Genomic evidence for supergene control of Darwin's "complex marriage arrangement" – the tristylous floral polymorphism

Haoran Xue[1,2], Yunchen Gong[1,3], Stephen I. Wright[1,3*], and Spencer C.H. Barrett[1*]

* These authors contributed equally to the work

1. Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario M5S 3B2, Canada

2. Institute for Biochemistry and Biology, University of Potsdam, 14476 Potsdam-Golm, Germany

3. Centre for Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5S 3B2, Canada

Corresponding author: Haoran Xue, haoran.xue@mail.utoronto.ca

Keywords: tristyly, plants, genomics, supergene

Running head: Supergene control of tristyly

Abstract

Tristyly is a sexual polymorphism characterized by three flower morphs with reciprocal stigma and anther heights controlled by two epistatically interacting diallelic loci (*S* and *M*), hypothesized to be supergenes. Chromosome-level genome assemblies of *Eichhornia paniculata* identified the *S*- and *M*-loci. The *S*-locus is a supergene consisting of two divergent alleles: The *S*-allele (2.51 Mb), with three *S*-allele specific genes hemizygous in most S-morph plants and the *s*-allele (596 kb) with five *s*-allele specific genes. Two of the *S*-allele specific genes, LAZY1-S and HRGP-S, were specifically expressed in styles and stamens, respectively, making them tristyly candidate genes. The *M*-locus contains one gene (LAZY1-M) present in the *M*-allele but absent from the *m*-allele and which is homologous to LAZY1-S. Estimates of allele age are consistent with theory predicting that the *S*-locus evolved before the *M*-locus. Re-use of the same gene family highlights the potential role of gene duplication in the evolution of epistatic multilocus polymorphisms.

Introduction

Flowering plants possess a remarkable diversity of floral strategies promoting cross-pollination and outcrossed mating. This variation attracted the attention of Charles Darwin who devoted most of his classic book "*The Different Forms of Flowers on Plants of the Same Species*"[1] to the floral polymorphism heterostyly in which populations are characterized by two (distyly) or three (tristyly) mating groups (floral morphs). The defining feature of heterostyly is that the floral morphs possess a reciprocal arrangement of stigma and anther heights[2], a condition referred to as reciprocal herkogamy[3]. However, the floral morphs usually also differ in a range of other traits involving pollen and stigmas and these collectively make up the heterostylous syndrome[2].

Darwin described heterostyly as a 'complex marriage arrangement' promoting cross-pollination between the floral morphs mediated by animal pollinators owing to segregated pollen placement on their bodies (see Fig. 1 in Ref. 4). Contemporary experimental evidence on the pollination biology of diverse heterostylous species generally supports this cross-promotion hypothesis, but with the additional benefit that reciprocal herkogamy also limits sexual interference between female and male function thus reducing pollen wastage[4-5]. Heterostyly is reported from 28 animal-pollinated angiosperm families[3, 6] and represents a remarkable example of adaptive convergence in form and function maintained in populations by negative frequency-dependent mating.

Studies on the genetics of heterostyly began with the classic early works of Hildebrand, Darwin, Bateson and Gregory, and Barlow. Later, controlled crosses on over a dozen unrelated species firmly established that distyly was governed by a single diallelic Mendelian locus (*S*) with dominance, whereas tristyly was controlled by two diallelic epistatically interacting loci (*S*, *M*), both with dominance[7] (Fig. 1). To resolve the apparent discrepancy between the complexity of the heterostylous syndrome and its relatively simple inheritance it was proposed that the heterostyly loci are in fact supergenes comprised of clusters of functionally related genes[8]. Evidence indicates that these genes are tightly linked to each other due to suppression of recombination and passed from generation to generation as a single unit[7].

Supergenes are also known to control several other complex polymorphic traits, including Batesian mimicry in *Heliconius* butterflies, social organizations of fire ants and plumage variation and behaviour in white-throated sparrows and ruff[9-11]. Of particular importance for understanding the evolution of supergenes is to determine their genetic architecture and the mechanisms (e.g., inversions, hemizygosity, epigenetic modifications) that cause suppression of

recombination and thus the maintenance of these linkage groups. Identification of the causal loci can also provide important insights into the age of the alleles, to assess the evidence for long-term balanced polymorphisms maintaining variation.

Recent efforts have been made to elucidate the genetic architecture of heterostylous supergenes in a handful of distylous systems (reviewed in[12-13]). Significantly, genomic studies on *Gelsemium*, *Linum*, *Nymphoides*, *Primula*, and *Turnera*, each belonging to a different angiosperm lineage, have demonstrated that distyly is controlled by several hemizygous genes present in the short-styled morph (S-morph) but absent from the long-styled morph (L-morph)[14-19]. For example, in *Primula*, the *S*-locus is a hemizygous region containing five genes controlling the expression of different components of the distylous polymorphism[14, 15, 20]. A feature shared by the five unrelated distylous genera is that the style in the S-morph is short in length because the S-morph carries a hemizygous gene that degrades brassinosteroid (BR). This finding thus serves as an excellent example of convergent evolution at both the molecular and phenotypic levels.

All molecular investigations of heterostyly have focused on distylous species and the supergene hypothesis has yet to be investigated in any of the six tristylous angiosperm families in which this more complex floral polymorphism occurs[21-22]. Indeed, Charlesworth[23] suggested that alternatives to the supergene hypothesis should be considered for tristyly because of the absence of direct empirical evidence for supergenes in trimorphic species. Thus, the molecular genetic basis and evolutionary assembly of this two-locus epistatically interacting genetic system remains unclear. The primary motivation for our study was therefore to investigate the genomic architecture of a tristylous species to evaluate the supergene hypothesis, determine whether

hemizygosity is also a feature of the loci governing sex-organ polymorphism, and to clarify the genetic basis of the epistatic interaction between the *S*- and *M*-locus.

Here, we generate chromosome-level *de novo* genome assemblies with long-read sequences and integrate them with data from population genomic and transcriptome analysis to investigate the genetic architecture of tristyly in *Eichhornia paniculata*, an annual aquatic native to the Neotropics. Earlier molecular studies using linkage mapping and a short-read genome assembly demonstrated that the *S*- and *M*-loci in this species are closely linked (separated by a map distance of 2.7 cM), and that the *M*-locus is located on linkage group 5 in a large region containing 204 genes[24]. Bulk segregation analysis (BSA) identified 334 genes containing SNPs potentially linked to the *M*-locus, 24 of which were in this interval. In this earlier work the genomic location of the *S*-locus was not investigated. These early findings therefore do not resolve whether the *M*-locus is a supergene or merely contains one pleiotropic gene[12].

To investigate this issue and provide novel insight into the molecular basis of tristyly here we addressed the following questions: 1) Where are the *S*- and *M*-loci located in the genome? 2) What is their genomic architecture and are they indeed supergenes? 3) What are the likely candidate genes at the two loci controlling different components of the tristylous floral polymorphism? 4) What is evolutionary timing of the origins of the *S*- and *M*-loci during the assembly of the tristylous polymorphism? We note that in the only theoretical analysis of the evolution of tristyly the origin of the *S*-locus was inferred to have preceded the origin of the *M*-locus[23]. As we show below, our results are consistent with this stepwise origin of the two loci, with recurrent gene duplication from the same ancestral gene playing a key role in the formation of the epistatic interaction between the two loci.

Results

Comparisons of chromosome-level genome assemblies

We generated chromosome-level genome assemblies from a heterozygous S-morph individual (*SsMm*) and an L morph individual (*ssmm*) combining long-read PacBio sequencing with proximity ligation (Omni-C and Hi-C) scaffolding (see methods and Fig. S1). In both assemblies, most of the two genomes were assembled into eight major chromosomal scaffolds (Table 1). Comparisons of the two assemblies demonstrated large-scale concordance, with small-scale differences likely reflecting inversions and/or small assembly errors (Fig. S2). There was clear evidence for a recent whole genome duplication event in which each chromosome showed strong synteny with one or more additional chromosomes. Based on the distribution of synonymous substitution rate (*Ks*) values between paralogs, the most recent whole genome duplication event has a mode of *Ks*=0.38 (Fig. S3). Assuming a substitution rate of 7×10^{-9} per year[25] and a whole genome duplication event, this would correspond to two sets of homeologous chromosomes that diverged roughly 30 million years ago. Because the base chromosome number in Pontederiaceae has been inferred to be n=8[26], the duplication event was likely shared across the family, with additional lineage-specific duplications due to additional rounds of polyploidy occurring in other lineages of the family[27]. Unless otherwise noted, our analysis below uses our *SsMm* assembly as the reference genome.

The Marey maps (plots of the linkage map against the physical map) indicated that there was a large (likely pericentromeric) region of low recombination in the middle of each chromosome (Fig. 2). In chromosomes 3 and 7, the region extended to one end of the

chromosome, likely corresponding to submetacentric or acrocentric chromosomes, whereas the remaining chromosomes appeared to be more metacentric. As expected, the regions of reduced recombination were associated with a high repeat density (particularly LTR retrotransposons) and low gene density, whereas the more highly recombining chromosome tips were gene-dense and comprised of fewer repeats.

Both tristyly loci occur in a large low-recombining region

Using a genome-wide association study (GWAS) we identified 4,640 SNPs that were significantly associated with the S-morph (*S*---) versus non-S-morphs (*ss*--, i.e. L- and M-morphs) ($p < 2.52397779 \times 10^{-8}$, i.e. Bonferroni correction with $\alpha = 0.05$, Fig. 3a). 96.8% of these SNPs occurred in a very large 63.1 Mb region on chromosome 2, embedded in an extensive region of reduced recombination on this chromosome (Fig. 3a). As expected from earlier genetic work[31], the *M*-locus region is partially linked to the *S*-locus. We found 4,541 SNPs significantly associated with the M- (*ssM*-) versus L- (*ssmm*) morphs ($p < 2.23826411 \times 10^{-8}$, i.e. Bonferroni correction with $\alpha = 0.05$, Fig. 3b). 98.0% of these SNPs were in a 35.1 Mb region on chromosome 2, at the other end of the region of reduced recombination from the *S*-locus region. Thus, both the *S*- and *M*-loci governing tristyly occur in a large, putatively pericentromeric region of low recombination on chromosome 2 causing extensive and significant SNP associations across many Mb of sequence.

Elevated nucleotide diversity (π) and Tajima's *D* are expected in regions under balancing selection[28]. However, contrary to this prediction, neither π nor Tajima's *D* were elevated in the *S*- or *M*-locus associated regions of chromosome 2 (Fig. S2). One possible explanation is that whereas recombination rates are low across the regions with the *S*- and *M*- loci creating

contemporary SNP associations over a large region, they may be high enough to have uncoupled diversity across most of these regions from the causal tristyly loci over long evolutionary timescales (see below).

The *S*-locus is a supergene consisting of divergent alleles with hemizygous genes

The *S*- and *M*-locus associated regions we identified were both extremely large with very low rates of recombination. To further narrow down possible candidate genes and to investigate the potential hemizygous regions associated with tristyly, we identified regions with significant differences in read coverage between the floral morphs.

We identified 50 consecutive 50-kb windows that were specific to the *S*-allele (total length: 2.5 Mb; genome coordinates: Chr2:114,300,000-116,800,000), with read coverage significantly higher in 20 S-morph (*S*---) plants than 40 non-S-morph (*ss*--) plants (criteria: the ratio between the average read coverage of non-S-morph plants and S-morph plants was lower than 0.4, with the *p* value of the *t*-test lower than 9.73217066×10^{-7}, i.e. Bonferroni correction with $\alpha = 0.05$, Fig. 4a). Downstream adjacent to this 2.5 Mb region were 13 consecutive 50-kb windows (total length: 650 kb; genome coordinates: Chr2:116,800,000-117,450,000) that were specific to the *s*-allele and had read coverage significantly lower in 20 S-morph (*S*---) plants than 40 non-S-morph (*ss*--) plants (criteria: the ratio between the average read coverage of non-S-morph plants and S-morph plants was higher than 1.6, with the *p* value of the *t*-test lower than $p < 9.73217066 \times 10^{-7}$, i.e. Bonferroni correction with $\alpha = 0.05$, Fig. 4a).

We hypothesized that given the high divergence between alleles, our collapsed haploid genome assembly may have created a chimeric assembly in this region, leaving the *s*-allele sequence adjacent to the *S*-allele region in our assembled scaffold and contributing to the lack of

a signal of high diversity when mapping to this reference genome. To investigate this hypothesis further, we compared our L-morph assembly (*ss*) to this assembly in this region using whole genome alignments. As expected, Anchorwave[29] alignments revealed the presence of a large 2.57 Mb deletion in the *ssmm* assembly relative to the *SsMm* assembly (Fig. S3), whereas the adjacent non-S morph specific region showed homology. Thus, we conclude that the 2.5 Mb region roughly corresponds to the *S*-allele and the 650 kb region corresponds to the *s*-allele. We refined the boundaries of the two alleles using the coordinates of the closest genes in their flanking regions (Fig. 4a).

Two genes (TIE/SPL and MPC2) were shared between the *S*- and *s*-alleles. The order of the two genes are opposite in the two alleles. In the *S*-allele, both of the genes are on the negative strand, whereas in the *s*–allele both are on the positive strand. This finding suggests the existence of an inversion containing the two genes (Fig. 4). The rest of the genes were allele specific: three genes (CYCD6;1-S, HRGP-S, and LAZY1-S) were specific to the *S*-allele, while five genes (Epg15546, Epg15547, CSU1-s, PHOS32-s, and CDC20-1-s) were specific to the *s*-allele. LAZY1 is a gene family that regulates the distribution of auxin[30], HRGP is a superfamily of genes that forms covalently cross-linked networks in primary cell walls[31], and CYCD6;1 is a gene that regulates the formative division of cells[32].

The *M*-locus contains one hemizygous gene

We identified five consecutive 50-kb windows (total length: 250 kb; genome coordinates: Chr2:13,000,000-13,250,000) that had read coverage significantly higher in the 20 M-morph (*ssM*-) plants than in the 20 L-morph (*ssmm*) plants (criteria: the ratio between the average read coverage of non-S-morph plants and S-morph plants was higher than 1.6, with the *p* value of the

t-test lower than $p < 9.73217066 \times 10^{-7}$, i.e. Bonferroni correction with $\alpha = 0.05$, Fig. 4b). There were nine coding genes in these five windows. However, among them only one gene was missing from the *ssmm* genome assembly suggesting that it is absent from the *m*-allele (Fig. 4b). Strikingly, this absent gene, LAZY1-M, also shows strong similarity to genes in the LAZY1 family, raising the possibility that this locus acts competitively in the same pathway as LAZY1-S. Consistent with the SNP association mapping, the hemizygous *M*-locus region was located on the other side of the low-recombing region on chromosome 2 (Fig. 3).

Genetic degeneration at the *S*-allele

In species with supergenes, signatures of genetic degeneration, including gene loss and accumulation of TEs, are expected for the dominant haplotypes of the supergenes, because dominant haplotypes are only present in heterozygotes and thus do not exhibit recombination[33]. The *S*-allele specific gene loss (the five *s*-allele specific genes) was in accord with this prediction. In addition, we found that the density of repetitive elements in the *S*-allele, which is only present in *Ss*-- heterozygotes, was significantly higher ($p = 0.000076$) than the flanking regions (Fig. 4f), whereas the differences between the *s*-allele or the *M*-allele and the flanking regions were not significantly different ($p < 0.911797$ and $p < 0.075267$, respectively) (Fig. 4g and h).

The expression pattern of hemizygous genes

Among the three *S*-allele specific genes, LAZY1-S and HRGP-S were expressed in floral organs of S-morph plants but not non-S-morph plants. LAZY1-S was expressed in styles but not stamens, whereas in contrast HRGP-S was expressed in the stamens but not styles (Figure 5a).

CYCD6;1-S has low expression levels in both style and stamen of S-morph plants. Therefore, it is not considered as a candidate gene in the following discussion. The *M*-allele specific gene LAZY1-M was expressed in both styles and stamens (Figure 5b).

Genes that were differentially expressed in different morphs

For genes that were differentially expressed genome-wide between floral morphs (see Methods for details of comparisons and Table S1 for the number of genes that were differentially expressed in each comparison), three sets of genes showed functional enrichment: **(**1) genes that were upregulated in S-morph styles (compared with non-S-morph styles) in early buds (enriched in three functions), (2) genes that were upregulated in S-morph styles (compared with non-S-morph styles) in late buds (enriched in one function), and (3) genes that were upregulated in M-morph styles (compared with L-morph styles) in late buds (enriched in one function) (Table S2; Note, functions that were subsets of one another were only counted once). However, none of the five functions inferred had a direct connection to growth or flower organs, suggesting that only a small number of genes downstream to the *S*- and *M*-locus candidate genes are involved in the phenotypic differences between the floral morphs. Other differentially expressed genes may simply be incidental associations.

The origins of the *S* and *M* loci

To further understand the origins of the *S*- and *M*-loci, we searched for additional members of the gene families for loci that were *S*- and *M*-specific. We identified one additional gene that was paralogous to LAZY1-S and LAZY1-M. Hereafter, we refer to it as “LAZY1-P1”. LAZY1-P1 was located 223 kb downstream to LAZY1-M (Fig. 4b). The closest paralog of HRGP-S was

found on the same chromosome but not in the same region, and the closest paralog of CYCD6;1-S was found on another chromosome (Fig. 4c). The distance between the three ancestral genes therefore suggests that the *S*-locus was formed via stepwise duplication rather than segmental duplication.

Based on our gene trees and assuming that the hemizygous loci represent the derived gene copies, LAZY1-S duplicated from LAZY1-P1 first, following by subsequent gene duplication from LAZY1-P1 to form LAZY1-M (Fig. 6). Phylogenetic dating with BEAST v2.7.6[34] suggested that the divergence between LAZY1-S and LAZY-P1 occurred approximately 52.4 million years ago (95% confidence interval: 44.9-59.7) and that the subsequent divergence between LAZY1-M and LAZY-P1 was around 43.0 million years ago (95% confidence interval: 36.4-50.0).

Discussion

We show that in tristylous *Eichhornia paniculata*, both the *S*- and *M*-loci are located in a low recombining region of Chromosome 2. The *S*-locus is comprised of two divergent alleles: the 2.51 Mb *S*-allele with three *S*-allele specific genes that are hemizygous in most S-morph plants, and the 596 kb *s*-allele with five *s*-allele specific genes. In the *S*-allele we identified two strong tristyly candidate genes: LAZY1-S and HRGP-S. LAZY1-S was expressed in styles but not stamens suggesting that it may inhibit style elongation in the S-morph by decreasing auxin concentration in styles. HRGP-S was expressed in stamens but not styles suggesting that it may be involved in filament elongation or possibly pollen competition. Our evidence for the *S*-locus therefore strongly supports its status as a supergene.

We found that the *M*-locus contained only one hemizygous gene – LAZY1-M – and this gene was hemizygous in --*Mm* plants. Strikingly, this gene is in the same gene family as LAZY1-S, and suggests the possible role of gene duplication and neofunctionalization in the origin of the epistatic interaction between the two loci. Unlike LAZY1-S, which was only expressed in styles, LAZY1-M was expressed in both styles and stamens and with a higher expression level in stamens. This suggests that while LAZY1-S may only be involved in the shortening of styles in the S-morph, LAZY1-M may also regulate auxin distribution and organ growth in both the styles and stamens of the M-morph. Given that LAZY-M is the only *M*-allele-specific hemizygous gene, the M-locus may not represent a supergene at this stage in its evolutionary history but that it simply contains LAZY1-M, a pleiotropic modifier.

LAZY1-M was also expressed in the stamens but not the styles of S-morph plants (Fig. 5b), suggesting that LAZY1-S and LAZY1-M may compete for the same transcriptional activator and that the style expression of LAZY1-M is shut down if LAZY1-S is expressed. The expression level of LAZY1-S in S-morph styles is also much higher than LAZY1-M in M-morph styles. For example, at early development stages the average normalized count of LAZY1-M in samples of M-morph styles is 0.02% of that of LAZY1-S in S-morph styles. At late development stages, the ratio is 1.9%, indicating that the promoter region of LAZY1-S may have stronger affinity to the transcriptional activator that they compete for, potentially driving differences in style length in the S- and M-morphs through different degrees of auxin suppression.

In a theoretical treatment of the evolutionary assembly of tristyly[23], Charlesworth predicted that the M-morph originated last spreading in dimorphic populations of the L- and S-morph and eventually giving rise to the tristylous polymorphism. Our results on the relative ages of LAZY1-S and LAZY1-M are consistent with the hypothesis that the *M*-locus evolved after the

S-locus providing empirical support for this prediction. Our results also provide evidence suggesting a major role for 're-use' through recurrent duplication of the same gene (LAZY1). Recruitment of this gene explains the origin of the *M*-locus and birth of the third floral morph in the tristylous system (Fig. 6) resulting in the morphological and mating complexity that characterizes Darwin's 'complex marriage arrangement'.

Methods

Study system. Plants used in our study originated from seed collected in wild populations of *E. paniculata* in N.E. Brazil, Cuba and Mexico. The localities of plant material and details of the five populations used in our study are presented in Table S4. Information on the reproductive ecology, mating systems and population genetics of *E. paniculata* are available in Ref. 35 and references cited therein. We germinated seeds using PRO-MIX Premium Potting Mix (Premier Tech, Rivière-du-Loup, QC, Canada) and seedlings were then transferred to individual 7 cm plastic pots filled with the same soil mix and placed in flooded trays in a glasshouse at the University of Toronto (Toronto, Canada) with temperatures of 25-28°C (day) and 18-21°C (night) and a 16-hour day length. Plants were then grown and flowered regularly for several years facilitating crosses and collection of leaves and floral buds for molecular analyses.

Genome assemblies. We conducted two chromosome-level genome assemblies of *E. paniculata.* The first assembly was obtained from an individual of genotype *SsMm* (inferred from selfed progeny, Table S3) obtained from an inter-population cross between an S-morph plant (*SsM*-) from N.E. Brazil and an M-morph plant (*ssMm*) from Cuba (see Table S4 for information on the two parents). From this individual, we obtained 79 Gb (105X coverage) PacBio CLR sequences

(sequenced by Arizona Genomics Institute, Tucson, AZ, USA), 55 Gb (45X coverage) Hi-C (sequenced by Phase Genomics, Seattle, WA, USA) sequences, and 117 Gb (156X coverage) 10X Chromium sequences (sequenced by The Hospital for Sick Children, Toronto, ON, Canada). The PacBio sequences were assembled with Canu[36] by the Canadian Center for Computational Genomics (C3G) (Montreal, QC, Canada). The genome was phased with Falcon-Unzip[37], then scaffolded into chromosome-level scaffolds with the Hi-C data by Phase Genomics (Seattle, WA, USA).

To identify misjoins and characterize the relationship between the physical distance and the genetic distance on each chromosome, we constructed a linkage map with genotyping-by-sequencing data obtained from a backcross pedigree from an L-morph plant and an M-morph plant, detailed in[24], accessible under Accession no. SRP069136 at the Sequence Read Archive (SRA), (http://www.ncbi.nlm.nih.gov/sra, accessed 17 January 2023), and the associated BioProject alias PRJNA310331). The reads were trimmed with Scythe[38] and Seqtk (https://github.com/lh3/seqtk). We aligned the trimmed reads to the *SsMm* genome assembly with NextGenMap V0.5.5[39], and the BAM files were processed with Picard (http://broadinstitute.github.io/picard) using the four functions FixMateInformation, CleanSam, SortSam, and AddOrReplaceReadGroups. SAMtools[40] was used to exclude the secondary alignments. We constructed the linkage map with the alignment files and the software program Lep-Map3[41]. Markers were split into eight large linkage groups using a logarithm of the odds (LOD) score limit of 70 and a Theta value of 0.25 for the initial split (SeparateChromosomes2). Most of the remaining single markers were put into these groups using a LOD score limit of 30 and a LOD difference of 5 (JoinSingles2All). We calculated marker order for each linkage group with OrderMarkers2 (with default settings). The first ten linkage groups contained 2,124

markers. We then used Juicebox[42] to correct the misjoins with the linkage map. The genome was polished with the 10X Chromium sequences, using the software Pilon[43]. The repetitive elements were annotated with RepeatModeller V1.0.11[44] and RepeatMasker V4.0.7[45].

The second assembly of *E. paniculata* used an L-morph (*ssmm*) plant from Mexico (see Table S4) using sequencing and assembly services from Dovetail. 59.6 Gb (~50X coverage) PacBio HiFi reads were generated and the reads were assembled with hifiasm[46]. The assembly was then scaffolded with Omni-C data using the HiRise software[47]. We used the linkage map previously generated to concatenate scaffolds that belong to the same linkage group. We also annotated TEs with RepeatModeller V1.0.11[44] and RepeatMasker V4.0.7[45]. For both of our genome assemblies, the distribution of scaffold length was assessed with QUAST[48], and the completeness of assemblies was assessed with BUSCO (Benchmarking Universal Single-Copy Orthologs)[49] (Table S5).

Genome-wide associations of single nucleotide polymorphisms (SNPs) and read coverage (RCs). We extracted genomic DNA from 60 *E. paniculata* plants (20 of each floral morph) from a population in N.E. Brazil (hereafter "Pop 1", see Table S4), using the QIAGEN DNeasy Plant Mini Kit (QIAGEN, Germantown, MD, USA). Library preparation and sequencing were conducted by the McGill University and Génome Québec Innovation Centre (Montréal, QC, Canada) using the 150-bp paired-end protocol on Illumina HiSeq 4000. The samples had an average sequencing depth of ~8X.

The Illumina sequencing reads were trimmed with Scythe[38] and Seqtk (https://github.com/lh3/seqtk). We aligned the trimmed reads to the *SsMm* genome assembly with NextGenMap V0.5.5[50] and the BAM files were processed with Picard

(http://broadinstitute.github.io/picard/) using the four functions FixMateInformation, CleanSam, SortSam, and AddOrReplaceReadGroups. We used SAMtools[40] to exclude the secondary alignments. We used freebayes-parallel[51,52] to call SNPs jointly for all samples. We then filtered the VCF file with VCFtools[53] and vcflib2[54] following a published guideline[55]. We removed sites if more than 50% of the individuals had missing data, had a minor allele count of <6, or had a quality score lower than 30. Genotypes with fewer than 4 reads were treated as "missing" and individuals with more than 50% missing data were not considered further. Then, sites with strong within-individual allelic bias at heterozygous sites ($0.01 < AB < 0.25$ or $AB > 0.75$), a discrepancy in the properly paired status for reads supporting reference or alternate alleles, a low quality relative to depth, a mean depth that was > 13, a quality score < 30, or that occurred within 5 bp of an indel, were excluded from our analysis. The resulting VCF file had 7,358,165 filtered variant sites.

We conducted a Genome-Wide Association Study (GWAS) with the variant sites using GEMMA[56], where a relatedness matrix and a linear mixed model were used and *p*-values of variant sites were calculated with the Wald test.

To detect hemizygous regions associated with the *S*- and *M*-loci, we conducted another GWAS with read-coverage, following the pipeline in Ref. 57. We split the reference genome into 25,688 windows, each of which was 50-kb, and calculated the average depth of each window in individuals using the software mosdepth[58]. For each individual, we normalized the coverage by dividing the coverage of each window with the average coverage across all windows, then multiplied it by 2. Therefore, the expected normalized coverage is 1 for hemizygous regions and 2 for the rest of the genome. We conducted two *t*-tests for normalized coverage for each window: one for the comparison between S-morph versus non-S-morph plants, and the other for the

comparison between the M-morph versus L-morph plants. We identified tristyly allele-specific windows with the following criteria:

1) *p* value of the *t*-test was smaller than $9.73217066 \times 10^{-7}$, i.e. $p < 0.05$ after the Bonferroni correction.

2) We set thresholds on the ratio between the average read coverage of plants of different morphs:

For *S*-allele specific windows: The read coverage of a *S*-allele is at least 4 times of a *s*-allele, i.e. the ratio between the average read coverage of non-S-morph plants and S-morph plants was lower than 0.4: $\frac{2RC(s)}{RC(S)+RC(s)} = \frac{2\times1}{4+1} = 0.4$.

For *s*-allele specific windows: The read coverage of a *s*-allele is at least 4 times of a *S*-allele, i.e. the ratio between the average read coverage of non-S-morph plants and S-morph plants was higher than 1.6: $\frac{2RC(s)}{RC(S)+RC(s)} = \frac{2\times4}{4+1} = 1.6$

For *M*-allele specific windows: The read coverage of an *M*-allele is at least 4 times of an *m*-allele, i.e. the ratio between the average read coverage of non-M-morph plants and M-morph plants was lower than 0.25, assuming all M-morph plants were homozygous: $\frac{2RC(m)}{RC(M)+RC(m)} = \frac{2\times1}{4+1} = 0.4$.

For *m*-allele specific windows: The read coverage of an *s*-allele is at least 4 times of a *M*-allele, i.e. the ratio between the average read coverage of non-M-morph plants and M-morph plants was higher than 1.6: $\frac{2RC(m)}{RC(M)+RC(m)} = \frac{2\times4}{4+1} = 1.6$

Only regions with at least two consecutive windows that were allele-specific were considered. As a result, a 2.5 Mb *S*-allele, a 650 kb *s*-allele, and a 250 kb *M*-allele were identified. We split each allele into 10 equal-sized windows (window sizes: *S*-allele: 250 kb; *s*-

allele: 65 kb; *M*-allele: 25 kb), calculated the repeat content of each window, then compared it with the 10 windows of the same size flanking each allele (5 windows upstream and 5 windows downstream of the locus) with Wilcoxon rank-sum tests.

Differential gene expression analysis and genome annotation with transcriptome data. To infer genes controlling different components of the tristyly syndrome, we conducted a floral organ-specific transcriptomic analysis. We dissected floral buds of the three morphs at two developmental stages: "early", with buds between 3-5 mm in length, and "late", with buds between 5-8 mm. We used sterilized forceps under a stereo dissecting microscope to open the floral buds and placed the styles and stamens into 2 mL centrifuge tubes which were then frozen by putting them into liquid nitrogen. We extracted RNA from the frozen tissues with Spectrum™ Plant Total RNA-Kit (Sigma-Aldrich, St. Louis, MO, USA) following the manufacturer's instructions. For each type of sample, we collected four biological replicates, each containing styles or stamens from 5-27 buds sampled from 5-17 individuals from each of two populations (Pop. 1, 2) originating from N.E. Brazil (Table S4, S6).

Sequencing libraries were prepared and sequenced on the Illumina HiSeq4000 platform (100-bp paired-end reads) by the McGill University and Génome Québec Innovation Centre (Montréal, QC, Canada). Then we aligned the Illumina sequences to the *SsMm* genome assembly with HISAT2[59] and assembled the transcripts with StringTie[60]. We conducted functional annotations of the floral organ transcriptomes with Mercator4[61]. We used StringTie[60] to estimate the abundance of each transcript in each sample. Then we used the R package DESeq2[62] to conduct differential gene expression analysis and plot the normalized count in each sample for the genes at the *S*- and *M*-loci. We conducted eight differential gene expression analyses for each

type of morph difference (*S* vs. non-S and M vs- L), each tissue (style vs. stamen), and each development stage (early bud vs. late bud). In each analysis, we identified the genes that are upregulated in each morph with false discovery rate < 0.05 and LFC (log fold change) > 2 or < -2. We then did a Mercator BIN enrichment analysis[61] with the list of upregulated genes for each comparison.

We also annotated the repeat-masked versions of the two genome assemblies with the floral organ transcriptomes and also leaf and floral bud transcriptomes from Ref. 63, available under accession number SRP049636 on the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra, accessed 20 January 2023) and the associated BioProject alias is PRJNA266681 (http://www.ncbi.nlm.nih.gov/bioproject/, accessed 20 January 2023), and also Ref. 64, available under accession number SRP060405 at the Sequence Read Archive and associated BioProject alias (PRJNA288861) using MAKER v3.01.03[65] and InterProScan 5[66]. This resulted in a total of 35,224 annotated genes for the *SsMm* genome and 29,592 genes for the *ssmm* genome.

Origin of the *S* and *M* loci. We aligned the coding sequences of LAZY1-S, HRGP-S, and LAZY1-M against the *SsMm* and *ssmm* genome assemblies with BLAST[67], to identify the closest homologs of the three genes. We identified a LAZY1 homolog (LAZY1-P1) that exists in both *SsMm* and *ssmm* genome assemblies.

We then downloaded the coding sequences of LAZY1 in *Oryza sativa* and *Zea mays* from NCBI (https://www.ncbi.nlm.nih.gov/, accession numbers: NM_001138862.1 and DQ855268.1), and two paralogs of LAZY1 in *Musa acuminata* 'DH-Pahang' from the website Banana Genome Hub (https://banana-genome-hub.southgreen.fr/). We conducted a multiple

sequence alignment of these sequences and the coding sequences of three LAZY1 genes in *E. paniculata*, using MAFFT[68].

We then estimated the time of divergence between LAZY1-S and LAZY1-P1 and between LAZY1-M and LAZY1-P1 with BEAST v2.7.6[34]. First, we did model selection of jModelTest[69] using Akaike information criterion (AIC) as the criteria. The GTR + I + Gamma model with nCat = 4 was selected. Then we set the time to the most recent ancestors of *Oryza sativa* and *Zea mays* (81.43 ± 0.3 MYA)[70], Zingiberales and Commelinales (79.8 ± 9.87 MYA)[71], and Poales and Commelinales (106.7 ± 8.3 MYA)[71] as calibration points. Ten independent Markov Chain Monte Carlo runs, each with 1 x 10^8 generations (first 10% was burn-in) were conducted. We combined the output of the eight runs using LogCombiner v2.7.6[34] with a sample frequency of 5,000. Then we used TreeAnnotator v2.7.6[34] to generate the maximum clade credibility tree, which was visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) (Fig. 6).

Synteny analyses, whole genome alignments and structural variants. We used COGE's SynMap2 algorithm[72] with the MAKER and InterProScan annotations to examine the synteny between the two genome assemblies and identify structural variation. To obtain finer-scale whole genome alignment of syntenic regions to localize putative insertion/deletion differences between the *S* and *s* alleles and between the *M* and *m* alleles, we used Anchorwave[29] to align the two genome assemblies, with the "minimap2" option and the "proali" function.

Population genomic statistics. We calculated four population genomic statistics with 1-Mb sliding windows for the 60 individuals (20 of each morph) used in the GWAS analyses: (1)

Nucleotide diversity (π), (2) Tajima's D, (3) F_{ST} between individuals of the S-morph and the non-S-morph, and (4) F_{ST} between individuals of the L- and M-morphs. We conducted these calculations with the filtered VCF file used in the SNP-based GWAS and the functions "window-pi", "TajimaD", and "weir-fst-pop" in the software program VCFtools[53].

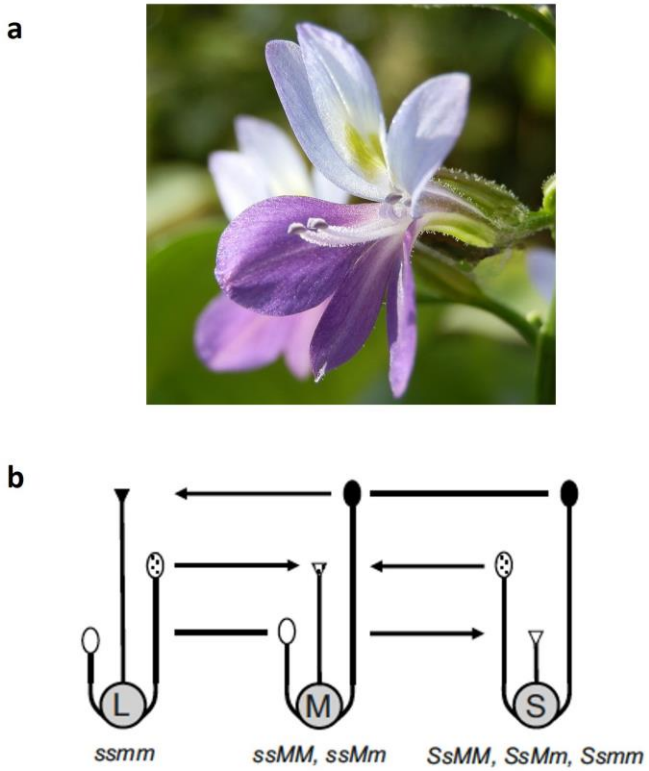


Figure 1 | Tristyly in *Eichhornia paniculata*. a, Flower of the S-morph illustrating exerted long- and mid-level stamens. The short style is concealed in the perianth tube. **b,** The arrangement of styles and stamens in the three floral morphs of a tristylous species. L, M and S refer to the long-, mid- and short-styled morphs, respectively. The arrows indicate "legitimate" cross-pollinations between anthers and stigmas of equivalent height. In most tristylous species these are the only compatible matings because of trimorphic incompatibility. However, in *E. paniculata* "illegitimate" pollinations are permitted as the species is self- and intramorph compatible[35]. Genotypes for the floral morphs of *E. paniculata* are indicated.

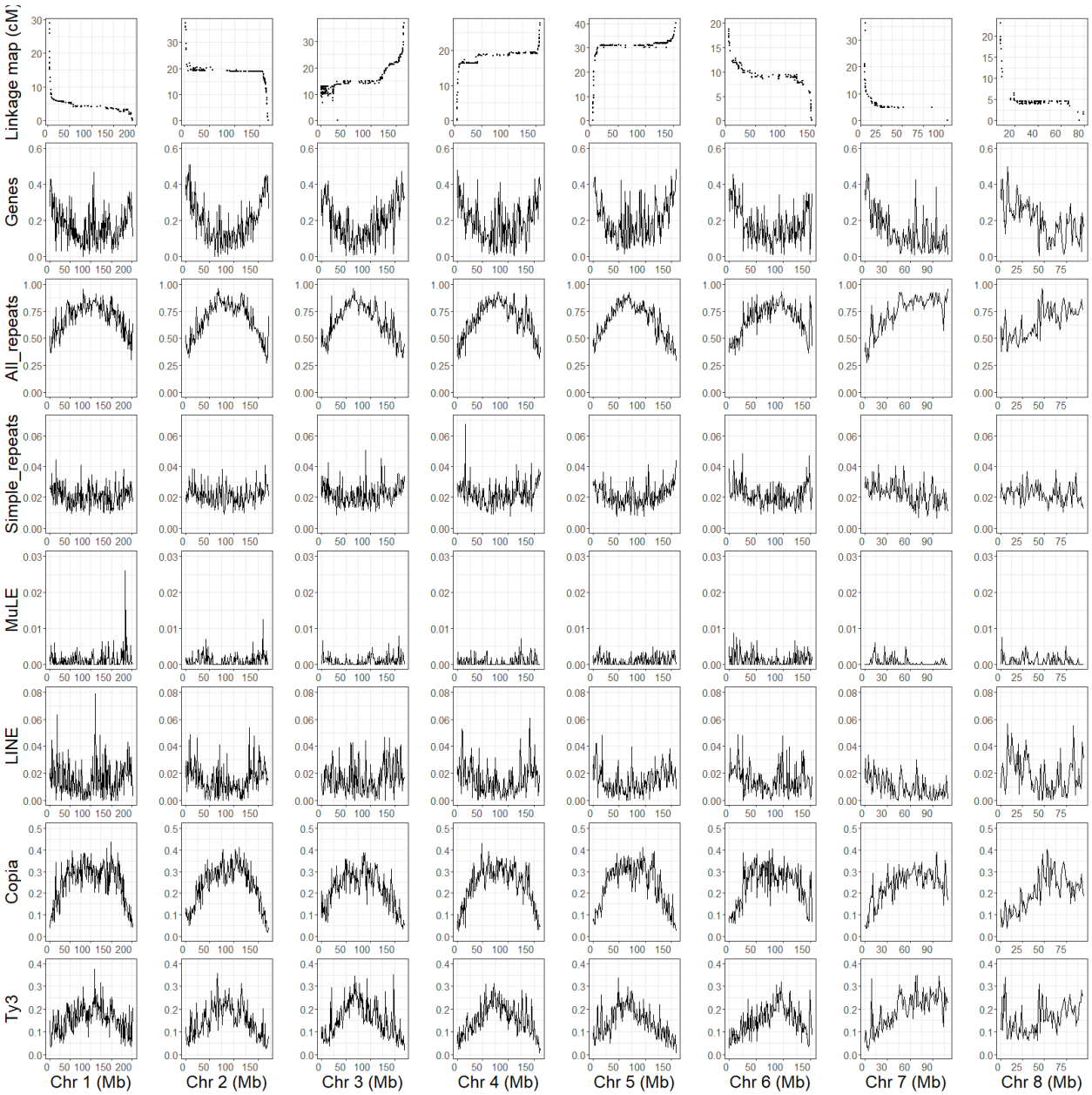
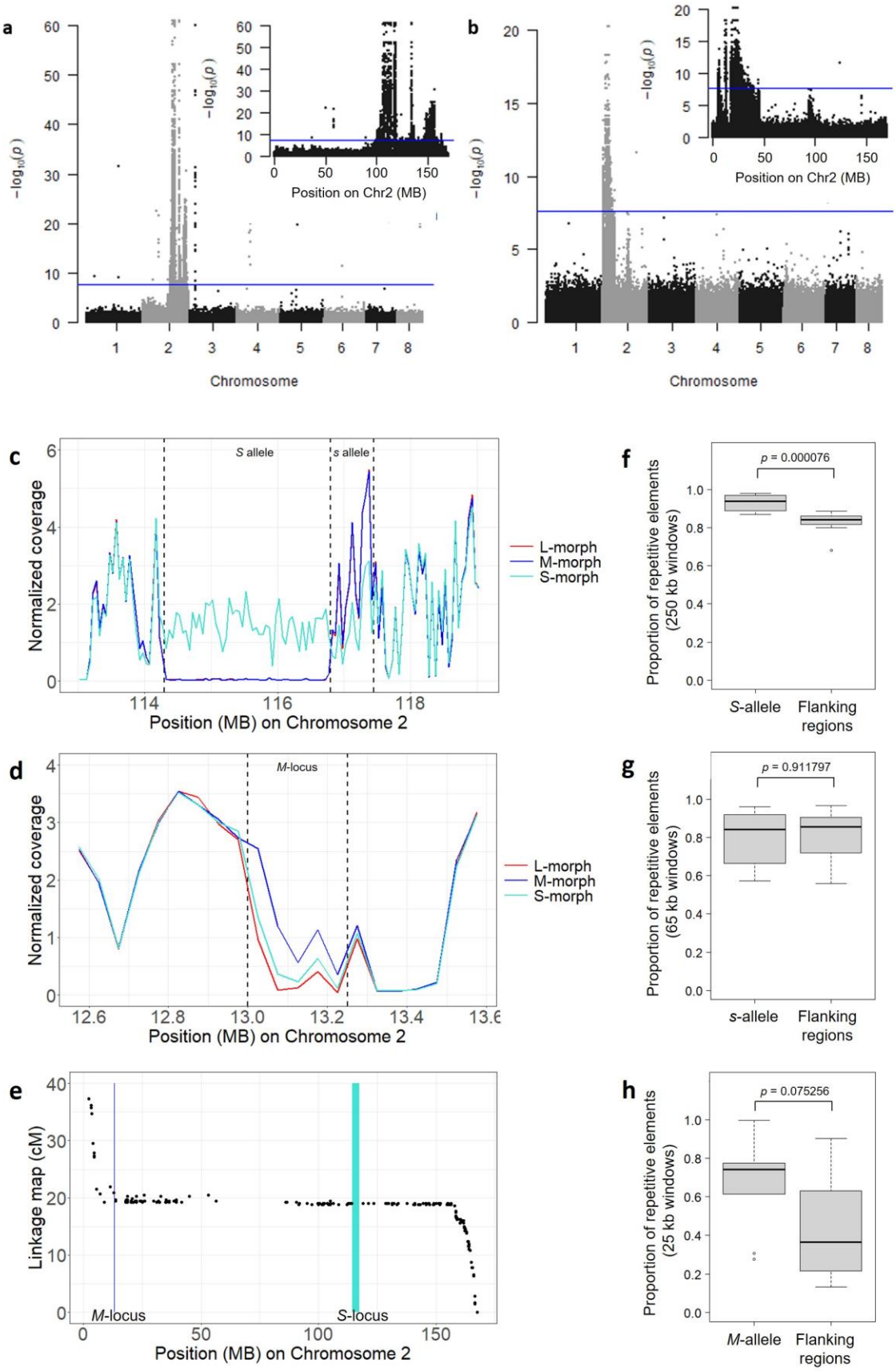
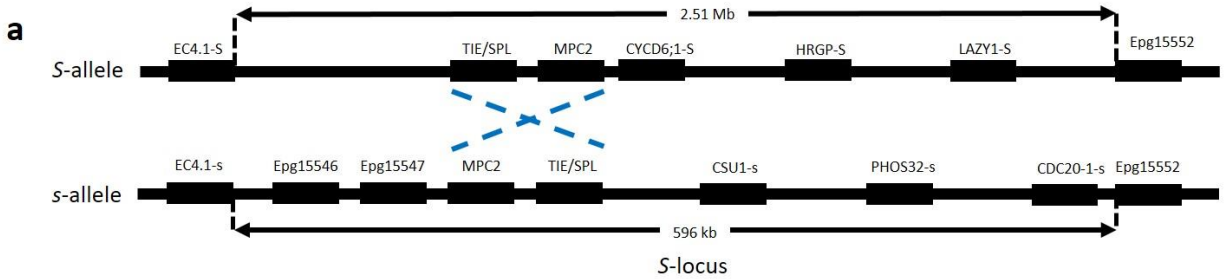


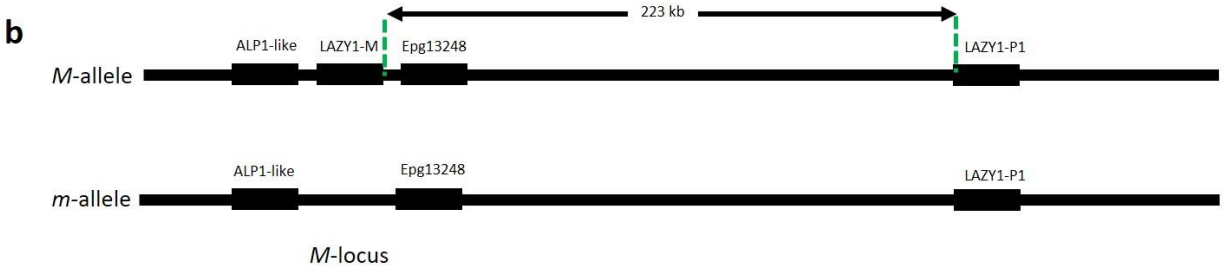
Figure 2 | The recombination landscape and genome content across the chromosomes of *Eichhornia paniculata*. a. Marey maps plotting the position on the linkage map (cM) against the position on the physical map (Mb). **b-h,** Densities of genes (**b**), all repetitive elements (**c**), simple repeats (**d**), Mutator-Like Element (MuLE) (**e**), Long Interspersed Nuclear Element (LINE) (**f**), Copia elements (**g**), and Ty3 elements (**h**), all measured as proportions of 1-Mb windows.

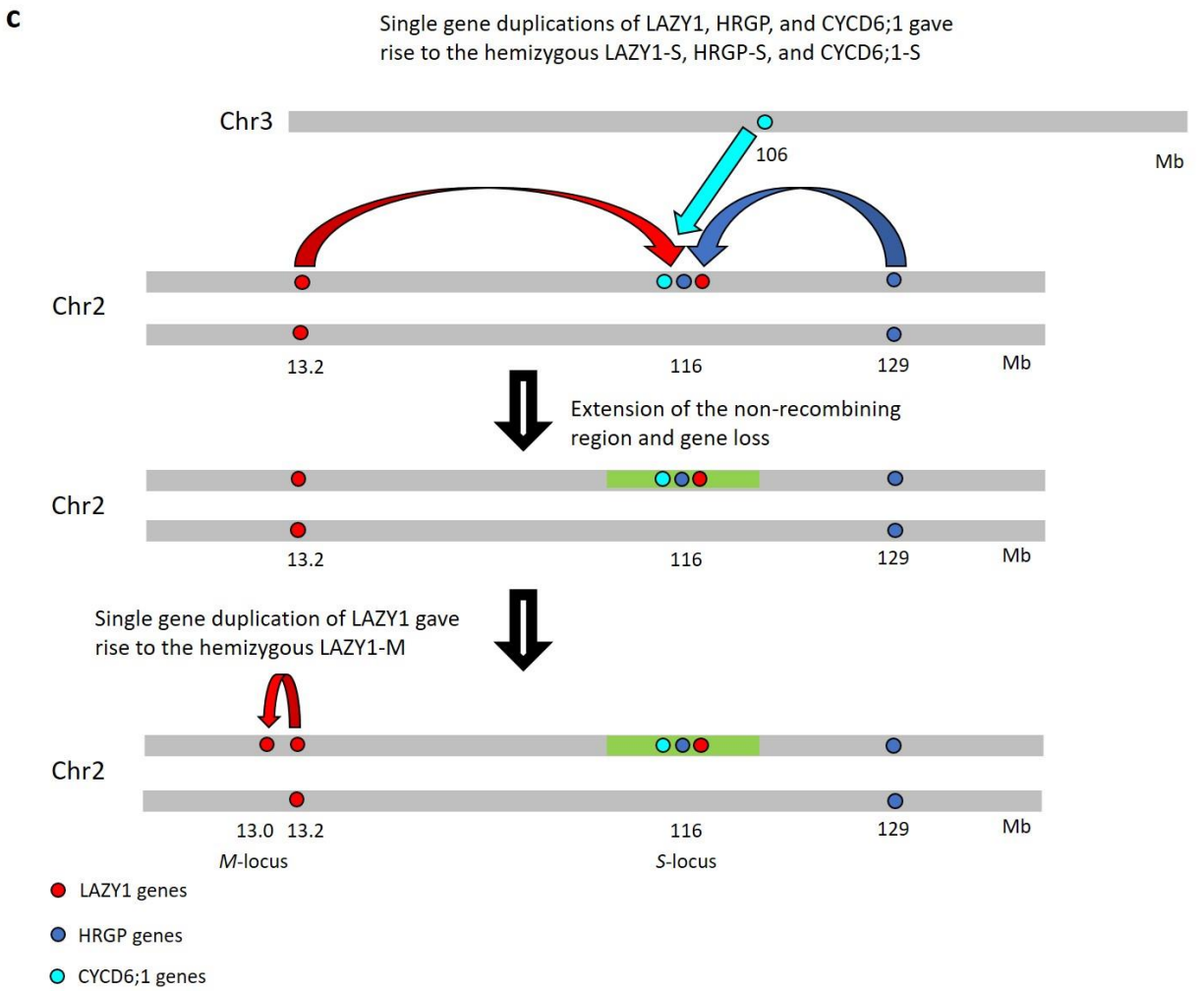

a
b
$-\log_{10}(p)$
Chromosome
Position on Chr2 (MB)
c
S allele
s allele
Normalized coverage
Position (MB) on Chromosome 2
L-morph
M-morph
S-morph
d
M-locus
e
Linkage map (cM)
M-locus
S-locus
f
p = 0.000076
Proportion of repetitive elements (250 kb windows)
S-allele
Flanking regions
g
p = 0.911797
Proportion of repetitive elements (65 kb windows)
s-allele
h
p = 0.075256
Proportion of repetitive elements (25 kb windows)
M-allele

499

Figure 3 | SNPs associated with tristyly in *Eichhornia paniculata* and morph-specific read coverage, recombination landscape, and repeat content of the *S*- and *M*-loci. a, SNPs associated with S- vs. non-S-morphs. The blue lines indicate $p = 2.52397779 \times 10^{-8}$. Inset panels zoom in on chromosome 2. **b,** SNPs associated with M- vs. L-morphs. The blue lines indicate $p = 2.23826411 \times 10^{-8}$. Inset panels zoom in on chromosome 2. **c**, Average normalized coverages of the L- (red), M- (blue), and S- (cyan) morph plants at the *S*-locus (*S*- and *s*-alleles assembled next to each other). The data were based on 20 plants per morph. **d,** The average normalized coverages of the L- (red), M- (blue), and S- (cyan) morph plants at the *M*-locus. The data were based on 20 plants per morph. **e.** The recombination landscape of chromosome 2. **f-h,** Comparisons of proportion of repetitive elements between the *S*-allele (ten 250-kb windows), *s*-allele (ten 65-kb windows), and *M*-allele (ten 25-kb windows), and their flanking regions (five windows upstream and five windows downstream to the locus, with window size equal to the allele being compared with).


a
2.51 Mb
S-allele
EC4.1-S
TIE/SPL
MPC2
CYCD6;1-S
HRGP-S
LAZY1-S
Epg15552
s-allele
EC4.1-s
Epg15546
Epg15547
MPC2
TIE/SPL
CSU1-s
PHOS32-s
CDC20-1-s
Epg15552
596 kb
S-locus


b
223 kb
M-allele
ALP1-like
LAZY1-M
Epg13248
LAZY1-P1
m-allele
ALP1-like
Epg13248
LAZY1-P1
M-locus


c
Single gene duplications of LAZY1, HRGP, and CYCD6;1 gave rise to the hemizygous LAZY1-S, HRGP-S, and CYCD6;1-S
Chr3
106
Mb
Chr2
13.2
116
129
Mb
Extension of the non-recombining region and gene loss
Chr2
13.2
116
129
Mb
Single gene duplication of LAZY1 gave rise to the hemizygous LAZY1-M
Chr2
13.0
13.2
116
129
Mb
M-locus
S-locus
LAZY1 genes
HRGP genes
CYCD6;1 genes

514

Figure 4 | The composition and proposed model of the evolution of the *S*- and *M*-loci in *Eichhornia paniculata*. *a,* Genomic organization of the *S*-locus genes. Black dashed lines represent the boundaries of the tristyly alleles inferred from the read coverage analysis. Blue dashed lines represent inferred boundaries of an inversion. **b,** Genomic organization of the genes around LAZY1-M. Green dashed lines represent the boundaries of LAZY1-M and its paralog LAZY1-P1. **c,** Hypothesized evolutionary pathway for the organization of the *S*- and *M*-loci.

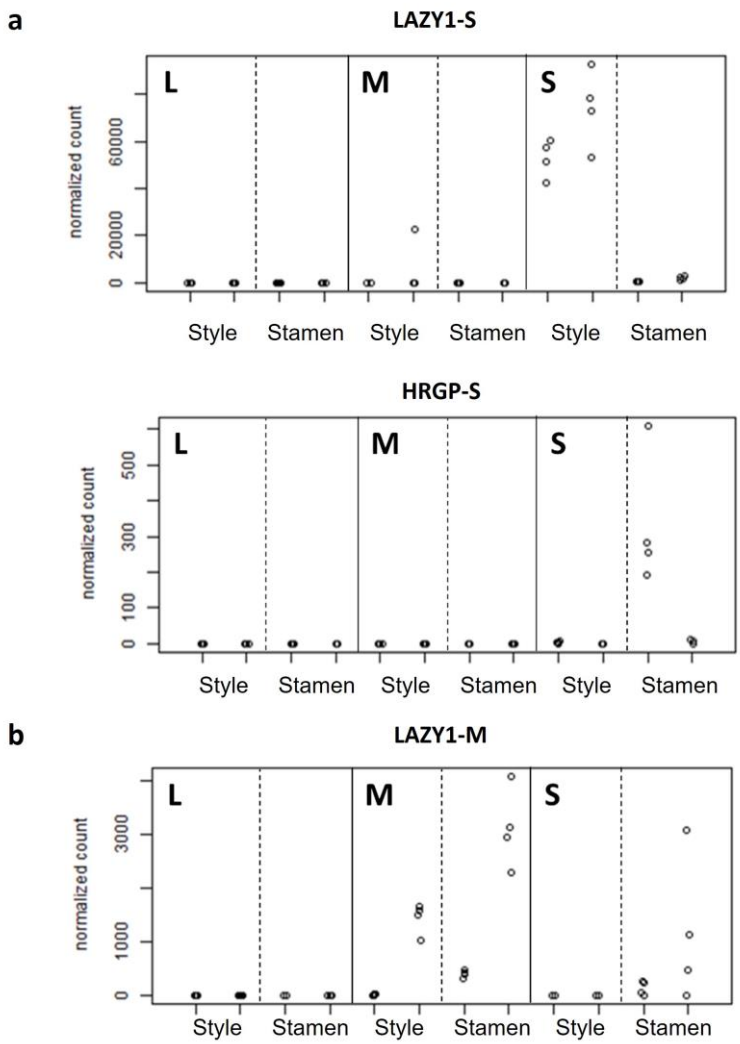


Figure 5 | Expression patterns of the three genes that are specific to the *S*- or *M*-allele and expressed in the styles or stamens of *Eichhornia paniculata*. a, Genes specific to the *S* allele. b, LAZY1-M, which is specific to the *M* allele. Samples of different morphs are separated by solid lines. For each morph, samples of different floral organs (styles or stamens) are separated by dashed lines. For each combination of morph and floral organ, the samples at the early and late development stages are on the left and right, respectively. Each type of sample had four biological replicates.

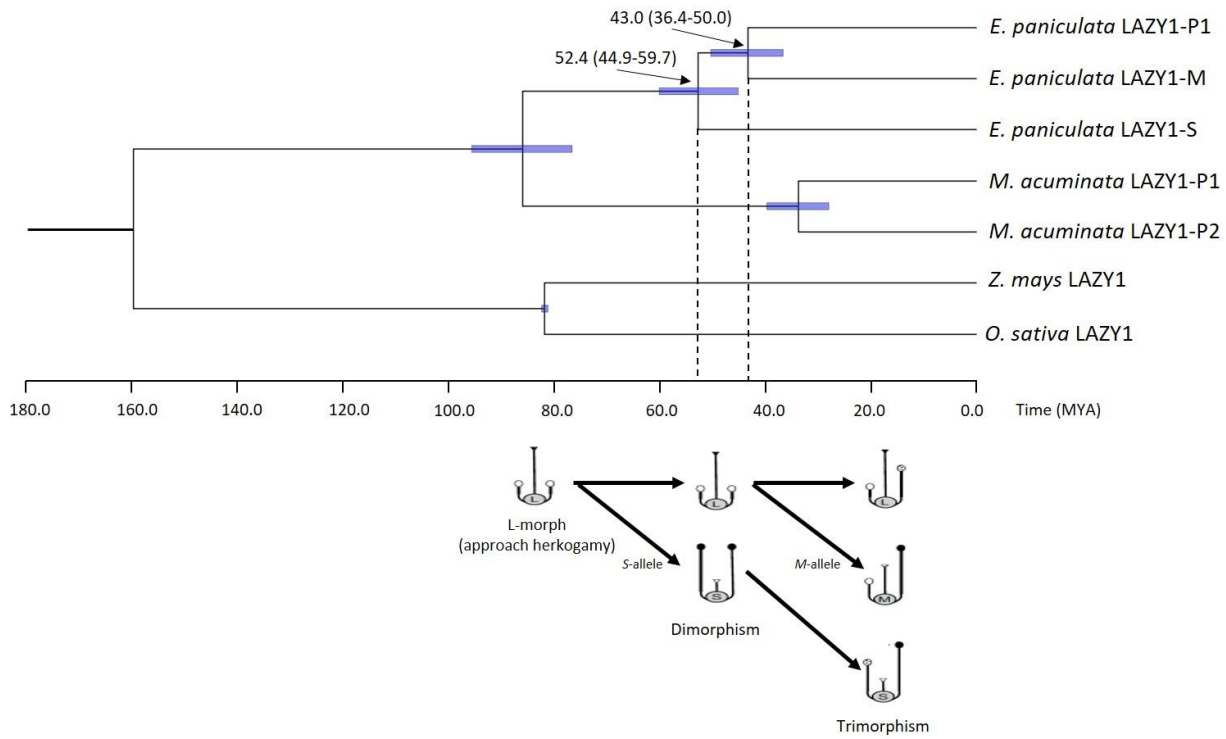


Figure 6 | Phylogenetic inference of the dates of divergence between LAZY1-P1 and LAZY1-S and between LAZY1 and LAZY1-M in *Eichhornia paniculata* and its ancestors. Bayesian tree of LAZY1 genes from *Oryza sativa*, *Zea mays*, *Musa acuminata*, and *E. paniculata*. Numbers outside of parentheses represent the estimated time of divergence. Blue rectangles and numbers in parentheses are the 95% confidence intervals of the estimated time of divergence.

References

1. Darwin, C. R. *The different forms of flowers on plants of the same species.* (John Murray, 1877).

2. Barrett, S. C. H. Heterostylous genetic polymorphisms: model systems for evolutionary analysis. In *Evolution and function of heterostyly* (ed. Barrett, S. C. H.) 1–29 (Springer, 1992).

3. Lloyd, D. G. & Webb, C. J. The evolution of heterostyly. In *Evolution and function of heterostyly* (ed. Barrett, S. C. H.) 151–178 (Springer, 1992).

4. Barrett, S. C. H. The evolution of plant sexual diversity. *Nat. Rev. Genet.* 3, 274-284 (2002).

5. Lloyd, D. G. & Webb, C. The selection of heterostyly. In *Evolution and function of heterostyly* (ed. Barrett, S. C. H.) 179–207 (Springer, 1992).

6. Barrett, S. C. H., Jesson, L. K. & Baker, A. M. The evolution and function of stylar polymorphisms in flowering plants. *Ann. Bot.* 85, 253-265 (2000).

7. Lewis, D. & Jones, D. The genetics of heterostyly. In *Evolution and function of heterostyly* (Ed. Barrett, S. C. H.) 129–150 (Springer, 1992).

8. Mather, K. The genetical architecture of heterostyly in *Primula sinensis*. *Evolution* 340–352 (1950).

9. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Current Biology* **24**, 288–294 (2014).

10. Thompson, M. J. & Jiggins, C. Supergenes and their role in evolution. *Heredity* **113**, 1-8 (2014).

11. Maney, D. L. & Küpper, C. Supergenes on steroids. *Philos. Trans. R. Soc. B* 377, 20200507 (2022).

12. Kappel, C., Huu, C. N. & Lenhard, M. A short story gets longer: recent insights into the molecular basis of heterostyly. *J. Exp. Bot.* 68, 5719-5730 (2017).

13. Barrett, S. C. H. 'A most complex marriage arrangement': recent advances on heterostyly and unresolved questions. *New Phytol.* 224, 1051-1067 (2019).

14. Huu, C. N. *et al.* Presence versus absence of CYP734A50 underlies the style-length dimorphism in primroses. *eLife* **5**, e17956 (2016).

15. Li, J. *et al.* Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. *Nat. Plants* **2**, 1-7 (2016).

16. Shore, J. S. *et al.* The long and short of the *S*-locus in *Turnera* (Passifloraceae). *New Phytol.* **224**, 1316–1329 (2019).

17. Gutiérrez-Valencia, J. *et al.* Genomic analyses of the *Linum* distyly supergene reveal convergent evolution at the molecular level. *Curr. Biol.* **32**, 4360–4371 (2022).

18. Zhao, Z. *et al.* Genomic evidence supports the genetic convergence of a supergene controlling the distylous floral syndrome. *New Phytol.* **237**, 601–614 (2023).

19. Yang, J. *et al.* Haplotype-resolved genome assembly provides insights into the evolution of *S*-locus supergene in distylous *Nymphoides indica*. *New Phytol.* **240**, 2058-2071(2023).

20. Potente, G. *et al.* Comparative genomics elucidates the origin of a supergene controlling floral heteromorphism. *Mol. Biol. Evol.* **39**, msac035 (2022).

21. Barrett, S. C. H. The evolutionary biology of tristyly. In *Oxford Surveys in Evolutionary Biology* (Eds. D. Futuyma & J. Antonovics) vol. 9, 283–326 (Oxford University Press, 1993).

22. Thompson, J. D., Pailler, T., Strasberg, D. & Manicacci, D. Tristyly in the endangered Mascarene Island endemic *Hugonia serrata* (Linaceae). *Am. J. Bot.* **83**, 1160–1167 (1996).

23. Charlesworth, D. The evolution and breakdown of tristyly. *Evolution* **33**, 489-498. (1979)

24. Arunkumar, R., Wang, W., Wright, S. I. & Barrett, S. C. H. The genetic architecture of tristyly and its breakdown to self-fertilization. *Mol. Ecol.* **26**, 752–765 (2017).

25. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).

26. Eckenwalder, J. E. & Barrett, S. C. H. Phylogenetic systematics of Pontederiaceae. *Syst. Bot.* **11**, 373–391 (1986).

27. Ness, R. W., Siol, M. & Barrett, S. C. H. *De novo* sequence assembly and characterization of the floral transcriptome in cross-and self-fertilizing plants. *BMC Genomics* **12**, 298 (2011).

28. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-595 (1989).

29. Song, B. *et al.* AnchorWave: Sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc. Natl. Acad. Sci.* **119**, e2113075119 (2022).

30. Jiao, Z., Du, H., Chen, S., Huang, W. & Ge, L. *LAZY* gene family in plant gravitropism. *Front. Plant Sci.* **11**, 606241 (2021).

31. Chen, Y. *et al.* Identification of the abundant hydroxyproline-rich glycoproteins in the root walls of wild-type *Arabidopsis*, an ext3 mutant line, and its phenotypic revertant. *Plants* **4**, 85–111 (2015).

32. Sozzani, R., Cui, H., Moreno-Risueno, M. A., Busch, W., Van Norman, J. M., Vernoux, T., Brady, S. M., Dewitte, W., Murray, J. A. H. and Benfey, P.N., 2010. Spatiotemporal regulation of cell-cycle genes by SHORTROOT links patterning and growth. *Nature*, 466, 128-132.

33. Gutiérrez-Valencia, J. *et al.* The genomic architecture and evolutionary fates of supergenes. *Genome Biology and Evolution* **13**, evab057 (2021).

34. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol***. 15**, e1006650 (2019).

35. Ness, R. W., Wright, S. I. & Barrett, S. C. H. Mating-system variation, demographic history and patterns of nucleotide diversity in the tristylous plant *Eichhornia paniculata*. *Genetics* **184**, 381–392 (2010).

36. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. **27**, 722–736 (2017).

37. Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**, 1050-1054 (2016).

38. Thress, K., Henzel, W., Shillinglaw, W. & Kornbluth, S. Scythe: a novel reaper-binding apoptotic regulator. *EMBO J.* **17**, 6135–6143 (1998).

39. Sedlazeck, F. J., Rescheneder, P. & Von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791 (2013).

40. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, (2021).

41. Rastas, P. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* **33**, 3726–3732 (2017).

42. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).

43. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).

44. Smit, A. F. A. & Hubley, R. RepeatModeler Open-1.0. 2008–2015. (2008).

45. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).

46. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

47. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).

48. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

49. Manni, M. et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647-4654 (2021).

50. Sedlazeck, F. J., Rescheneder, P. & Von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790-2791 (2013).

51. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *ArXiv Prepr.* ArXiv12073907 (2012).

52. Tange, O. Gnu parallel: the command-line power tool. *Usenix Mag*. **36**, 42 (2011).

53. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

54. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput. Biol.* **18**, e1009123 (2022).

55. Puritz, J. B., Hollenbeck, C. M. & Gold, J. R. dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ* **2**, e431 (2014).

56. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet*. **44**, 821–824 (2012).

57. Xue, L. *et al.* Evidences for a role of two Y-specific genes in sex determination in *Populus deltoides*. *Nature Communications* **11**, 5893 (2020).

58. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).

59. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

60. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

61. Lohse, M. *et al.* Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant, Cell & Environment* **37**, 1250-1258 (2014).

62. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

63. Arunkumar, R., Ness, R. W., Wright, S. I. & Barrett, S. C. H. The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics* **199**, 817–829 (2015).

64. Arunkumar, R., Maddison, T. I., Barrett, S. C. H. & Wright, S. I. Recent mating-system evolution in *Eichhornia* is accompanied by cis-regulatory divergence. *New Phytol.* **211**, 697–707 (2016).

65. Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).

66. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

67. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

68. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).

69. Posada, D. jModelTest: Phylogenetic Model Averaging. *Mol. Biol. Evol.* **25**, 1253-1256 (2008).

70. Huang, W. *et al.* A well-supported nuclear phylogeny of Poaceae and implications for the evolution of C4 photosynthesis. *Mol. Plant* **15**, 755-777 (2022).

71. Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 437-453 (2015).

72. Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E. & Lyons, E. SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* **33**, 2197–2198 (2017).

Acknowledgements

This work was supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) to S.C.H.B. and S.I.W. and student fellowships from the University of Toronto to H.X. The Canu assembly of the *SsMm* genome was conducted by the Canadian Center for Computational Genomics (C3G), which is a Genomics Technology Platform (GTP) supported by the Canadian Government through Genome Canada. We thank William Cole and Tom Gludovacz for glasshouse maintenance, Joanna Tran, SungA Choe, Ryan Ruan, and Jianuo Feng for help with growing plants, Claire Ellis and Po-Yuan Huang for help with tissue collection, Baharul Choudhury, Iuliia Pelikh, and Yihe Zheng for help with DNA

extraction, and Joanna Rifkin, Christian Kappel, Mathias Scharmann, Julia Kreiner, Jingshan Yang, Liangjiao Xue, Zoë Humphries, Tyler Kent, and Ting Liu for advice on data analysis.

Author contributions

H.X., S.I.W., and S.C.H.B. conceived and designed the study and wrote the manuscript. S.C.H.B. obtained the seeds used for the study. H.X. grew the plants, collected the tissues, and extracted DNA and RNA. Y.G. performed the genome annotation. S.I.W. performed the Anchorwave alignment. H.X. performed all remaining analyses. All authors read and approved the final manuscript.

Competing interests

None declared.