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

-- Supplemental Information

A. Supplementary Notes

Supplementary Note 1. Description of the Smithsonian Institution (SI)-ForestGeo long-term forest dynamics plots. One of the fundamental aspects in forest ecology is to study the structure and dynamics of tree communities *in situ*, and the functions and services of forest ecosystems. In order to pursue this long-term ecological research agenda, SI-ForestGeo has established a worldwide network of permanent forest dynamics plots ranging in size from 2-ha to 120-ha[1]. Trees in the permanent plots ≥ 1 cm diameter at breast height have been tagged and identified and re-censusing is carried out for these plots every five years to document species gained or lost. In this study, we sampled as many identified and undetermined *Syzygium* spp. as possible from the two long-term ecological plots listed below.

(i) Bukit Timah Nature Reserve (BTNR), Singapore
The Bukit Timah Nature Reserve is a 1.64 km^2 nature reserve situated in the central region of Singapore. The name is derived from the highest hill in Singapore, Bukit Timah (about 164 m at sea level), which is also located within the nature reserve. Prior to its conversion to a nature reserve, part of the site was an active granite quarry until the mid 1900s. Today, Bukit Timah Nature Reserve has the last remaining patch of intact hill-dipterocarp forest in Singapore and is an important site for preserving and protecting native flora and fauna. A total of 34 *Syzygium* species are recorded to occur in the nature reserve, but only 23 species are enumerated in the 2-ha primary forest and 2-ha secondary forest plots[2].

(ii) Danum Valley Conservation Area (DVCA), Lahad Datu, Sabah, Malaysia
The Danum Valley Conservation Area is located in the interior of the Malaysian state of Sabah, on its east coast; it covers an area of 438 km^2 encompassing undisturbed lowland dipterocarp forest. A total of 24 *Syzygium* spp. is recorded from the 50-ha ForestGEO plot, although some taxa require further attention due to the lack of flowering and fruiting materials needed for species identification..

Supplementary Note 2. 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.

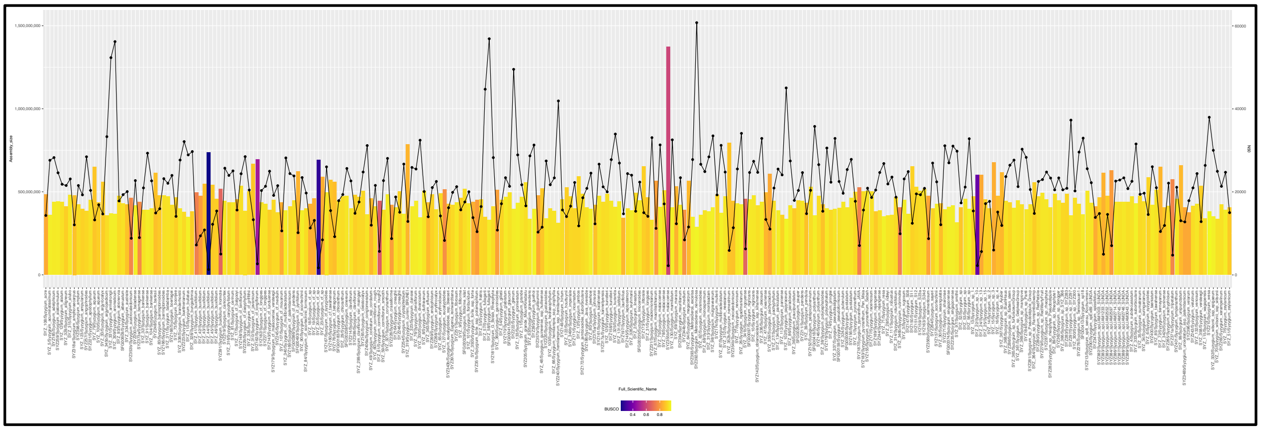
(i) Inflorescence habit of *Syzygium* can in general be categorised into two groups based on the orientation of the inflorescences being displayed and presented on branchlets. The group of taxa with erect inflorescences generally have inflorescences presented in an upright position, possibly influenced by pollination syndromes. The other group of taxa have pendulous inflorescences.

(ii) The *Syzygium* perianth can be divided into two main categories, the first category based on morphological traits of the calyx, and the other based on the corolla. In general, calyx lobes are free (Supplementary Figure S18; B1, B2, C1 and C2), but they can also be fused in the bud and eventually split free into equal portions along a suture as the stamens expand (Supplementary Figure S18; D1 and D2: *S. fibrosum*). Calyx lobes have also been recorded as fused into a true calyptra that tears irregularly as the stamens expand (Supplementary Figure S18; A1 and A2: *S. paradoxum*). Apart from the calyx, petals have been recorded to be free, spreading, and persistent, such as in *S. grande* (Supplementary Figure S18; C1 and C2). However, the corolla can also form a pseudocalyptra, in which the petals are tightly folded above the stamens to form a cap that tears along the attachment at the base, the cohered petals shedding like a calyptra at anthesis, as seen in *S. cumini*.

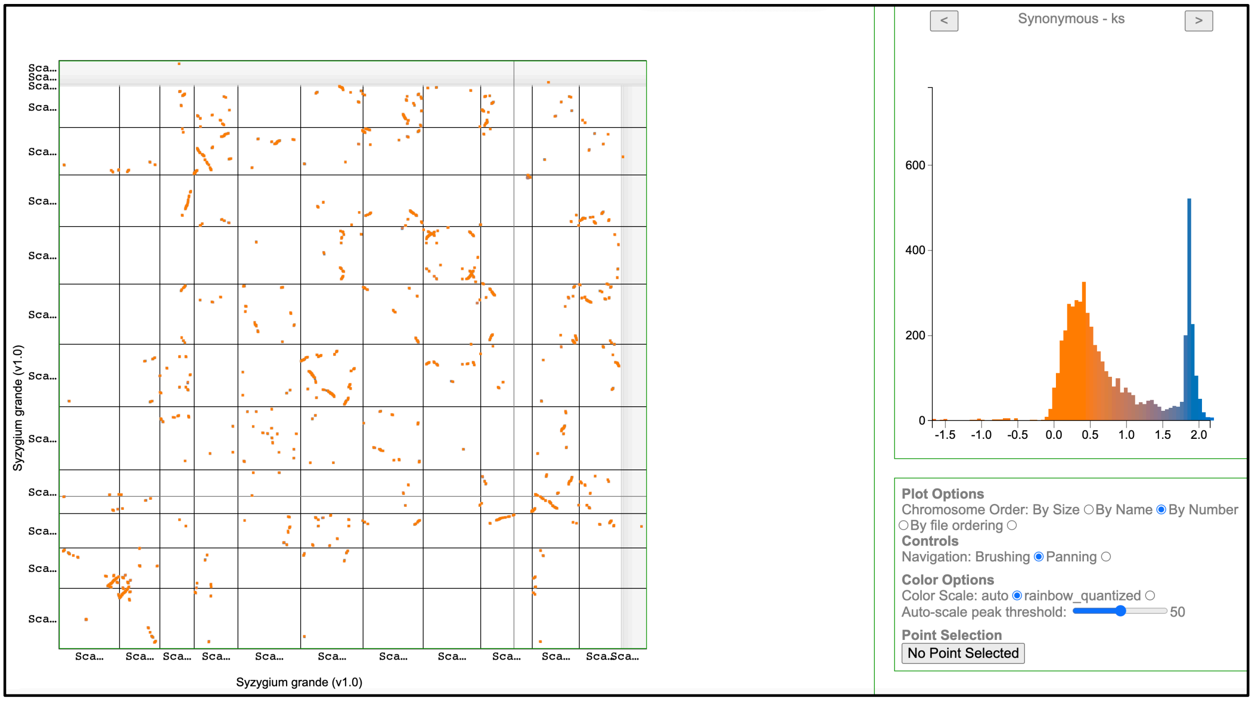
(iii) Mature fruit colour of *Syzygium* is extremely diverse and broad in colour spectrum, ranging from very bright hues to dark-coloured fruits that maximise visual detection by specific dispersers (Figs. 4D-E). It has been shown that fruits dispersed by birds tend to be in the red part of the spectrum, while mammalian dispersed fruits display the green part of the spectrum.

B. Supplementary Figures

Supplementary Figure S1. Resequenced MaSuRCA assemblies, coloured by BUSCO completeness; left y-axis (histogram) is assembly size, right y-axis (black line) is contig N50.



Supplementary Figure S2. A SynMap self:self syntenic dotplot (left) for *Syzygium grande* with coloration by Ks (histogram) reveals internal paralogy suggesting one ancestral WGD in *Syzygium* following the *gamma* paleohexaploidy event. The analysis can be regenerated at https://genomevolution.org/r/1gh12.

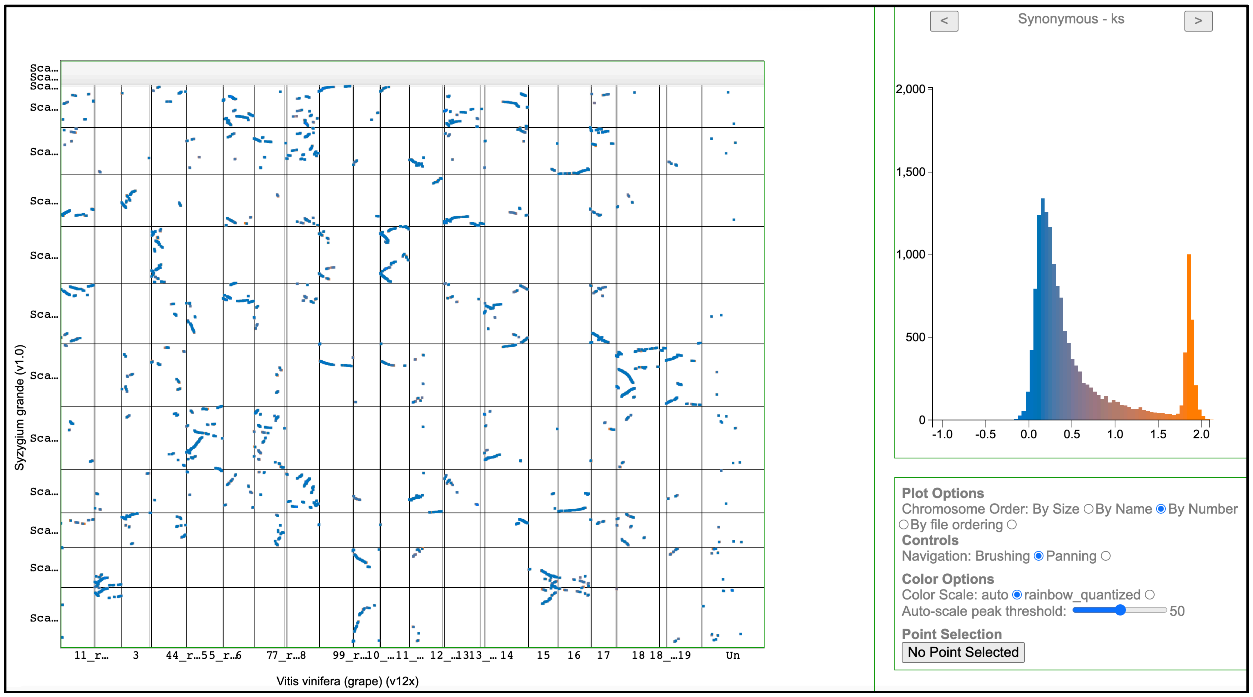


Plot Options
Chromosome Order: By Size ○By Name ◉By Number
○By file ordering ○
Controls
Navigation: Brushing ◉Panning ○

Color Options
Color Scale: auto ◉rainbow_quantized ○
Auto-scale peak threshold: 50

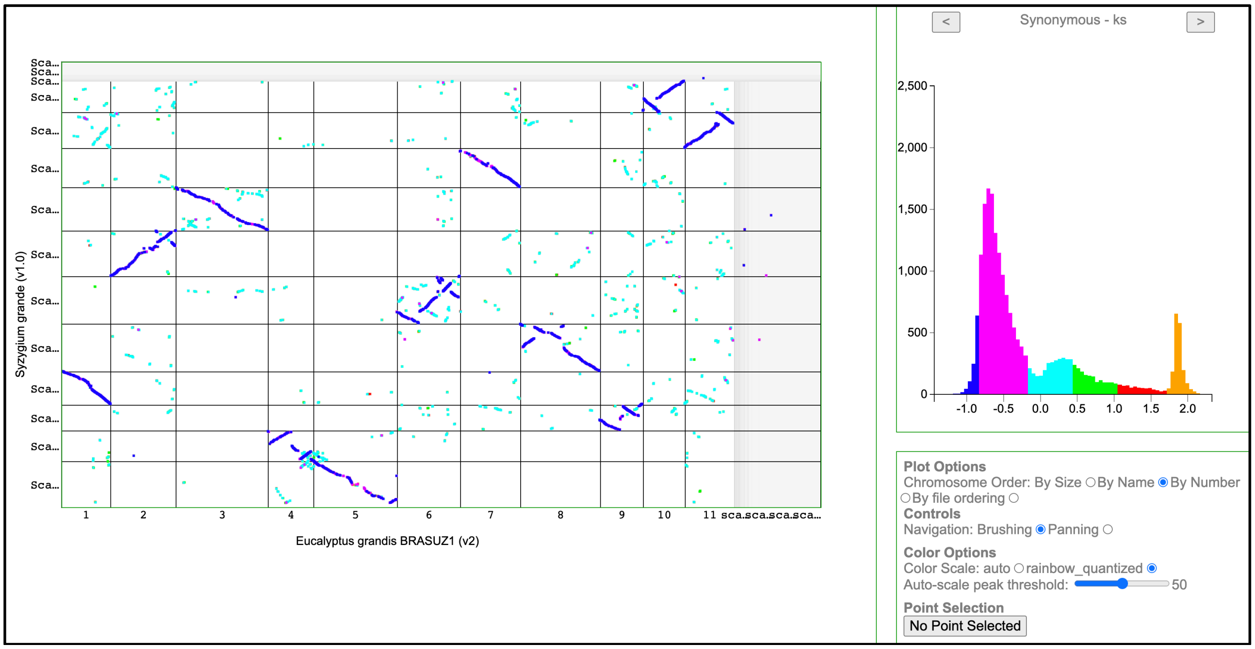
Point Selection
No Point Selected

Supplementary Figure S3. A SynMap syntenic dotplot (left) for *Syzygium grande*:*Vitis vinifera* with coloration by Ks (histogram) reveals a 2:1 relationship, indicating one WGD in *Syzygium* following its species split with Vitis, which otherwise only contains the ancient *gamma* hexaploidy event. The medium-blue blocks are 2:1 syntenic orthologs, and the more dispersed and fractionated cyan blocks are syntenic paralogs dating from the ancient *gamma* event. The orange peak represents irrational Ks values from poor CDS alignments. The analysis can be regenerated at https://genomevolution.org/r/1i4rm.



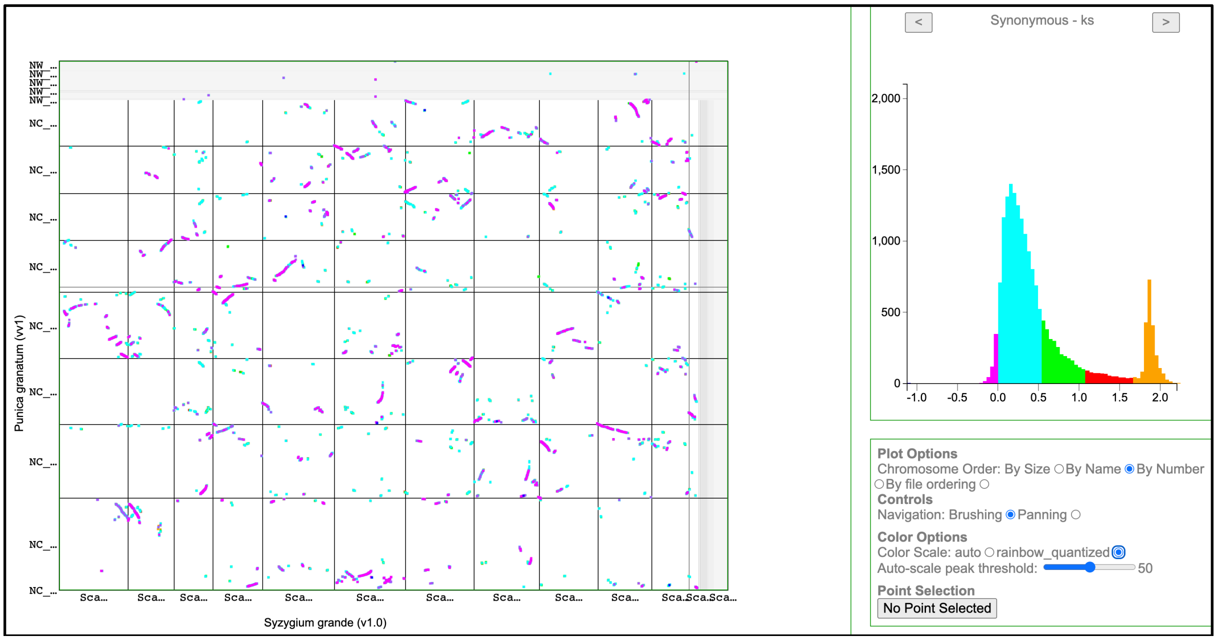
Plot Options
Chromosome Order: By Size ○ By Name ● By Number ○ By file ordering ○
Controls
Navigation: Brushing ● Panning ○
Color Options
Color Scale: auto ● rainbow_quantized ○
Auto-scale peak threshold: 50
Point Selection
No Point Selected

<u>Supplementary Figure S4</u>. A SynMap syntenic dotplot (left) of *Syzygium grande* against *Eucalyptus grandis* with coloration by Ks (histogram). The violet blocks are 1:1 syntenic orthologs, and the more dispersed and fractionated cyan blocks are syntenic paralogs. The orange peak represents irrational Ks values from poor CDS alignments. The analysis can be regenerated at https://genomevolution.org/r/1i4rm.



Plot Options
Chromosome Order: By Size ○ By Name ◉ By Number ○ By file ordering ○
Controls
Navigation: Brushing ◉ Panning ○
Color Options
Color Scale: auto ○ rainbow_quantized ◉
Auto-scale peak threshold: 50
Point Selection
No Point Selected

Supplementary Figure S5. SynMap syntenic dotplot (left) of *Syzygium grande* against *Punica granatum* with coloration by Ks. The pink blocks are 1:1 syntenic orthologs, and the more dispersed and fractionated cyan blocks are syntenic paralogs. The orange peak represents irrational Ks values from poor CDS alignments. The analysis can be regenerated at https://genomevolution.org/r/1hxo0.



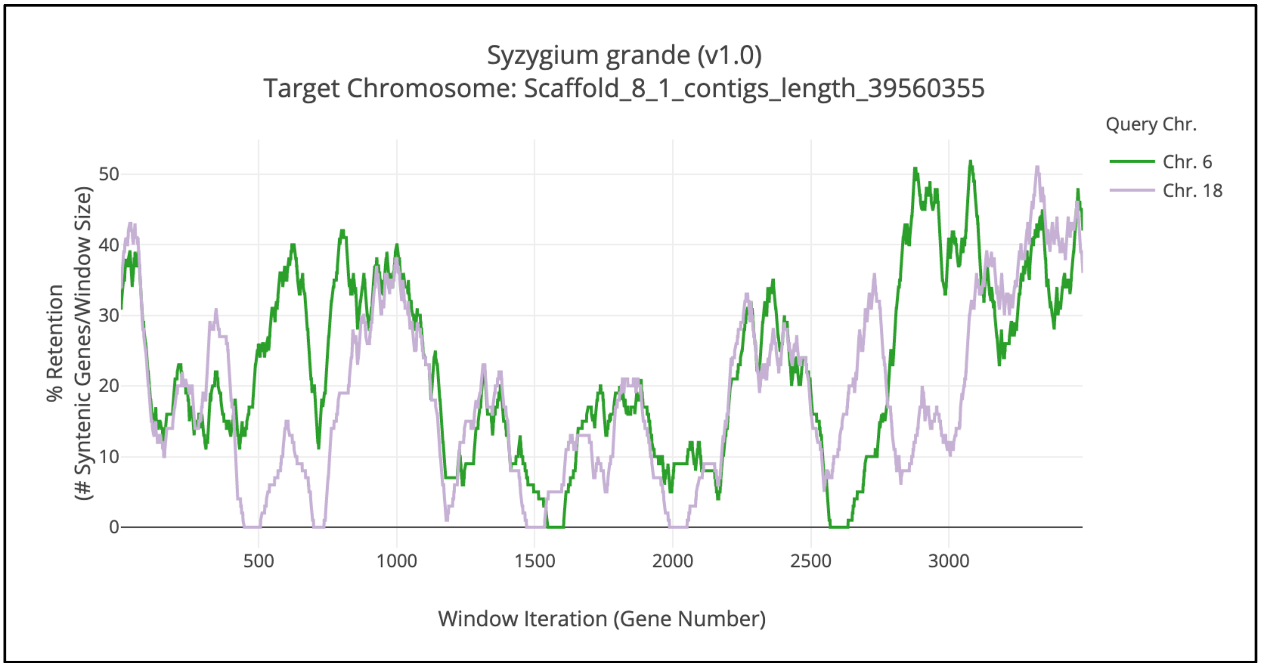
Plot Options
Chromosome Order: By Size ○ By Name ● By Number ○ By file ordering ○

Controls
Navigation: Brushing ● Panning ○

Color Options
Color Scale: auto ○ rainbow_quantized ●
Auto-scale peak threshold: 50

Point Selection
No Point Selected

Supplementary Figure S6. FractBias mapping of the *Populus trichocarpa* genome against *Syzygium grande* shows a 2:2 relationship between the species, confirming independent polyploidy events in the two lineages. This analysis can be regenerated at https://genomevolution.org/r/1ig9q.



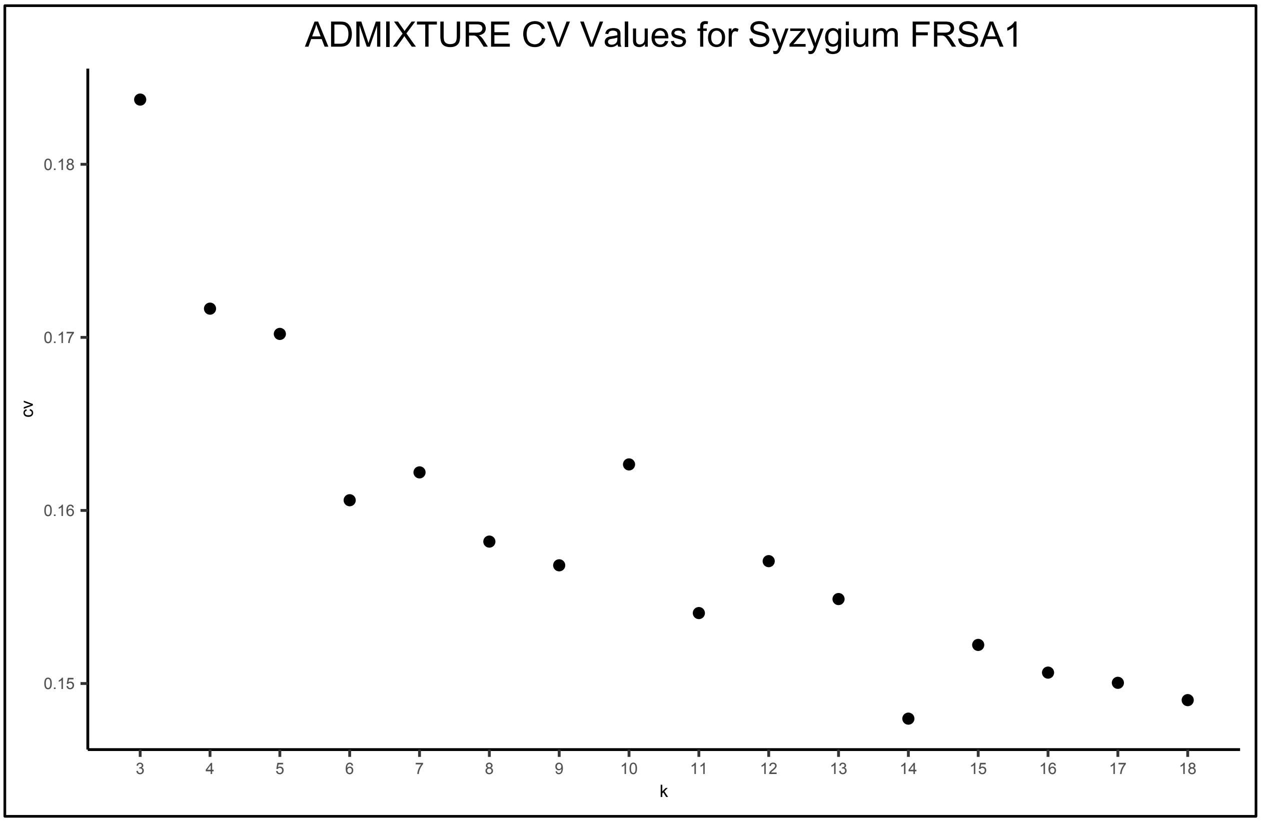
Supplementary Figure S7 (separate file). Tanglegram comparing the BUSCO- and SNP-based phylogenetic trees. The R package dendextend[3] (version 3.5.1) was used. Specifically, the tanglegram() function was used. Branches that contribute to unique sub-trees are marked with black dotted lines; incongruent relationships are shown with red lines. The BUSCO species tree and genome-wide SNP tree are both well-resolved, with robust support throughout. Five major clades, *Syzygium* subgenus *Acmena*, *S.* subgenus *Perikion*, *S.* subgenus *Sequestratum*, *S.* subgenus *Syzygium*, and a yet-to-be-named clade (*S.* cf. *attenuatum-rugosum*-SULAWESI2 clade), are consistently present in both trees. Minor discordances between these two phylogenies are present, but these do not affect the positions of the five clades recognised in this study.

Supplementary Figure S8 (separate file). Tanglegram comparing the plastome- and BUSCO-based phylogenetic trees. The R package dendextend[3] (version 3.5.1) was used. Specifically, the tanglegram() function was used. Branches that contribute to unique sub-trees are marked with black dotted lines; incongruent relationships are shown with red lines. These results show that *Syzygium* phylogeny inferred from plastome data is well-resolved when compared to a previous study[4], but still with some internal and external branches having low bootstrap supports distributed throughout the tree. Five major clades are recognised on our plastome tree, namely *Syzygium* subgenus *Acmena*, *S.* subgenus *Perikion*, *S.* subgenus *Sequestratum*, *S.* subgenus *Syzygium*, and a yet-to-be-named clade (*S.* cf. *attenuatum-rugosum*-SULAWESI2 clade), while the placement of *Syzygium wesa* is poorly supported, although robustly nestled in the *S.* subgenus *Acmena* clade in both the BUSCO genes and genome-wide SNP trees. The most significant finding is that relationships within *Syzygium* subgenus *Syzygium*, the largest of the five clades recognised, are well-resolved. One disparate placement is the position of *Syzygium jambos*; in the plastome tree, *S. jambos* is embedded in a clade otherwise comprised of all *Syzygium buxifolium* taxa, whereas in the BUSCO genes and genome-wide SNP trees, *S. jambos* is sister to *S. filiforme* in a small clade that comprised of 11 other *Syzygium* individuals. One possible explanation for the incongruent placements for *S. jambos* and *S. wesa* on the plastome tree against the nuclear inferred trees could be chloroplast capture through ancient hybridisation events, or possibly even deep ILS.

Supplementary Figure S9 (separate file). ADMIXTURE results for all *K* = 5-15.

Supplementary Figure S10. Time-calibrated BUSCO species tree for the 292 resequenced Myrtaceae accessions with ADMIXTURE results for *K*=14.

Supplementary Figure S11. ADMIXTURE cross-validation scores for dataset FRSA1 indicate that *K*=14 is the best supported number of clusters.



Supplementary Figure S12 (separate file). Local PCA analyses.

Supplementary Figure S13 (separate file). PCA analysis of main SNP dataset; different pairwise PC's are shown.

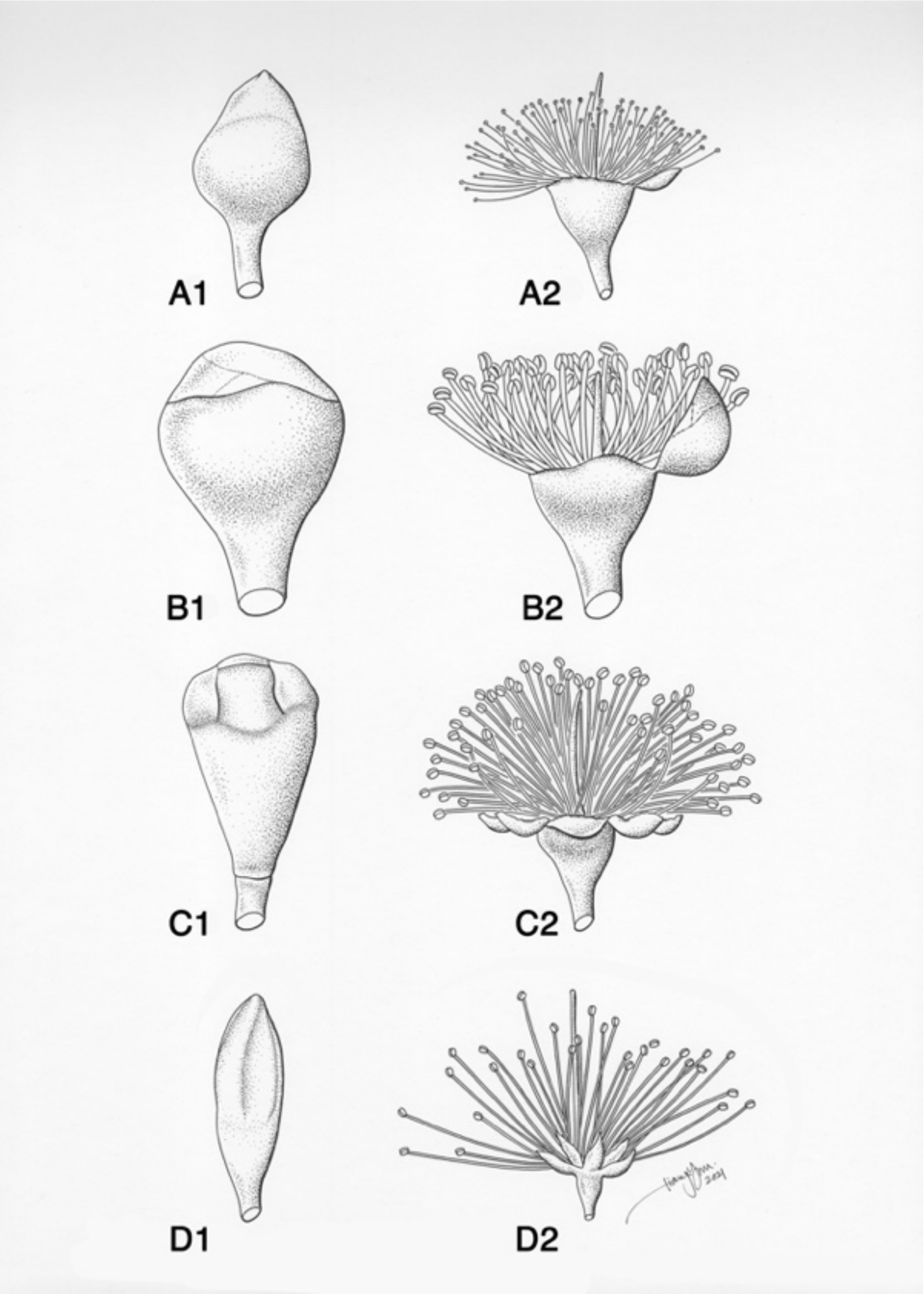
Supplementary Figure S14 (separate file). PCA analysis of alternatively-filtered SNP datasets.

Supplementary Figure S15 (separate file). PCA analysis of the *Syzygium grande* group.

Supplementary Figure S16 (separate file). Biogeographic reconstruction for the genus *Syzygium* based on RASP software.

Supplementary Figure S17 (separate file). Biogeographic reconstruction for the genus *Syzygium* based on BioGeoBEARS software.

Supplementary Figure S18. Perianth traits used for morphological character-state optimisations with Mesquite; see Supplementary Note 2 for details.

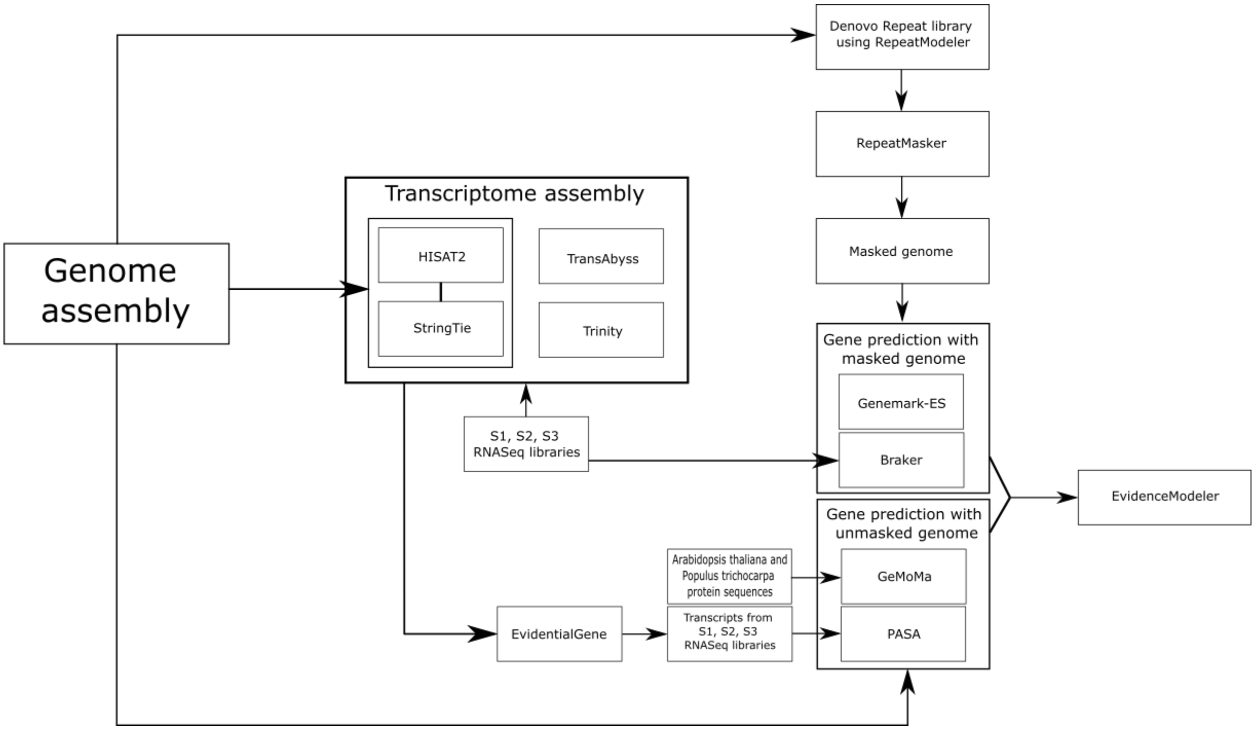


Supplementary Information, Figure S19 (separate file). Mesquite parsimony optimisation for perianth types.

Supplementary Information, Figure S20 (separate file). Mesquite parsimony optimisation for mature fruit colour.

Supplementary Information, Figure S21 (separate file). Mesquite parsimony optimisation for inflorescence habit.

Supplementary Figure S22. Workflow diagram to describe genome annotation pipeline.


Genome assembly
Denovo Repeat library using RepeatModeler
RepeatMasker
Masked genome
Transcriptome assembly
HISAT2
StringTie
TransAbyss
Trinity
S1, S2, S3 RNASeq libraries
Gene prediction with masked genome
Genemark-ES
Braker
EvidenceModeler
Gene prediction with unmasked genome
Arabidopsis thaliana and Populus trichocarpa protein sequences
GeMoMa
EvidentialGene
Transcripts from S1, S2, S3 RNASeq libraries
PASA

C. Supplementary Tables

Supplementary Table S1. Summary of repetitive DNA content annotated in the *Syzygium grande* reference genome assembly.

Class	Count	bpMasked	%masked
=====			
DNA			
CMC-EnSpm	1899	2202678	0.54%
MULE-MuDR	5294	2055064	0.51%
PIF-Harbinger	614	574480	0.14%
hAT-Ac	2290	1570501	0.39%
LINE			
L1	5521	5378442	1.33%
LTR			
Caulimovirus	1203	1965098	0.49%
Copia	25199	22088800	5.45%
Gypsy	40940	48814122	12.05%
RC			
Helitron	2996	1874534	0.46%
Unknown	278456	89889423	22.19%
	-------------	-------------	-------------
total interspersed	364412	176413142	43.55%
Low_complexity	15055	733371	0.18%
Simple_repeat	100059	5516132	1.36%
	-------------	-------------	-------------
Total	479526	182662645	45.09%

Supplementary Table S2 (separate file). Species identity, collection location, and voucher information on the 292 *Syzygium* and outgroup accessions sequenced using Illumina HiSeqX technology.

Supplementary Table S3 (separate file). Assembly and BUSCO information for the 292 Illumina-resequenced accessions.

Supplementary Table S4 (separate file). Results from Patterson's f3 three-population admixture analysis.

Supplementary Table S5. SNP datasets used for various population genomic and phylogenetic analyses.

Dataset	Samples	Filters	Downstream analyses	# of sites
FRSA-GATK	ALL	Indels removed, GATK best guidelines filtering	Heterozygosity	56,164,035
FRSA-1	ALL	Indels removed, no missing data, Depth 5-500, mac 2, min GQ 30, GATK best guidelines filtering	RAxML,PCA, ADMIXTURE, Heterozygosity	1,867,173
FRSA-2	ALL but 3 outgroups	Indels removed, no missing data, Depth 5-500, mac 2, min GQ 30, GATK best guidelines filtering	RAxML, PCA, ADMIXTURE	2,273,619
FRSA-3	Only subgenus Syzygium with rugosum SYZ3 outgroup	Indels removed, no missing data, Depth 5-500, mac 2, min GQ 30, GATK best guidelines filtering	RAxML, PCA, ADMIXTURE	2,343,383
FRSA-5	Only subgenus Syzygium with rugosum SYZ3 outgroup	Indels removed, Depth 5-500, mac 2, min GQ 30, GATK best guidelines filtering	f3 statistics	22,601,275

Supplementary Table S6 (separate file). F3 3-population results including Z-scores.

D. Supplementary Data Files

Supplementary Data S1 (separate file). Reference genome assembly of *Syzygium grande*.

Supplementary Data S2 (separate file). Annotation for *Syzygium grande* genome.

Supplementary Data S3 (separate files). Draft genome assemblies of 292 *Syzygium* and outgroup accessions.

Supplementary Data S4 (separate files). BUSCO gene alignments from the 292 assembled genomes, gene trees (Newick format) for these alignments, and the ASTRAL species tree (Newick format).

Supplementary Data S5 (separate files). SNP data sets used for various population genomic analyses; see Supplementary Table S5.

Supplementary Data S6 (separate files). Plastid DNA alignment and phylogenetic tree in Newick format.

Supplementary Data S7 (separate files). NEXUS files containing morphological and biogeographical codings.

References

1. Davies, S.J. *et al.* ForestGEO: Understanding forest diversity and dynamics through a global observatory network. *Biological Conservation* **253**, 108907 (2021).
2. Ho, B. *et al.* The plant diversity in Bukit Timah Nature Reserve, Singapore. *Gard. Bull. Singap* **71**, 41-144 (2019).
3. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718-3720 (2015).
4. Craven, L.A. & Biffin, E. An infrageneric classification of Syzygium (Myrtaceae). *Blumea-Biodiversity, Evolution and Biogeography of Plants* **55**, 94-99 (2010).