Article

Current Biology

Genomic analyses of the *Linum* distyly supergene reveal convergent evolution at the molecular level

Highlights

- The genome sequence of *Linum tenue* enabled the identification of the distyly supergene
- The *Linum* distyly supergene harbors a 260-kb region unique to thrum individuals
- The supergene is enriched for repeats and harbors a style-length candidate gene
- Distyly supergenes exhibit convergent genetic architectures and evolution

Authors

Juanita Gutiérrez-Valencia, Marco Fracassetti, Emma L. Berdan, ..., Adrian C. Brennan, Juan Arroyo, Tanja Slotte

Correspondence

tanja.slotte@su.se

In brief

Gutiérrez-Valencia et al. identify the supergene that governs the iconic floral polymorphism of distyly in *Linum*. They show that the supergene harbors a 260-kb thrum-specific hemizygous region carrying a style-length candidate gene. These findings reveal striking convergence in the genetic architecture of independently evolved distyly supergenes.

Gutiérrez-Valencia et al., 2022, Current Biology *32*, 4360–4371
October 24, 2022
https://doi.org/10.1016/j.cub.2022.08.042

CellPress

CellPress
OPEN ACCESS

Current Biology

Article

Genomic analyses of the *Linum* distyly supergene reveal convergent evolution at the molecular level

Juanita Gutiérrez-Valencia,[1] Marco Fracassetti,[1] Emma L. Berdan,[1,9] Ignas Bunikis,[2] Lucile Soler,[3] Jacques Dainat,[3] Verena E. Kutschera,[4] Aleksandra Losvik,[1] Aurélie Désamoré,[1] P. William Hughes,[1,10] Alireza Foroozani,[1] Benjamin Laenen,[1] Edouard Pesquet,[1] Mohamed Abdelaziz,[5] Olga Vinnere Pettersson,[2] Björn Nystedt,[6] Adrian C. Brennan,[7] Juan Arroyo,[8] and Tanja Slotte[1,11,12,*]

[1]Department of Ecology, Environment and Plant Sciences, Science for Life Laboratory, Stockholm University, 10691 Stockholm, Sweden
[2]Uppsala Genome Center, Department of Immunology, Genetics and Pathology, Uppsala University, Box 815, 751 08 Uppsala, Sweden
[3]Department of Medical Biochemistry and Microbiology, Uppsala University, National Bioinformatics Infrastructure Sweden (NBIS), Science for Life Laboratory, Uppsala University, 752 37 Uppsala, Sweden
[4]Department of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University, Box 1031, 171 21 Solna, Sweden
[5]Department of Genetics, University of Granada, Granada, Spain
[6]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, 752 37 Uppsala, Sweden
[7]Department of Biosciences, Durham University, Durham DH1 3LE, UK
[8]Department of Plant Biology and Ecology, University of Seville, Seville, Spain
[9]Present address: Tjärnö Marine Laboratory, Department of Marine Sciences, University of Gothenburg, 45296 Strömstad, Sweden
[10]Present address: BASF Vegetable Seeds, Napoleonsweg 152, 6083AB, Nunhem, the Netherlands
[11]Twitter: @tanjaslotte
[12]Lead contact
*Correspondence: tanja.slotte@su.se
https://doi.org/10.1016/j.cub.2022.08.042

SUMMARY

Supergenes govern multi-trait-balanced polymorphisms in a wide range of systems; however, our understanding of their origins and evolution remains incomplete. The reciprocal placement of stigmas and anthers in pin and thrum floral morphs of distylous species constitutes an iconic example of a balanced polymorphism governed by a supergene, the distyly *S*-locus. Recent studies have shown that the *Primula* and *Turnera* distyly supergenes are both hemizygous in thrums, but it remains unknown whether hemizygosity is pervasive among distyly *S*-loci. As hemizygosity has major consequences for supergene evolution and loss, clarifying whether this genetic architecture is shared among distylous species is critical. Here, we have characterized the genetic architecture and evolution of the distyly supergene in *Linum* by generating a chromosome-level genome assembly of *Linum tenue*, followed by the identification of the *S*-locus using population genomic data. We show that hemizygosity and thrum-specific expression of *S*-linked genes, including a pistil-expressed candidate gene for style length, are major features of the *Linum* *S*-locus. Structural variation is likely instrumental for recombination suppression, and although the non-recombining dominant haplotype has accumulated transposable elements, *S*-linked genes are not under relaxed purifying selection. Our findings reveal remarkable convergence in the genetic architecture and evolution of independently derived distyly supergenes, provide a counterexample to classic inversion-based supergenes, and shed new light on the origin and maintenance of an iconic floral polymorphism.

INTRODUCTION

Supergenes control complex phenotypic polymorphisms under balancing selection through the preservation of advantageous allelic combinations.[1] They play a key role in multi-trait adaptation and occur in a wide range of systems, including ants, butterflies, and plants.[2,3] Although supergenes have been identified in several eukaryotic lineages, we still do not understand how they evolve and how frequently similar genetic architectures underlie convergent multi-trait phenotypic polymorphisms.

Distyly in flowering plants is one of the most emblematic instances of a multi-trait polymorphism controlled by a supergene. Both the adaptive significance and inheritance of distyly[4,5] have received sustained interest ever since Darwin's studies.[6,7] In distylous lineages, individuals exhibit one of two types of flowers that differ primarily in the positions of their sexual organs (Figure 1) and with respect to pollen and stigma traits.[8] Pin (L-morph) individuals are long styled and present low anthers, whereas thrum (S-morph) individuals are short styled with anthers at a high level in the flower. Distyly has been suggested to increase the efficacy of pollination[7,9,10] through the deposition of pollen

Figure 1. Distyly in *Linum tenue*

(A) Thrum individuals (also termed S-morph, to the left) are short styled, whereas pin individuals are long styled (L-morph, to the right).

(B) Reciprocal location of male (anther) and female (stigma) reproductive organs in thrum (left) and pin (right) flowers. Petals and sepals have been removed to better visualize reproductive structures, and the scale is indicated by the scale bar (5 mm). The expected *S*-locus genotype is indicated below each morph.

grains from each morph on different regions of pollinating insects' bodies.[7] Distyly has evolved convergently multiple times[11,12] and is frequently associated with a heteromorphic self-incompatibility system,[8,9,13] which prevents inbreeding and guarantees disassortative mating.

Early geneticists showed that distyly was inherited as if it was governed by a single Mendelian locus, the *S*-locus, with a short style allele that was dominant over the long style allele in several distylous species.[4,11,14] The classic model of the *Primula* distyly *S*-locus supergene[5] included at least three genes affecting the traits that differ between morphs, with thrum individuals heterozygous (*S*/*s*) and pin individuals homozygous (*s*/*s*) at the *S*-locus. However, contrary to the classic model, genomic studies in *Primula* revealed that the dominant haplotype is defined by a 278-kb region exclusively inherited by thrum individuals in hemizygosity.[15–17] The hemizygous region harbors *CYP734A50*, a gene that simultaneously controls the position of the style[18] and female self-incompatibility,[19] and *GLO*^{T}, which determines anther position.[20] The presence of paralogs of *S*-linked genes at different genomic locations suggests stepwise assembly of the genes at the *S*-locus.[20,21] Thrum-specific hemizygosity and *S*-linked candidate genes for style elongation have also been documented in distylous *Turnera*[22,23] and *Fagopyrum*.[24,25]

Presence-absence variation resulting in hemizygosity at *S*-loci has important implications for our understanding of the origin, maintenance, and breakdown of distyly.[3] Although inversions are arguably the most common type of structural variation associated with recombination suppression at supergenes,[3] hemizygosity as a result of insertion-deletion (indel) polymorphism has been suggested to be a common feature of distyly supergenes.[13,22,26] Hemizygosity has major consequences for the evolutionary trajectories of supergenes and modifies expectations for modes of supergene breakdown.[3] Determining the prevalence of this genomic architecture, particularly among distylous systems where it may be common, is critical for our understanding of the mechanisms and processes leading to supergene formation and disruption. Although distyly has evolved convergently in multiple lineages, we have only just begun to elucidate the genomic architecture and evolutionary processes shaping *S*-loci. Genomic and functional studies in additional distylous systems are therefore key to shed light on whether distyly *S*-loci exhibit convergent genetic architecture, gene function, and evolution.

Wild flaxseed species (*Linum*) present a remarkable opportunity to study the evolution of distyly supergenes because they exhibit dynamic evolution of mating systems, including multiple independent origins and losses of distyly.[27–29] Distyly in *Linum* has been studied since Darwin,[6,7] revealing dimorphism not only in flower structure[6,7,27] but also in pollen color and exine sculpturing, and stigma surface.[8,30] Although molecular studies have led to the identification of candidate *S*-locus genes,[31,32] the distyly *S*-locus in *Linum* has not been fully sequenced nor characterized.

Here, we uncover the genetic architecture and evolution of the supergene that governs distyly in *L. tenue* (Figure 1). By first building a chromosome-level genome assembly, we identify the *Linum S*-locus and show that it is characterized by the presence of a thrum-specific hemizygous ~260-kb region that harbors nine predicted genes, including a candidate gene for style length. By extending the study of distyly supergenes to a novel system, we demonstrate remarkable molecular convergence in supergene architecture and evolution across widely diverged systems and shed light on the origin and maintenance of an iconic floral polymorphism.

RESULTS

A chromosome-level genome assembly of *L. tenue*

We produced a high-quality *de novo* genome assembly of a *L. tenue* thrum individual to aid the identification of the distyly *S*-locus. We first generated a contig assembly based on high-coverage (~170×) PacBio long-read data, scaffolded with chromatin conformation capture (Hi-C) data, and polished with 10x Genomics linked reads (Figure S1). We generated a genetic map to correct and anchor primary scaffolds to 10 linkage groups (LGs), corresponding to the 10 haploid chromosomes of *L. tenue*.[33] The resulting 702.1 Mb haploid genome assembly had a scaffold N50 of 69.3 Mb and was highly complete (complete BUSCOs = 94.2%, flow cytometry-based estimate of genome size ~690 Mb) (Table S1). Annotation of the assembly using both *ab initio* prediction tools and RNA sequencing (RNA-seq) evidence led to the identification of 52,826 coding sequences (complete BUSCOs = 93.8%, based on the longest

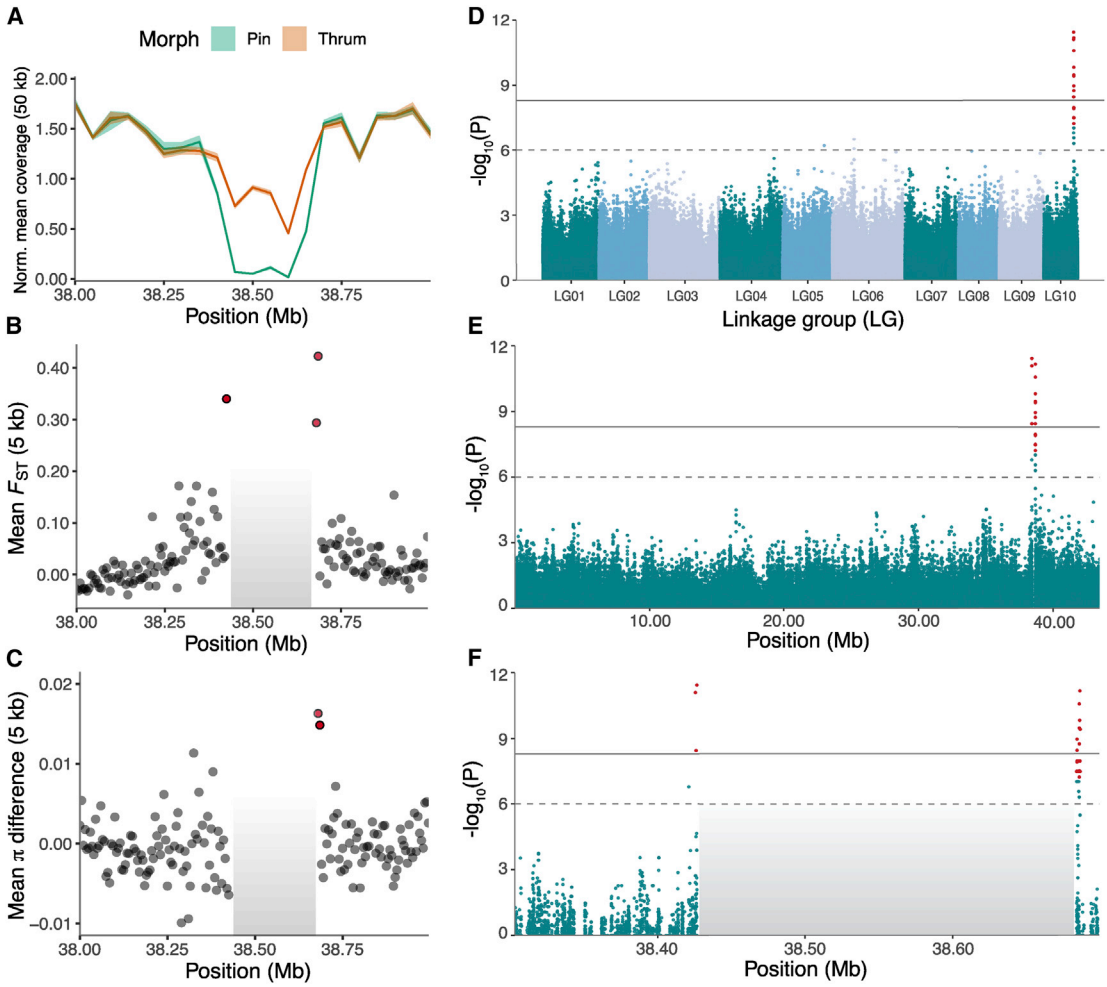

Figure 2. Identification of the distyly *S*-locus based on population genomic analyses

(A) Differences in coverage between thrum (n = 26) and pin (n = 25) samples indicates the existence of an ~260-kb hemizygous region in LG10 (shaded regions indicate 95% CI), as supported by the existence of six 50 kb adjacent windows that are absent in pin samples.

(B) Genetic differentiation (F_{ST}) between morphs in 5-kb windows flanking the ~260-kb hemizygous region, using samples from population SMT (thrum = 13 and pin = 13) (see estimates for population CL; Figure S4). Windows with statistically significant estimates are highlighted in red ($p < 0.01$, after 1,000 replicates of permutation test per window, followed by FDR correction with the Benjamini-Hochberg method).

(C) Differences in mean π between thrum and pin individuals, also in 5 kb windows flanking the ~260 kb hemizygous region for population SMT and with statistically significant estimates highlighted in red.

(D) Manhattan plot depicting the association between SNP genotype and floral morph in *L. tenue* (n = 42 individuals, 21 of each morph) across the entire genome. Dashed and contiguous horizontal lines denote suggestive ($-\log_{10}(1 \times 10^{-6})$) and significant association ($-\log_{10}(1 \times 10^{-9})$) prior to multiple testing correction. SNPs colored in red highlight loci that were significantly associated with floral morph following FDR correction using the Benjamini-Hochberg procedure, $p < 0.01$.

(E) Manhattan plot depicting GWAS results for LG10.

(F) Manhattan plot of the region neighboring the masked ~260 kb hemizygous region.

The masked ~260 kb hemizygous region is shaded in gray in (B), (C), and (F). See also Figures S2 and S3.

isoform per gene; Table S1) and 595,563 non-overlapping repetitive sequences that make up ca. 49.36% of the genome.

The distyly supergene is hemizygous in thrum individuals

Presence-absence variation is an important feature of distyly supergenes in *Primula* and *Turnera*.[15,17,22] To identify genomic regions harboring presence-absence variation between floral morphs, we analyzed genome coverage for 21 pin and 22 thrum individuals sequenced with short reads. We identified an ~300-kb region between 38.40 and 38.70 Mb on LG10 with significantly elevated coverage in thrums relative to pins (Figure 2A) (Wilcoxon signed-rank test followed by Bonferroni correction, W = 0.0, $p < 0.001$ for windows between LG10: 38.40–38.70 Mb, and *NS* for all remaining 50 kb windows across the entire genome; mean coverage across samples: mean = 29.45, SD = 9.73, n = 43) (Figure S2A). In contrast, there were no windows where pin individuals had significantly higher coverage than thrums. Compared with flanking regions, coverage in the 38.40–38.70 Mb region of LG10 was significantly reduced in thrum individuals, as expected if thrums are hemizygous (Figures 2A and S2A). The presence of an ~260-kb

thrum-specific region on LG10 was confirmed by the alignment of 10x Genomics linked-read Supernova genome assemblies of thrum (n = 4) and pin (n = 5) individuals to the *L. tenue* reference genome (Figure S2E) and linked-read *S*-locus haplotype reconstruction for the reference genome individual (Figures S5C and S5D). This ~260-kb insertion-deletion (indel) harbors a gene encoding a homolog of the *L. grandiflorum* style-expressed thrum-specific protein TSS1 (see section distyly candidate genes at the *S-locus* show thrum-specific expression) suggested to be *S*-linked in *L. grandiflorum*.[31,32] We therefore considered the ~260-kb thrum-specific region on LG10 a candidate region for the distyly *S*-locus in *L. tenue*.

To conclusively identify the distyly supergene in *L. tenue*, we performed a genome-wide association study (GWAS) of 8.7 million SNPs in 21 thrum and 21 pin individuals grown from different seed families collected in two natural populations in Andalusia (STAR Methods). This analysis intended to pinpoint SNPs strongly associated with floral morph, coded as a binary trait. We identified a single genomic region on LG10 (positions 38,425,470–38,686,519) as significantly associated with floral morph (Figures 2D and S2B) and thus likely to harbor the *L. tenue* distyly supergene. There were 79 SNPs immediately flanking (within 3.5 kb) the previously identified ~260-kb hemizygous region that were strongly associated with floral morph (Fisher's exact test, followed by false discovery rate [FDR] control with the Benjamini-Hochberg procedure, $p < 0.01$) (Figures 2E and 2F). High genetic differentiation (F_{ST}) between thrum and pin individuals extended ~15 kb outside of the hemizygous region (Figures 2B, 2C, and S2C) ($p < 0.01$ after FDR/BH correction, permutation test, 1,000 replicates per window), and 5-kb windows immediately flanking the hemizygous region had higher nucleotide diversity (π) in thrum than that in pin individuals, as expected if these regions were heterozygous in thrum individuals (Figures 2C and S2D) ($p < 0.01$ after FDR/BH correction, permutation test, 1,000 replicates per window).

Together, coverage analyses, GWAS, haplotype reconstruction, and analyses of a high-quality thrum reference assembly and nine additional draft assemblies (including five pin and four thrum assemblies) indicate that the *L. tenue S*-locus is characterized by the presence of an ~260 kb hemizygous region exclusive to the dominant *S*-haplotype and missing from the recessive *s*-haplotype. Sequence differentiation between the dominant and recessive haplotype is limited to ~15 kb immediately flanking the hemizygous region. Thus, the distyly *S*-locus in *L. tenue* is predominantly hemizygous in thrums (98.64% of its sequence).

Lack of broad-scale recombination suppression around the distyly supergene

Highly localized sequence differentiation between pins and thrums surrounding the hemizygous region suggests that the distyly *S*-locus is not located in a genomic region with broad-scale recombination suppression. Accordingly, map-based broad-scale recombination rates were not lower in the region around the *S*-locus (4.09 cM/Mb) than the rest of the genome (Figure S3A), and linkage disequilibrium (LD) decay in natural populations (n = 43) showed no evidence for broad recombination rate suppression outside of the hemizygous region (Figure S3B). Thus, structural variation at the *S*-locus is unlikely to be a consequence of reduced efficacy of selection against structural variation in a broad non-recombining region, and integrity of floral and reproductive trait combinations in pin and thrum morphs is upheld by local recombination suppression at the supergene.

Evolutionary genetic signatures of relaxed selection at the distyly supergene

Hemizygosity of the *S*-locus in thrums can elegantly explain the absence of recombination at distyly supergenes. Morph-limited inheritance and lack of recombination could lead to genetic degeneration and accumulation of repetitive elements, but decay can be slowed down by hemizygous exposure of recessive deleterious alleles to selection.[3] To better understand the evolution of the non-recombining dominant *S*-haplotype, we asked whether the *S*-locus harbors more transposable element (TE) insertions and deleterious mutations than its genomic neighborhood.

First, we compared TE content at the *S*-locus with that of its flanking regions and the rest of LG10. We found that the distyly *S*-locus was enriched in TEs relative to its immediate genomic neighborhood (Figure 3C) (*S*-locus proportion TEs in 25-kb windows: median = 62.50%, n = 9; flanking regions: median = 31.8%, n = 18) (Wilcoxon rank-sum test, W = 5.0, $p < 0.001$) and compared with other windows in LG10 (Figure 3B). The vast majority of *S*-locus TEs that could be classified were retrotransposons (81.73%), consisting of 63.65% long terminal repeat (LTR) retroelements and 18.08% long interspersed repeats (LINEs), followed by DNA transposons (17.18%) and finally Helitrons (1.08%), but there was no marked difference in the TE composition of the *S*-locus relative to its linkage group or the whole genome. Accumulation of TE insertions in distyly supergenes might be facilitated by the joint action of lack of recombination and reduced effective population size, as for other self-incompatibility loci.[34]

Next, using polymorphism data from thrum samples, we tested for a signature of relaxed purifying selection and found no significant difference in the ratio of nonsynonymous to synonymous polymorphism (π_N/π_S) between *S*-linked and neighboring genes (Figure 3E; Wilcoxon rank-sum test, W = 12, NS). These results suggest that *S*-linked genes are not accumulating deleterious SNPs, consistent with recent simulations showing that hemizygosity at supergenes might slow this process.[3]

Distyly candidate genes at the *S*-locus show thrum-specific expression

As a basis for further functional and evolutionary genetic studies, we conducted detailed annotation of the *S*-locus and analyzed RNA-seq data from floral buds, individual mature floral organs (pistils, stamens, and petals), and leaves to delineate candidate genes with floral organ-specific or biased expression between morphs.

There were nine protein-coding genes in the dominant *S*-haplotype, excluding six genes with TE-related functions (Figure 3A; Data S1A). Five of the *S*-locus genes had no known function (Data S1A). Genes encoding proteins with assigned functions included: a protein of the vascular-related unknown protein1 (VUP1) family involved in regulating tissue growth modulated by hormone signaling (see discussion of *LtTSS1*

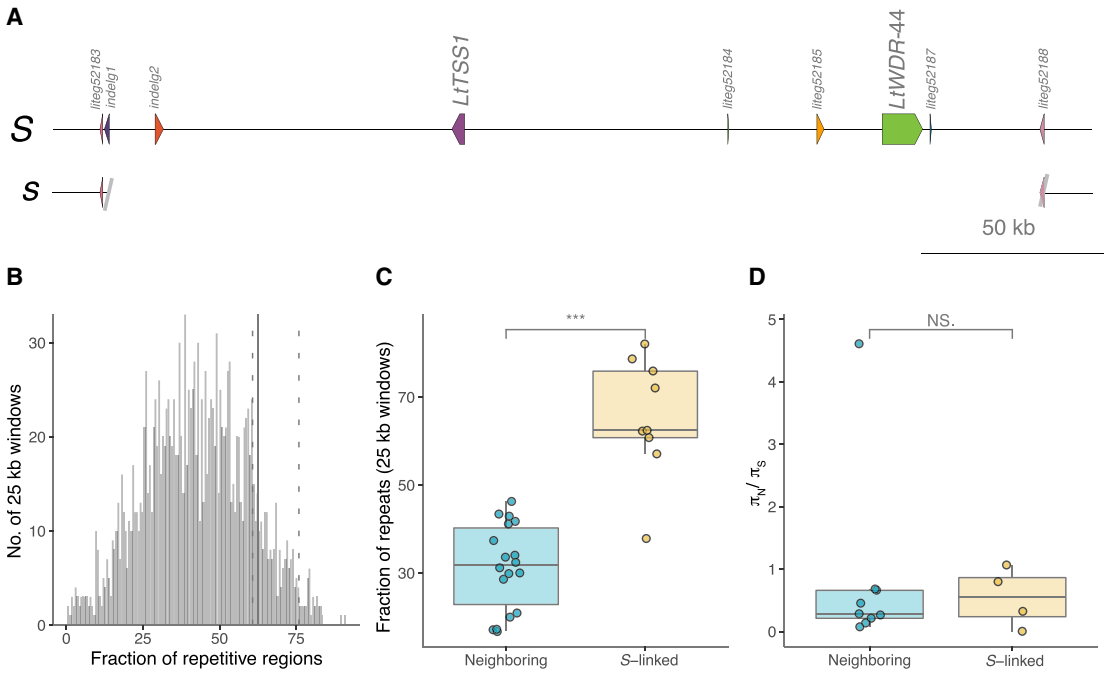

Figure 3. Haplotype structure and patterns of molecular evolution at the *S*-locus

(A) Schematic representation of the dominant *S* and recessive *s* alleles of the distyly *S*-locus, indicating the location of the two distyly candidate genes *LtTSS1* and *LtWDR-44*, only present in the dominant *S* allele.

(B) Distribution of the fraction of repetitive regions in 25-kb windows across LG10. The median value of the fraction of repeats in 25-kb windows ($n = 9$) for the *S*-locus falls in the third quartile of the distribution (solid line represents the *S*-locus median = 62.50%; dotted lines represent the first and third quartiles at the *S*-locus = 60.77% and 75.90%).

(C) Comparison of the fraction of repetitive sequences linked to the *S*-locus (n = 9, 25 kb windows) and neighboring loci (n = 9, 25 kb windows for each up- and down-stream regions) (***$p \leq 0.001$, Wilcoxon rank-sum test). Box edges indicate the lower and upper quartiles, respectively, and the middle line the median, whereas whiskers extend to 1.5 times beyond the upper or lower quartile.

(D) Comparison of π_N/π_S estimates between *S*-linked (n = 4 genes with $\pi_s > 0$) and neighboring genes (n = 8 genes with $\pi_s > 0$ on both sides of the *S*-locus) (mean π_N/π_s: *S*-locus = 0.597, neighboring = 0.300, 13 individuals from population SMT) ($p > 0.05$, Wilcoxon rank-sum test). Boxplots are defined as in (C).

See also Figure S2.

below), an amino acid-binding protein (WD repeat-containing protein 44, WDR44—see discussion of *LtWDR44* below), a tetratricopeptide repeat-containing protein, and a gene encoding a protein belonging to a gene family that includes genes with functions in pollen development (PANTHER family: PTHR34190, GO term: GO:0009555) (Data S1A). The latter gene (*LITEG00000052183*) was the only *S*-locus gene that was also present in the recessive *s*-haplotype, ~1-kb downstream of the ~260-kb indel (Figure 3A). At the upstream limit of the indel, we further annotated a complete gene (*LITEG00000052188*) in the dominant *S*-haplotype that was truncated in the recessive *s*-haplotype (Figure 3A).

The dominant *S*-haplotype harbored a strong candidate gene for style length, *L. tenue TSS1* (*LtTSS1*). This gene is homologous to the *L. grandiflorum* gene *TSS1* (significant alignment with Blastp: E value = 3×10^{-6}, length = 173, gaps = 8%) (Figure 3A; Data S1A), a thrum-specific *S*-linked gene in *L. grandiflorum*.[32] *LtTSS1* shows pistil- and bud-specific expression in thrum and negligible expression in pin (Figures 4A and 4B; Wilcoxon rank-sum test, $W_{bud} = 0$, $p < 0.01$; $W_{pistil} = 0$, $p < 0.05$). The amino acid sequence of *LtTSS1* shows significant similarity to *A. thaliana* VUP1 (AT3G21710; best match = IPR039280, E value = 7.8×10^{-11}, BLAST search with UniProt).[35] VUP1 is a differentially spliced nuclear protein associated with development that suppresses cell expansion leading to reductions in organ size, including shorter floral organs, when overexpressed.[36] Interestingly, the effect of VUP1 overexpression was conserved also for orthologs from other vascular plants such as poplar, *Brachypodium* and *Selaginella*.[36] As cell length is shorter in thrum than pin styles of both *L. tenue* and *L. grandiflorum*,[32,37] *LtTSS1* is a strong candidate for a locus regulating style length.

The *S*-linked gene *LtWDR-44* is an interesting candidate gene for floral differentiation and pollen function between morphs due to its annotation suggesting functions in developmental growth and hormone-mediated signaling. *LtWDR-44* is also exclusively expressed in thrum tissues (Figures 4A, 4B, and 4E), showing primary but not exclusive expression in pistils (median: pistils = 3.72 transcripts per million [TPM], stamen = 0.13 TPM) (Data S1A). Three additional *S*-linked genes (*INDELG00000000001*, *LITEG00000052183*, and *LITEG00000052188*) were expressed in both petals and stamens and are potential candidate genes for anther position or pollen specificity, respectively, whereas the remaining *S*-linked genes were not detectably expressed in our samples. These results help delineate candidate genes at the *S*-locus, confirm thrum-specific expression *S*-linked genes,

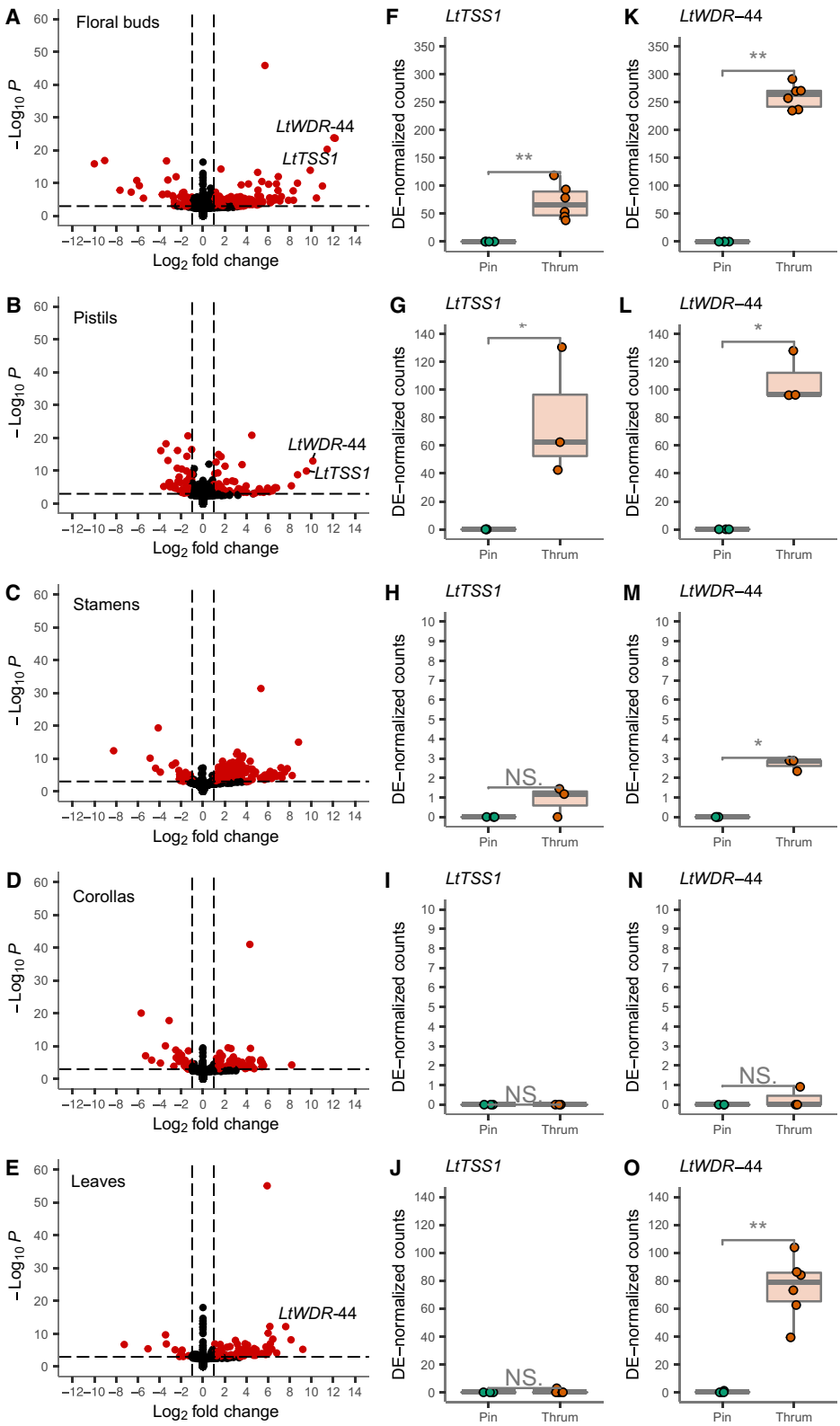

Figure 4. Differential expression in thrum and pin floral organs and leaves demonstrates thrum-biased expression of *S*-linked candidate genes

Volcano plots of log_2-fold change of expression versus significance for (A) entire floral buds (n thrum = 6, pin = 4), (B) pistils (n thrum = 3, pin = 3), (C) stamens (n thrum = 3, pin = 3), (D) corollas (n thrum = 3, pin = 3), and (E) leaves (n thrum = 6, pin = 4). Differentially expressed genes (DEGs) are highlighted in red (horizontal dashed line, cutoff for adjusted $p < 0.001$; vertical dashed lines, log2-fold change > |1.5|). Since *LtTSS1* and *LtWDR-44* were identified as DEG upregulated in

(legend continued on next page)

Table 1. Summary of differentially expressed genes in thrum and pin floral organs and leaves

Tissue	Analyzed	Significant DEG (%)	LFC > 0 (%)	LFC < 0 (%)
Bud	37,001	123 (0.33)	73 (0.20)	50 (0.14)
Pistil	28,615	101 (0.35)	32 (0.11)	69 (0.24)
Stamen	26,908	223 (0.83)	200 (0.74)	23 (0.09)
Petal	23,803	52 (0.22)	23 (0.10)	29 (0.12)
Leaf	30,439	54 (0.18)	37 (0.12)	17 (0.06)

Significant differentially expressed genes (adjusted $p < 0.01$) were classified as up- (LFC > 0) or downregulated (LFC < 0) in thrum relative to pin samples. The analyzed sets exclude genes with zero counts across all samples, extreme count outliers (detected by Cook's distance), and low mean normalized counts. See Data S1B for gene set enrichment results.

and demonstrate that the style-length candidate gene *LtTSS1* has pistil-specific expression in *L. tenue*.

Cell-wall-related genes are differentially regulated in thrum and pin pistils and stamens

To investigate sets of genes and pathways that might be regulated by *S*-linked genes, we performed differential expression analyses. We were specifically interested in differential expression of cell-wall-related genes between floral morphs, as altered expression of such genes is an effect of overexpression of the *TSS1* homolog *VUP1*.[36]

Floral buds showed a higher proportion of differentially expressed genes (DEGs) between morphs compared with leaves, and among mature floral organs, stamens had the largest fraction of DEGs, followed by pistils and petals (Table 1). Cell-wall-related genes (GO:0071555; GO:0042545), mostly related to pectins, were significantly enriched among DEG ($p \leq 0.05$; Data S1E) in both pistils and stamens, which differ in length and cell size between floral morphs,[37] as well as in buds, but not in petals and leaves (Data S1B). In pistils, GO terms showed significant enrichment in cell wall organization/composition through oxido-reductive activity mainly located in the cell walls and anchored to membrane compartments. Thrum-specific expression of *LtTSS1* in pistils is thus associated with altered regulation of genes involved in cell wall organization/composition, potentially resulting in shorter styles in thrum flowers.

Evidence for stepwise assembly of the *Linum S*-locus gene set

We investigated the genomic position of paralogs of *S*-linked genes in our genome assembly. If the distyly supergene formed via segmental duplication, most paralogs of *S*-linked genes should stem from the same genomic region, but if they are scattered across the genome, this would support stepwise assembly of the *S*-linked gene set. Reciprocal BLAST analyses identified putative paralogs of three *S*-linked genes (*LITEG00000052184*, *LITEG00000052185*, and *LtWDR-44*) in two regions of LG1 separated by > 60 Mb, with > 2 Mb between genes at the upstream end of the chromosome (Data S1D). The closest putative paralogs of the *S*-locus genes *LITEG00000052183* and *LITEG00000052188* were found on LG9 separated by 33 Mb. Synonymous divergence (d_S) between *S*-linked genes and their putative paralogs, a proxy for the time since gene duplication, varied greatly and suggests duplication at different times (Figure 5).

Finally, to investigate whether the scattered distribution is ancestral, we conducted similar analyses based on the high-quality reference genomes of two outgroups within the Malphigiales, cassava (*Manihot esculenta*)[38] and poplar (*Populus trichocarpa*).[39] There were four *S*-linked genes with significant matches in both genomes (Data S1E), located on different chromosomes, with the exception of matches to *LtTSS1* and *LtWDR-44*, which were co-located in both genomes, although more than 0.8 and 1.5 Mb apart in the case of cassava and poplar, respectively. Taken together, these results suggest stepwise assembly of the gene set at the *Linum S*-locus.

DISCUSSION

Complex phenotypic polymorphisms have long fascinated biologists; however, until very recently, the genetic architecture and evolution of supergenes remained uncharacterized. Here, we expand detailed characterization of distyly supergenes to *Linum*, a classic system for the study of distyly.[6] Building on a chromosome-level genome assembly of *L. tenue*, we identify the distyly supergene and show that the dominant *S*-haplotype harbors an ~260 kb thrum-specific hemizygous region. Although the dominant allele has accumulated TEs, *S*-linked genes do not show signatures of relaxed purifying selection, in line with theoretical predictions.[3,40,41] Finally, we identified a pistil-specific candidate gene for style length. Our results show that independently evolved *Linum*, *Primula*, and *Turnera S*-loci, which are the distyly supergenes studied in most detail so far, exhibit remarkable convergence in their genetic architecture with hemizygosity in thrum individuals as a shared feature. In ants, convergent genetic architectures involving supergenes underlie independently evolved colony social form polymorphisms.[42] Likewise, convergent evolution of recombination suppression resulting in mating-type supergenes has previously been documented in anther smut fungi.[43] These examples highlight that strong selection can result in convergent evolution of supergene architectures.

Hemizygosity in thrums could elegantly explain both the lack of recombination and dominance of the thrum *S*-locus allele, but structural variation could potentially constitute either a cause or a consequence of suppressed recombination. In *Primula*, the location of the *S*-locus in a pericentromeric region[21,26,44] might have favored supergene formation if recombination rates were ancestrally low in this region. In contrast, at the *Linum* distyly supergene, we observed localized recombination suppression largely coinciding with the boundaries of the thrum-hemizygous

thrum floral buds and pistils, we compared normalized counts of both genes between pin and thrum samples in (F and K) floral buds, (G and L) pistils, (H and M) stamens, (I and N) corollas, and (J and O) leaves (*$p < 0.05$, **$p < 0.01$, Wilcoxon rank-sum test). The lower and upper edge of the boxes show the location of the lower and upper quartiles, respectively, whereas the median is indicated by the thick line within the box. The whiskers extend 1.5 times the interquartile range beyond the lower and upper quartiles, respectively. See also Data S1B.

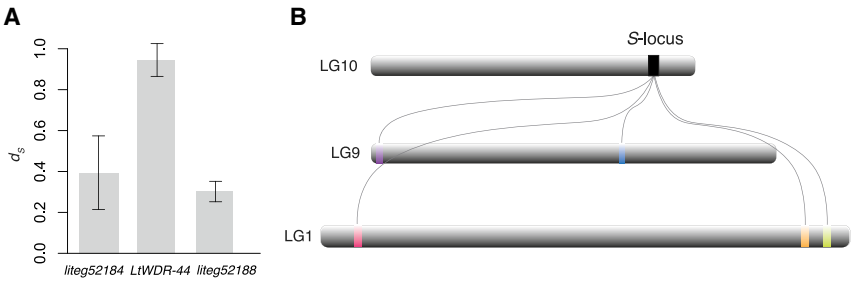

Figure 5. Stepwise assembly of the gene set at the *Linum* distyly *S*-locus

(A) Synonymous divergence between *S*-linked genes and their closest putative paralogs for three *S*-linked genes varies greatly, suggesting gene duplication at different times. Error bars represent standard error estimates based on 500 bootstrap replicates.

(B) Putative paralogs of *S*-linked genes are located in widely separated parts of the *L. tenue* genome assembly, suggesting stepwise assembly of the gene set at the *S*-locus. See Data S1D and S1E for all significant BLAST hits to the *L. tenue* genome and Malphigiales outgroups.

region, without broad-scale recombination suppression surrounding the supergene. This observation suggests that structural variation is not a consequence of suppressed recombination and that local recombination suppression coinciding with structural variation at the *S*-locus preserves the integrity of pin and thrum trait combinations.

The joint effects of suppressed recombination and morph-limited inheritance can precipitate accumulation of deleterious mutations. This process is expected to be slower for supergenes harboring indels instead of inversions because recessive deleterious mutations in hemizygous regions are directly exposed to selection.[3] At the *Linum* distyly supergene, we observed an enrichment of TEs, but no signature of relaxed purifying selection on *S*-linked genes. This result mirrors recent findings in *Primula*,[17,21] and although efficient selection against recessive deleterious mutations in hemizygous regions could contribute to this pattern, sequence constraints on *S*-linked genes resulting from their role in distyly could also be involved. The *Linum* distyly supergene thus provides a counterexample to classic inversion supergenes such as the *P* locus in *Heliconius* which is accumulating both TEs and deleterious mutations.[45]

How then did the *Linum* distyly supergene originate? Although detailed comparative studies will be required to answer this question, our results provide some information on the likely age and sequence of events. The finding that the *L. tenue* *S*-locus harbors *LtTSS1*, an ortholog of the *L. grandiflorum* thrum-specific and *S*-linked gene *TSS1*,[31,32] suggests that aspects of distyly such as style-length polymorphism governed by presence-absence polymorphism at the *S*-locus could have arisen once and been retained for at least 30–40 million years, since the divergence of *L. tenue* and *L. grandiflorum*.[29,46] Our finding that paralogs of *S*-linked genes are found in disparate genomic locations both in *L. tenue* and in outgroups suggests that the *S*-locus gene set evolved in a stepwise process involving several gene duplications and translocations, similar to the *Primula* *S*-locus.[20,21]

The independent evolution of distyly supergenes with hemizygosity in thrums inevitably raises the question: why did evolution repeatedly favor this genetic architecture? One possibility is that hemizygosity facilitates establishment of new advantageous mutations, as they will be directly exposed to efficient selection.[47] Under the "selfing avoidance" model, the first mutation would have led to the origin of pollen incompatibility in a monomorphic and self-compatible population.[48] In contrast, the "pollen transfer" hypothesis suggests that the first mutation would modify the length of the style in a population of approach herkogamous flowers.[12] Although it is unknown how distyly evolved in *Linum*, ancestral state reconstruction studies suggest several transitions from approach herkogamy to distyly.[29] Determining whether mutations affecting pistil length were the first step in the formation of the *Linum* distyly *S*-locus will shed further light on this question.

A particularly strong candidate gene for the control of style elongation in *L. tenue* is the *S*-linked gene *LtTSS1*, due to its pistil-specific expression only in thrums, its inferred biological function, and the presence of an ortholog, *TSS1* as a thrum-specific and *S*-linked gene in *L. grandiflorum*, a species which exhibits style-length polymorphism but not anther height polymorphism.[6] Previous morphometric work in *Linum* reported longer style cells in pin compared with thrum flowers, indicating that style length dimorphism is caused by differences in cell elongation.[32,37] Importantly, because *LtTSS1* significantly resembles genes from the VUP1 family, which have been linked to organ size reduction through repressed cell expansion,[36] it is likely that the expression of *LtTSS1* in thrum pistils limits cell expansion leading to shorter styles. The expression of *LtTSS1* in floral buds suggests the early onset of developmental processes underlying differences in style length between morphs, and the significant enrichment of genes involved in cell wall modification among floral bud and pistil DEG between morphs further supports this idea. Cell length reduction leading to shorter thrum pistils is controlled by hemizygous *CYP734A50* in *Primula*[18] and *TsBAHD* in *Turnera*.[22,23] Interestingly, both genes encode brassinosteroid-inactivating enzymes,[18,23] which not only determine pistil length but also female incompatibility function.[19,23] Heteromorphic self-incompatibility has been documented in *L. tenue*.[49] As overexpression of *VUP1* downregulates both brassinosteroid-responsive and auxin-responsive genes,[36] it will be relevant to investigate the molecular pathways involved in determining pistil length and female incompatibility in *Linum*. Further functional studies of *LtTSS1* and other *S*-linked genes that we have identified here will be required to conclusively elucidate the genetic and developmental mechanisms underlying distyly in *Linum*.

Taken together, our findings in combination with those of earlier studies indicate remarkable convergence with respect to the genetic architecture, origin, and evolution of distyly supergenes in *Linum*, *Primula*,[15,17,20,21] and *Turnera*.[22] Our results motivate further work to better understand the conditions favoring similar mechanisms and processes for the origin and maintenance of supergenes and suggest that convergent molecular mechanisms may underlie a classic case of phenotypic convergence.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Plant material
- METHOD DETAILS
  - DNA extraction and sequencing
  - RNA extraction and sequencing
  - Genome assembly
  - *S*-locus haplotype phasing
  - Genome annotation
  - Population genomic analyses
  - Recombination rate and linkage disequilibrium estimates
  - Genome-wide association mapping
  - Comparison of TE content between the distyly *S*-locus and neighboring windows
  - Comparison of π_N/π_S between *S*-linked and neighboring genes
  - Differential expression, gene set enrichment and patterns of expression of *S*-linked genes
  - Identification of putative paralogs of *S*-linked genes and divergence estimates
- QUANTIFICATION AND STATISTICAL ANALYSIS

ACKNOWLEDGMENTS

We thank Sara Mehrabi for assistance with lab work, José Ruíz-Martin for assistance with field work, and Jerker Eriksson for technical assistance. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 757451), from the Swedish Research Council (grant no. 2019-04452) and the Erik Philip-Sörensen foundation to T.S., and from the Bergströms foundation to J.G.-V. E.L.B. was funded through a Carl Tryggers grant to T.S. B.N. and V.E.K. are financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. The authors acknowledge support from the National Genomics Infrastructure (NGI) in Stockholm and Uppsala (Uppsala Genome Center, SNP&SEQ), funded by the Knut and Alice Wallenberg Foundation, the Swedish Research Council, and Science for Life Laboratory. We acknowledge support of the Swedish National Infrastructure for Computing (SNIC) and Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for assistance with massively parallel sequencing and access to computational infrastructure. SNIC was partially funded by the Swedish Research Council through grant agreement no. 2018-05973. Support by NBIS (National Bioinformatics Infrastructure Sweden) is gratefully acknowledged.

AUTHOR CONTRIBUTIONS

J.G.-V., A.L., A.D., P.W.H., A.F., B.L., and M.A. conducted experimental work. J.G.-V., M.F., E.L.B., I.B., L.S., J.D., A.F., and T.S. performed bioinformatic analyses. O.V.P., B.N., V.E.K., and E.P. advised on sequencing and bioinformatic analyses. J.-G.V., A.D., and B.L. conducted field work. J.A. and A.C.B. advised on study system and field work; J.G.-V. and T.S. wrote the manuscript, with extensive input by M.F. and E.L.B. All co-authors commented on the manuscript. T.S. designed the study and supervised the work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in science. One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community. One or more of the authors of this paper self-identifies as living with a disability.

Received: July 11, 2022
Revised: August 11, 2022
Accepted: August 15, 2022
Published: September 9, 2022

REFERENCES

1. Thompson, M.J., and Jiggins, C.D. (2014). Supergenes and their role in evolution. Heredity *113*, 1–8.
2. Schwander, T., Libbrecht, R., and Keller, L. (2014). Supergenes and complex phenotypes. Curr. Biol. *24*, R288–R294.
3. Gutiérrez-Valencia, J., Hughes, P.W., Berdan, E.L., and Slotte, T. (2021). The genomic architecture and evolutionary fates of supergenes. Genome Biol. Evol. *13*, evab057.
4. Bateson, W., and Gregory, R.P. (1905). On the inheritance of heterostylism in *Primula*. Proc. R. Soc. Lond. B *76*, 581–586.
5. Ernst, A. (1936). Heterostylie-forschung. Versuche zur genetischen Analyse eines Organisations-und "Anpassungs" Merkmales. Z. induktive Abstammungs Vererbungsl. *71*, 156–230.
6. Darwin, C. (1863). On the existence of two forms, and on their reciprocal sexual relation, in several species of the genus *Linum*. Bot. J. Linn. Soc. *7*, 69–83.
7. Darwin, C. (1877). The Different Forms of Flowers on Plants of the Same Species (Cambridge University Press).
8. Dulberger, R. (1992). Floral polymorphisms and their functional significance in the heterostylous syndrome. In Evolution And Function Of Heterostyly. Monographs on Theoretical and Applied Genetics, S.C.H. Barrett, ed. (Springer), pp. 41–84.
9. Barrett, S.C.H. (2002). Sexual interference of the floral kind. Heredity *88*, 154–159.
10. Keller, B., Thomson, J.D., and Conti, E. (2014). Heterostyly promotes disassortative pollination and reduces sexual interference in Darwin's primroses: evidence from experimental studies. Funct. Ecol. *28*, 1413–1425.
11. Ganders, F.R. (1979). The biology of heterostyly. N. Z. J. Bot. *17*, 607–635.
12. Lloyd, D.G., and Webb, C.J. (1992). The evolution of heterostyly. In Evolution And Function Of Heterostyly. Monographs on Theoretical and Applied Genetics, S.C.H. Barrett, ed. (Springer), pp. 151–178.
13. Barrett, S.C.H. (2019). 'A Most Complex Marriage Arrangement': recent advances on heterostyly and unresolved questions. New Phytol. *224*, 1051–1067.
14. Laibach, F. (1923). Die Abweichungen vom "mechanischen" Zahlenverhältnis der Long- under Kurz- griffel bei heterostylen Pflanzen. Biol. Zentralbl. *43*, 148–157.
15. Li, J., Cocker, J.M., Wright, J., Webster, M.A., McMullan, M., Dyer, S., Swarbreck, D., Caccamo, M., van Oosterhout, C.V., and Gilmartin, P.M. (2016). Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*. Nat. Plants *2*, 16188.
16. Burrows, B.A., and McCubbin, A.G. (2017). Sequencing the genomic regions flanking *S*-linked *PvGLO* sequences confirms the presence of two

GLO loci, one of which lies adjacent to the style-length determinant gene *CYP734A50*. Plant Reprod. *30*, 53–67.

17. Cocker, J.M., Wright, J., Li, J., Swarbreck, D., Dyer, S., Caccamo, M., and Gilmartin, P.M. (2018). *Primula vulgaris* (primrose) genome assembly, annotation and gene expression, with comparative genomics on the heterostyly supergene. Sci. Rep. *8*, 17942.
18. Huu, C.N., Kappel, C., Keller, B., Sicard, A., Takebayashi, Y., Breuninger, H., Nowak, M.D., Bäurle, I., Himmelbach, A., Burkart, M., et al. (2016). Presence versus absence of *CYP734A50* underlies the style-length dimorphism in primroses. eLife *5*, e17956.
19. Huu, C.N., Plaschil, S., Himmelbach, A., Kappel, C., and Lenhard, M. (2022). Female self-incompatibility type in heterostylous *Primula* is determined by the brassinosteroid-inactivating cytochrome P450 *CYP734A50*. Curr. Biol. *32*, 671–676.e5.
20. Huu, C.N., Keller, B., Conti, E., Kappel, C., and Lenhard, M. (2020). Supergene evolution via stepwise duplications and neofunctionalization of a floral-organ identity gene. Proc. Natl. Acad. Sci. USA *117*, 23148–23157.
21. Potente, G., Léveillé-Bourret, É., Yousefi, N., Choudhury, R.R., Keller, B., Diop, S.I., Duijsings, D., Pirovano, W., Lenhard, M., Szövényi, P., and Conti, E. (2022). Comparative genomics elucidates the origin of a supergene controlling floral heteromorphism. Mol. Biol. Evol. *39*, msac035.
22. Shore, J.S., Hamam, H.J., Chafe, P.D.J., Labonne, J.D.J., Henning, P.M., and McCubbin, A.G. (2019). The long and short of the *S*-locus in *Turnera* (Passifloraceae). New Phytol. *224*, 1316–1329.
23. Matzke, C.M., Hamam, H.J., Henning, P.M., Dougherty, K., Shore, J.S., Neff, M.M., and McCubbin, A.G. (2021). Pistil mating type and morphology are mediated by the brassinosteroid inactivating activity of the *S*-locus gene *BAHD* in heterostylous *Turnera* species. Int. J. Mol. Sci. *22*, 10603.
24. Yasui, Y., Mori, M., Aii, J., Abe, T., Matsumoto, D., Sato, S., Hayashi, Y., Ohnishi, O., and Ota, T. (2012). S-Locus early flowering 3 is exclusively present in the genomes of short-styled buckwheat plants that exhibit heteromorphic self-incompatibility. PLoS One *7*, e31264.
25. Yasui, Y., Hirakawa, H., Ueno, M., Matsui, K., Katsube-Tanaka, T., Yang, S.J., Aii, J., Sato, S., and Mori, M. (2016). Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes. DNA Res. *23*, 215–224.
26. Kappel, C., Huu, C.N., and Lenhard, M. (2017). A short story gets longer: recent insights into the molecular basis of heterostyly. J. Exp. Bot. *68*, 5719–5730.
27. Armbruster, W.S., Pérez-Barrales, R., Arroyo, J., Edwards, M.E., and Vargas, P. (2006). Three-dimensional reciprocity of floral morphs in wild flax (*Linum suffruticosum*): a new twist on heterostyly. New Phytol. *171*, 581–590.
28. McDill, J., Repplinger, M., Simpson, B.B., and Kadereit, J.W. (2009). The phylogeny of *Linum* and Linaceae subfamily Linoideae, with implications for their systematics, biogeography, and evolution of heterostyly. Syst. Bot. *34*, 386–405.
29. Ruiz-Martín, J., Santos-Gally, R., Escudero, M., Midgley, J.J., Pérez-Barrales, R., and Arroyo, J. (2018). Style polymorphism in *Linum* (Linaceae): a case of Mediterranean parallel evolution? Plant Biol. (Stuttg.) *20* (suppl 1), 100–111.
30. Rogers, C.M. (1980). Pollen dimorphism in distylous species of *Linum* sect. Linastrum (Linaceae). Grana *19*, 19–20.
31. Ushijima, K., Nakano, R., Bando, M., Shigezane, Y., Ikeda, K., Namba, Y., Kume, S., Kitabata, T., Mori, H., and Kubo, Y. (2012). Isolation of the floral morph-related genes in heterostylous flax (*Linum grandiflorum*): the genetic polymorphism and the transcriptional and post-transcriptional regulations of the S locus. Plant J. *69*, 317–331.
32. Ushijima, K., Ikeda, K., Nakano, R., Matsubara, M., Tsuda, Y., and Kubo, Y. (2015). Genetic control of floral morph and petal pigmentation in *Linum grandiflorum* Desf., a heterostylous flax. The Hortic. J. *84*, 261–268.
33. Pastor, J., Diosdado, J.C., Santa Bárbara, C., Vioque, J., and Pérez, E. (1990). Números cromosómicos flora Española. Lagascalia *15*, 556–619.
34. Goubet, P.M., Bergès, H., Bellec, A., Prat, E., Helmstetter, N., Mangenot, S., Gallina, S., Holl, A.C., Fobis-Loisy, I., Vekemans, X., and Castric, V. (2012). Contrasted patterns of molecular evolution in dominant and recessive self-incompatibility haplotypes in *Arabidopsis*. PLoS Genet. *8*, e1002495.
35. UniProt Consortium (2021). UniProt: the universal protein KnowledgeBase in 2021. Nucleic Acids Res. *49*, D480–D489.
36. Grienenberger, E., and Douglas, C.J. (2014). *Arabidopsis* VASCULAR-RELATED UNKNOWN PROTEIN1 regulates xylem development and growth by a conserved mechanism that modulates hormone signaling. Plant Physiol. *164*, 1991–2010.
37. Foroozani, A. (2018). Structural and molecular analyses of heterostyly in *Linum tenue* (Linaceae). PhD thesis (Durham University).
38. Bredeson, J.V., Lyons, J.B., Prochnik, S.E., Wu, G.A., Ha, C.M., Edsinger-Gonzales, E., Grimwood, J., Schmutz, J., Rabbi, I.Y., Egesi, C., et al. (2016). Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. Nat. Biotechnol. *34*, 562–570.
39. Tuskan, G.A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science *313*, 1596–1604.
40. Charlesworth, B., Morgan, M.T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics *134*, 1289–1303.
41. Birky, C.W., and Walsh, J.B. (1988). Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. USA *85*, 6414–6418.
42. Purcell, J., Brelsford, A., Wurm, Y., Perrin, N., and Chapuisat, M. (2014). Convergent genetic architecture underlies social organization in ants. Curr. Biol. *24*, 2728–2732.
43. Branco, S., Carpentier, F., Rodríguez de la Vega, R.C., Badouin, H., Snirc, A., Le Prieur, S., Coelho, M.A., de Vienne, D.M., Hartmann, F.E., Begerow, D., et al. (2018). Multiple convergent supergene evolution events in mating-type chromosomes. Nat. Commun. *9*, 2000.
44. Li, J., Webster, M.A., Wright, J., Cocker, J.M., Smith, M.C., Badakshi, F., Heslop-Harrison, P., and Gilmartin, P.M. (2015). Integration of genetic and physical maps of the *Primula vulgaris S* locus and localization by chromosome *in situ* hybridization. New Phytol. *208*, 137–148.
45. Jay, P., Chouteau, M., Whibley, A., Bastide, H., Parrinello, H., Llaurens, V., and Joron, M. (2021). Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. Nat. Genet. *53*, 288–293.
46. Maguilla, E., Escudero, M., Ruíz-Martín, J., Arroyo, J., and Schneeweiss, G. (2021). Origin and diversification of flax and their relationship with heterostyly across the range. J. Biogeogr. *48*, 1994–2007.
47. Haldane, J.B.S. (1927). A mathematical theory of natural and artificial selection, Part V: Selection and mutation. Math. Proc. Camb. Philos. Soc. *23*, 838–844.
48. Charlesworth, D., and Charlesworth, B. (1979). A model for the evolution of distyly. Am. Nat. *114*, 467–498.
49. Murray, B.G. (1986). Floral biology and self-incompatibility in *Linum*. Botanical Gazette *147*, 327–333.
50. Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods *13*, 1050–1054.
51. Kronenberg, Z.N., Rhie, A., Koren, S., Concepcion, G.T., Peluso, P., Munson, K.M., Porubsky, D., Kuhn, K., Mueller, K.A., Low, W.Y., et al. (2021). Extended haplotype-phasing of long-read *de novo* genome assemblies using Hi-C. Nat. Commun. *12*, 1935.

52. Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics *19*, 460.

53. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arXiv.1303.3997.

54. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

55. Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat. Plants *5*, 833–845.

56. Rastas, P. (2020). Lep-Anchor: automated construction of linkage map anchored haploid genomes. Bioinformatics *36*, 2359–2364.

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

58. Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., Fu, Y., Yuan, S., Chen, S., and Xu, A. (2012). HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. *22*, 1581–1588.

59. Li, H. (2018). minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

60. Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J.M., Ouellette, S.B., Azhir, A., Kumar, N., et al. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. *19*, 125.

61. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

62. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat. Protoc. *8*, 1494–1512.

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

64. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

65. Smit, A.F.A., and Hubley, R. (2008). RepeatModeler. http://www.repeatmasker.org.

66. Smit, A.F.A., Hubley, R., and Green, P. (2013). RepeatMasker. http://www.repeatmasker.org.

67. Smith, C.D., Edgar, R.C., Yandell, M.D., Smith, D.R., Celniker, S.E., Myers, E.W., and Karpen, G.H. (2007). Improved repeat identification and masking in dipterans. Gene *389*, 1–9.

68. Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12*, 491.

69. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics *24*, 637–644.

70. Bushnell, B. (2015). BBMap/BBTools. sourceforge.net/projects/bbmap/.

71. Broad Institute (2019). Picard tools (Broad Institute, GitHub Repository). http://broadinstitute.github.io/picard/.

72. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

73. Danecek, P., and McCarthy, S.A. (2017). BCFtools/csq: haplotype-aware variant consequences. Bioinformatics *33*, 2037–2039.

74. Korunes, K.L., and Samuk, K. (2021). Pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. Mol. Ecol. Resour. *21*, 1359–1368.

75. Fox, E.A., Wright, A.E., Fumagalli, M., and Vieira, F.G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics *35*, 3855–3856.

76. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.

77. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

78. Alexa, A., and Rahnenfuhrer, J. (2022). topGO: Enrichment Analysis for Gene Ontology (Bioconductor Version: Release (3.15)).

79. Löytynoja, A., and Goldman, N. (2010). webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics *11*, 579.

80. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. *35*, 1547–1549.

81. Stecher, G., Tamura, K., and Kumar, S. (2020). Molecular evolutionary genetics analysis (MEGA) for macOS. Mol. Biol. Evol. *37*, 1237–1239.

82. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

83. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped Blast And PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

84. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). Blast+: architecture and applications. BMC Bioinformatics *10*, 421.

85. Workman, R.E., Timp, W., Fedak, R., Kilburn, D., Hao, S., and Liu, K.J. (2019). High molecular weight DNA extraction from recalcitrant plant species for third generation sequencing. Protocols.io. https://doi.org/10.17504/protocols.io.4vbgw2n.

86. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

87. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963.

88. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. Preprint at arXiv. https://doi.org/10.48550/arXiv.1207.3907.

89. Magrane, M.; UniProt Consortium (2011). UniProt KnowledgeBase: a hub of integrated protein data. Database (Oxford) *2011*, bar009.

90. Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using De Bruijn graphs. Genome Res. *18*, 821–829.

91. Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics *28*, 1086–1092.

92. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

93. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.

94. Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. *40*, D306–D312.

95. Lowe, T.M., and Eddy, S.R. (1997). TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

96. Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., and Finn, R.D. (2015). Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. *43*, D130–D137.

97. Nawrocki, E.P., and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics *29*, 2933–2935.

98. Broman, K.W., Wu, H., Sen, S., and Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics *19*, 889–890.

99. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B Methodol. *57*, 289–300.

100. Zhu, A., Ibrahim, J.G., and Love, M.I. (2019). Heavy-tailed prior distributions for sequence Count Data: removing the noise and preserving large differences. Bioinformatics *35*, 2084–2092.

101. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

102. Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., and Lewis, S.; AmiGO Hub, and Web Presence Working Group (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288–289.

103. Gene Ontology Consortium (2021). The gene ontology resource: enriching a Gold Mine. Nucleic Acids Res. *49*, D325–D334.

104. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. *3*, 418–426.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Linum tenue plants	This study	N/A
Critical commercial assays		
Genomic-tip 100/G	QIAGEN	Cat#10243
Isolate II Plant DNA	Bioline	Cat#52070
RNeasy Plant Mini kit	QIAGEN	Cat#74904
Deposited data		
Raw Illumina sequences	This study	ENA: PRJEB52918 (https://www.ebi.ac.uk/ena/)
PacBio sequences	This study	ENA: PRJEB52918 (https://www.ebi.ac.uk/ena/)
Linum tenue genome	This study	ENA: PRJEB52918 (https://www.ebi.ac.uk/ena/)
Software and algorithms		
FALCON Unzip	Chin et al. [50]	https://pb-falcon.readthedocs.io/en/latest/about.html
FALCON-Phase v1.2.0	Kronenberg et al.[51]	https://github.com/phasegenomics/FALCON-Phase
PurgeHaplotigs	Roach et al.[52]	https://bitbucket.org/mroachawri/purge_haplotigs/src/master/
BWA-MEM v07.17	Li and Durbin[53]	https://github.com/lh3/bwa
SAMtools v1.9	Li et al.[54]	https://github.com/samtools/samtools
ALLHiC (v0.9.13)	Zhang et al.[55]	https://github.com/tanghaibao/allhic
Long Ranger	10x Genomics	https://github.com/10xGenomics/longranger
Pilon	Walker et al.[51]	https://github.com/broadinstitute/pilon
Lep-MAP3	Rastas[56]	https://sourceforge.net/p/lep-map3/wiki/LM3%20Home/
Lep-Anchor	Rastas[57]	https://sourceforge.net/p/lep-anchor/wiki/Home/
HaploMerger	Huang et al.[58]	https://github.com/mapleforest/HaploMerger
Minimap2	Li[59]	https://github.com/lh3/minimap2
HiGlass	Kerpedjiev et al.[60]	https://higlass.io/
Trinity (2.9.1)	Grabherr et al.,[61] Haas et al.[62]	https://github.com/trinityrnaseq/trinityrnaseq/wiki
HISAT2	Kim et al.[63]	https://github.com/DaehwanKimLab/hisat2
StringTie	Pertea et al.[64]	https://github.com/gpertea/stringtie
RepeatModeler	Smit and Hubley[65]	https://github.com/Dfam-consortium/RepeatModeler
RepeatMasker	Smit et al.[66]	https://github.com/rmhubley/RepeatMasker
RepeatRunner	Smith et al.[67]	https://github.com/Yandell-Lab/RepeatRunner
MAKER2	Holt and Yandell[68]	https://www.yandell-lab.org/software/maker.html
Augustus	Stanke et al.[69]	N/A
BBMap/BBTools	Bushnell[70]	https://github.com/BioInfoTools/BBMap
Picard tools v2.0.1	Broad Institute[71]	https://github.com/broadinstitute/picard
BEDTools	Quinlan and Hall[72]	https://github.com/arq5x/bedtools2
Bcftools	Danecek and McCarthy[73]	N/A
Pixy	Korunes and Samuk[74]	https://github.com/ksamuk/pixy/blob/master/docs/index.rst
ngsLD	Fox et al.[75]	https://github.com/fgvieira/ngsLD
PLINK v1.90b4.9	Chang et al.[76]	https://github.com/chrchang/plink-ng
DESeq2 R package	Love et al.[77]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
TopGO 2.46.0 R package	Alexa and Rahnenfuhrer[78]	https://bioconductor.org/packages/release/bioc/html/topGO.html
webPRANK	Löytynoja and Goldman[79]	https://www.ebi.ac.uk/goldman-srv/webprank/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MEGA X	Kumar et al.,[80] Stecher et al.[81]	https://www.megasoftware.net/
BLAST	Altschul et al.,[82] Altschul et al.,[83] Chamacho et al.[84]	https://blast.ncbi.nlm.nih.gov/Blast.cgi
UniProt (2021)	UniProt Consortium[35]	https://www.uniprot.org/
Other		
Scripts and code for analyses	Zenodo	https://doi.org/10.5281/zenodo.6786883

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Tanja Slotte (tanja.slotte@su.se).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All sequencing data generated in this study has been uploaded to ENA: PRJEB52918 (https://www.ebi.ac.uk/ena/). Accession numbers are listed in the key resources table.
- All original code is publicly available at Zenodo, with doi listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Plant material

Mature fruits of the distylous *Linum tenue* Desf. which occurs in southern Spain and north Africa were sampled at two localities in Andalusia, Spain: iii) in Castillo de Locubín, Valdepeñas de Jaén (Jaén) (37°33'50.5"N 3°53'22.9"W) (CL onwards), and) Santa María de Trassierra, Sierra Morena (Córdoba) (37°55'55.1"N, 4°52'59.9"W) (SMT onwards). Fruits from each mother plant were collected in individual paper bags and treated with silica gel for 24 to 48 hours to reduce moisture. For plant propagation, we sterilized seeds using a treatment of 10% bleach solution with liquid detergent, followed by washes with 70% ethanol and sterile, distilled water. Seeds were germinated in Murashige-Skoog medium (Sigma Aldrich, USA) and covered with a thin layer of top agar. After 25 days of stratification at 4°C, we moved petri dishes to a controlled climate chamber at Stockholm University (Stockholm, Sweden) set to 16 h light at 20°C: 8 h dark at 18°C, 60% maximum humidity, 122 μE light intensity. We transplanted germinated seeds to pots containing a mixture of equal parts of vermiculite and perlite for two parts of regular soil (Hasselfors Garden, Sweden).

We collected and snap froze leaves of one thrum individual (SMT-34-3) for the production of a genome assembly, leaves of 43 individuals with known morph (CL: pin=8, thrum=8; SMT: pin=13, thrum=13) for population genomic analyses, and F1 (pin=F1.11 and thrum=F1.5) and 180 F2 individuals for the production of a linkage map.

For annotation of our genome assembly, we sampled and snap froze leaves, stems, floral buds and mature flowers of two thrum individuals (SMT-3-1 and SMT-38-2). Samples were taken with sterilized forceps and placed in 2 mL Eppendorf tubes, then flash frozen in liquid nitrogen. For differential expression analyses, we sampled and snap froze floral buds (*n* thrum=6, pin=4, three replicates per individual), leaves (*n* thrum=6, pin=4, three replicates per individual), pistils (*n* thrum=3, pin=3), stamens (*n* thrum=3, pin=3) and petals (*n* thrum=3, pin=3) (Data S1C).

For generation of chromatin conformation and capture data, we sampled leaves from a thrum individual (individual SMT-24-4). For a subset of three individuals, fresh leaves were collected and tested for absolute DNA content at Plant Cytometry Services (Didam, The Netherlands) using the propidium iodide (PI) staining method and flow cytometry using *Allium schoenoprasum* as standard (DNA content=15.03 pg/2C).

We produced F1 plants by performing legitimate crosses between individuals from populations CL and SMT grown in our growth chambers at Stockholm University. After performing multiple hand-pollinations, we selected a single pair of parental plants (SMT.2.1 - Pin x CL.1.1 - Thrum) based on the number of successfully produced fruits harvested and stored as previously described. Two individuals from this F1 were used to produce an F2 mapping population. As before, we performed multiple crosses of a pair of F1 individuals (F1.11 - Pin x F1.5 - Thrum) and collected ripe fruits from plants that were grown and sampled following the same experimental settings and procedures previously described. Finally, we grew 180 F2 offspring from this cross and sampled leaves for the production of a genetic map. In total, our study included 244 *L. tenue* individuals (62 individuals from natural populations, 2 F1s and 180 F2s; Data S1C).

METHOD DETAILS

DNA extraction and sequencing

To generate a reference genome for the thrum morph of *L. tenue*, we produced a variety of data sets for which different extraction and sequencing methods were used. To obtain long-read sequencing data using the Single Molecule, Real-Time (SMRT) Sequencing technology from Pacific Biosciences (PacBio) (Sequel sequencing instrument, V3.0 chemistry), young leaves from a thrum individual (SMT-34-3) were snap frozen and disrupted by high-speed shaking with stainless steel beads. High molecular-weight (HMW) genomic DNA was extracted from approximately 100 mg of lysed material using the kit Genomic-tip 100/G (QIAGEN, Germany) following manufacturer's instructions. DNA concentration and purity were measured using Qubit and NanoDrop respectively. Sample integrity and fragment size distribution were determined through pulsed-field capillary electrophoresis using a Femto Pulse system. DNA was sheared to 60 Kb, and a SMRTbell library of size 31 Kb was generated and further sequenced on 16 SMRT cells. The resulting data was merged using DatasetMeger 0.3.0, and imported into the SMRT Analysis software suite (v2.3.0) (Pacific Biosciences, CA) to filter out sequences shorter than 500 bp or with a quality lower than 80, and generate subreads. This led to a data set of 119 Gb represented in 10,217,936 subreads.

To generate chromosomal conformation and capture data, nuclei from young leaves of a thrum individual were extracted using the protocol presented by Workman et al.[85] Chromatin conformation capture data was generated following the manufacturer's instructions for the commercially available kit of the Hi-C method[86] Proximo Hi-C 2.0 of Phase Genomics (Seattle, WA). First, intact nuclei were crosslinked using a formaldehyde solution, digested using the DpnII restriction enzyme, and proximity ligated with biotinylated nucleotides to create chimeric molecules composed of fragments from different regions of the genome that were physically proximal *in vivo*. Molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library.

To polish the genome assembly, an HMW DNA extract from the individual SMT-34-3 (previously sequenced with SMRT PacBio) was used to create a 10X Chromium genomic library for linked-read sequencing. HMW DNA was extracted using the kit Genomic-tip 100/G (QIAGEN, Germany). Sequencing libraries were prepared from 0.6 ng of DNA using the Chromium Genome Library preparation kit (cat# 120∼260/58/61/62) following the manufacturers' protocol (#CG00043 Chromium Genome Reagent Kit v2 User Guide). The library was sequenced on a HiSeqX system (paired-end 150bp read length, v2.5 sequencing chemistry). Linked-read sequences were also obtained for additional samples from natural populations (thrum=5 and pin=5). Both library preparation and sequencing were conducted at the SNP&SEQ Technology Platform in Uppsala (Sweden).

To aid scaffolding and correction of potential misassembles, we generated data for the production of a genetic linkage map. Thus, we sequenced the pair of parental plants (SMT.2.1 - Pin x CL.1.1 - Thrum), the F1 individuals (F1.11 - Pin and F1.5 - Thrum) and 180 F2 samples. Genomic DNA from young frozen leaves was extracted using the kit Isolate II Plant DNA (Bioline, UK). DNA concentration and purity were measured using Qubit and NanoDrop respectively. For parental individuals and F1s, sequencing libraries were prepared from 100 ng of DNA using the TruSeq Nano DNA sample preparation kit from Illumina (cat# 20015964/5) targeting an insert size of 350 bp, following the manufacturers' instructions. For F2s, indexed sequencing libraries were generated using the Nextera DNA Flex protocol, according to the manufacturers' instructions. For the sequencing of the parental individuals, libraries were sequenced on Illumina's HiSeqX sequencer (150-bp paired-end reads, v2.5 sequencing chemistry). For the sequencing of the F1 individuals, a NovaSeq 6000 system from Illumina was used with a S4 flowcell (150-bp paired-end reads, v1 sequencing chemistry). Finally, DNA from 180 F2 individuals was processed on a NovaSeq6000 (NovaSeq Control Software 1.6.0/RTA v3.4.4) sequencer with a 2x151 setup using 'NovaSeqXp' workflow in 'S4' mode flowcell. The Bcl to FastQ conversion was performed using bcl2fastq_v2.20.0.422 from the CASAVA software suite. For parental and F1 individuals, library preparation and sequencing were executed at the SNP&SEQ Technology Platform in Uppsala (Sweden) and for F2 individuals at the Genomics Applications Platform of the National Genomics Infrastructure (Sweden).

Finally, we generated a population genomic data set by sequencing 43 individuals with known morph (Data S1B). Genomic DNA extraction and short-read sequencing of these individuals was done as described above for parental plants.

RNA extraction and sequencing

For annotation of the genome assembly, total RNA was extracted from leaves, stems, floral buds and mature flowers using RNAeasy Plant Mini kit (QIAGEN, Germany) as per the manufacturer's instructions. RNA quantity and quality were evaluated by running aliquots of all samples on an Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, USA) using RNA Plant Nano microfluidic chips. Illumina TruSeq RNA Library v2 prep kits (Illumina, San Diego) were used for library construction. Two technical replicates derived from independently generated sequencing libraries were sequenced for each biological replicate (individual). Sequencing libraries were prepared using the TruSeq stranded mRNA library preparation kit (Illumina) including polyA selection according to the manufacturers' protocol. Sequencing was then performed on an Illumina NovaSeq S1 Sequencing System to produce paired-end 150bp read length sequences using the v1 chemistry at the SNP&SEQ Technology Platform in Uppsala (Sweden).

Genome assembly

We produced a high-quality *de novo* genome assembly of a single outbred *L. tenue* thrum individual based on high-coverage (∼170x) PacBio long-read data (V3.0 chemistry on Sequel), with scaffolding using Hi-C data and a genetic linkage map. We first generated a partially phased diploid assembly (i.e. a diploid assembly with primary contigs and haplotigs) using FALCON and FALCON Unzip[50] that was polished using Arrow. The resulting primary assembly was 879.49 Mb (No. sequences=1222, contig

N50=1.43 Mb; Table S1) and had an associated set of haplotigs of 161.13 Mb (No. sequences=1540). We used PurgeHaplotigs[52] to reassign misclassified primary contigs as haplotigs based on read-depth analyses and repeat annotations (Figure S1).

We next scaffolded the assembly using chromatin conformation and capture (Hi-C) data. The purged version of the primary assembly (615.74 Mb, No. sequences=553), the set of haplotigs originally identified with FALCON-Unzip concatenated to those reassigned by PurgeHaplotigs (420.59 Mb, No. sequences=2,128), and the Hi-C data were used to phase and scaffold the genome assembly. First, the partially phased PacBio-based assembly was processed together with the Hi-C data using FALCON-Phase v1.2.0[51] available as the pb-assembly package (v0.0.8) from Bioconda (https://bioconda.github.io/recipes/pb-assembly/) to produce two full length pseudo-haplotypes for phase0 and 1. For scaffolding, Hi-C reads were aligned to phase0 of the FALCON-Phase assembly using BWA-MEM (v0.7.17)[53] followed by filtering with SAMtools (v1.9)[54] and the script PreprocessSAMs.pl (https://github.com/tangerzhang/ALLHiC/blob/master/scripts/) to keep only links with strong signals in the Hi-C data set, and the scaffolds of phase0 were chained into ten pseudochromosomes using ALLHiC (v0.9.13).[55] This step was followed by a rescue step to assign unplaced contigs into partitioned clusters as implemented in ALLHiC (v0.9.13).[55] After a second round of FALCON-Phase using the scaffolded version of phase0 as primary assembly and the non-scaffolded version of phase1 as associated contigs, misplaced scaffolds were reassigned to each phase. This yielded a chromosome-scale (scaffold N50 64.2 Mb) phased pseudo-diploid assembly of ten pseudochromosomes totaling 703.83 Mb for phase0 and 702.16 Mb for phase1 (Table S1). To polish the assembly, 10x Genomics linked reads were aligned to each phase of the reference using Long Ranger (https://github.com/10xGenomics/longranger), followed by two rounds of Pilon.[87] We further edited phase1 to retain a region (contig 000228F_arrow) preliminarily identified as associated with floral morph, that was removed during the first step of the Hi-C scaffolding procedure (Figure S4). Assembly statistics for each step of the assembly were obtained using the script gaas_fasta_statistics.pl (https://github.com/NBISweden/GAAS/tree/master/bin).

We anchored scaffolds to pseudochromosomes and corrected mis-joins using Lep-Anchor.[56] As Lep-Anchor requires a haploid assembly, we processed a haploid assembly representing one of our phases of our scaffolded pseudo-diploid assembly (phase1). First, we generated a linkage map in Lep-MAP3[57] based on short-read sequencing data from a mapping population of 180 F2 individuals. Second, we generated a chains file made in HaploMerger.[58] Third, we generated a paf file mapping the FALCON phase round 1 assembly to the scaffolded pseudo-diploid assembly (phase1) with minimap2.[59] We input these data in Lep-Anchor to obtain 10 pseudochromosomes.

Inspection of Lep-Anchor Marey maps identified regions, where the recombination map position and the physical position of the marker did not match (Figure S5A). Further inspection of the Hi-C data based on a HiGlass visualization[60] supported this conclusion, and thus we manually edited two regions on LG1 and LG5 (Figure S5B). In the first case we moved the region to the end of LG1 and in the second we inserted it at position 56,376,345 of LG5. The final haploid genome assembly contained 10 pseudochromosomes with a total length of 702.08 Mb and was highly contiguous (scaffold N50 69.33 Mb) and complete (BUSCO complete 94.2%) (Table S1).

S-locus haplotype phasing

To phase haplotypes at the *S*-locus we mapped 10x Genomics linked-read sequencing data from thrum individual SMT-34-3 to the reference genome using Long Ranger (https://github.com/10xGenomics/longranger) which through FreeBayes[88] calls phased variants. Within the region identified as associated with floral morph in GWAS analyses, we verified the presence of two divergent haplotypes differing with regard to an indel, a ~260 kb region (LG10:38,426,515-38,684,012) within a 1 Mb phased block (LG10:37,829,718-38,865,685) (Figure S5C). The haplotype information was used to modify the reference genome sequence in this region to represent the dominant *S*-haplotype prior to further analyses (Figure S5D).

Genome annotation

The identification of coding sequences relied on the usage of curated and custom protein sequences databases, and the assembly of the transcriptomes from different tissues. We used the Uniprot Swiss-Prot database[89] to produce a non-redundant protein sequence database that consisted of 563,972 proteins (downloaded on December of 2020). We also identified and downloaded taxon-specific protein data bases from UniProt and Ensembl, including the proteomes of Rosids (rosids_swissprot.fasta with 22,046 proteins, and Vitis_vinifera.12X.pep.all.fa, with 29,927 proteins), the order Malpighiales (Populus_trichocarpa.Pop_tri_v3.pep.all.fa with 73,012 proteins) and the genus *Linum* (Lusitatissimum_200_v1.0.protein.fa with 43,484 proteins).

We generated *de novo* transcriptome assemblies for annotation using RNA-Seq data from mature flowers, floral buds, leaves and stems from two thrum individuals. After error correction, adapter and quality trimming and removal of ribosomal reads, we generated *de novo* assemblies with Trinity (2.9.1) using default settings.[61,62] Additional assemblies were produced by processing the *in silico* normalized files created through Trinity with Velvet[90] and Oases.[91] The transcriptomes obtained with the pipelines of Trinity and Velvet/Oases were merged using EvidentialGene (http://eugenes.org/EvidentialGene/) to identify the primary transcripts and alternate transcripts/isoforms accepted as valid. In addition, we also used a reference-guided assembly approach. RNA-Seq data and the chromosome-level reference genome were processed through the pipeline TranscriptAssembly (https://github.com/NBISweden/pipelines-nextflow/blob/master/subworkflows/transcript_assembly/README.md) that uses fastp[92] for quality control and preprocessing of raw files, HISAT2[63] for the alignment of RNA reads, and StringTie[64] to assembly transcripts from each tissue independently.

To identify repeats and mask the assembly prior to structural annotation, we created a custom repeat library modelled using RepeatModeler[65] and RepeatMasker.[66] Since protein-coding genes can contain repetitive sequences, the library of repeats was

vetted against the protein set (after transposon removal) to exclude any nucleotide motif present in low-complexity coding sequences. The final identification of repetitive sequences in the genome was conducted using RepeatMasker[66] and RepeatRunner,[67] allowing the identification of highly divergent repeats and protein coding portions.

Gene models were constructed using MAKER2[68] guided by evidence from both aligned transcript sequences and reference proteins, and were then used to train *ab initio* prediction tools (https://github.com/NBISweden/pipelines-nextflow/blob/master/subworkflows/abinitio_training/README.md). Once the *ab initio* tools were trained, a new run of MAKER2 was conducted. Two prediction strategies were conducted and gene models were compared. The approach using only Augustus[69] lead to a higher percentage of genes identified compared with the Augustus + Snap[93] gene build (93.8 vs 91.6% BUSCO v4.0.2 complete, respectively) (Table S1), and the visual inspection of the results showed a clear reduction in false positive prediction. Hence, the Augustus only-based was retained for further analyses.

We manually annotated the *S*-locus region (LG10: 38,426,515-38,684,012) which was of particular interest for this study. Specifically, we first predicted gene models using Augustus[69] with softmasking=0 and then sought additional evidence supporting the existence of the predicted protein coding sequences based on RNA sequencing data from leaves, stems, floral buds and mature flowers. If a certain protein coding sequence was present both in the automatic and manual annotations, we kept only one of them for downstream analyses.

Functional annotation of the translated CDS features was performed using the pipeline FunctionalAnnotation (https://github.com/NBISweden/pipelines-nextflow/blob/master/subworkflows/functional_annotation/README.md). This pipeline uses Blast and InterProScan[94] to retrieve information on protein function from 20 different sources, which was associated to each mRNA feature. To infer gene and protein names, protein sequences were blasted against the Uniprot/Swissprot reference data set, and hits with the best score (E-values < $1x10^{-6}$) were kept. The annotation of tRNA sequences relied on tRNAscan (1.3.1)[95] followed by the removal of features with AED scores equal to 1. Finally, other ncRNAs were predicted using the database Rfam[96] using only highly conserved eukaryotic ncRNA families, and the co-variance models provided by Rfam were then processed using Infernal.[97]

Population genomic analyses

To identify regions with coverage differences between thrum and pin individuals, we analyzed whole-genome sequences (Illumina short reads) from 43 individuals with known morph from two populations (Data S1B). After adapter and quality trimming using 'bbduk' from BBMap/BBTools,[70] reads were mapped to the chromosome-scale assembly with BWA-MEM (v0.7.17) and duplicated reads were removed using MarkDuplicates from Picard tools v2.0.1.[71] Sites in repetitive regions were identified using RepeatMasker[66] and removed from consideration. We estimated per-window genome coverage in 50 kb windows separately for pin and thrum samples using BEDTools.[72] We tested for a difference in median coverage between thrum and pin samples for each 50 kb window using a two-sided Wilcoxon signed-rank test followed by Bonferroni correction.

We estimated polymorphism and divergence using the same short-read data as for coverage analyses. BAM files were processed with SAMtools/bcftools[73] to produce genotype likelihoods from sequence alignments using the function 'mpileup', and the VCF file containing variants (SNPs/INDELs) and invariant sites was created with the function 'call'. The VCF file was further processed to only keep biallelic SNPs and invariant sites, and filtered based on quality (QUAL > 20), coverage (AVG(FMT/DP) > 10 & AVG(FMT/DP) < 50) and data missingness (F_MISSING < 0.2) using the function filter from bcftools. We removed repetitive regions and excluded the thrum-specific hemizygous region (LG10 38,426,515-38,684,012) prior to estimating polymorphism and differentiation/divergence statistics. The resulting repeat-masked VCF file was processed with pixy[74] to estimate F_{ST} (between thrum and pin samples) and π in thrum and pin samples in 5 kb windows. Each population (CL=17 and SMT=26 samples) was analyzed independently, and sample-morph association files were provided to pixy in the argument '–populations'. To identify which windows show significant F_{ST} and difference in π between morphs, we conducted permutation tests in R. Approximate test-statistic distributions for each window were obtained by permuting the list of sample ID – morph association 1000 times, and the observed value was compared with the distributions to calculate the *P*-values, followed by adjustment for multiple testing with the FDR method using the Benjamini-Hochberg approach.

Recombination rate and linkage disequilibrium estimates

We converted the genetic map obtained from Lep-MAP3 in the Rqtl format.[98] After polarization and filtering to retain markers with monotonically increasing genetic distance with physical distance, the genetic map contained 2,471 markers. The recombination rate for each chromosome was calculated with the 'est.recrate' function of the xoi R package (https://github.com/kbroman/xoi). Linkage disequilibrium decay estimates were obtained using biallelic SNPs previously identified using data from both populations SMT and CL (n=43) as described in *Population genomic analyses*. LD decay analyses were performed with ngsLD[75] calculating the LD on SNPs with a minor allele frequency higher than 0.1 and at a maximum distance of 20 kb between each other. These estimates of r^2 decay were obtained for the ~260 kb indel, three down- and three upstream 250 kb windows neighboring the region, and all LG10-linked windows to obtain a mean estimate.

Genome-wide association mapping

We conducted a GWAS to identify loci associated with floral morph. Only biallelic SNPs derived from the VCF file produced in *Population genomic analyses* were included. We tested for an association between morph and SNP in PLINK (v1.90b4.9, https://www.cog-genomics.org/plink/1.9/)[76] using Fisher's exact test, with significance adjustment using the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR).[99]

Comparison of TE content between the distyly *S*-locus and neighboring windows

We estimated and compared the percentage of TEs between *S*-linked and neighboring regions. For this, we relied on the chromosome-level genome assembly and its corresponding annotation of repetitive elements. Microsatellites (STRs and simple repeats) and regions of low complexity were not considered in these analyses, and the fraction of TEs in 25 kb windows were compared between the *S*-locus and the neighboring regions using a Wilcoxon rank sum test. Finally, the annotation of repetitive elements was summarized to characterize the most abundant types of TEs (DNA transposons, LINE, LTR, Helitron or other) for all LG, LG10 and the *S*-locus.

Comparison of π_N/π_S between *S*-linked and neighboring genes

To investigate if π_N/π_S estimates differ between distyly *S*-linked and neighboring genes, we produced a thrum-only VCF file containing biallelic SNPs and invariant sites (SMT=13 samples). We identified 0-fold and 4-fold degenerate sites (here assumed as non-synonymous and synonymous loci respectively) by using the chromosome-level thrum assembly and its corresponding annotation with the script NewAnnotateRef.py (https://github.com/fabbyrob/science/tree/master/pileup_analyzers). Nucleotide diversity was estimated separately for each gene at 0-fold non-synonymous and 4-fold degenerate synonymous sites using pixy. We compared π_N/π_S between *S*-linked (LG10: 38,425,470-38,686,519) and neighboring genes (π_N/π_S: n=4 *S*-linked and on each side of the *S*-locus after keeping only genes with π at 4-fold sites > 0) using a Wilcoxon rank sum test.

Differential expression, gene set enrichment and patterns of expression of *S*-linked genes

We analyzed differential expression between thrum and pin samples of floral buds (n thrum=6, pin=4), leaves (n thrum=6, pin=4), pistils (n thrum=3, pin=3), stamens (n thrum=3, pin=3) and petals (n thrum=3, pin=3) in the R package DESeq2,[77] with logarithmic fold change correction using Approximate Posterior Estimation in the R package apeglm.[100] We controlled FDR using the Benjamini-Hochberg method[99] and considered genes with adjusted $P \leq 0.01$ as significantly differentially expressed. Normalized counts of RNAseq reads mapped to *S*-linked genes were compared between thrum and pin samples with a Wilcoxon rank-sum test. Gene set enrichment analyses were conducted in TopGO 2.46.0[78] using the weighted Fisher exact test. We additionally investigated patterns of expression of *S*-linked genes in mature floral organs by calculating their abundance as Transcript per Million (TPM), which allowed us to determine if transcripts from *S*-linked genes are expressed in pistils, stamens and petals. Gene expression was considered detected in a sample if TPM $\geq$ 0.5 percentile of TPM values for the sample, and genes were determined as expressed in the organ if present in two or more biological replicates. Finally, to better inform our understanding of the involvement of all genes clustered within the *S*-locus in controlling distyly, we leveraged the results of the functional annotation obtained with InterProscan and Blast to classify genes based on PANTHER family membership, and this information was used to manually retrieve GO terms using the AmiGO database.[101–103]

Identification of putative paralogs of *S*-linked genes and divergence estimates

We investigated the genomic distribution of the paralogs of *S*-linked genes by conducting Blast analyses.[84] Using the command 'makeblastdb', we created a custom data base including all protein-coding genes present in the annotation of the *L. tenue* assembly. Next, the longest isoforms of *S*-linked genes sequences were used as queries against the data base using the command 'blastp', and matches with an E-value $\leq$ 0.01 were deemed significant. We kept the first ten best matches for each Blast result, and conducted reciprocal Blast analyses for each entry. Genes with the best matching results in both Blast and reciprocal Blast analyses were determined as the most likely paralogs of *S*-linked genes.

We estimated synonymous divergence between *S*-linked genes and their paralogs by extracting the coding sequence of the longest isoforms and aligned the sequences based on their translated amino acid sequences in webPRANK.[79] *S*-linked genes for which we could not reliably align their sequences with their corresponding putative paralogs were excluded from the analyses. We also excluded gene *LITEG00000052185* as the closest paralog was uncertain with nine highly similar paralogs detected across the genome (Data S1D). The number of synonymous substitutions per synonymous site was estimated in MEGA X[80,81] using the Nei-Gojobori method,[104] excluding all alignment sites with gaps. Standard errors of synonymous divergence estimates were obtained based on 500 bootstrap replicates.

Finally, to investigate if homologs of *S*-linked genes are closely located or scattered throughout the genomes of other two species of the order Malpighiales, we used Phytozome's Blast tool (https://phytozome-next.jgi.doe.gov/blast-search) to find the best matches of *L. tenue* *S*-linked genes in *Manihot esculenta* (v8.1)[38] and *Populus trichocarpa* (v4.1).[39] Finally, we conducted reciprocal Blast using the *L. tenue* protein data base for each of the matches previously obtained for both outgroup species.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses are described in the STAR Methods, methods details section, in the main text and Figure/Table legends. Scripts for all analyses are available as detailed in the data and code availability statement.