Genetic architecture and evolution of the *S* locus supergene in *Primula vulgaris*

Jinhong Li[1,2,*], Jonathan M. Cocker[1,2,*], Jonathan Wright[3], Margaret A. Webster[1,2], Mark McMullan[3], Sarah Dyer[3,†], David Swarbreck[3], Mario Caccamo[3,†], Cock van Oosterhout[4] & Philip M. Gilmartin[1,2,‡]

[1] School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK.

[2] John Innes Centre, Norwich Research Park, Norwich NR4, 7UH, United Kingdom.

[3] The Earlham Institute, Norwich Research Park, Norwich, NR4 7UH, United Kingdom.

[4] School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK.

[*] These authors contributed equally.

[†] Current address: National Institute for Agricultural Botany, Huntingdon Road, Cambridge, CB3 0LE UK.

[‡] Corresponding author

Summary

Darwin's studies on heterostyly in *Primula* described two floral morphs, pin and thrum, with reciprocal anther and stigma heights that promote insect-mediated cross-pollination. This key innovation evolved independently in several angiosperm families. Subsequent studies on heterostyly in *Primula* contributed to the foundation of modern genetic theory and the neo-Darwinian synthesis. The established genetic model for *Primula* heterostyly involves a diallelic *S* locus comprising several genes, with rare recombination events that result in self-fertile homostyle flowers with anthers and stigma at the same height. Here we reveal the *S* locus supergene as a tightly-linked cluster of thrum-specific genes that are absent in pins. We show that thrums are hemizygous not heterozygous for the *S* locus, which suggests that homostyles do not arise by recombination between *S* locus haplotypes as previously proposed. Duplication of a floral homeotic gene 51.7 MYA, followed by its neofunctionalisation, created the current *S* locus assemblage which led to floral heteromorphy in *Primula.* Our findings provide new insights into the structure, function and evolution of this archetypal supergene.

Introduction

Heterostyly evolved independently in at least 28 families of animal-pollinated angiosperms[1]. In the Primulaceae the majority of species[2] produce dimorphic flowers[3], a characteristic inherited as a simple Mendelian trait; alleles are defined as *S* (*Short style*) and *s* (*long style*)[4]. The two floral forms are known as pin and thrum; thrums behave as heterozygous *S*/*s* and pins homozygous *s*/*s*[5]. Classical genetic studies on mutation, linkage and recombination by Bateson[5], Bridges[6], Ernst[7,8], Haldane[9], Darlington[10], and others, established *Primula* as an early genetic model, and led to the definition of a co-adapted linkage group of three genes at the *S* locus, *G* (*Griffel* (style) *length*), *P* (*Pollen*) and *A* (*Antheren* (anther) *position*)[7], which control distinct aspects of heteromorphic flower development; this locus defined the archetypal supergene[11]. Studies of heterostyly in *Primula* contributed significantly to the foundation of modern genetic theory and the neo-Darwinian synthesis. Supergenes have subsequently been shown to control other multi-trait complex phenotypes in plants, animals and fungi[12].

Pin flowers have a long style and low anthers, thrum flowers have a short style and high anthers (Fig.1a). This reciprocal herkogamy promotes insect-mediated cross-pollination between floral morphs, which actively enhances efficiency of reciprocal pollen transfer[13]; such biotic pollination is associated with an elevated speciation rate in angiosperms[14]. Differences in stigma shape, papillae length, pollen size and corolla mouth diameter characterise dimorphic *Primula* flowers[3,15]; a sporophytic self-incompatibility system[16] inhibits intra-morph pollination[1], with different efficacy in different *Primula* species[17]. Self-fertile homostyle flowers (Fig.1a) occasionally occur[13]; although originally considered mutants[8], later studies led to the widely-accepted view that self-fertile homostyles arise by recombination in heterozygous thrums between dominant (*GPA*) and recessive (*gpa*) haplotypes, with associated disruption of coupling between male and female self-incompatibility functions (e.g. *gPA* and *Gpa*)[18,19]. This interpretation defined the order of genes at the *Primula S* locus, and has formed the backdrop to the last 60 years of research into the *S* locus supergene, including models on the evolution of heterostyly[20,21] and population genetic analyses[22-24] in natural homostyle populations[25,26].

More recent studies aimed at identifying *S* locus genes involved examination of flower development[27], analysis of differentially-expressed floral genes[28], characterisation of *S* locus-linked sequences[29,30], molecular genetic analysis of *S*-linked mutant phenotypes[31-33], creation of genetic and physical maps[34,35], assembly of a partial genome sequence[36], and construction of BAC contigs spanning the *S* locus[35]. Despite these extensive investigations, the genetic architecture of the *S* locus has, until now, been an unresolved enigma. Here we compare the *S* haplotype sequences from pin, thrum, long homostyle and short homostyle plants. The *s* haplotype lacks a 278 kb sequence containing five thrum-specific genes present in thrum and homostyles; thrums are therefore hemizygous not heterozygous for the *S* locus. We demonstrate that this 278 kb region is the only thrum-specific genomic region transcribed in flowers, and by genetic and natural population analyses demonstrate complete linkage to the *S* locus; our data indicate that homostyles cannot occur by recombination as proposed. We also provide an estimate of the evolutionary-age of assembly for the *S* locus supergene.

Identification and assembly of the *S* locus

We previously used four *S*-linked probes to assemble two BAC contigs flanking the *S* locus; these were integrated into a genetic map with the gap between contigs predicted to contain some, or all, of the *S* locus genes[35]. A fifth *S*-linked probe[33], GLO^T, also identified a BAC clone which we could not position relative to our *S* locus map[35]. In parallel, we initiated the *de novo* assembly of a *P. vulgaris* reference genome using a self-fertile homozygous long homostyle (S^{LH1}/S^{LH1}) from the Somerset population identified by Crosby[25]. We also generated genome sequence data from individual pin (*s*/*s*) and thrum (*S*/*s*) plants, pools of their pin and thrum progeny, and a short homostyle (S^{SH1}/*s*)[35] (Supplementary Table 1a). Fig. 1a shows relevant floral phenotypes and genotypes.

Using the GLO^T BAC (BAC70F11) we searched a long homostyle genome assembly (Supplementary Table 1b) to identify and link two genome sequence contigs. This step initiated the assembly of a contiguous 455,881 bp sequence encompassing the entire S^{LH1} haplotype from this highly homozygous inbred line (Supplementary Fig. 1a). This assembly contains a 278,470 bp sequence which is absent from pins and flanked by a ~3 kb tandem repeat that is present only as a single copy in the *s* haplotype

(Fig. 1); each repeat contains a *Cyclin-like F box* (*CFB*) gene. We therefore focused on this region as the presumptive *S* locus. Sequences flanking the S^{LH1} 278,470 bp region on the left (75,084 bp), and right (96,327 bp) share extensive similarity to the *s* haplotype (Fig. 1b, Supplementary Fig. 1b).

Next, we designed PCR primers for left- and right-border regions of S^{LH1}, and separate *s* haplotype-specific primers (Supplementary Table 2). Analyses with pin, thrum, long and short homostyle genomic DNA confirmed pin as *s/s*, and long homostyle as homozygous S^{LH1}/S^{LH1} (Fig. 1c). Supported by sequence alignment (Supplementary Sequence Analysis 1, 2), these data also show that thrum and the short homostyle share the same left- and right-border sequences as the long homostyle, and that they are both heterozygous for the *s* and S^{LHI} flanking markers (Fig. 1c). The established model defines homostyles as recombinants between *S* and *s* haplotypes; if this is the case, long and short homostyles should possess reciprocal combinations of *s* and *S* haplotype left and right border sequences, but they do not (Fig. 1c).

Comparative analysis of *S* haplotypes

We then focused on the 278 kb region from S^{LH1} that is absent from the *s* haplotype. This region contains five predicted gene models, CCM^T, GLO^T, CYP^T, PUM^T and KFB^T which where manually curated and are supported by RNA-Seq data as thrum-specific in expression; four other models identify transposon sequences which were discounted as functional *S* locus genes and excluded from further analysis (Supplementary Fig. 2a, b). CCM^T (*Conserved Cysteine Motif*) encodes a protein with a C-terminal domain that is conserved in monocots and dicots (Supplementary Sequence Analysis 3); proteins containing this novel domain are rich in either proline or negatively charged amino acids. One of these, PIG93 from *Petunia x hybrida*, is a partner of PSK8, a protein involved in brassinolide signalling[37]. A second *CCM*-like gene with 90% sequence similarity is found in both pin and thrum genomes. GLO^T was originally defined as a thrum-specific allele[33] of *P. vulgaris GLO*, a floral homeotic gene responsible for the *S* locus-linked mutant phenotype *Hose in Hose*[32,35]. These data show *GLO* and GLO^T as distinct loci; the encoded proteins share 82% sequence identity but the *Hose in Hose* mutation, in which *GLO* is dominantly up-regulated, does not affect heterostyly[32]. CYP^T encodes a cytochrome P450 similar to *Arabidopsis* CYP72B1, a brassinolide 26-hydroxylase[38]. CYP^T is one of four *CYP72* class

genes in the *P. vulgaris* thrum genome, the other three are present in both pin and thrum; the closest encodes a protein with 65% sequence identity to CYP^T. PUM^T encodes a Pumilio-like[39] RNA-binding protein, and KFB^T encodes a protein with similarity to the *Arabidopsis* Kiss-Me-Deadly Kelch repeat F Box protein involved in regulating cytokinin activity[40]; Both PUM^T and KFB^T are unique to the 278 kb region with no homologues found in our pin genome sequence. The tandemly duplicated sequences flanking the *S* locus contain *Cyclin-like F Box* genes, CFB^{TL} and CFB^{TR} (Supplementary Fig. 2a); in pin, a single CFB^P exists. Gene model predictions also identified seven genes in the 75 kb to the left of CFB^{TL}, and eight genes in the 96 kb to the right of CFB^{TR}, designated S Flanking Gene Left (SFG^L) and Right (SFG^R) (Supplementary Fig. 2a); these genes are present in both pin and thrum.

To further investigate *S* haplotype differences, we aligned thrum and the short homostyle genome contigs to the 455 kb S^{LH1} region (Supplementary Fig. 3a, b); although *S* and S^{SH1} assemblies are not contiguous, they show homology across the 278 kb region. We also aligned genome sequence reads from pin, thrum, long[25] and short[35] homostyle to the S^{LH1} assembly and plotted sequence read depth across the 455 kb region (Fig. 2a). Sequences flanking the 278 kb insertion show a read depth of ~60 in all four genomes. However, between CFB^{TL} and CFB^{TR} we see differences; the long homostyle (S^{LH1}/S^{LH1}) behaves as a homozygote, but both thrum (*S*/*s*) and the short homostyle (S^{SH1}/*s*) have half this read depth, and pin (*s*/*s*) lacks this region (Fig. 2a); they behave genetically as heterozygotes but our data show they are hemizygous for a region that is absent in pin. Alignment of all four genomes over this region further show that thrum, long and short homostyles share the same boundary regions (Figs. 1c, 2a, Supplementary Figs. 2, 3). This detail, coupled to the presence of all five *S* locus genes in thrum, long and short homostyle (Fig. 2a) show that these homostyles did not arise by recombination as proposed[18,19], and that *S*, S^{LH1} and S^{SH1} haplotypes all reside within an equivalent region that is absent from pins.

To determine whether the 278 kb region is the only thrum-specific region in the genome, we searched for additional thrum-specific genome sequences encoding genes. In two parallel analyses we identified transcripts that were only expressed in thrums, and also mapped pin genomic sequencing reads to a thrum genome assembly. We then examined the depth and breadth of pin genome reads mapped to the

thrum genome in the regions defined by the thrum-specific transcripts; *k*-means clustering analysis resolved the transcribed regions into two clusters (Fig. 2b); deep and broad read coverage defined presence of the region in both pin and thrum genomes, low read depth or low coverage identified a region as thrum-specific, with pin sequence alignments representing erroneously mapped sequence reads (Supplementary Table 3). Nine thrum-specific regions were thus identified; these define four of the five thrum-specific genes from the 278 kb region; GLO^T, CYP^T, PUM^T and KFB^T (Fig. 2b). CCM^T is expressed at a low level (see below) and is the only gene from the cluster not represented. Identification of three contigs for GLO^T and KFB^T, and two for CYP^T, is due to the use of a non-scaffolded thrum genome assembly (Supplementary Table 1b), and the length of GLO^T and CYP^T (see below). We conclude that there are no other flower-expressed genes unique to thrums and that the 278 kb sequence is the only thrum-specific genomic region. Significantly, these data show that the thrum *S* haplotype does not contain any additional genes compared to the long homostyle S^{LH1} haplotype. These analyses revealed 391 gene models that are uniquely expressed in thrums, and 270 gene models that are uniquely expressed in pins, but present in both pin and thrum genomes; these are candidates for direct or indirect targets of the *S* locus genes that control pin and thrum flower development.

Next we investigated whether the sequences flanking the thrum-specific 278 kb region could also be part of the *S* locus that contained pin- and thrum-specific alleles of genes involved in the control of heterostyly. If sequences flanking the thrum-specific region contain genes that also contribute to S locus function, restriction of recombination between pin and thrum alleles would be required to maintain integrity and functionality of the locus. However, if these flanking regions are freely recombining this would indicate that the thrum-specific region alone contains the entire *S* locus gene cluster. We therefore undertook a recombination analysis investigating the pattern of nucleotide polymorphisms (SNPs) across the flanking sequences, comparing the alleles present in a pin and a thrum plant. These data (Supplementary Fig. 4) reveal that sequences flanking the thrum-specific region contain blocks of significantly reduced polymorphism, which is consistent with recent recombination events. The sequences flanking the thrum-specific 278 kb region thus seem to be homogenised by

recombination between pin and thrum alleles and suggest that they are not involved in the control of the heterostyly phenotype.

Linkage of *GLO^T* and the *S* locus

GLO^T was initially identified as thrum-specific in a small segregating population[33]. To demonstrate unequivocal linkage of *GLO^T*, and therefore the *S^{LH1}* assembly, to the *S* locus, we revisited a three-point-cross with 2075 progeny[35] used previously to place *Oakleaf*[31] (<1.7 cM) and *Hose in Hose*[41] (<1.6 cM) on either side of the *S* locus[35] (Fig. 3a, Supplementary Table 4). This cross also yielded the short homostyle *Hose in Hose* plant[35] used here (Fig. 1a). We analysed DNA from pin and thrum parents, pools of pin and thrum non-recombinant progeny, and two double-recombinant (*Oakleaf-S-Hose in Hose*) thrum progeny by PCR analysis with *GLO^T* and *GLO* specific primers (Fig. 3, Supplementary Table 2); *GLO* is present in both pin and thrum[32].

The parent plants show the original linkage profiles (Fig. 3b); we found no linkage disruption between *GLO^T* and thrum phenotype using pools of 100 non-recombinant progeny. Furthermore, double-recombinants show that recombination between *Oakleaf* and *S*, or *S* and *Hose in Hose*, does not disrupt linkage between *GLO^T* and thrum phenotype (Fig. 3b). These data place the 455 kb assembly between *Oakleaf* and *Hose in Hose*, within the *S* locus BAC assembly[35] (Fig. 3a); previous studies did not identify any BACs that link the 455kb region to BAC contigs S-left and S-right[35]. To increase mapping resolution, we analysed natural populations of *P. vulgaris* and *P. veris*. Pooled genomic DNA from 200 pin plants of each species was analysed by PCR using *GLO^T* and *GLO* specific primers (Fig. 3c). A single thrum plant was used as control because loss of a dominant marker in one individual would not be detected in a thrum pool. In total, 500 pin plants were analysed (Figs. 3b,c), none showed recombination; these data demonstrate that *GLO^T* and the surrounding region is in tight thrum-specific linkage (<0.2 cM) with the *S* locus in both *P. vulgaris* and *P. veris* (Fig. 3a).

S locus gene expression and function

Having shown that homostyles did not occur by recombination we sought to determine their molecular basis by comparing gene expression across the four haplotypes. Expression analysis of genes within, and flanking, the *S* locus was undertaken by mapping four replicate RNA-Seq datasets from pin and thrum flowers to the S^{LH1} assembly (Fig. 4a and Supplementary Table 5). GLO^T, CYP^T, PUM^T, KFB^T and CCM^T all show thrum-specific expresion (Figs. 2b, 4a). CFB^{TL} is expressed at a low level in both pin and thrum flowers; CFB^{TR} is not expressed. Genes flanking the *S* locus are expressed in both pin and thrum flowers, except SFG^R6 which has low expression in thrum and is not detected in pin; SFG^L1 is expressed at a low level in both pin and thrum (Fig. 4a, Supplementary Table 5). These analyses reveal the 278 kb region as an island of thrum-specific gene expression.

Gene model predictions (Supplementary Fig. 2a) for CCM^T, GLO^T, CYP^T, PUM^T, KFB^T, and three *CFB* alleles were confirmed by alignment to RNA-Seq data to define intron-exon boundaries. Two *S* locus genes are surprisingly large, GLO^T spans 25 kb with two introns over 10 kb; CYP^T spans 68 kb with 10, 20 and 30 kb introns (Supplementary Fig. 2c). Interestingly, the GLO^T S^{SH1} allele contains a 2.5 kb retro-transposon in exon 2 which disrupts and severely truncates the encoded protein; mutation of GLO^T in the short homostyle is associated with loss of anther elevation; style length and pollen size are unaffected. The long homostyle (S^{LH1}) CYP^T allele has a single base insertion in exon 3 that introduces a disruptive premature stop codon, and is associated with loss of style length suppression; anther height and pollen size are unaffected. We also sequenced an independent long homostyle (S^{LH2}) from the Chiltern Hills[25] which represents a second CYP^T mutant allele with a G-C transversion in exon 2 that results in an Asp126His substitution. CFB^{TR} has an 11 bp deletion compared to CFB^{TL} and CFB^P that introduces a premature stop codon. The architecture of *S* and s haplotypes is summarised in Fig. 4b. Comparison of alleles is presented in Supplementary Sequence Analysis 3.

Date of the GLO^T duplication

The first indication that GLO^T was a discrete locus from *GLO* came when we identified distinct BAC clones for each gene, together with insight from other studies of B function MADS box genes which suggested duplication could underpin diversification of novel floral morphologies[42]. The short homostyle GLO^T mutation is not complemented by *GLO*, or by ectopic expression of *GLO*; the short

homostyle is in the *Hose in Hose* background[32]. The recent report of a partial *P. veris* genome sequence[36] noted the duplication of *GLO* and referred to the genes as *GLO1* and *GLO2* but could not show linkage of GLO^T (*GLO2*) to the *S* locus. Demonstration that these genes represent distinct loci with GLO^T at the *S* locus provides the opportunity to date the duplication event associated with assembly of the *S* locus supergene. To determine the age of duplication we isolated *GLO* and GLO^T sequences from six *Primula* species, and used these with sequences from other species to conduct a Bayesian relaxed-clock phylogenetic analysis with a combination of secondary calibrations (Fig. 5, Supplementary Tables 6a,b). The index of substitution saturation value[43] for *GLO* and GLO^T sequences (0.1187) was significantly lower than the Iss critical value (0.7318, $p<0.0001$) indicating low saturation between these sequences. These analyses yielded a mean (5-95% Highest Posterior Density) age estimate of 51.7 (33.1-72.1) MYA for the duplication leading to the divergence of *GLO* and GLO^T lineages.

The duplication and neofunctionalisation of GLO^T represents a landmark evolutionary event at the *S* locus, and precedes estimates for the Primula-Androsace divergence; estimates for this node are 32 (20-51) MYA[44], and 44 (33-54) MYA[44] with fossil priors being set with a log normal distribution, and 40 (30-51) MYA with fossils modelled as exponential priors[45]. The Androsace were predicted to be the first taxon within the Primulaceae to exhibit heteromorphy[46], our data indicate that the *GLO*-GLO^T duplication predates this divergence, which implies heterostyly evolved following a single duplication event in the Primulaceae. Two models have been proposed for the evolution of *Primula* heterostyly, the first postulates a long homostyle[21], and the other an approach herkogamous pin-form flower[20], as the original floral form. The duplication and neofunctonalization of GLO^T would be consistent with both models if this was the first gene at the (Fig. 5). The *S* locus sequence, structure and timing of the GLO^T duplication, and analysis of other genes at the *S* locus genes, will inform further evolutionary genetic analysis of primary and secondary homostyly in *Primula* and help to determine the sequence of events leading to the establishment of the *S* locus gene cluster.

Conclusions

We show that the *S* locus supergene is a tightly-linked cluster of five thrum-specific genes, spanning a 278 kb sequence that is absent in pins (Fig. 2a), this finding defines the basis for Bateson and Gregory's *S* haplotype dominance[5]. The annotation *S*/*s* and *s*/*s* for thrum and pin could be represented by *S*/- and -/-, but we suggest retention of the traditional nomenclature with recognition of *s* as a null haplotype. Floral heteromorphy in *Primula* has evolved after duplication of a floral homeotic gene 51.7 MYA, followed by its neofunctionalisation, creating the current *S* locus assemblage. This insight has profound implications for our understanding of a key evolutionary innovation of flowering plants. The molecular basis of the *Primula S* locus supergene appears to be different from those proposed for the control of butterfly mimicry, and avian and insect social behaviour[47-49]. It is also unlike the mating-type locus in ascomycete fungi which comprises two distinct idiomorphs[50]. Ernst originally proposed that *Primula* homostyles arose by mutation[8], he was correct, and mutations in CYP^T and GLO^T homostyle alleles earmark these genes as candidates for the style length suppression (*G*), and anther elevation (*A*), functions[7] respectively. Darwin suggested the primary function of heterostyly evolved to promote out-crossing[13], generating novel variation that is the substrate of natural selection. The parallel evolution of heterostyly in diverse angiosperm families[1] has exploited insect-mediated pollination, which in turn is associated with an accelerated rate of speciation in angiosperms[14]. Deciphering the genetic architecture of the *Primula S* locus as the first heterostyly supergene provides a blueprint for the comparative evolutionary genetic analysis of this key adaptation in other angiosperm families, as well as for the molecular characterisation of other pollination syndromes underpinning both biodiversity and food security.

Methods

Plant Material

The long homostyle plant (S^{LH1}) used for DNA sequencing was a homozygote derived from a population originally described by Crosby in 1940[25] at Wyke Champflower, Somerset, UK, which had undergone several generations of selfing to generate a homozygous line which greatly facilitated assembly of the

genome sequence. The independent long homostyle population in the Chiltern Hills discovered by Crosby in 1944[26] provided our second long homostyle (S^{LH2}) from Hawridge, Buckinghamshire, UK. Pin and thrum *P. vulgaris* were grown from seed (http://www.wildseed.co.uk) as described previously[27]. Pin and thrum plants selected for genome sequencing were crossed to generate an F1 population. The short homostyle was originally identified in a mapping population of *P. vulgaris* plants[35] *P. veris* for genome sequencing were grown from seed collected at the Durham University Mountjoy site. *P. elatior* leaf material was collected from Bull's Wood (http://www.suffolkwildlifetrust.org) with permission of Suffolk Wildlife Trust. *P. farinosa* were obtained from Kevock Garden Plants (http://www.kevockgarden.co.uk/), *P. vialii* and *P. denticulata* were from the laboratory collection. The population of *P. veris* used for *S* locus linkage analysis was sampled from Lolly Moor (http://www.norfolkwildlifetrust.org.uk) with permission of Norfolk Wildlife Trust. *P. vulgaris* used for *S* locus linkage analysis was sampled with permission of Norfolk County Council from the B1135 roadside verge between Ketteringham and Browick, near Wymondham, Norfolk. Plants from the three-point mapping cross have been described previously[35].

Preparation of sequencing libraries

DNA and RNA preparation was as described previously[30,33]. All genomic DNA and RNA-Seq libraries for sequencing were prepared at The Genome Analysis Centre using standard Illumina protocols. Genomic paired-end libraries: An Illumina TruSeq library was prepared using a protocol optimized for 1μg of input genomic DNA (Illumina 15026486 Rev. C). Mate-pair libraries: The protocol was optimized for 4-10 μg of high molecular weight DNA; following fragmentation, samples were size fractionated to enable generation of mate-pair libraries of 5, 7 and 9 kb (Illumina 15035209 Rev. D). RNA paired-end libraries: Libraries were constructed using the Illumina TruSeq RNA protocol (Illumina 15026495 Rev.B). The *S* locus region was assembled as outlined in Supplementary Methods and Supplementary Fig. 1. The assembly was validated by comparison of independent assemblies from different Illumina paired-end and and mate-paired sequencing libraries of thrum, long and short homostyle individuals, gaps between contigs and regions of Ns were resolved by PCR amplification and Sanger sequencing of the products.

Genomic DNA PCR analysis

Genomic DNA was isolated as described previously[30]. Primers are shown in Supplementary Table 2. PCR was performed in 50 µl reactions with 100-200 ng genomic DNA using Promega GoTaq G2 Green Master mix (cat# M7822). *Pfu*Turbo DNA polymerase (Agilent Technologies, cat#600250) was used when PCR products were to be sequenced. Amplification conditions for primers TRB-F and -R and CFB-F and –R: 95°C 3 min; 95°C 30 sec, 61°C 30 sec, 72°C 1 min x35 cycles; 72°C 5 min. Amplification cycles for primers PRB-F and -R, TLB-F and -R, GLOT-F and -R: 95°C 3 min; 95°C 30 sec, 58°C 30 sec, 72°C 1 min, x35 cycles : 72°C 5 min. Amplification cycles for primers GLO-F and –R: 95°C 3 min; 95°C 30 sec, 65°C 30 sec, 72°C 1 min, x35 cycles; 72°C 5 min.

Bioinformatic and evolutionary analyses

Bioinformatic and evolutionary analyses are described in Supplementary Methods due to space constraints.

Data availability

Sequence data are available through Genbank accessions KT257663-KT257681, under Bioproject PRJEB9683 http://www.ebi.ac.uk/ena/data/view/PRJEB9683, and http://opendata.earlham.ac.uk/primula/

Correspondence and requests for materials should be addressed to P.M.G. (p.gilmartin@uea.ac.uk).

Acknowledgements We thank Martin Lappage, Mike Hughes and Pam Wells for horticultural support; colleagues at TGAC for Illumina sequencing; Anil Thanki for TGAC Browser support; Olivia Kent for *P. elatior GLO* and *GLO*^{T} sequences; Norfolk Wildlife Trust, Suffolk Wildlife Trust and Norfolk County Council for permission to sample *P. veris*, *P. elatior* and *P. vulgaris* respectively; Matt Gage, Brendan Davies and Dianna Bowles for comments on the manuscript; Wenjia Wang for advice on *k*-means analysis; BBSRC for funding via grant BB/H019278/2, and prior awards G11027 and P11021; The Gatsby Foundation for early stage funding; University of Leeds, Durham University and University

of East Anglia for support over several years of the project to PMG. PMG's lab is hosted at the John Innes Centre under the UEA-JIC Norwich Research Park collaboration.

Author Contributions J.L. contributed to project design, performed all molecular analyses, generated the *S* locus assembly, manually annotated *S* locus gene structures, and undertook data analysis. J.M.C. contributed to bioinformatic analyses, including automated annotation of the *S* locus region, undertook *in silico* gene expression and *k*-means clustering analyses, assembled genome sequences and library scaffolds, generated the molecular phylogeny, undertook recombination analysis of the *S* locus flanking regions and contributed to project design. J.W. assembled genome sequences and library scaffolds, contributed to genome annotation and generated the automated gene model predictions across the *S* locus, aligned sequence reads to the *S* locus assembly and contributed to project design. M.A.W. contributed the inbred long homostyle line, other genetic resources and classical genetics, identified the short homostyle mutant, and generated the three-point cross used to demonstrate linkage. M.M. and C.v.O. contributed to the molecular phylogeny construction, evolutionary data analysis and recombination analysis. S.A., D.S. and M.C. contributed to the genome sequencing strategy, assembly and annotation that underpins this project. P.M.G. conceived, designed and directed the project, contributed to data analysis, prepared the figures and drafted the manuscript, with revision input from C.v.O.; all authors contributed to editing the manuscript.

Figure Legends

Figure 1 *P. vulgaris* floral phenotypes and genotypes.

a, Heterostyly phenotypes and genotypes with respect to *s*, *S*, S^{LH1} and S^{SH1} haplotypes; the short homostyle carries the *Hose in Hose* mutation[32]. Anther (A) and stigma (S) **b,** Comparison of *s*, *S*, S^{LH1} and S^{SH1} haplotypes, sequence present in the S^{LH1}, S^{SH1} and *S*, but absent from *s* (red); duplicated flanking sequence present as a single copy in the *s* haplotype (yellow); flanking sequences common to all haplotypes left (blue), and right (green); not to scale. The GLO^{T} BAC location is shown. PCR primers used for amplification of flanking regions (→←) (Supplementary Table 2). **c,** PCR analysis of genomic DNA from pin (P), thrum (T), long (LH) and short homostyle (SH) plants (shown in **a**), using primers

(as in **b)**, that distinguish left (LB) and right (RB) borders of S^{LH1}, S^{SH1} and *S* haplotypes from the *s* haplotype; sizes as indicated. See also Supplementary Sequence Analysis 1.

Figure 2 Organisation of *S* locus haplotypes. a, The S^{LH1} haplotype showing *S* locus genes (red) duplicated flanking *CFB* loci (yellow), left and right flanking genes SFG^{L} 1-7 and SFG^{R} 1-8 (black) (see also Supplementary Fig. 2a). Illumina sequence read depth from pin (black), thrum (blue), short (red) and long homostyle (yellow) genomic DNA. **b,** Scatter plot analysis showing breadth of read coverage (%) and $\log_{10}$ depth of read coverage for pin progeny pool genome sequence reads mapped to thrum genome contigs encoding genes with thrum-specific expression. Two clusters, defined by *k*-means analysis are shown; transcript regions in contigs where reads map with low depth and breadth (red): 1, KFB^{T}; 2, GLO^{T}; 3, CYP^{T}; 4, GLO^{T}; 5, KFB^{T}; 7, PUM^{T}; 8, GLO^{T}; 8, CYP^{T}; 9, KFB^{T} (Supplementary Table 3). Transcript regions in contigs to which pin progeny pool genome sequence reads map with high depth and breadth (blue)**.**

Figure 3 Linkage of the *S* haplotype to the thrum phenotype. a, Map of the *S* locus region, distances in cM[35]. The *S* haplotype 278 kb region (red) between duplicated ~3 kb *CFB* loci (yellow) is shown relative to sequenced BAC contigs[35] with 75 kb left flanking (blue) and 96 kb right flanking (green) sequences (Fig. 1b). **b**, PCR analysis of GLO^{T} linkage using Pin (P) and thrum (T) plants and 100 pooled non-recombinant (no x-over) progeny from a three-point cross[35] (Supplementary Table 4) compared to non-*S* locus *GLO* as control. Two thrum plants (T1 and T2) from double-recombination events (xx-over)[35], *Oakleaf* to *S* and *S* to *Hose in Hose*, are also shown. **c**, PCR analyses of GLO^{T} linkage to thrum in natural populations using 200 pooled pin (P) *P. vulgaris*, and 200 pooled pin *P. veris* plants, compared to individual thrum (T) plants, with *GLO* as control; sizes in kb.

Figure 4 Expression and genomic organisation of *S* locus genes. a, Gene expression from the *S* (red) and *s* (blue) haplotypes using pin and thrum RNA-Seq data represented as Log_{10} of the number of fragments per kb of transcripts per million fragments mapped (FPKM) +1; gene models as defined in Supplementary Fig. 2a. **b,** Pictorial representation of genes within the *s*, *S*, S^{LH1} and S^{SH1} haplotypes shown alongside stylized flowers. The base insertion in S^{LH1} CYP^{T} (red +), G-C transversion in S^{LH2}

(red ı) and transposon insertion in S^{SH1} GLO^{T} (red Δ) are indicated. Sequences of mutant alleles are compared in Supplementary Sequence Analysis 3.

Figure 5 Phylogenetic analysis and the date of duplication of GLO^{T} from *GLO*. Phylogram of B function MADS box genes from *Antirrhinum* (*Am.*), *Petunia* (*Pe.*), *Arabidopsis* (*A.*) and *Primula* (*P.*) species presented against an evolutionary time scale in millions of years (MYA); see Supplementary Tables 6a,b. Thick blue lines represent the time scale range estimates at divergence branch points; the thick red line defines the same for duplication of GLO^{T} from *GLO.*

References

1 Barrett, S. C. H. The evolution of plant sexual diversity. *Nature Reviews, Genetics* **3**, 274-284 (2002).

2 Richards, A. J. *Primula 2nd edition*. (2002).

3 Darwin, C. R. On the two forms or dimorphic condition in the species of *Primula*, and on their remarkable sexual relations. *Journal of the Proceedings of the Linnean Society, Botany* **6**, 77-96 (1862).

4 Gregory, R. P., De Winton, D. & Bateson, M. A. Genetics of *Primula sinensis*. *Journal of Genetics* **13**, 219-253 (1923).

5 Bateson, W. & Gregory, R. P. On the inheritance of heterostylism in *Primula*. *Proceedings of the Royal Society of London B Series* **76**, 581-586 (1905).

6 Bridges, C. B. The chromosome hypothesis of linkage applied to cases in sweetpeas and *Primula*. *American Naturalist* **48**, 524-534 (1914).

7 Ernst, A. Weitere untersuchungen zur Phänanalyse zum Fertilitätsproblem und zur Genetik heterostyler Primeln. II. *Primula hortensis*. *Archive der Julius Klaus Stiftung für Vererbungsforschung Sozialanthropologie und Rassenhygiene* **11**, 1-280 (1936).

8 Ernst, A. Heterostylie-Forschung Versuche zur genetischen analyse eines organisations und 'Anpassungs ' merkmales. *Zeitschrift für Induktive Abstammungs und Vererbungslehre* **71**, 156-230 (1936).

9 De Winton, D. & Haldane, J. B. S. The genetics of *Primula sinensis*. III.Linkage in the diploid. *Journal of Genetics* **31**, 67-100 (1935).

10 Darlington, C. D. Meiosis in diploid and tetraploid *Primula sinensis*. *Journal of genetics* **24**, 65-95 (1931).

11 Mather, K. The genetical architecture of heterostyly in *Primula sinensis*. *Evolution* **4**, 340-352 (1950).

12 Schwander, T., Libbrecht, R. & Keller, L. Supergenes and Complex Phenotypes. *Current Biology* **24**, R288-R294 (2014).

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

14 Dodd, M. E., Silvertown, J. & Chase, M. W. Phylogenetic analysis of trait evolution and species diversity variation among angiosperm families. *Evolution* **53**, 732-744 (1999).

15 Webster, M. A. & Gilmartin, P. M. Analysis of late stage flower development in *Primula vulgaris* reveals novel differences in cell morphology and temporal aspects of floral heteromorphy. *New Phytologist* **171**, 591-603 (2006).

16 Shivanna, K. R., Heslop-Harrison, J. & Heslop-Harrison, Y. Heterostyly in *Primula* .2. Sites of pollen inhibition, and effects of pistil constituents on compatible and incompatible pollen tube growth. *Protoplasma* **107**, 319-337 (1981).
17 Richards, A. J. & Ibrahim, H. B. The Breeding System in *Primula veris* L .2. Pollen-Tube Growth and Seed-Set. *New Phytologist* **90**, 305-314 (1982).
18 Lewis, D. Comparative incompatibility in angiosperms and fungi *Advances in Genetics Incorporating Molecular Genetic Medicine* **6**, 235-285 (1954).
19 Dowrick, V. P. J. Heterostyly and homostyly in *Primula obconica*. *Heredity* **10**, 219-236 (1956).
20 Lloyd, D. G. & Webb, C. J. in *Evolution and Function of Heterostyly* (ed S.C.H. Barrett) 151-175 (Springer Verlag, 1992).
21 Charlesworth, D. & Charlesworth, B. Model for the evolution of distyly. *American Naturalist* **114**, 467-498 (1979).
22 Bodmer, W. F. The genetics of homostyly in populations of *Primula vulgaris*. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences* **242**, 517-549 (1960).
23 Fisher, R. A. A Theoretical system of selection for homostyle *Primula*. *Sankhya* **9**, 325-342 (1949).
24 Piper, J. G., Charlesworth, B. & Charlesworth, D. A high-rate of self-fertilization and increased seed fertility of homostyle primroses. *Nature* **310**, 50-51 (1984).
25 Crosby, J. L. High proportions of homostyle plants in populations of *Primula vulgaris*. *Nature* **145**, 672-673 (1940).
26 Crosby, J. L. Selection of an unfavourable gene complex. *Evolutionary Ecology Research* **3**, 212-230 (1949).
27 Webster, M. A. & Gilmartin, P. M. A comparison of early floral ontogeny in wild-type and floral homeotic mutant phenotypes of *Primula*. *Planta* **216**, 903-917 (2003).
28 McCubbin, A. G., Lee, C. & Hetrick, A. Identification of genes showing differential expression between morphs in developing flowers of *Primula vulgaris*. *Sexual Plant Reproduction* **19**, 63-72 (2006).
29 Li, J., Webster, M. A., Furuya, M. & Gilmartin, P. M. Identification and characterization of pin and thrum alleles of two genes that co-segregate with the *Primula S* locus. *Plant Journal* **51**, 18-31 (2007).
30 Manfield, I. W. *et al.* Molecular characterization of DNA sequences from the *Primula vulgaris S* locus. *Journal of Experimental Botany* **56**, 1177-1188 (2005).
31 Cocker, J. *et al. Oakleaf:* an *S* locus-linked mutation of *Primula vulgaris* that affects leaf and flower development *New Phytologist 10.1111/nph.13370* (2015).
32 Li, J. *et al. Hose in Hose,* an *S* locus-linked mutant of *Primula vulgaris* is caused by an unstable mutation at the *Globosa* locus. *PNAS* **107**, 5664-5668 (2010).
33 Li, J. *et al.* The *S* locus-linked *Primula* homeotic mutant *sepaloid* shows characteristics of a B-function mutant but does not result from mutation in a B-function gene. *Plant Journal* **56**, 1-12 (2008).
34 Yoshida, Y. *et al.* QTL analysis of heterostyly in *Primula sieboldii* and its application for morph identification in wild populations. *Annals of Botany* **108**, 133-142 (2011).
35 Li, J. *et al.* Integration of genetic and physical maps of the *Primula vulgaris S* locus and localization by chromosome *in situ* hybridisation *New Phytologist 10.1111/nph.13373* (2015).
36 Nowak, M. D. *et al.* The draft genome of *Primula veris* yields insight into the molecular basis of heterostyly. *Genome Biology* **16**, 16 (2015).
37 Verhoef, N. *et al.* Brassinosteroid biosynthesis and signalling in Petunia hybrida. *Journal of Experimental Botany* **64**, 2435-2448 (2013).
38 Turk, E. M. *et al.* CYP72B1 inactivates brassinosteroid hormones: An intersection between photomorphogenesis and plant steroid signal transduction. *Plant Physiology* **133**, 1643-1653 (2003).
39 Abbasi, N., Park, Y.-I. & Choi, S.-B. Pumilio Puf domain RNA-binding proteins in *Arabidopsis*. *Plant signaling & behavior* **6**, 364-368 (2011).

40 Kim, H. J., Chiang, Y.-H., Kieber, J. J. & Schaller, G. E. SCFKMD controls cytokinin signaling by regulating the degradation of type-B response regulators. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 10028-10033 (2013).
41 Webster, M. A. & Grant, C. J. The inheritance of calyx morph variants in *Primula vulgaris* (Huds). *Heredity* **64**, 121-124 (1990).
42 Viaene, T. *et al.* Pistillata-Duplications as a Mode for Floral Diversification in (Basal) Asterids. *Molecular Biology and Evolution* **26**, 2627-2645 (2009).
43 Xia, X. DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution. *Molecular Biology and Evolution* **30**, 1720-1728 (2013).
44 Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist*, 10.1111/nph.13264 (2015).
45 Bell, C. D., Soltis, D. E. & Soltis, P. S. The age and diversification of the angiosperms re-revisited. *American Journal of Botany* **97**, 1296-1303 (2010).
46 Mast, A. R. *et al.* Phylogenetic relationships in *Primula* L. and related genera (Primulaceae) based on noncoding chloroplast DNA. *International Journal of Plant Sciences* **162**, 1381-1400 (2001).
47 Joron, M. *et al.* Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203-206 (2011).
48 Thomas, J. W. *et al.* The chromosomal polymorphism linked to variation in social behavior in the white-throated sparrow (Zonotrichia albicollis) is a complex rearragement and suppressor of recombination. *Genetics* **179**, 1455-1468 (2008).
49 Wang, J. *et al.* A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493**, 664-668 (2013).
50 Turgeon, B. G. & Yoder, O. C. Proposed Nomenclature for Mating Type Genes of Filamentous Ascomycetes. *Fungal Genetics and Biology* **31**, 1-5 (2000).