that at least in one case, the *Syzygium grande* group, appears to have been marked by significant ILS.

We thereafter sought, using Mesquite[70] parsimony optimisations, to discover morphological and accompanying ecological variables that might correlate with these East-to-West migrations. An interesting trait in this regard is the presence of a pseudocalyprate (or "calyprate" in *Syzygium barringtonioides* and *S. perspicuinervium*) versus free corolla (Figs. 4A-C and Supplementary Information, Note 2 and Figs. S18 and S19). A pseudocalyprate corolla, which is relatively common among genera of Myrtaceae, describes a perianth that is variously fused into a cap-like structure that may protect developing stamens from predation, degradation by desiccation, or fungal rot[71]. As determined previously, based on PCR marker phylogenies[17,18] and ontogenetic studies[71], pseudocalyptrate corollas evolved convergently in several *Syzygium* groups. One remarkable transition from free to pseudocalyprate corollas appears at the base of the *Syzygium grande* group; indeed, it was apparently fixed first in its outgroup taxa (Supplementary Information, Fig. S19). Several evolutionary reversals thereafter to free corolla lobes occurred, including one reversal that marks a large sublineage of the *Syzygium grande* group including 75 species as well as *S. grande* itself. The *Syzygium creaghii* and *S. jambos* groups have free corollas, but the *S. pustulatum* group may have been primitively pseudocalyptrate. Regardless, this trait seems highly labile within *Syzygium*, and other than the possible exception of the *S. grande* group's ancestral state, there is no clear correlation with Sahul-to-Sunda diversification. However, one remarkable correlation accompanying diversification of the *Syzygium grande* group into Sunda is the most-parsimonious resolution of green fruits as ancestral to this clade and some of its outgroup species (Fig. 4D and Supplementary Information, Fig. S20). Later transitions from green to purplish-black fruit are also noteworthy in the group. We speculate that this combination of traits - pre-anthesis protection by pseudocalyptrae and bearing of green to purplish-black fruits that attract far-flying birds or bats[72-74] - may have together pre-adapted this group to broad migration.

One other trait of note that marks large diversifications is the presence of pendulous inflorescences, which characterises the *Syzygium creaghii* group and largely marks the *S. longipes* group (Fig. 4E and Supplementary Information, Note 2 and Fig. S21). This trait is correlated with large fruits, often fleshy, which are known to reflect a specialised fruit display and dispersal strategy called flagellichory that increases fruit display for echolocating bats[75], other flying/arboreal vertebrates[76] or large vertebrate browsers (e.g., cassowaries[77]).

Conclusions

Here, we have explored species diversification patterns and their drivers in the world's largest tree genus, *Syzygium*. We generated a high-quality reference genome for *Syzygium grande*, the sea apple, and shotgun sequenced more than 15% of the species of this large genus to study their phylogenomic relationships. Through this extensive sampling of *Syzygium* diversity, we were able to solidify major clade relationships within the genus, currently recognised as subgenera, and, within *Syzygium* subg. *Syzygium*, provided unprecedented clarity on subclades that may become sectional units in the future. We discovered that many *Syzygium* species, particularly within *Syzygium* subg. *Syzygium,* likely branched from one another in rapid succession, yielding true radiations of morphological and ecological diversity. One example was a group of species containing *Syzygium grande* itself that was marked by extremely short coalescence time intervals in our BUSCO species tree; this result was matched by highly networked edges at the base of the group in our NeighborNet analysis, which reflects underlying incongruence in the data. Since none of the f_3 tests showed admixture, we interpret such webbed stem lineages in the NeighborNet network to reflect incomplete lineage sorting during rapid species radiation. PCA analysis of our samples illustrated clines of fixed allelic variation arrayed by *Syzygium* sublineage, possibly reflecting that a simple process of neutral geographic speciation predominated during most of the group's cladogenesis. However, plotting occurrences of species native to Singapore's Bukit Timah Nature Reserve and East Malaysia's Danum Valley Conservation Area illustrated that while large-scale lineage diversification occurred before sympatric occupation of these habitats, some closely related species groups may have evolved *in situ* through adaptive ecological speciation. As such, the immense radiation of the world's largest tree genus may serve as a model for further detailed research, for example at the population level - integrating transcriptomic, proteomic, and metabolomic data - to explore actual mechanisms underlying morphological and ecological specialisation during a diversification that rivals any others under current study.

Methods

Oxford Nanopore sequencing of *Syzygium grande*

Young leaf tissue and twigs of *Syzygium grande* from a cultivated individual (Gleneagles Hospital, along Napier Road, Singapore; *Low s.n.* [SING]) were gathered, cleaned and flash frozen in liquid nitrogen, and then stored in -80 °C prior to extraction. About 10 grams of flash frozen tissue was used for high-molecular-weight (HMW) genomic DNA isolation. The first step followed the BioNano NIBuffer nuclei isolation protocol in which frozen leaf tissue was homogenised in liquid nitrogen, followed by a nuclei lysis step using IBTB buffer with spermine and spermidine added and filtered just before use. IBTB buffer consists of Isolation Buffer (IB; 15 mM Tris, 10 mM EDTA, 130 mM KCl, 20 mM NaCl, 8%(m/V) PVP-10, pH9.4) with 0.1% Triton X-100, and 7.5% (V/V) β-Mercaptoethanol (BME) mixed in and chilled on ice. The mixture of homogenised leaf tissue and IBTB buffer was strained to remove undissolved plant tissue. 1% Triton X-100 was added to lyse the nuclei before centrifugation at 2000 x g for 10 min to pellet the nuclei. Once the

nuclei pellet was obtained, we proceeded with CTAB (Cetyltrimethylammonium Bromide) DNA extraction with modifications for Oxford Nanopore sequencer as described in Michael et al. (2018)[78]. The quality and concentration of HMW genomic DNA was checked using Thermo Scientific™ NanoDrop™ Spectrophotometer, as well as on agarose gel electrophoresis following standard protocols. Genomic DNA obtained was further purified with a Qiagen® Genomic-Tip 500/G following the protocol provided by the developer.

The purified genomic DNA sample obtained was sequenced on Oxford Nanopore Technologies (ONT) PromethION platform. We generated 60,136,770,518 bp of Nanopore reads with a read length N50 of 9,382 bp and an average read quality score of 6.5. Raw ONT reads (fastq) of *Syzygium grande* were filtered prior to assembly using seqtk[79] such that only reads 35 kb or longer were used for genome assembly, which was performed using wtdbg2[22] version 2.2 with flags -p19 -AS2 -e2. The genome consensus was also generated with wtdbg2. Consensus correction was performed was performed with the input ONT reads and three rounds of racon[80]. The assembly generated was polished with Pilon[81] using 30 Gb of 2x150 paired Illumina HiSeqX reads of *Syzygium grande* that were trimmed and filtered. The assembly of *Syzygium grande* comprised 1669 contigs with an N50 length of 556,915 bp. The assembly was filtered for organellar and contaminating contigs using the blobtools[82] pipeline, resulting in the removal of 30 out of 1669 contigs. Next, *purge haplotigs*[83] was used to identify 744 contigs contributing to a diploid peak, which were then removed. These contigs comprised less than 40 Mb of the genome assembly. This filtered primary assembly was thereafter scaffolded into chromosomes by Dovetail HiC technology[23]. The final scaffolded assembly size was 405,179,882 bp.

Transcriptome assembly and annotation of the *Syzygium grande* genome

Transcriptome assembly was carried out for 3 RNASeq libraries (S1: young leaves, S2: mature leaves, S3: twig tips; sequencing performed by NovogeneAIT) separately using an in-house custom assembly pipeline. The first step involved *de novo* assembly for multiple kmer values – 51, 61, 71, 81, 91, 101 using TransAbyss[84] v2.0.1, and for kmer value 25 using Trinity[85] v2.8.5. The second step comprised genome-guided assembly using StringTie[86] v2.0. The input for this second step involved aligning the RNASeq reads against the reference genome using HISAT2[87] v2.1.0. The third step encompassed combining all the results from the first and the second steps using EvidentialGene[88] v2018.06.18 to obtain a final high confidence transcriptome assembly. S1 produced 57,746 transcripts (BUSCO completeness 92.9%), S2 produced 56,536 transcripts (BUSCO completeness 94.6%) and S2 produced 64,163 transcripts (BUSCO completeness 94.1%).

The genome annotation of the reference *Syzygium grande* genome was carried out using an in-house custom annotation pipeline. The first step involved preparation of a *de novo* repeat library using RepeatModeler v1.0.11. This library was used to mask the repetitive regions in the genome assembly using RepeatMasker[89] v4.0.9 resulting in 45.09% of the genome being masked. The second step was the gene prediction step, based on a modular approach using three different gene predictors genemark-es, braker (using the three RNASeq libraries) and GeMoMa[90] (using gene models from the model

species *Arabidopsis thaliana* [TAIR10] and *Populus trichocarpa* [v3.1]). Additionally, the spliced transcript aligner PASA[91] (using transcripts from the three RNASeq libraries) was used to generate evidence for gene structures. These results were then combined using the combiner tool EvidenceModeler[92] to produce a single high confidence final prediction of 39,903 gene models with a BUSCO completeness score of 86.6%. A graphic workflow of these procedures is presented in Supplementary Information, Fig. S22.

Genome structural analyses

The chromosome-level *Syzygium grande* genome assembly and annotation was uploaded to the online CoGe comparative genomics platform (https://genomevolution.org/coge/GenomeInfo.pl?gid=60239)[93]. Syntenic dotplots and data for synonymous substitution rate (Ks) calculations were derived from CoGe SynMap[93] calculations using default settings, with CodeML set to "Calculate syntenic CDS pairs and color dots: Synonymous (Ks) substitution rates". Ks data were collected from corresponding downloads at the "Results with synonymous/non-synonymous rate values" tabs. Each pairwise SynMap analysis (including self:self) was performed for the following species and the CoGe genome IDs: *Syzygium grande* (id60239), *Eucalyptus grandis* (id28624), *Punica granatum* (id61248), *Populus trichocarpa* (id25127), *Vitis vinifera* (id19990). Syntenic dotplots from SynMap were further investigated for synteny relationships within and between species using the FractBias tool[94]. FractBias mappings for fractionation profiles between species were generated using Quota Align syntenic depth of 2:1 for *Syzygium*, *Eucalyptus* and *Punica* against *Vitis* (analyses can be regenerated at https://genomevolution.org/r/1ig9p, https://genomevolution.org/r/1ig9r, and https://genomevolution.org/r/1ig9o, respectively), max query chromosomes =100, max target chromosomes = 25, and "Use all genes in target genome". For *Populus* against *Syzygium* (which can be regenerated at https://genomevolution.org/r/1ig9q), mapping of the former assembly against the latter used a Quota Align syntenic depth of 2:2 and the same options as previously. Density plots (both histogram and smoothed curve) of Ks values for syntenic paralogs were generated in R[95] using the tidyverse[96], ggplot2[97], RColorBrewer[98], ggridges[99], and ggpmisc[100] packages. Ks peaks were calibrated by their shared *gamma* hexaploidy event using the method described by Wang et al., 2015[35].

Illumina sequencing of Syzygium and outgroup individuals

A total of 289 *Syzygium* individuals were selected to represent the six subgenera recognised by Craven & Biffin (2010)[19], across its natural distribution from Africa to the Indian subcontinent, through the Indomalaya region and into the Pacific. Three outgroup taxa in Myrtaceae, *Metrosideros excelsa*, *M. nervulosa* (tribe Metrosidereae) and *Eugenia reinwardtiana* (tribe Myrteae), were also sampled. Most of the 292 samples used in this study were freshly collected in the field, utilising the silica gel teabag method for preserving plant DNA[101], between 2017–2019 either from collecting expeditions conducted in Singapore, Australia, Brunei, Indonesia (West Papua and Papua provinces) and Malaysia, or from cultivated specimens in the Singapore Botanic Gardens (Singapore), Bogor Botanical Garden (Bogor, Indonesia), Cairns Botanic Gardens (Queensland, Australia) and Royal Botanic Gardens, Kew (UK).

Approximately 20 mg of silica-dried leaf tissue were sampled for genomic sequencing. Plant tissue was ground to a fine powder using Omni International Bed Rupture Homogeniser. DNA isolation was carried out at the molecular lab of the Singapore Botanic Gardens using the Qiagen DNeasy® Plant Mini Kit, following the protocol provided by the manufacturer. In rare cases, DNA yields were low when obtained from Qiagen DNeasy® Plant Mini Kit; hence for these problematic samples, the Qiagen DNeasy® Plant Maxi Kit was used instead. Quality and concentration of DNA aliquots were checked using a Thermo Scientific™ NanoDrop™ Spectrophotometer before submission to NovogeneAIT (Singapore) for QC, library construction and sequencing of 30 Gb each (150x150 paired ends) on an Illumina HiSeqX.

Assembly, BUSCO QC, and species tree phylogeny of the resequenced *Syzygium* and outgroup accessions

The 292 Illumina resequenced accessions were assembled using MaSuRCA[25] v3.3.1 with library insert average length of 350 bp and standard deviation of 100 bp. The genome completeness percentages were estimated using BUSCO v4.0.2 based on eudicots_odb10 database.

The phylogeny for the *Syzygium* and outgroup species was estimated using the BUSCO genes. Two species tree versions were estimated. The first tree was estimated using 229 BUSCO genes that were complete and found in all 292 species. The second tree was estimated using 1227 BUSCO genes that were present in 286 species and above.

The species tree generation was constructed using an in-house phylogeny pipeline. The first step involved extraction of BUSCO genes from all resequenced individuals, generating a multi-fasta file for each BUSCO gene containing a representation of that gene from the available species. The second step involved performing multiple sequence alignment (MSA) for each BUSCO multi-fasta file using MAFFT[102] v7.407. The resulting MSA files were used to generate gene trees using RAxML[38] v8.2.12. These gene trees were concatenated and sent as a single input to ASTRAL[39] v5.15.1 to generate the final species tree.

Mapping the resequenced individuals to the *Syzygium grande* reference genome

The 30 Gb each of raw Illumina reads were trimmed to remove adapters using default settings of Trimmomatic[103] version 0.38. Following trimming, the samples were mapped using bwa mem[104] (version 0.7.17), and the subsequent bam file was filtered for a quality score of 20 using samtools[105] view, and sorted using samtools sort. Picard MarkDuplicates (version 2.7.1; https://broadinstitute.github.io/picard/) was used to remove PCR duplicates from the mapped reads. Depth and width of mapping coverage was calculated using BEDTools[106] version 2.23.0.

SNP calling and statistics

SNP calling was performed using GATK version 3.8 in ERC mode for each sample. GenotypeGVCFs was used to call joint genotypes; due to RAM and time limitations this

was split into 70 intervals using the –L flag. To combine the 70 files, GatherVcfs was used to generate a VCF file. As a quality control, GATK VariantFiltration was used with the following filter expression based on GATK recommendations: 'QD < 2.0 || FS > 60.0 || MQ < 50.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0 || SOR > 4.0'. Further filtrations were carried out in VCFtools[107] (version 0.1.13) to create various datasets for downstream analyses. The –plink flag was applied to generate .ped and .map files, and also –recode was used to generate a filtered vcf file. SNP statistics were calculated through vcftools options –het and –singletons for dataset FRSA-1. The output was subsequently plotted using the R package ggplot (https://github.com/tidyverse/ggplot2).

RAxML SNP tree

A pseudo-alignment of SNPs was generated for phylogenetic reconstruction for dataset FRSA1. The plink .ped file was used to convert into fasta input for RAxML. Only variable SNPs were retained for a total of 2,384,277 SNPs. A maximum likelihood tree was generated using RAxML version 8 including adjustments for ascertainment bias (--asc-corr lewis) and 500 bootstraps. Trees were viewed and edited using FigTree[108].

Plastome assembly and phylogenetic tree

We filtered and removed nuclear reads from the *Syzygium grande* Nanopore assembly and constructed a complete chloroplast genome of 158,980 bp in length. This genome was used as a reference to examine phylogenetic relationships of *Syzygium* based on the plastome. Before mapping Illumina reads of 289 *Syzygium* individuals and three outgroup taxa (two *Metrosideros* and one *Eugenia*) to the reference, one of the Inverted Repeat regions (IRs) was removed to prevent sequence calling bias. The combined DNA alignment file of the 292 individuals was then subjected to an ML analysis using RAxML with 1000 bootstrap replicates.

NeighborNet analysis

We used the NeighborNet[41,109] approach to assess incongruence within our SNP data set for *Syzygium* subg. *Syzygium* and *S. rugosum* as an outgroup taxon (FRSA5). We used SplitsTree[110] version 4.17.0 to calculate the network with LogDet[111] distances.

ADMIXTURE analysis

We ran ADMIXTURE[45] (version 1.3) for *K* values 5-15 and used the –cv option to find the best (lowest) *k* value for the number of ancestral populations (Supplementary Information, Fig. S9). The results were then plotted using the barplot function in R.

Local PCA

Single nucleotide polymorphisms (SNPs) called against the draft assembly were transferred to the Hi-C scaffolded assembly through the use of Minimap2[112], transanno (https://github.com/informationsea/transanno), and LiftoverVcf[113]. The BED file needed to remove SNPs from repeat regions was generated using convert2bed[114]. SNPs from repeat regions were removed using the VCFtools --exclude-bed option[107]. The VCF file was divided by the eleven pseudomolecules using HTSlib[115], and converted to BCF format and indexed using BCFtools[116]. Local PCA was carried out using the R[95] lostruct

package[48] and window size was chosen to include about 1,000 SNPs per window as recommended by the authors.

PCA

The eigensoft[117] package (version 6.1.3) (ref) was used to convert plink .map and .ped files into .ind, .geno, and .snp files. Thereafter, the smartpca.perl script was used to run PCA for PC1 to PC10 under default parameters for datasets FRSA1 (all *Syzygium*) and FRSA3 (*Syzygium* subg. *Syzygium*). Taxa that were removed using the smartpca default five rounds of outlier removal shown in red (Supplementary Information, Fig. S13). In addition, three separate PCA checks were performed to confirm that clinal results were not artefactual. PCA was run with SNPs using (i) a more stringent minimum depth of coverage of 20 (Supplementary Information, Fig. S14), (ii) homozygous sites only (Supplementary Information, Fig. S14), and (iii) LD correction turned on using the nsnpldregress option in the smartpca programme to control for linkage. In order to search for possible correlations, the PCA was coloured in numerous ways, by geography, by ecoplots, and also according to ADMIXTURE *K*=14 ancestral groups (Supplementary Information, Fig. S11), using the scatterpie package in R.

F3 statistics

Dataset FRSA5 was used to formally test for admixture using the f3[44] statistic implemented in the qp3pop (version 650) function of the AdmixTools package (https://github.com/DReichLab/AdmixTools). A total of 7,195,530 SNPs were used to test 3,889,44 triplets, every possible combination within the *Syzygium grande* group. We then applied a FDR correction to the Z-scores using a custom R function developed as part of the silver birch genome project[118]. Next, heatmaps were plotted in R for each target.

Biogeography

An ultrametric dated BUSCO phylogeny was generated both for phylogenetic dating and biogeographic reconstruction. The function chronos() in the R package ape[119] (version 3.5.2) was used to create the ultrametric tree. The model used was correlated, and the calibration applied a minimum of 20,900,000 and maximum of 22,100,000 years at the node for which *Syzygium* subg. *Acmena* and *S.* subg. *Syzygium* share a common ancestor. This calibration is based on a *Syzygium* fossil (*S. christophelii*) found in New South Wales, Australia[63]. The ultrametric tree was also used as input to search for and execute the best model using RASP[61] and BioGeoBEARS[60]. Each of the 292 samples were assigned to one or more of eight geographic regions (Africa, India, Mainland Asia, Sunda, Sahul, Wallacea, Zealandia, and Pacific islands) based on their distribution patterns). The best model was found to be DEC+J, which was then plotted using the built-in plotting function plot_BioGeoBEARS_results().

Character evolution with Mesquite

States for three morphological characters – specifically (i) inflorescence habit (erect vs. pendent), (ii) shedding fused corolla present as a true calyptra, a pseudocalyptra, vs. corolla free at anthesis, and (iii) mature fruit colour (green, white or cream, black, pink, purple, red, brown, orange, yellow, blue, or grey) – were gathered from living material, herbarium specimens, published flora accounts, and species protologues. The

categorical morphological characters were coded into the form of numbers 0-9 and/or letters a-z. The ASTRAL tree of BUSCO genes, and selected traits, were loaded into Mesquite[70] version 3.61 and the Trace Character Evolution option with parsimony was selected to predict ancestral states.

References

1. Gavrilets, S. & Losos, J.B. Adaptive radiation: contrasting theory with data. *Science* **323**, 732-737 (2009).
2. Darwin, C., Wallace, A.R., Lyell, S.C. & Hooker, J.D. On the tendency of species to form varieties: and on the perpetuation of varieties and species by natural means of selection. (Linnean Society of London, 1858).
3. Wallace, A.R. *Alfred Russel Wallace: Letters from the Malay Archipelago*, (Oxford University Press, 2013).
4. Ashton, P.S. & Seidler, R. *On the forests of tropical Asia: lest the memory fade*, (Kew Publishing, 2014).
5. Cámara-Leret, R. *et al.* New Guinea has the world's richest island flora. *Nature* **584**, 579-583 (2020).
6. Givnish, T.J. & Sytsma, K.J. *Molecular evolution and adaptive radiation*, (Cambridge University Press, 2000).
7. Beech, E., Rivers, M., Oldfield, S. & Smith, P. GlobalTreeSearch: The first complete global database of tree species and country distributions. *Journal of Sustainable Forestry* **36**, 454-489 (2017).
8. Govaerts, R. *et al. World checklist of Myrtaceae*, (Royal Botanic Gardens, 2008).
9. McVaugh, R. Nomenclatural notes on Myrtaceae and related families (continuation). *Taxon*, 162-167 (1956).
10. Nair, K.N. *The genus Syzygium: Syzygium cumini and other underutilized species*, (CRC Press, 2017).
11. Kochummen, K. & Ng, F. Tree Flora of Malaya (Volume 3). *Kuala Lumpur*, 119-134 (1978).
12. Boo, C.M., Omar-Hor, K., Ou-Yang, C.L. & Ng, C.K. *1001 garden plants in Singapore*, (National Parks, 2003).
13. Parnell, J., Craven, L.A. & Biffin, E. Matters of scale: dealing with one of the largest genera of angiosperms. in *Reconstructing the tree of life: taxonomy and systematics of species rich taxa* (CRC Press LLC, 2007).
14. Lee, H. *et al.* Floristic and structural diversity of mixed dipterocarp forest in Lambir Hills National Park, Sarawak, Malaysia. *Journal of Tropical Forest Science*, 379-400 (2002).
15. Craven, L. Unravelling knots or plaiting rope: What are the major taxonomic strands in Syzygium sens. lat.(Myrtaceae) and what should be done with them? (2001).
16. Schmid, R. A resolution of the Eugenia–Syzygium controversy (Myrtaceae). *American Journal of Botany* **59**, 423-436 (1972).
17. Biffin, E., Craven, L.A., Crisp, M.D. & Gadek, P.A. Molecular systematics of Syzygium and allied genera (Myrtaceae): evidence from the chloroplast genome. *Taxon* **55**, 79-94 (2006).
18. Biffin, E., Harrington, M.G., Crisp, M., Craven, L.A. & Gadek, P. Structural partitioning, paired-sites models and evolution of the ITS transcript in Syzygium and Myrtaceae. *Molecular Phylogenetics and Evolution* **43**, 124-139 (2007).
19. Craven, L.A. & Biffin, E. An infrageneric classification of Syzygium (Myrtaceae). *Blumea-Biodiversity, Evolution and Biogeography of Plants* **55**, 94-99 (2010).
20. Jain, M., Olsen, H.E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology* **17**, 1-11 (2016).
21. Soltis, D.E. *et al.* Polyploidy and angiosperm diversification. *American journal of botany* **96**, 336-348 (2009).
22. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature methods* **17**, 155-158 (2020).

23. Putnam, N.H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research* **26**, 342-350 (2016).
24. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
25. Zimin, A.V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669-2677 (2013).
26. Stevens, P.F. Angiosperm Phylogeny Website. Version 17. *Angiosperm Phylogeny Website. Version 17.* (2017).
27. Chanderbali, A.S., Berger, B.A., Howarth, D.G., Soltis, D.E. & Soltis, P.S. Evolution of floral diversity: genomics, genes and gamma. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20150509 (2017).
28. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *nature* **449**, 463 (2007).
29. Jiao, Y. *et al.* A genome triplication associated with early diversification of the core eudicots. *Genome biology* **13**, 1-14 (2012).
30. Myburg, A.A. *et al.* The genome of Eucalyptus grandis. *Nature* **510**, 356-362 (2014).
31. Qin, G. *et al.* The pomegranate (Punica granatum L.) genome and the genomics of punicalagin biosynthesis. *The Plant Journal* **91**, 1108-1128 (2017).
32. Yuan, Z. *et al.* The pomegranate (Punica granatum L.) genome provides insights into fruit quality and ovule developmental biology. *Plant biotechnology journal* **16**, 1363-1374 (2018).
33. Feng, C. *et al.* A chromosome-level genome assembly provides insights into ascorbic acid accumulation and fruit softening in guava (Psidium guajava). *Plant biotechnology journal* **19**, 717-730 (2021).
34. Leebens-Mack, J.H. *et al.* One thousand plant transcriptomes and the phylogenomics of green plants. (2019).
35. Wang, X. *et al.* Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Molecular plant* **8**, 885-898 (2015).
36. Maurin, O. *et al.* A nuclear phylogenomic study of the angiosperm order Myrtales, exploring the potential and limitations of the universal Angiosperms353 probe set. *American Journal of Botany* (2021).
37. Albert, V.A. *et al.* The Amborella genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
38. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
39. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-i548 (2014).
40. Maddison, W.P. Gene trees in species trees. *Systematic biology* **46**, 523-536 (1997).
41. Bryant, D. & Moulton, V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution* **21**, 255-265 (2004).
42. Burgon, J.D. *et al.* Phylogenomic inference of species and subspecies diversity in the Palearctic salamander genus Salamandra. *Molecular Phylogenetics and Evolution* **157**, 107063 (2021).
43. Suh, A., Smeds, L. & Ellegren, H. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biol* **13**, e1002224 (2015).
44. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093 (2012).
45. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome research* **19**, 1655-1664 (2009).

46. Falush, D., van Dorp, L. & Lawson, D.J. A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. *BioRxiv*, 066431 (2016).
47. Zhang, X. *et al.* Genomes of the Banyan Tree and Pollinator Wasp Provide Insights into Fig-Wasp Coevolution. *Cell* **183**, 875-889. e17 (2020).
48. Li, H. & Ralph, P. Local PCA Shows How the Effect of Population Structure Differs Along the Genome. *Genetics* **211**, 289 (2019).
49. Reich, D., Price, A.L. & Patterson, N. Principal component analysis of genetic data. *Nature genetics* **40**, 491-492 (2008).
50. Slatkin, M. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47**, 264-279 (1993).
51. Wright, S. Isolation by distance. *Genetics* **28**, 114 (1943).
52. Seeholzer, G.F. & Brumfield, R.T. Isolation by distance, not incipient ecological speciation, explains genetic differentiation in an Andean songbird (Aves: Furnariidae: Cranioleuca antisiensis, Line-cheeked Spinetail) despite near threefold body size change across an environmental gradient. *Molecular Ecology* **27**, 279-296 (2018).
53. Barton, N.H. Natural selection and random genetic drift as causes of evolution on islands. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **351**, 785-795 (1996).
54. Choi, J.Y. *et al.* Ancestral polymorphisms shape the adaptive radiation of Metrosideros across the Hawaiian Islands. *Proceedings of the National Academy of Sciences* **118**(2021).
55. Gavrilets, S. Perspective: models of speciation: what have we learned in 40 years? *Evolution* **57**, 2197-2215 (2003).
56. Rundle, H.D. & Nosil, P. Ecological speciation. *Ecology letters* **8**, 336-352 (2005).
57. Schluter, D. Evidence for ecological speciation and its alternative. *Science* **323**, 737-741 (2009).
58. Shafer, A.B. & Wolf, J.B. Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. *Ecology letters* **16**, 940-950 (2013).
59. Kooyman, R.M. *et al.* Origins and assembly of Malesian rainforests. *Annual Review of Ecology, Evolution, and Systematics* (2019).
60. Matzke, N.J. BioGeoBEARS: BioGeography with Bayesian (and likelihood) evolutionary analysis in R Scripts. *R package, version 0.2* **1**, 2013 (2013).
61. Yu, Y., Blair, C. & He, X. RASP 4: ancestral state reconstruction tool for multiple genes and characters. *Molecular Biology and Evolution* **37**, 604-606 (2020).
62. Thornhill, A.H., Ho, S.Y., Kulheim, C. & Crisp, M.D. Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. *Mol Phylogenet Evol* **93**, 29-43 (2015).
63. Tarran, M., Wilson, P.G., Paull, R., Biffin, E. & Hill, R.S. Identifying fossil Myrtaceae leaves: the first described fossils of Syzygium from Australia. *American Journal of Botany* **105**, 1748-1759 (2018).
64. Hall, R. Cenozoic geological and plate tectonic evolution of SE Asia and the SW Pacific: computer-based reconstructions, model and animations. *Journal of Asian Earth Sciences* **20**, 353-431 (2002).
65. Hall, R. The palaeogeography of Sundaland and Wallacea since the Late Jurassic. *Journal of Limnology* **72**, e1 (2013).
66. Toussaint, E.F. *et al.* The towering orogeny of New Guinea as a trigger for arthropod megadiversity. *Nature communications* **5**, 1-10 (2014).
67. Cannon, C.H., Morley, R.J. & Bush, A.B. The current refugial rainforests of Sundaland are unrepresentative of their biogeographic past and highly vulnerable to disturbance. *Proceedings of the National Academy of Sciences* **106**, 11188-11193 (2009).
68. Stelbrink, B. Humboldt-Universität zu Berlin, Lebenswissenschaftliche Fakultät (2015).

69. PAGES, P.I.W.G.o. Interglacials of the last 800,000 years. *Reviews of Geophysics* **54**, 162-219 (2016).
70. Maddison, W. & Maddison, D. Mesquite: A modular system for evolutionary analysis. Version 3.61. 2019. (2019).
71. Vasconcelos, T.N.C., Lucas, E.J., Conejero, M., Giaretta, A. & Prenner, G. Convergent evolution in calyptrate flowers of Syzygieae (Myrtaceae). *Botanical Journal of the Linnean Society* **192**, 498-509 (2019).
72. Gorchov, D.L., Cornejo, F., Ascorra, C.F. & Jaramillo, M. Dietary overlap between frugivorous birds and bats in the Peruvian Amazon. *Oikos*, 235-250 (1995).
73. Hodgkison, R., Balding, S.T., Zubaid, A. & Kunz, T.H. Fruit Bats (Chiroptera: Pteropodidae) as seed dispersers and pollinators in a lowland malaysian rain Forest1. *Biotropica* **35**, 491-502 (2003).
74. Teixeira, R.C., Corrêa, C.E. & Fischer, E. Frugivory by Artibeus jamaicensis (Phyllostomidae) bats in the Pantanal, Brazil. *Studies on Neotropical Fauna and Environment* **44**, 7-15 (2009).
75. Kalko, E.K. & Condon, M. Echolocation, olfaction and fruit display: how bats find fruit of flagellichorous cucurbits. *Functional Ecology* **12**, 364-372 (1998).
76. Biffin, E. *et al.* Evolution of exceptional species richness among lineages of fleshy-fruited Myrtaceae. *Annals of Botany* **106**, 79-93 (2010).
77. Stocker, G. & Irvine, A. Seed dispersal by cassowaries (Casuarius casuarius) in North Queensland's rainforests. *Biotropica*, 170-176 (1983).
78. Michael, T.P. *et al.* High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature communications* **9**, 1-8 (2018).
79. Li, H. seqtk Toolkit for processing sequences in FASTA/Q formats. *GitHub* **767**, 69 (2012).
80. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**, 737-746 (2017).
81. Walker, B.J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).
82. Laetsch, D.R. & Blaxter, M.L. BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287 (2017).
83. Roach, M.J., Schmidt, S.A. & Borneman, A.R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics* **19**, 1-10 (2018).
84. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nature methods* **7**, 909-912 (2010).
85. Haas, B.J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494-1512 (2013).
86. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290-295 (2015).
87. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907-915 (2019).
88. Gilbert, D.G. Genes of the pig, Sus scrofa, reconstructed with EvidentialGene. *PeerJ* **7**, e6374 (2019).
89. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **25**, 4.10. 1-4.10. 14 (2009).
90. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods in molecular biology (Clifton, NJ)* **1962**, 161-177 (2019).
91. Haas, B.J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* **31**, 5654-5666 (2003).

92. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1-22 (2008).
93. Lyons, E. *et al.* Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant physiology* **148**, 1772-1781 (2008).
94. Joyce, B.L. *et al.* FractBias: a graphical tool for assessing fractionation bias following polyploidy. *Bioinformatics* **33**, 552-554 (2017).
95. Team, R.C. R: A language and environment for statistical computing. (2013).
96. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of open source software* **4**, 1686 (2019).
97. Hadley, W. *Ggplot2: Elegrant graphics for data analysis*, (Springer, 2016).
98. Neuwirth, E. & Neuwirth, M.E. Package 'RColorBrewer'. *ColorBrewer Palettes* (2014).
99. Wilke, C.O. Ridgeline Plots in 'ggplot2'[R Package Ggridges Version 0.5. 3]. *January. https://cran. r-project. org/web/packages/ggridges/index. html* (2021).
100. Aphalo, P. ggpmisc: Miscellaneous Extensions to "ggplot2"(R package version 0.3. 6). (2020).
101. Wilkie, P., Poulsen, A.D., Harris, D. & Forrest, L.L. The collection and storage of plant material for DNA extraction: the teabag method. *Gardens' Bulletin Singapore* **65**, 4 (2013).
102. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).
103. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
104. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
105. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
106. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
107. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).
108. Rambaut, A. FigTree v1. 4. (2012).
109. Levy, D. & Pachter, L. The neighbor-net algorithm. *Advances in Applied Mathematics* **47**, 240-258 (2011).
110. Huson, D. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68-73. *Bioinformatics (Oxford, England)* **14**, 68-73 (1998).
111. Lockhart, P.J., Steel, M.A., Hendy, M.D. & Penny, D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular biology and evolution* **11**, 605-612 (1994).
112. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
113. Toolkit, P. Broad institute, GitHub repository. *See http://broadinstitute. github. io/picard* (2019).
114. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919-1920 (2012).
115. Bonfield, J.K. *et al.* HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10**, giab007 (2021).
116. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
117. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909 (2006).

118. Salojärvi, J. *et al.* Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. *Nature genetics* **49**, 904-912 (2017).
119. Popescu, A.-A., Huber, K.T. & Paradis, E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536-1537 (2012).

Figures and Figure Captions

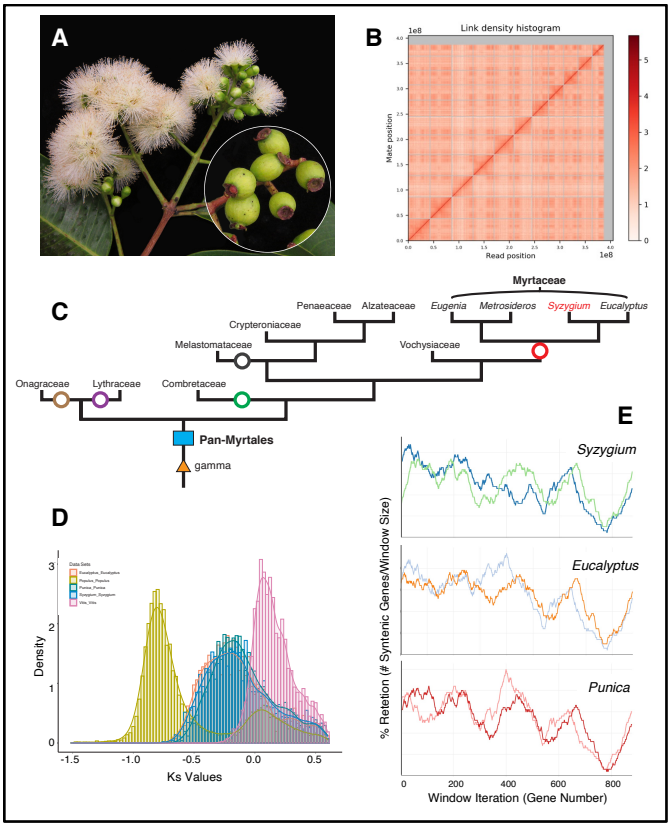

Fig. 1. Assembly and structural evolution of the *Syzygium grande* reference genome. **A.** *S. grande* inflorescence, flowers and fruits; the latter evoke the common name "sea apple"; **B.** HiC contact map for the scaffolded genome, showing 11 assembled chromosomes; **C.** Phylogeny of major lineages of Myrtales, following Maurin et al. (2021)[36]. Genera of Myrtaceae used in genome structural and phylogenetic analyses are also depicted. *Punica* (Lythraceae) was also examined for structural evolution. Open circles represent the multiple, independent polyploidy events predicted by the 1KP study[34]; our results here suggest instead a single Pan-Myrtales whole genome duplication (blue rectangle) which followed the gamma hexaploidy (orange triangle) present in all core eudicots. **D.** Synonymous substitution rate (Ks) density plots for internal polyploid paralogs within *Syzygium*, *Eucalyptus*, *Punica* and *Vitis*. Modal peaks in these three Myrtales species suggest a single underlying polyploidy event. Ks asymmetries were calibrated using the *gamma* event present in each species. Both histograms and smoothed curves are shown. **E.** Fractionation bias mappings of Myrtales chromosomal scaffolds, 2 each (different colours), onto *Vitis vinifera* chromosome 2 show similar patterns for all three Myrtales species (excluding cases of chromosomal rearrangements among the three, which are discernible as different scaffold colour switchings compared to the *Vitis* chromosome). X-axis shows percent retention of fractionated gene pairs following polyploidization; Y-axis shows position of gene pairs along the *Vitis* chromosome. Photograph credit: WHL (**A**).

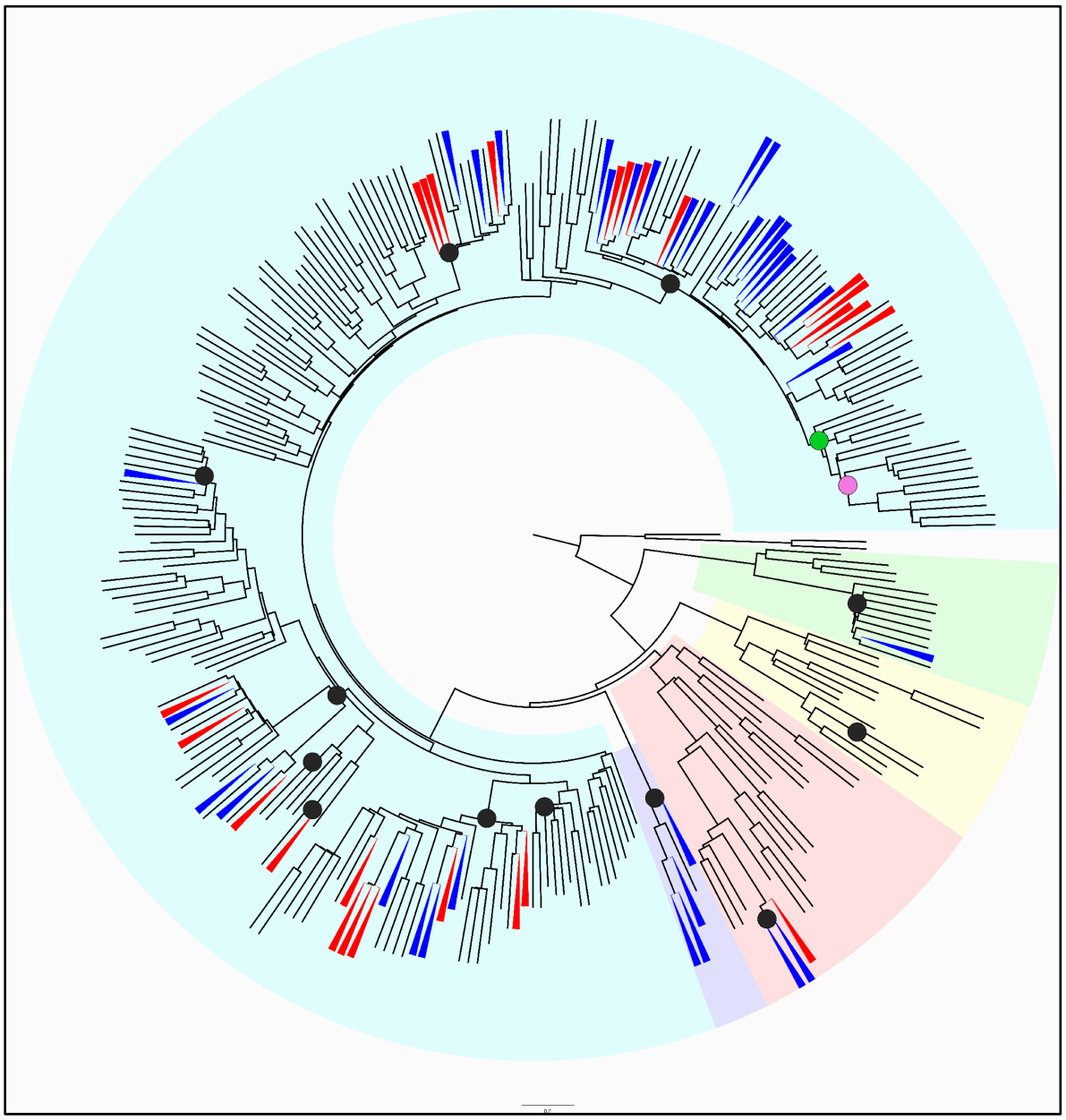

Fig. 2. BUSCO species tree for all 292 resequenced Myrtaceae accessions. Internal branch lengths are interpretable in terms of time in generations (*g*) divided by effective population size (N_e); external branches are shown only for visualisation purposes. Black circles represent the at-least 12 parallel invasions of Sunda from Sahul. The green circle represents migration from Sunda to the Indian subcontinent, and the purple circle denotes further migration from there to Africa. Blue versus red triangles at leaves of the tree represent *Syzygium* accessions from Bukit Timah Nature Reserve and Danum Valley Conservation area, respectively. Background colors represent the recognised subgenera, (clockwise from the root, excluding the outgroup taxa) *Syzygium* subg. *Sequestratum* (green), *S.* subg. *Perikion* (yellow), *S.* subg. *Acmena* (red), the unnamed clade related to *Aphanomyrtus* (purple) and *S.* subg. *Syzygium* (cyan).

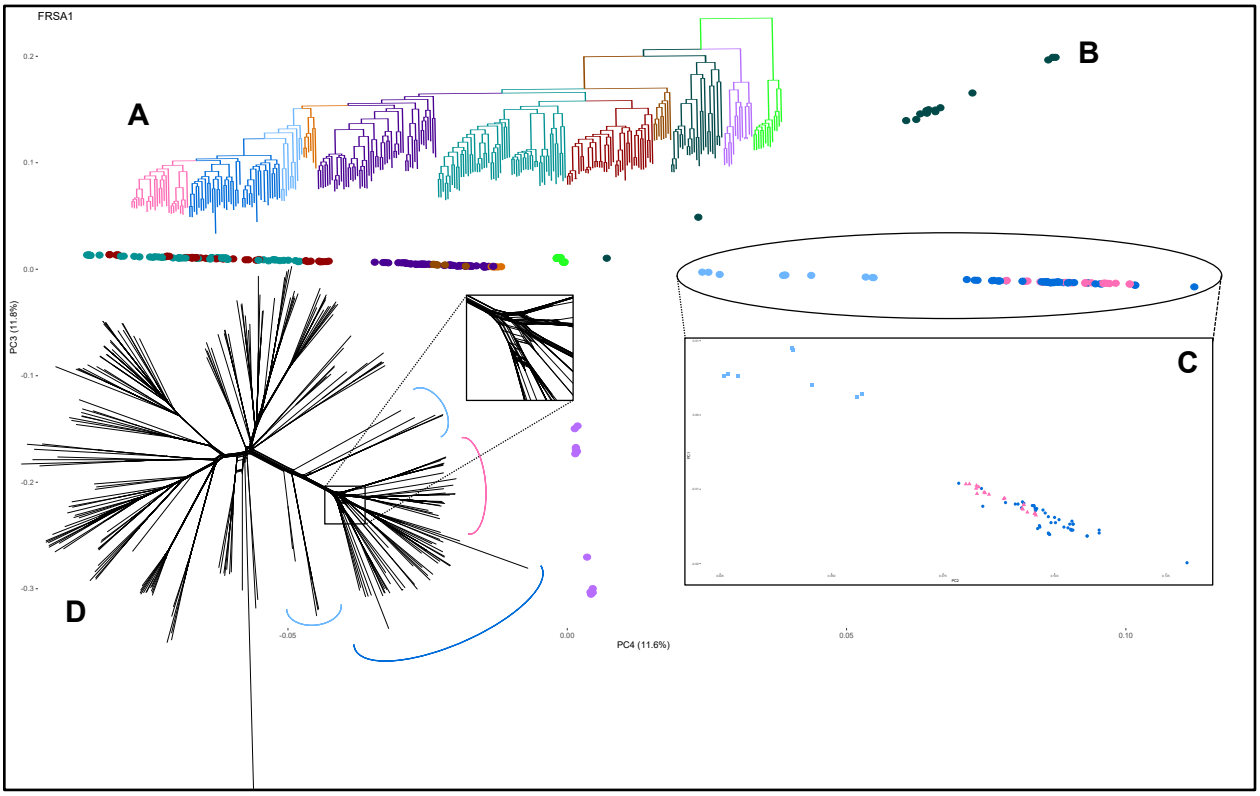

Fig. 3. PCA and NeighborNet analyses of single nucleotide polymorphism (SNP) variation within *Syzygium*. **A.** The RAxML SNP tree of *Syzygium* (outgroup removed) is colour-coded by phylogenetic progression, with the principal component analysis (PCA) below matching these colours. **B.** PCA of principal components 3 and 4 of all *Syzygium* individuals included in the RAxML tree above. Clinal patterns are readily observed; the accessions to the right surrounded by the ellipse represent the *Syzygium grande* group, which is comprised of two clades (medium blue and pink) and its immediate outgroups (light blue). **C.** The two clades in the *Syzygium grande* group are interdigitated, as is discernible also from an independent PCA analysis of just this group (PC1 and PC2 shown). **D.** A NeighborNet analysis shows considerable character discordance among the genome-wide SNPs that may be indicative of incomplete lineage sorting (ILS). This discordance is particularly noteworthy at the highly webbed base of the *Syzygium grande* group (close-up view in square inset). Coloured brackets match the colour coding on the tree and PCA plots.

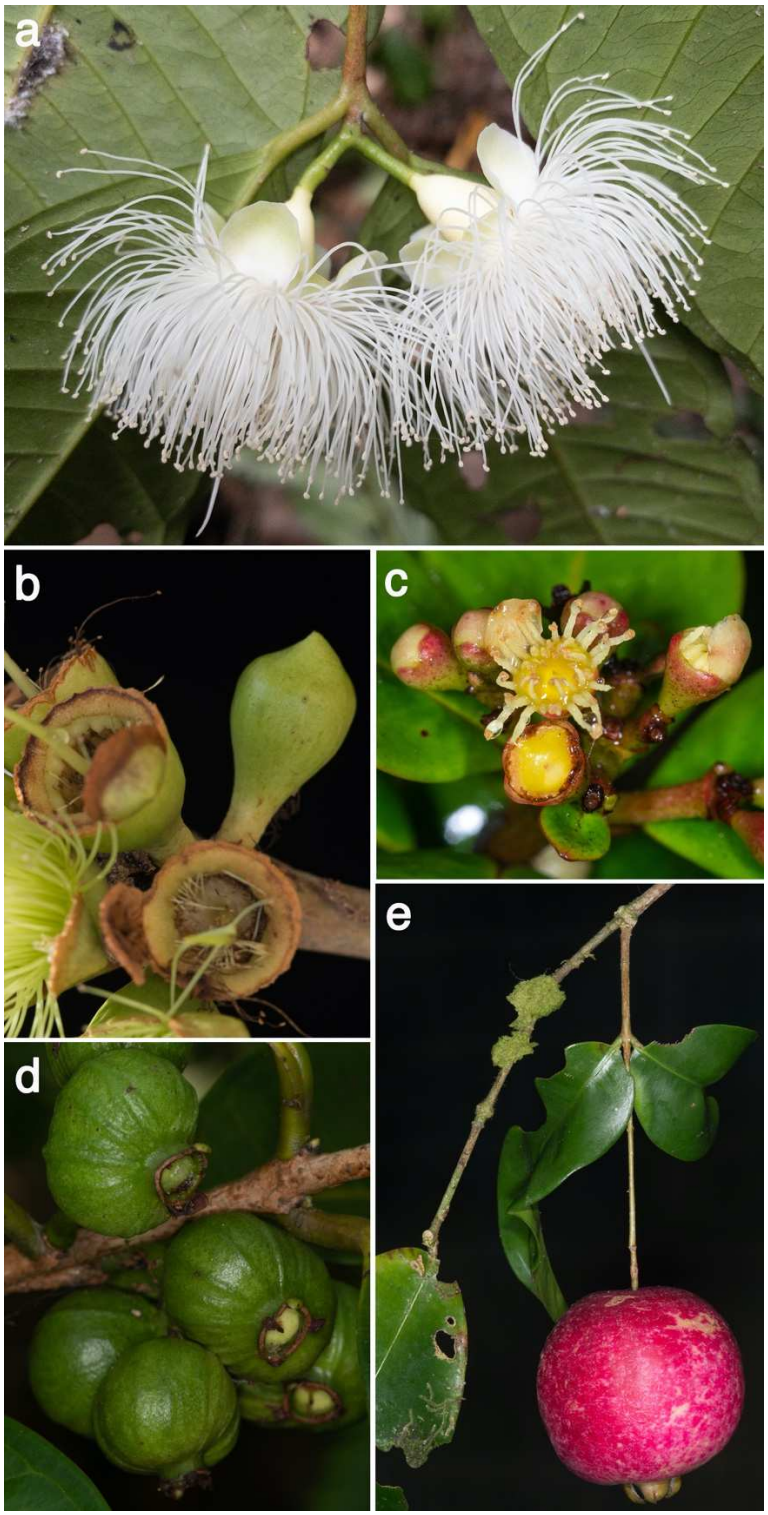

Fig. 4. Reproductive trait diversity in the genus *Syzygium,* as examined to reconstruct ancestral states using Mesquite. **A.** Free petals (*Syzygium pendens*); **B.** Calyptrate calyx (*Syzygium paradoxum*); **C.** Pseudocalyptrate corolla (*Syzygium adelphicum*); **D.** Fruits maturing green (*Syzygium* cf. *dyerianum*); **E.** Pendulous inflorescence or infructescence (*Syzygium boonjee*). Photograph credits: YWL (**A**-**E**).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SyzygiumSupplementalsubm.pdf