




Comparative Genomics Elucidates the Origin of a Supergene Controlling Floral Heteromorphism

Giacomo Potente ^{*,1,2,3} Étienne Léveillé-Bourret,^{1,4} Narjes Yousefi,¹ Rimjhim Roy Choudhury,¹ Barbara Keller,¹ Seydina Issa Diop,^{1,2,3} Daniël Duijsings,² Walter Pirovano,² Michael Lenhard ⁵, Péter Szövényi,^{*,1,3} and Elena Conti ^{*,1,3}

¹Department of Systematic and Evolutionary Botany, University of Zurich, Zurich, Switzerland

²BaseClear BV, Leiden, The Netherlands

³Zurich-Basel Plant Science Center, Zurich, Switzerland

⁴Département de Sciences Biologiques, Institut de Recherche en Biologie Végétale, Université de Montréal, Montréal, Canada

⁵Institute for Biochemistry and Biology, University of Potsdam, Potsdam-Golm, Germany

*Corresponding authors: E-mails: giacomo.potente@uzh.ch; peter.szovenyi@systbot.uzh.ch; elena.conti@systbot.uzh.ch.

Associate editor: Juliette de Meaux

Abstract

Supergenes are nonrecombining genomic regions ensuring the coinheritance of multiple, coadapted genes. Despite the importance of supergenes in adaptation, little is known on how they originate. A classic example of supergene is the *S* locus controlling heterostyly, a floral heteromorphism occurring in 28 angiosperm families. In *Primula*, heterostyly is characterized by the cooccurrence of two complementary, self-incompatible floral morphs and is controlled by five genes clustered in the hemizygous, ca. 300-kb *S* locus. Here, we present the first chromosome-scale genome assembly of any heterostylous species, that of *Primula veris* (cowslip). By leveraging the high contiguity of the *P. veris* assembly and comparative genomic analyses, we demonstrated that the *S*-locus evolved via multiple, asynchronous gene duplications and independent gene translocations. Furthermore, we discovered a new whole-genome duplication in Ericales that is specific to the *Primula* lineage. We also propose a mechanism for the origin of *S*-locus hemizygosity via nonhomologous recombination involving the newly discovered two pairs of *CFB* genes flanking the *S* locus. Finally, we detected only weak signatures of degeneration in the *S* locus, as predicted for hemizygous supergenes. The present study provides a useful resource for future research addressing key questions on the evolution of supergenes in general and the *S* locus in particular: How do supergenes arise? What is the role of genome architecture in the evolution of complex adaptations? Is the molecular architecture of heterostyly supergenes across angiosperms similar to that of *Primula*?

Key words: genome architecture, supergene, heterostyly, evolutionary genomics, chromosome-scale genome assembly, *primula*.

Introduction

Understanding the genetic basis of adaptation and the molecular mechanisms underlying the emergence and maintenance of adaptive polymorphisms is central to evolutionary biology (Yeaman 2013; Purcell et al. 2014; Schwander et al. 2014; Llaurens et al. 2017). Complex adaptive polymorphisms are characterized by the coexistence of different phenotypes with contrasting trait combinations (Thompson and Jiggins 2014; Wellenreuther and Bernatchez 2018). The alternative allelic arrangements controlling complex polymorphisms are often maintained through the clustering of coadapted genes in a nonrecombining region inherited as a single Mendelian locus, termed supergene (Darlington and Mather 1949; Thompson and Jiggins 2014). Supergenes, identified in most eukaryotic lineages, including plants (Okada et al. 2011; Kotani et al. 2014; Li et al. 2016), animals (Wang et al. 2013;

Lamichhaney et al. 2016; Tuttle et al. 2016) and fungi (Sun et al. 2017; Branco et al. 2018), vary in size, number of genes and mechanism of recombination suppression (Gutiérrez-Valencia et al. 2021). Although most supergenes are protected from recombination by genomic inversions, other mechanisms, such as hemizygosity, association with genomic regions with restricted recombination (e.g., centromeres), enrichment of small-scale structural variants and epigenetic modifications are also known (Schwander et al. 2014; Gutiérrez-Valencia et al. 2021).

Despite the ubiquitous role of supergenes in adaptation, knowledge of their evolutionary origins remains limited (Schwander et al. 2014; Thompson and Jiggins 2014). Three general models have been proposed to explain the origin of supergenes (reviewed in Gutiérrez-Valencia et al. 2021): 1) colocalized genes undergo mutations, thus forming a region containing multiallelic polymorphisms that can increase in

size via subsequent mutations in the two haplotypes with antagonistic effects, which we term “colocalization first” model (reviewed in Charlesworth 2016); 2) functionally interacting but not colocalized genes are brought into physical linkage via genomic rearrangements or transposition, which we term “colocalization later” model (Turner 1967; Yeaman 2013); 3) a DNA segment already characterized by clustered, coadapted genes is acquired via introgression from another species and maintained as a polymorphism, which we term “introgression” model (Jay et al. 2018).

One of the best-studied supergenes is the S locus controlling heterostyly, a complex floral polymorphism occurring in at least 28 angiosperm families (Barrett 2019). *Primula* (primrose) has served as the main model for heterostyly since Darwin (Darwin 1877; Mast et al. 2006; Gilmartin 2015; Kappel et al. 2017). In primroses, heterostylous species produce two types of flower differing in the reciprocal positions of male and female sexual organs: L-morph (pin) flowers with long style and low anthers, S-morph (thrum) flowers with short style and high anthers (Darwin 1877; fig. 1A). In *Primula*, this dimorphism is accompanied by differences in the size of pollen grains and stigma papillae and associated with a diallelic self-incompatibility system preventing self- and intramorph-fertilization (Darwin 1877; Shivanna et al. 1981). The adaptive advantage of heterostyly lies in promoting outcrossing in two ways: 1) the reciprocal positioning of sexual organs facilitates pollen transfer between flowers of different morphs, thus reducing pollen wastage and favoring disassortative mating; 2) self-incompatibility prevents self-fertilization, protecting from inbreeding depression (Lloyd and Webb 1992; Barrett 2002; Keller et al. 2014; Barrett 2019).

In *Primula vulgaris*, the heterostyly supergene is a 278-kb region comprising five genes that is hemizygous in S-morphs and absent in L-morphs, hence recombination in this region is suppressed via hemizyosity (Li et al. 2016; fig. 1B). Hemizyosity of the S locus has been additionally confirmed in the heterostylous *P. veris*, *P. farinosa*, and *P. forbesii* (Cocker et al. 2018). Because of its hemizyosity, the S locus represents a peculiar example of supergene, for in most other supergenes recombination is prevented by inversions (Gutiérrez-Valencia et al. 2021). Consequently, expectations regarding the evolutionary origins of hemizygous supergenes differ from those proposed for supergenes maintained by inversions. For instance, gene duplications are expected to play a key role in the origins of hemizygous supergenes, because they can create at once both the genetic substrate for evolution to act in order to produce phenotypic novelty, and presence–absence polymorphism (Kappel et al. 2017; Li et al. 2020). Additionally, hemizyosity can stem from either deletion or insertion and the specific mechanism through which hemizyosity originates in supergenes remains unknown.

Two main models, both involving a key role for gene duplications, have been proposed for the evolution of the heterostyly supergene. One model posits that a large genomic segment with the clustered precursors of the S-locus genes was duplicated, allowing for the neofunctionalization of the gene duplicates into S-locus genes, whereas intervening regions were subsequently lost (segmental duplication model

[Kappel et al. 2017]; fig. 1C). Alternatively, S-locus genes might have arisen via multiple duplications and been independently translocated to the same genomic region (stepwise duplication model [Huu et al. 2020]; fig. 1C). Indeed, a recent study showed that two (CYP^T and GLO^T) of the five S-locus genes duplicated asynchronously and that their paralogs are not physically linked, providing initial support for the latter model (Huu et al. 2020; fig. 1C). However, because previous evidence in favor of the stepwise duplication model was limited to two of the five S-locus genes, a “hybrid” build-up of the S locus involving one large segmental duplication and additional, independent gene duplications cannot be discarded until the age of all S-locus genes and the genomic location of their paralogs have been established. Both goals require a highly contiguous genome assembly for a heterostylous species, which was not available until the present study. Finally, the possibility that S-locus genes originated via whole-genome duplications (WGDs) has never been proposed, even though the role of WGDs in the origin of phenotypic novelty has been amply demonstrated, especially in plants (Panchy et al. 2016; Ren et al. 2018).

Here, we present the chromosome-scale, haplotype-phased genome assembly of the heterostylous *Primula veris* (cowslip), which, combined with comparative genomic analyses across angiosperms, enabled us to test whether: 1) any of the S-locus gene paralogs colocalized and duplicated synchronously; 2) S-locus gene duplications stemmed from WGDs; 3) S-locus gene duplications preceded or cooccurred with the origin of heterostyly; 4) S-locus genes showed signatures of degeneration compared with their closest paralogs and to the rest of the genome. Finally, we were able to propose the first model for the origin of hemizyosity as a mechanism of recombination suppression in supergenes. This study generates new knowledge on the evolutionary build-up of genomic architectures underlying adaptive polymorphisms.

Results and Discussion

Genome Assembly and Annotation

We combined 51.5 Gb of nanopore data and 28.5 Gb of Illumina data (corresponding to 114× and 63× coverage, respectively; supplementary tables S14 and S15, Supplementary Material online) with the trio binning approach (Koren et al. 2018) to assemble the two haplotypes of a *P. veris* S-morph ($2n = 2x = 22$ [Nowak et al. 2015]; haploid genome size estimated by flow cytometry = 452 Mb, see Materials and Methods). The two resulting draft assemblies were polished with short and long reads and scaffolded using long-range information obtained from chromatin conformation capture methods (i.e. Chicago and Hi-C libraries; supplementary table S16, Supplementary Material online; Belton et al. 2012; Putnam et al. 2016), followed by *in silico* gap-closure with nanopore reads. Scaffolds representing contaminants, mitochondrial and plastid genomes were removed and misassemblies manually corrected (supplementary figs. S1–S9, Supplementary Material online). The final maternal haplotype assembly comprised 421.37 Mb ($N50 = 34.03$ Mb), corresponding to 93.2% of the genome size estimated

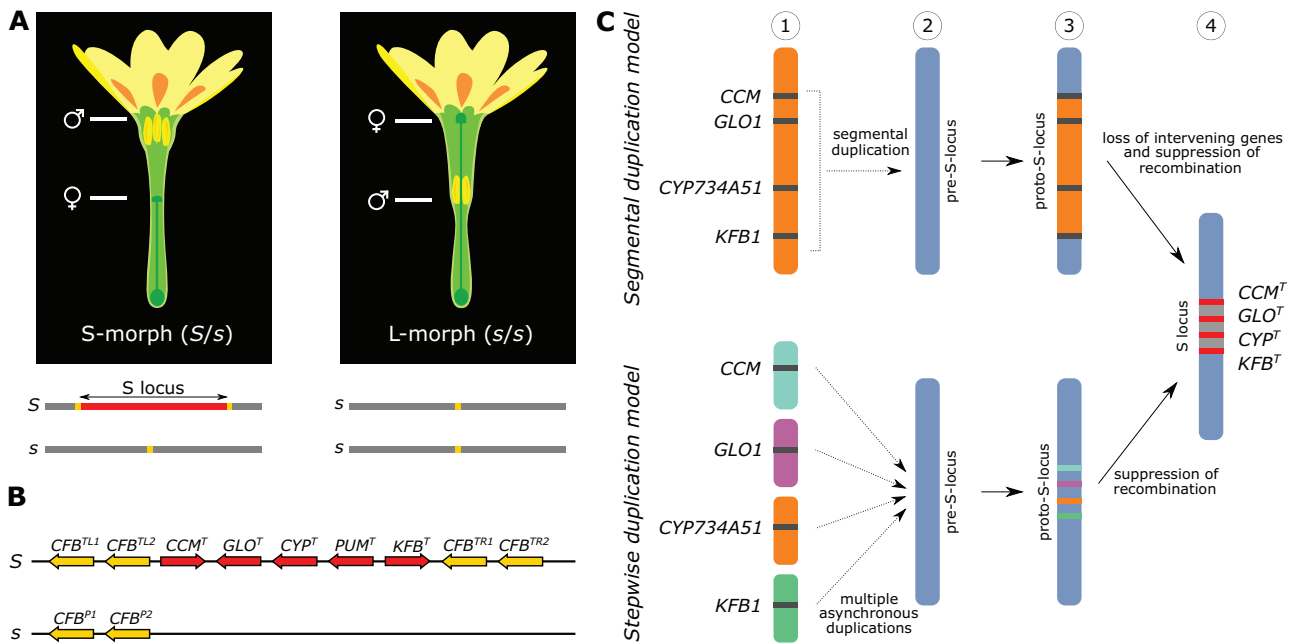


Fig. 1. Heterostyly in *P. veris* and models for the origin of the S locus. (A) Top: short-styled (S) and long-styled (L) morphs differ by having male (anthers) and female (stigma) sexual organs reciprocally positioned in their flowers. Bottom: The S locus (red) is hemizygous in S-morphs (S haplotype), absent in L-morphs (s haplotype); the location of *CFB* genes is indicated in yellow. (B) Structure of the S locus in the S-morph, with gene orientations indicated by pointed ends. Top: dominant S haplotype containing five genes (red) and two copies of *CFB* in each flanking region (yellow); bottom: recessive s haplotype with only two copies of *CFB*. Superscripts on genes stand for: T, thrum (S-morph); P, pin (L-morph); L, left; R, right; 1, gene copy 1; 2, gene copy 2. (C) Two models for the origin of the S locus involving its four duplicated genes. Segmental duplication model (top): the paralogs of S-locus genes were originally clustered (1), then duplicated as a single segment and inserted into a different genomic region, here called pre-S-locus region (blue) (2), forming a proto-S-locus (yellow) (3); intervening genes were then lost and recombination suppressed, forming the S locus (4). Stepwise duplication model (bottom): the paralogs of S-locus genes were originally unlinked (1), duplicated asynchronously and independently inserted into the pre-S-locus (blue) (2), forming the proto-S-locus (3; different colors indicate that the paralogs derived from different genomic locations); recombination among these genes was then suppressed, forming the S locus (4).

via flow cytometry, and the paternal haplotype comprised 419.58 Mb (N50 = 34.35 Mb), corresponding to 92.8% of the estimated genome size. Each haplotype assembly contained 11 chromosome-sized scaffolds, matching the actual number of *P. veris* haploid chromosomes. In the maternal haplotype, these scaffolds ranged from 31.81 to 48.64 Mb and comprised 397.14 Mb, corresponding to 94.3% of the assembly length; in the paternal assembly, the 11 largest scaffolds comprised 399.46 Mb, corresponding to 95.2% of the assembly length. The high quality of both genome assemblies was confirmed by BUSCO (91.9% complete genes; Manni et al. 2021) and k-mer analysis (Mapleson et al. 2017) (table 1 and supplementary tables S1–S4 and figs. S10 and S11, Supplementary Material online).

Using a combination of ab initio, evidence-based and comparative gene-prediction approaches, we identified 34,581 and 34,009 protein-coding genes in the maternal and paternal haplotypes, respectively. Transposable elements (TEs) comprised 46.07% and 46.10% (193.97 and 193.25 Mb) of the maternal and paternal haplotype assemblies, respectively (supplementary table S5, Supplementary Material online). The majority of annotated TEs belonged to class I, with long terminal repeat (LTR) retrotransposons being the most abundant TE order, covering ~27% of both assemblies. Regions with high LTR density and low gene density identified

in each chromosome (fig. 2A) likely represent pericentromeric regions, as often found in plant genomes (supplementary fig. S12, Supplementary Material online; Kejnovsky et al. 2012). Altogether, these results imply that the 11 largest scaffolds of each assembly correspond to the 11 chromosomes of *P. veris*.

Inter-Haplotype Comparison

The identification of within-species structural genomic rearrangements has long been precluded by the scarcity of highly contiguous genome assemblies (Goel et al. 2019; Jiao and Schneeberger 2020; Mérot et al. 2020). The chromosome-scale assembly of both haplotypes enabled us to investigate structural variability in *P. veris*. We compared the two haploid assemblies and identified 267 inversions (totaling 18.1 Mb), 2,830 translocations (15.5 Mb), and 16,925 duplications (49.5 Mb) that cumulatively comprised 83.2 Mb, corresponding to 20.9% of the assembled genome (fig. 2B and supplementary tables S6 and S7, Supplementary Material online). These rearrangements ranged in size from a few base pairs to several megabase-pairs: all rearrangements >500 kb were inversions, with the longest spanning 4.49 Mb on chromosome 4 (fig. 2B). Visual inspection of the detected large structural variants using Chicago and Hi-C read mapping demonstrated that they are not the product of assembly errors (supplementary fig. S13, Supplementary Material online). Among the

Table 1. Statistics for the *Primula veris* Genome Assembly and Gene Annotation.

	Maternal Haplotype	Paternal Haplotype
Assembly size	421.38 Mb	419.58 Mb
% of the genome size	93.2%	92.8%
Cumulative length of the 11 largest scaffolds	397.14 Mb	399.47 Mb
% of the assembly in the 11 largest scaffolds	94.25%	95.21%
Number of scaffolds	648	640
Scaffold N50	34.03 Mb	34.35 Mb
BUSCO complete genes—genome mode	91.90%	92.30%
Number of genes	34,581	34,009

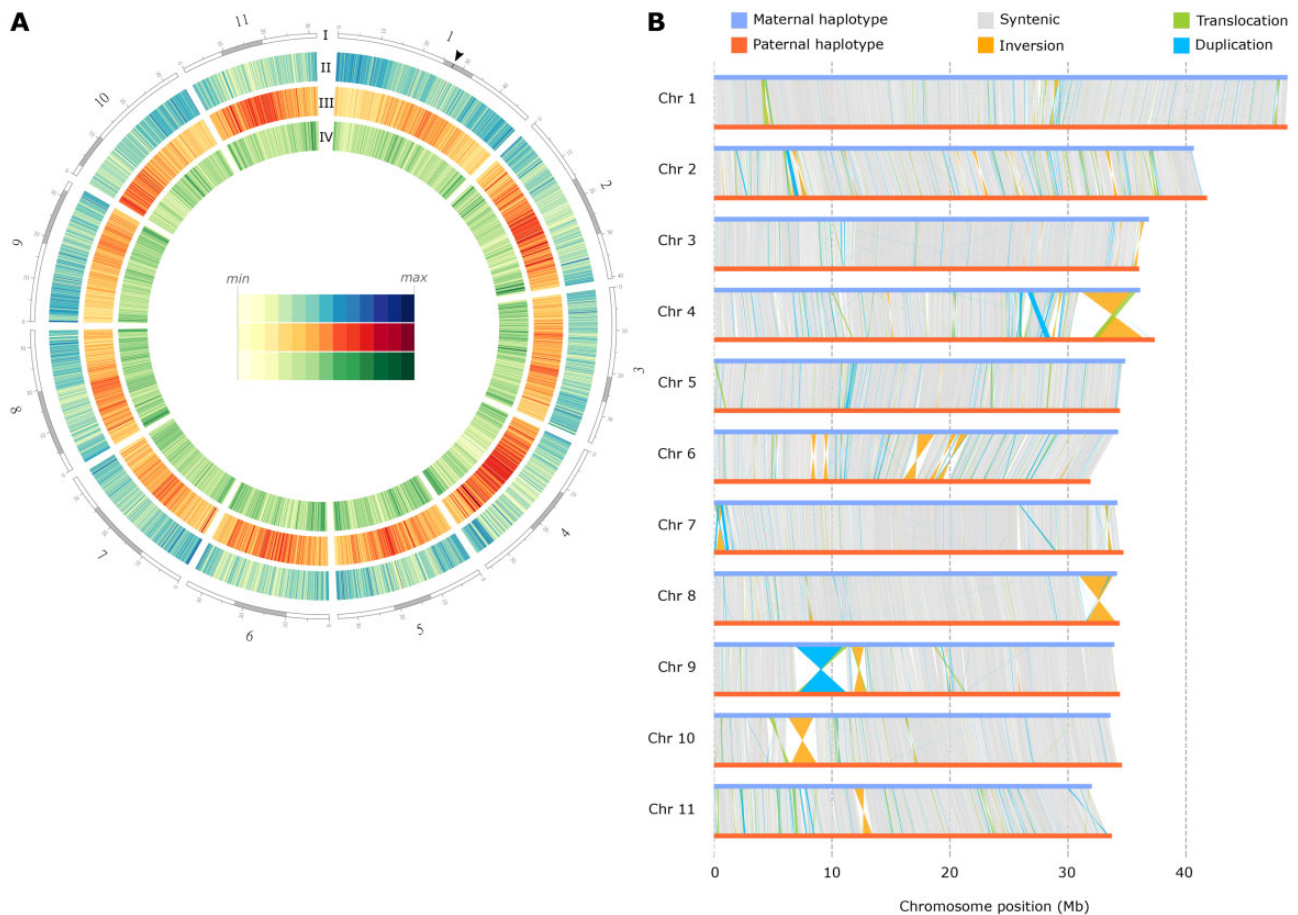


Fig. 2. Overview of the *P. veris* genome and comparison between haplotypes. (A) Circle plot of the *P. veris* genome assembly (maternal haplotype). Tracks from outside to inside correspond to: (I) the 11 chromosome-scale scaffolds, with the putative centromeric and putative pericentromeric regions shown in gray; position of the S locus is marked by a black arrow in chromosome 1; (II) gene density (blue); (III) LTR retrotransposons (red); (IV) DNA transposons (green). Tracks II, III, and IV are calculated in 100-kb nonoverlapping windows. (B) Structural rearrangements are represented by colored lines (orange for inversions, green for translocations, blue for duplications) connecting regions of the maternal and paternal chromosomes (blue and red horizontal lines, respectively); syntenic regions are connected by gray lines.

observed structural variants is the ca. 280 kb hemizygous S-locus supergene.

The large structural variants detected between the two *P. veris* haplotypes are equally or more abundant than those previously identified in five other model species (Goel et al. 2019; Jiao and Schneeberger 2020; supplementary table S6, Supplementary Material online). However, it is not possible to disentangle whether this result stems from the high divergence between the two haplotypes of the *P. veris* individual used for the present study or from a more general, highly

dynamic nature of *Primula* genomes, for example, mediated by elevated TE activity.

Position, Structure, and Flanking Genes of the S Locus

We identified the *P. veris* S locus as a 260-kb genomic region present only in the maternal haplotype (inherited from the S-morph parent), confirming its hemizygosity in S-morphs (Nowak et al. 2015; Huu et al. 2016; Li et al. 2016). The heterostyly supergene is located within the putative pericentromeric region of chromosome 1 (27.43–27.70 Mb; arrow in

fig. 2A), consistent with cytogenetic observations in the closely related *P. vulgaris* (Li et al. 2015). The *P. veris* S locus contains the same five genes in the same order as those of *P. vulgaris* (Li et al. 2016): CCM^T encodes a protein with a conserved cysteine motif in the C-terminal domain with unknown function; GLO^T (*GLOBOSA2*) is a B-class floral homeotic gene that determines higher anthers in S-morphs; CYP^T (*CYP734A50*) encodes a cytochrome P450 and determines shorter styles in S-morphs; PUM^T encodes a Pumilio-like RNA-binding protein; KFB^T encodes a Kelch domain-containing F-box protein (Huu et al. 2016, 2020; Li et al. 2016). Genes of the dominant S and recessive s haplotypes are designated by the superscripts T and P for thrum (i.e., S-morph) and pin (i.e., L-morph), respectively (fig. 1B). Although the functions of GLO^T and CYP^T mentioned above have been experimentally demonstrated (Huu et al. 2016, 2020), the potential roles of CCM^T , PUM^T , and KFB^T in heterostyly remain unknown.

Previous studies in *P. vulgaris* suggested that the S locus was flanked by one copy of a *Cyclin-like F box* gene at each side (CFB^{TL} , CFB^{TR}), whereas the s haplotype contained a single *CFB* copy (CFB^P) (Li et al. 2016; Cocker et al. 2018). Our results revealed that the S haplotype of *P. veris* contains four *CFB* copies, two at each side of the S locus (here named CFB^{TL1} , CFB^{TL2} , and CFB^{TR1} , CFB^{TR2}), whereas the s haplotype contains only two *CFB* copies (CFB^{P1} and CFB^{P2} ; fig. 1B). All *CFB* copies have the same orientation. We showed that the difference in *CFB* copy number between *P. veris* and *P. vulgaris* stems from both incomplete assembly and erroneous gene annotation of the *P. vulgaris* genome (Li et al. 2016; Cocker et al. 2018; supplementary figs. S14 and S15, Supplementary Material online).

Whole-Genome Duplication in *Primula*

WGDs are known to play a fundamental role in the evolution of novel functions in angiosperms, including floral structures (Panchy et al. 2016; Ren et al. 2018). *Primula* belongs to the order Ericales, in which several WGDs have been identified, including a WGD (named *Ad-β*) at the root of Ericales (Shi et al. 2010; Larson et al. 2020). A more recent WGD within Primulaceae was tentatively suggested by a study based on transcriptomic data, although no conclusive evidence was provided (Larson et al. 2020; fig. 3A). Although previous studies in *Primula* identified paralogs for three (CCM^T , GLO^T , and CYP^T) out of five S-locus genes (Li et al. 2016; Kappel et al. 2017; Huu et al. 2020), the hypothesis that these duplicates might stem from WGD has never been investigated. To test it, we used the newly generated *P. veris* genome assembly and comparative genomic analyses.

First, we performed an intragenomic synteny analysis, revealing that a substantial portion of the *P. veris* genome is duplicated, with 12,942 (37.4% of the total gene number), 544 (1.6%), and 10 (<0.1%) genes showing one, two or three paralogs, respectively (fig. 3B). Thus, 39.0% of the genes in the *P. veris* genome are present in two or more copies, similarly to other Ericales known to have experienced both the *Ad-β* and more recent WGDs (Larson et al. 2020; Wang et al. 2020; supplementary fig. S16, Supplementary Material online). Furthermore,

1,561 paralogous gene pairs occurred in 134 collinear genomic blocks ranging from 5 to 70 genes (fig. 3C), further corroborating the hypothesis of a WGD more recent than *Ad-β*.

To estimate the number and approximate timing of WGDs in the evolutionary history of *Primula*, we calculated the number of synonymous substitutions per synonymous site (K_S) for paralogous gene pairs contained in syntenic blocks of the *P. veris* genome and plotted their distribution (fig. 3D). We identified four statistically significant K_S peaks (supplementary fig. S17, Supplementary Material online) but discarded one of them (at $K_S = 3.66$; $SD = 0.20$) as an artifact (Tiley et al. 2018; see Materials and Methods). We assigned the remaining three peaks to three putative WGDs: the oldest at $K_S = 2.24$ ($SD = 0.63$), corresponding to previously reported K_S values for the γ triplication shared by all eudicots (Qiao et al. 2019); the second at $K_S = 1.25$ ($SD = 0.30$), compatible with the *Ad-β* WGD at the root of Ericales (Shi et al. 2010; Larson et al. 2020); the third at $K_S = 0.87$ ($SD = 0.17$), representing a WGD more recent than *Ad-β* (fig. 3A). To test whether the detected WGD at $K_S = 0.87$ is shared between *P. veris* and other Ericales, we identified collinear blocks between the *P. veris* genome and five other highly contiguous genomes of Ericales (*Actinidia chinensis*, *Camellia sinensis*, *Diospyros oleifera*, *Rhododendron delavayi*, and *Vaccinium corymbosum*), and calculated K_S between orthologous syntenic gene pairs. The resulting K_S plots imply that the peak at $K_S = 0.87$ is more recent than the split between *P. veris* and the Ericales species listed above (fig. 3D and supplementary fig. S18, Supplementary Material online). We further tested this result by mapping the number of gene duplications inferred from the gene trees of 20,770 orthogroups onto the species tree of 13 angiosperm species with high-quality genome assemblies (see Materials and Methods for details), including *P. veris* and *P. vulgaris* and five additional Ericales (supplementary fig. S19, Supplementary Material online). The highest numbers of gene duplications were located at the following nodes, from oldest to most recent: the base of angiosperms (1,055; ε WGD in fig. 3A), the base of eudicots (951; γ triplication in fig. 3A), between *Vaccinium corymbosum* and *Rhododendron delavayi* (1,238) and between *P. veris* and *P. vulgaris* (3,763), suggesting a WGD at each of these four nodes.

Taken together, the results of self-syntenic, K_S , and phylogenetic analyses support a WGD, here named *Pv-α* (fig. 3A), likely corresponding to a previously hypothesized, but undemonstrated Primulaceae-specific WGD (Larson et al. 2020). Using a neutral substitution rate of 6.15×10^{-9} (95% $CI = 5.60 \times 10^{-9} - 6.62 \times 10^{-9}$) substitutions per synonymous site per year between paralogous gene pairs (see Materials and Methods), *Pv-α* was dated at 70.57 Ma (95% $CI = 65.51 - 77.51$ Ma), significantly predating the origin of heterostyly in *Primula*, previously estimated at 15–35 Ma (de Vos et al. 2014).

The S Locus Originated via Stepwise Duplications

Leveraging the high contiguity of the *P. veris* genome assembly, we tested the previously proposed segmental versus stepwise duplication models for the origin of the heterostyly supergene. If the supergene originated via segmental

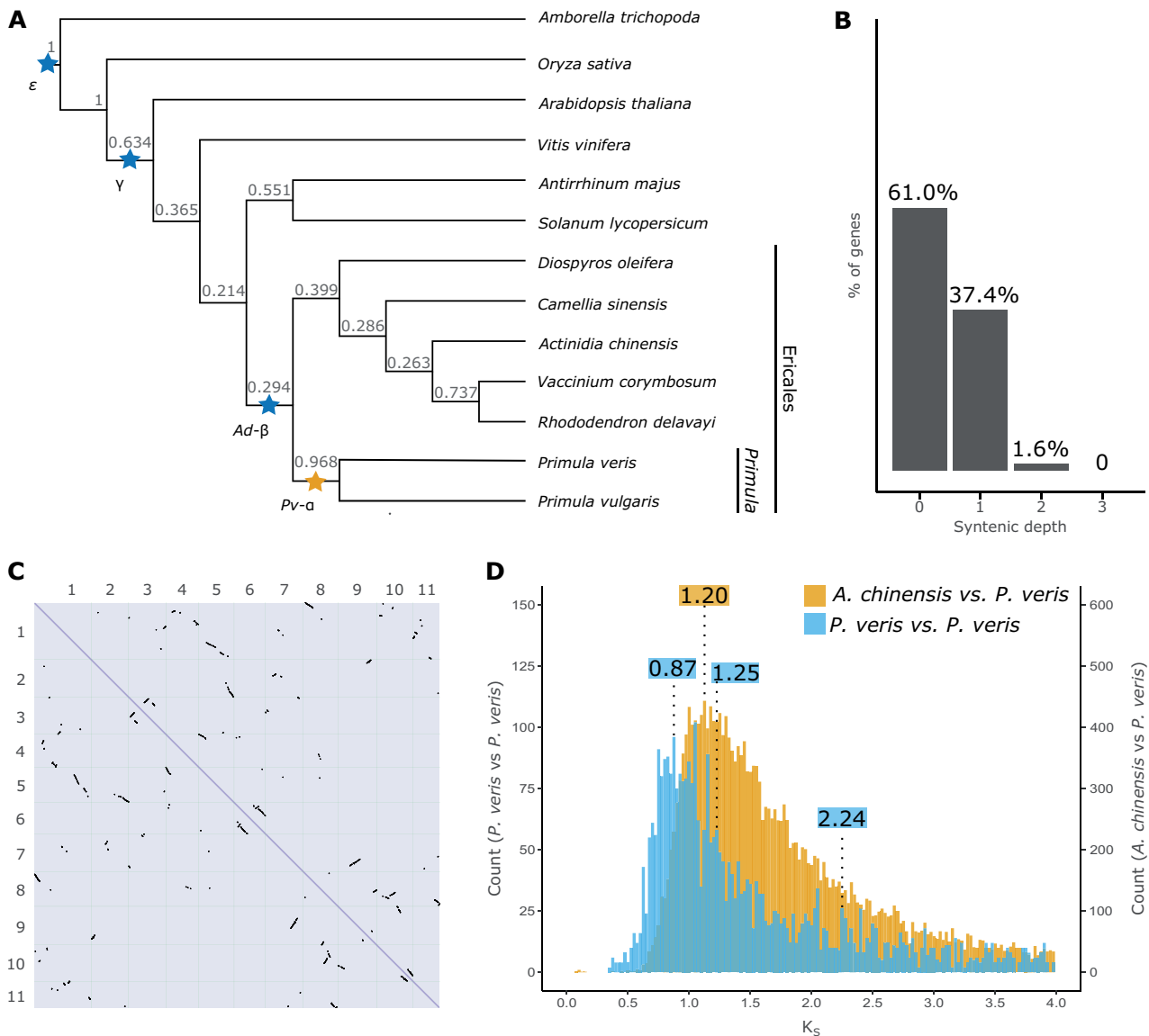


Fig. 3. Evidence of a WGD in the *Primula* lineage. (A) Phylogeny of 13 angiosperm species inferred by OrthoFinder using the STAG algorithm and rooted using STRIDE; numbers at each node represent STAG support values, that is, the fraction of orthogroup trees supporting each bipartition (see Materials and Methods for details); WGDs inferred in previous studies are marked by blue stars; the yellow star represents a WGD newly demonstrated here. (B) Proportion of genes with different syntenic depths in the *P. veris* genome. (C) Dotplot obtained by aligning the *P. veris* maternal haplotype against itself; self-syntenic (i.e., duplicated) regions containing >5 collinear genes are represented by black marks. (D) Density distribution of K_S in paralogous gene pairs within *P. veris* (blue) and in orthologous gene pairs between *P. veris* and *A. chinensis* (yellow), representing the sister clade of *Primula*; to ease visualization, the columns of the blue histogram were increased four times in height. The three statistically significant peaks in the blue distribution and the peak representing the divergence between *P. veris* and *A. chinensis* are marked with the respective K_S values (in blue and yellow, respectively).

duplication, all S-locus genes should have the same age and their paralogs should colocalize elsewhere in the genome. Conversely, if the supergene originated via stepwise duplications, S-locus genes should have different ages and their paralogs should be scattered throughout the genomes of *P. veris* and other Ericales species (fig. 1C). Additionally, we investigated whether the S-locus genes originated via any of the WGDs detected above.

We first searched for paralogs of S-locus genes in the *P. veris* genome. We identified *CYP734A51*, *GLO1*, and *CCM1* as the phylogenetically closest paralogs of S-locus *CYP^T*, *GLO^T*, and *CCM^T*, respectively, confirming previous

results in *P. vulgaris* and *P. veris* (Huu et al. 2016; Li et al. 2016; Burrows and McCubbin 2017). Differently from previous studies, we discovered two rather than just one *KFB^T* paralogs and named them *KFB1* and *KFB2*, with *KFB1* having the highest similarity to *KFB^T*. As in previous studies (Li et al. 2016), no close paralog was identified for *PUM^T*.

We then estimated the relative duplication ages of the four duplicated S-locus genes by first assuming substitution rate constancy, then by relaxing this assumption, as described below. We calculated K_S between S-locus genes and their paralogs in the maternal haplotype of the *P. veris* plant used for genome assembly and in the genomes of ten

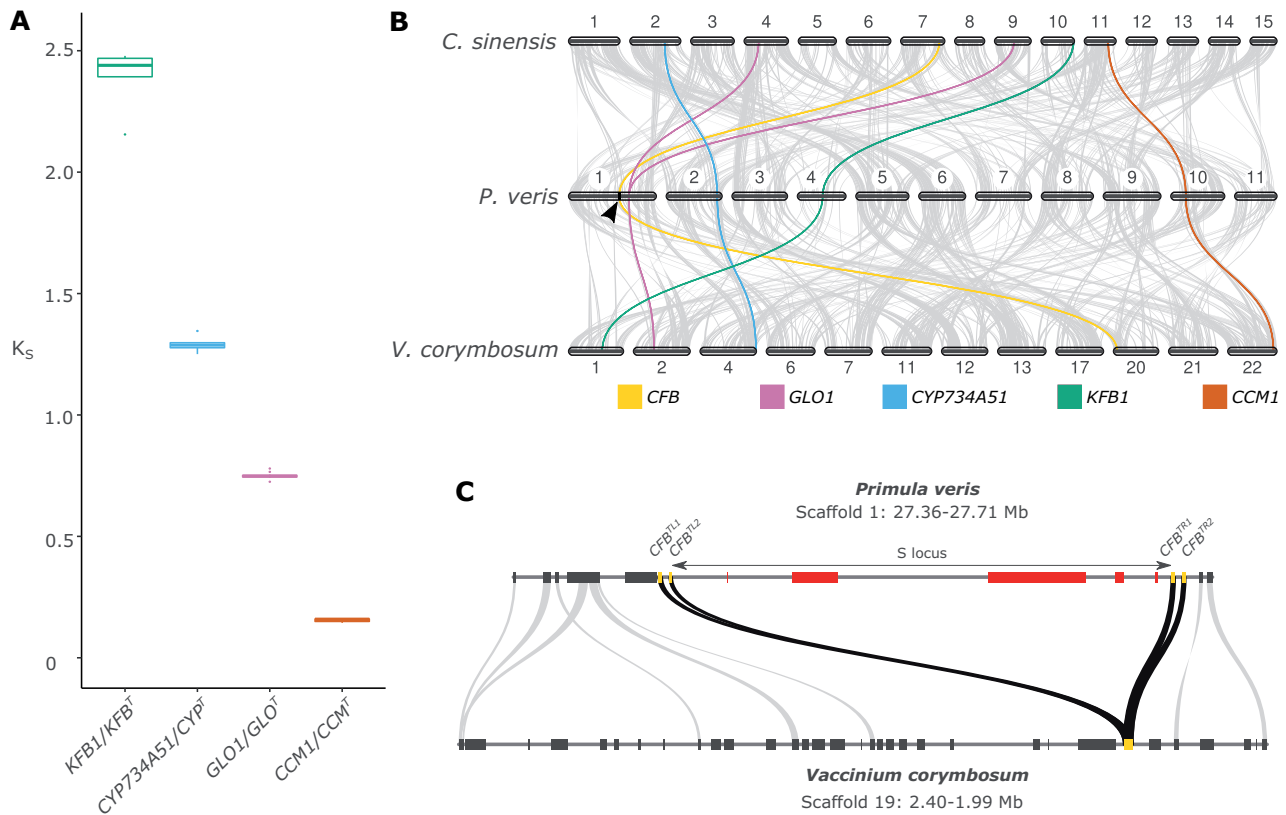


Fig. 4. The S locus originated in a stepwise manner. (A) Boxplot of K_s distributions between each duplicated S-locus gene and its closest paralog calculated in 11 S-morph individuals of *P. veris*. (B) Synteny plot between *C. sinensis*, *P. veris*, and *V. corymbosum*. Regions containing >5 collinear genes are connected with gray lines, whereas the paralogs of S-locus genes in the three species are connected with color lines: GLO1 (pink), CYP734A51 (blue), KFB1 (green), CCM1 (orange), plus the flanking CFB (yellow). The S locus is marked by a black arrow on the *P. veris* chromosome 1. (C) Microsynteny plot between the region containing the S locus in *P. veris* and its collinear region in *V. corymbosum*: S-locus genes are represented as red boxes; CFB genes are represented as yellow boxes; genes outside the S locus are represented as gray boxes. Gray lines connect orthologous gene pairs; black lines connect the four CFB copies in *P. veris* with their orthologs in *V. corymbosum*.

additional S-morph individuals of *P. veris* from different geographic regions (supplementary fig. S20, Supplementary Material online). The inferred K_s distributions for the four gene pairs did not overlap with each other, with mean values of 2.47 for KFB1/KFB^T, 1.29 for CYP734A51/CYP^T, 0.75 for GLO1/GLO^T, and 0.15 for CCM1/CCM^T (fig. 4A and table 2). Assuming synonymous substitution rate constancy, these results support the hypothesis that S-locus genes originated via multiple, asynchronous duplications involving KFB1 first, CYP734A51 second, GLO1 third, and CCM last. However, the assumption of substitution rate constancy is often violated, leading to potentially erroneous age estimates, especially when using K_s values between only four paralogous gene pairs (Tiley et al. 2018; Huu et al. 2020), as opposed to using K_s values inferred from thousands of gene pairs, as we did for dating WGD events (see above). Therefore, we also employed molecular dating analyses that account for substitution-rate variation to phylogenetically estimate the duplication ages of S-locus genes by generating calibrated gene trees for all orthogroups containing the four duplicated S-locus genes (supplementary figs. S21–S24, Supplementary Material online). The mean duplication ages thus inferred were: 104.0 Ma for KFB1/KFB^T, 42.7 Ma for CYP734A51/CYP^T, 37.4 Ma for GLO1/GLO^T (compatible with previous reports; Li et al.

2016), and 10.3 Ma for CCM1/CCM^T (table 2), confirming the asynchronous origins of S-locus genes in the same chronological sequence inferred by assuming rate constancy (see above).

To further compare the segmental versus stepwise hypotheses for the origin of the heterostyly supergene, we tested whether the paralogs of the S-locus genes colocalize or not. We found that the paralogs are scattered throughout the *P. veris* genome, with GLO1 on chromosome 1 (6.49 Mb away from the CFB^{TR2} gene flanking the S locus in the *P. veris* genome assembly, confirming its proximity to the S locus documented in *P. vulgaris*; Li et al. 2016, 2015), CYP734A51 on chromosome 2, KFB1 on chromosome 4, and CCM1 on chromosome 10 (fig. 4B). Because these results do not exclude the possibility that the paralogs of the S-locus genes were initially linked when a segmental duplication occurred, but were subsequently separated via genomic rearrangements, we additionally tested whether they colocalized in the genomes of nonheterostylous Ericales. We discovered that paralogs of the S-locus genes are located on different chromosomes also in the highly contiguous genome assemblies of *A. chinensis*, *C. sinensis*, *D. oleifera*, and *V. corymbosum* (fig. 4B and supplementary fig. S25, Supplementary Material online). Taken together, the evidence presented here

Table 2. Duplication Ages Estimated for S-Locus Genes.

	Synonymous Substitutions per Synonymous Site (K_S)		Phylogenetically Inferred Duplication Age (Ma)	
	Mean	Standard deviation	Mean	95% HPD
<i>KFB1/KFB^T</i>	2.47	0.115	104.0	91.0–116.9
<i>CYP734A51/CYP^T</i>	1.29	0.024	42.7	30.4–56.9
<i>GLO1/GLO^T</i>	0.75	0.013	37.4	26.9–49.2
<i>CCM1/CCM^T</i>	0.15	0.006	10.3	4.34–16.94

NOTE.—Ma, million years ago; HPD, highest posterior density.

conclusively supports the stepwise duplication model for the origin of the S locus, with S-locus genes originating via asynchronous duplications and likely being clustered via independent translocations.

We then tested whether any S-locus gene duplicated via WGD by comparing their duplication ages with the ages of the *Pv- α* and *Ad- β* WGDs. We found that *CYP^T*, *GLO^T*, and *CCM^T* originated via duplication significantly after *Pv- α* WGD (dated at 65.51–77.51 Ma). Conversely, the duplication age of *KFB^T* overlaps with *Ad- β* WGD at the root of Ericales (~110 Ma [Rose et al. 2018]; fig. 3A), thus *KFB^T* is the only S-locus gene that might have originated via WGD. We also note that *CYP^T*, *GLO^T*, and *CCM^T* must have acquired their new function in controlling heterostyly soon after they duplicated, for their duplication ages (10.3–42.7 Ma) overlap with the age inferred for the emergence of heterostyly in Primulaceae (15–35 Ma [de Vos et al. 2014]). This age overlap is consistent with the prediction that duplicate genes should neofunctionalize soon after duplication, lest they quickly pseudogenize (Lynch and Conery 2000).

The inferred stepwise origin of the S locus implies that all duplicated S-locus genes were translocated independently to the same region, which we term pre-S-locus region (fig. 1C). We thus tested whether the only nonduplicated S-locus gene (*PUM^T*) and the *CFB* genes flanking the S locus were originally present in the pre-S-locus region by performing pairwise synteny analyses between the genomes of *P. veris* and five Ericales species. The regions flanking the S locus were collinear within Ericales and contained a copy of the *CFB* gene in *C. sinensis*, *R. delavayi*, and *V. corymbosum* (fig. 4C and supplementary fig. S26, Supplementary Material online), whereas *PUM^T* was absent. Therefore, the heterostyly supergene formed via translocations of the five S-locus genes to the pre-S-locus region, which already included one copy of *CFB*.

Our results suggest a scenario by which the S-locus genes originated via duplication at different times, then clustered together via independent translocations, likely after acquiring a novel function connected with the control of heterostyly. Thus, the S locus appears to fit the “colocalization later” general model for the origin of supergenes presented above, according to which separate, but functionally interacting genes get physically linked via genomic rearrangement or transposition (Turner 1967; Yeaman 2013).

Origin of S-Locus Hemizygosity

A key feature of the S locus in *Primula* is hemizygosity, ensuring the inheritance of the heterostyly genes as a single

Mendelian locus. A hemizygous region originates via either insertion or deletion. The stepwise origin of the S locus makes it unlikely that each of its five genes independently maintained a hemizygous state since their duplication and insertion in the pre-S-locus region. Thus, it is more likely that the S locus was initially present in both haplotypes and became hemizygous following its deletion in one haplotype. The presence of a repeated region (containing two *CFB* copies) at each side of the S locus favors the latter hypothesis, as it could have provided the substrate for a nonhomologous recombination that caused the deletion of the intervening region (i.e., the S locus) from one of the haplotypes (fig. 5C). Specifically, a recombination breakpoint between *CFB^{TL1}/CFB^{TR1}* and *CFB^{TL2}/CFB^{TR2}* would result in the two internal *CFB* copies (*CFB^{TL2}* and *CFB^{TR1}*) being unique to the S haplotype and the two external *CFB* copies of the S haplotype (*CFB^{TL1}* and *CFB^{TR2}*) being homologous to the *CFB* copies of the s haplotype (*CFB^{P1}* and *CFB^{P2}*). In this event, we would expect *CFB^{TL1}* and *CFB^{TR2}* to be most closely related to *CFB^{P1}* and *CFB^{P2}*, respectively.

We tested this hypothesis by estimating molecular divergence via K_S and phylogenetic relationships of the six *CFB* copies, four from the S haplotype, and two from the s haplotype (fig. 5A and B and supplementary figs. S27 and S28, Supplementary Material online). The results show that the two internal *CFB* copies (*CFB^{TL2}* and *CFB^{TR1}*) are indeed unique to the S haplotype, whereas the two external copies (*CFB^{TL1}* and *CFB^{TR2}*) are homologous to those present in the s haplotype (*CFB^{P1}* and *CFB^{P2}*). Moreover, the topology of the *CFB* gene tree implies that the ancestral single-copy *CFB* gene underwent first a tandem duplication resulting in two *CFB* copies (e.g., *CFB^{TL1}* and *CFB^{TL2}*) that were then segmentally duplicated forming the second *CFB* pair (e.g., *CFB^{TR1}* and *CFB^{TR2}*; fig. 5C). Our analyses also located the nonhomologous recombination breakpoint within *CFB^{TL1}/CFB^{TR1}*, close to their 5' end (supplementary fig. S29, Supplementary Material online).

We then tested whether S-locus hemizygosity evolved once concomitantly with or repeatedly after the origin of the supergene. Given the mechanistic model for S-locus hemizygosity proposed above (fig. 5C), a nonhomologous recombination producing hemizygosity could occur only after the second *CFB* duplication, that is, hemizygosity in *P. veris* should be more recent than 2.28–4.28 Ma (fig. 5B). This time interval overlaps with the previously inferred age (0.80–3.74 Ma) of the clade containing *P. veris* and *P. vulgaris* (de Vos et al. 2014), but postdates the divergence between *P. veris* and

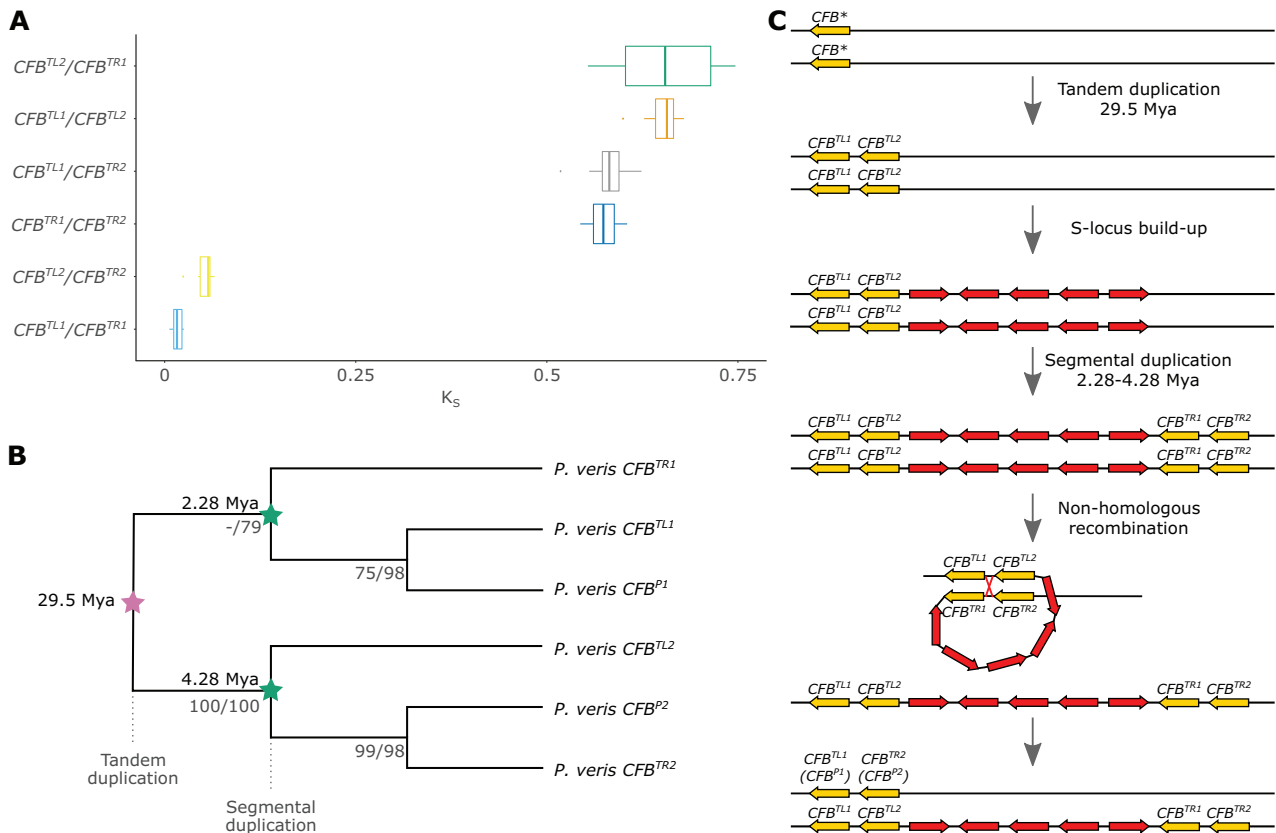


Fig. 5. The hemizygosity of the S locus originated via non-homologous recombination between CFB copies. (A) Boxplot of K_S values distributions for all six pairwise comparisons of the four CFB copies in the *P. veris* S haplotype. (B) Gene tree topology of CFB sequences from *P. veris* (S and s haplotypes), representing a subset of the larger CFB phylogeny of [supplementary figure S28, Supplementary Material online](#); branch labels represent ML/parsimony bootstrap support values, inferred with RAxML and PAUP, respectively (see [supplementary methods, Supplementary Material online](#) for details). CFB underwent two duplication rounds: a tandem duplication 29.5 Ma (pink star), then a segmental duplication 2.28–4.28 Ma (green stars). (C) Schematic model for the origin of hemizygosity of the *P. veris* S locus via nonhomologous recombination reflecting the inferred temporal sequence of the CFB tandem and segmental duplications; the ancestral copy of CFB prior to duplication is indicated by an asterisk.

the two other heterostylous *Primula* species known to have a hemizygous S locus (*P. farinosa* and *P. forbesii*; [Cocker et al. 2018](#)), dated at 12.22–18.26 and 21.16–29.59 Ma, respectively ([de Vos et al. 2014](#)). The time intervals above imply that hemizygosity evolved independently multiple times after the emergence of the heterostyly phenotype in *Primula* (15–35 Ma; [de Vos et al. 2014](#)), likely following an initial diallelic stage for the S locus. However, the possibility that homologous recombination between CFB genes of the S and s haplotypes and/or gene conversion between CFB^{TL} and CFB^{TR} copies homogenized the left and right CFB pairs in the S haplotype cannot be dismissed. The resulting lower sequence divergence between left and right CFB pairs of the S haplotype could cause the underestimation of the age of the second duplication event, implying that hemizygosity might not have originated multiple times independently after the origin of the heterostyly supergene. Whole-genome sequencing data across *Primula* would help to resolve whether S-locus hemizygosity evolved once concomitantly with the origin of the supergene or multiple times after its origin.

Selection on S-Locus Genes

Contrasting processes shape the evolution of supergenes. For example, suppression of recombination ([Cutter and Payseur](#)

[2013](#); [Corbett-DeTig et al. 2015](#); [Becher et al. 2020](#)) and reduced effective population size caused by hemizygosity ([Gossmann et al. 2011](#)) should decrease selection efficiency on the S locus. Consequently, the S locus should accumulate slightly deleterious mutations at a higher rate than the rest of the genome, ultimately leading to genetic degeneration ([Charlesworth and Charlesworth 2000](#)). Conversely, hemizygosity makes every mutation at the S locus effectively dominant, increasing the efficacy of selection within the S locus, potentially slowing down degeneration. Indeed, a recent study based on forward simulations concluded that, contrary to genes located in supergenes maintained by inversions, genes in hemizygous supergenes should exhibit only weak or no signs of degeneration ([Gutiérrez-Valencia et al. 2021](#)).

To test whether S-locus genes accumulate slightly deleterious mutations at an accelerated rate, d_N/d_S (nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site) was calculated between *P. veris* and *P. vulgaris* for all S-locus genes and between three additional *Primula* species for CYP^T and GLO^T (*P. forbesii*, *P. maximowiczii*, *P. oreodoxa*; [table 3](#) and [supplementary table S8, Supplementary Material online](#)) and compared with the d_N/d_S values obtained for the respective paralogs. The obtained values were also compared with empirical d_N/d_S

null distributions calculated between *P. veris* and *P. vulgaris* by randomly sampling *P. veris* genes with average expression levels matching those of S-locus genes from either the putative pericentromeric regions or the entire genome (see [supplementary methods, Supplementary Material online](#) for details). Knowing that d_N/d_S analyses may be misleading when genetic divergence between species is low, we also used a more sensitive approach by testing for accelerated evolution on S-locus genes using the clade model of the CodeML program in PAML (Yang 2007).

All S-locus genes were characterized by d_N/d_S values comparable with those of genes in putative pericentromeric regions ([supplementary fig. S30, Supplementary Material online](#)) and in the entire genome ([supplementary fig. S31, Supplementary Material online](#)), suggesting that selection does not appreciably differ between S-locus genes and the rest of the genome. Additionally, we compared d_N/d_S values for S-locus genes with those of their respective paralogs. The S-locus genes *CYP^T* and *GLO^T* showed significantly or marginally significantly higher d_N/d_S values than their respective paralogs, *CYP734A51* and *GLO1* ([table 3](#) and [supplementary table S9](#) and [fig. S32, Supplementary Material online](#)). This result was confirmed by the more sensitive clade-model approach of CodeML, implying an accelerated d_N/d_S for the clades including *CYP^T* and *GLO^T* compared with the clades including their paralogs (likelihood-ratio test $P < 0.05$; [supplementary fig. S33](#) and [table S10, Supplementary Material online](#)). Conversely, *KFB^T* and *CCM^T* showed lower pairwise d_N/d_S than their paralogs (*KFB1* and *CCM1*; [table 3](#) and [supplementary table S9, Supplementary Material online](#)), although significance of these results could not be tested due to insufficient number of comparisons. Finally, no significant differences in substitution rates were detected between the *KFB^T* and *KFB1* clades on CodeML, nor between the *CCM^T* and *CCM1* clades (likelihood-ratio test $P = 0.76$ and $P = 0.13$, respectively; [supplementary fig. S33](#) and [table S10, Supplementary Material online](#)). To sum up, *CYP^T* and *GLO^T* appear to accumulate mutations slightly faster than their paralogs, whereas results are inconclusive for *KFB^T* and *CCM^T*.

To clarify whether the elevated d_N/d_S ratios inferred for *CYP^T* and *GLO^T* versus their paralogs were indicative of positive or relaxed selection, we sequenced ten S-morphs of

P. veris to generate polymorphism data and perform McDonald–Kreitman tests (McDonald and Kreitman 1991) using *P. vulgaris* as an outgroup. We found insufficient within-population variation to obtain statistically significant results for the McDonald–Kreitman test ([supplementary table S11, Supplementary Material online](#)). Nevertheless, the low within-population variation of *CYP^T* and *GLO^T* supports the conclusion that these two S-locus genes exhibit only weak or no signs of degeneration, in line with the prediction that hemizygosity should increase efficacy of selection on the S-locus. Altogether, our results imply that S-locus genes do not evolve significantly differently from the rest of the genome, thus corroborating Gutiérrez-Valencia et al. (2021)'s prediction that hemizygosity should slow down the S-locus degeneration that would be expected given suppression of recombination in this region.

Conclusions

We assembled the first chromosome-scale, haplotype-phased genome of any heterostylous species by combining short- and long-read sequencing data under the trio binning approach (Koren et al. 2018) with Chicago and Hi-C scaffolding ([fig. 2](#) and [table 1](#)). The high quality of the *P. veris* haploid assemblies, in combination with comparative analyses of high quality genomes from other Ericales, allowed us to test whether the segmental or stepwise duplication models best describe the build-up of the S locus (Kappel et al. 2017; Huu et al. 2020; [fig. 1](#)). For the first time, we determined that all paralogs of the four duplicated S-locus genes are unlinked in *Primula* and other Ericales, refuting the segmental duplication model ([fig. 4B](#)). Furthermore, we proved that the S-locus genes duplicated asynchronously, confirming the stepwise duplication model ([fig. 4A](#) and [table 2](#)). Using comparative genomic analyses we also revealed that none of the five S-locus genes stemmed from the recent WGD at the base of *Primula* (*Pv- α* ; [fig. 3A](#) and [table 2](#)), and that only the oldest S-locus gene (*KFB^T*) might have duplicated through the older WGD shared by all Ericales (*Ad- β* : 91.0–116.9 Ma).

Finally, we propose the first mechanistic model for the origin of hemizygosity in any supergene with the four *CFB* copies flanking the S locus serving as substrates for the non-homologous recombination causing S-locus deletion from one haplotype ([fig. 5B](#)). The resulting S-locus hemizygosity,

Table 3. Selection on S-Locus Genes and Their Paralogs.

	<i>n</i>	N Substitutions (\pm SD)	S Substitutions (\pm SD)	d_N/d_S (\pm SD)	d_N/d_S P Value
<i>CYP^T</i>	10	126.69 \pm 61.34	130.01 \pm 62.71	0.27 \pm 0.07	0.064
<i>CYP734A51</i>	10	82.98 \pm 39.73	133.13 \pm 59.03	0.18 \pm 0.05	
<i>GLO^T</i>	10	18.54 \pm 9.32	35.56 \pm 18.36	0.16 \pm 0.06	0.014
<i>GLO1</i>	10	14.01 \pm 7.05	32.81 \pm 14.04	0.10 \pm 0.03	
<i>CCM^T</i>	1	2.95	4.05	0.337	NA
<i>CCM1</i>	1	25.18	13.82	0.711	
<i>KFB^T</i>	1	1.80	7.20	0.079	NA
<i>KFB1</i>	1	9.91	10.09	0.349	
<i>PUM^T</i>	1	16.92	8.08	0.848	NA

NOTE.—*n*, number of pairwise comparisons (when 1, only *P. veris* and *P. vulgaris*; when 10: all pairwise combinations of *P. forbesii*, *P. maximowiczii*, *P. oreodoxa*, *P. veris*, *P. vulgaris*); S, synonymous; N, nonsynonymous; SD, standard deviation; *d*, nucleotide substitutions; P values of Wilcoxon matched-pairs signed-rank tests (which do not assume independence between samples) between the d_N/d_S distribution of each S-locus gene and its respective paralog; NA, not available; ^T, S-locus genes.

whereas ensuring the coinheritance of the genes controlling heterostyly, could either increase or decrease the efficacy of selection on the S locus, depending on the strength of contrasting evolutionary processes. Altogether, our results suggest that hemizyosity might effectively counteract the tendency to degeneration potentially caused by suppression of recombination in supergenes. This conclusion is in line with the results of previous simulation analyses showing that degeneration in hemizygous supergenes is weaker than in supergenes where recombination is suppressed via inversions (Gutiérrez-Valencia et al. 2021; table 3).

This is the first study that elucidates the key stages in the build-up of the heterostyly supergene. Thus, it provides a useful resource for future research addressing key questions on the evolution of supergenes in general and the S locus in particular: What is the role of genome architecture in the evolution of complex adaptations? How common is hemizyosity as a mechanism for suppression of recombination in supergenes? Is the genetic composition and molecular architecture of heterostyly supergenes across angiosperms similar to that of *Primula* or not?

Materials and Methods

Plant Material

The individual for the genome assembly was obtained by crossing a short-styled *P. veris* ssp. *veris* (accession: T2DB3; female parent) raised from seeds collected in a natural population in the lake Thun region (Switzerland) with a long-styled *P. veris* ssp. *columnae* (accession: XX-0-Z-20031402; male parent) which was raised from seeds received by the Botanical Garden Jardin Alpin, Meyrin, Switzerland (ex. BG München, Germany). From the F1 population obtained by crossing T2DB3 and P20031402, a short-styled individual (T78) was selected for creating the reference genome. T78 leaf tissue was used for: nanopore sequencing, Illumina sequencing, Chicago, and Hi-C libraries preparation. Ten additional S-morph *P. veris* individuals coming from ten geographical regions were grown from seeds (supplementary fig. S20, Supplementary Material online). The plant material used for RNA-seq experiments is described in supplementary methods, Supplementary Material online. Details on the origins of all the samples used for DNA and RNA sequencing can be found in supplementary tables S12 and S13, Supplementary Material online, respectively.

Genome Size Estimation

The size of the *P. veris* genome was estimated by flow cytometry. A long-styled and a short-styled *P. veris* plants (accessions GR-0-JENA-7758020 and HU-0-Z-20100271, respectively) were measured. We followed a previously published protocol (Temsch et al. 2010), with slight modifications. Briefly, fresh leaf material of each sample was co chopped with a reference (*Solanum pseudocapsicum* 2C = 2.59 pg; Dolezel et al. 1992, 1998; Temsch et al. 2010) in Otto I buffer, the suspension filtrated, mixed with Otto II buffer, digested with RNase, and stained with propidium iodide in the dark at 4 °C for 1–24 h. At least 10,000 nuclei were analyzed on a Cyflow

Space (Sysmex-Partec) flow cytometer. Only nuclei peaks with coefficients of variation below 2% were analyzed.

DNA Isolation, Sequencing, and Genome Assembly

DNA isolation and sequencing are described in supplementary methods, Supplementary Material online. To generate a haplotype-phased genome assembly, we ran the TrioCanu module of the Canu v.1.8 assembler (Koren et al. 2018, 2017) using nanopore reads from T78 and Illumina reads from the two parents. In brief, k-mers specific to each parent were identified in the maternal and paternal Illumina data sets and were then used to sort the nanopore reads from T78 into the maternal and paternal haplotypes. Then, the maternal and paternal nanopore data sets were assembled separately, resulting in two haploid assemblies. An overview of the assembly strategy is schematized in supplementary fig. S1, Supplementary Material online. Further details on genome assembly can be found in supplementary methods, Supplementary Material online.

Repetitive Element Annotation

To identify and classify repetitive elements, we used both de novo and homology-based approaches, generating two repeat annotations, one used to mask the genome assembly for the gene annotation (RepeatModeler annotation), and one used to generate the repeats annotation (GTF files (EDTA annotation)). For the RepeatModeler annotation, the repetitive DNA sequences were identified by running RepeatModeler v1.0.11 (<http://www.repeatmasker.org>; last accessed February 14, 2022) with default parameters and the resulting repeat library was merged with the RepBase (Kapitonov and Jurka 2008) plant library, generating a concatenated library. Finally, RepeatMasker v4.0.9 (<http://www.repeatmasker.org>; last accessed February 14, 2022) was run with default parameters using the concatenated repeats library to annotate the assemblies. For the EDTA annotation, a second repeat library was built for the maternal haplotype assembly using EDTA v1.9.4 (Ou et al. 2019), combining structure- and homology-based approaches for de novo TE identification. Structural discovery of TEs was achieved using LTRharvest (Ellinghaus et al. 2008) and LTR_retriever (Ou and Jiang 2018) for LTR-RTs, TIR-Learner (Su et al. 2019) for TIR transposons, and HelitronScanner (Xiong et al. 2014) for helitrons. Other repetitive elements were identified using RepeatModeler v2.0.1 (Flynn et al. 2020). A filtered nonredundant de novo TE library was produced by concatenating the structurally intact and fragmented elements and were further classified by searching for conserved protein domains using TESorter v.1.2.5 (Zhang et al. 2019). Finally, we used the library to annotate the reference assembly using RepeatMasker v4.0.9 with RM-BLAST as search engine.

RNA Isolation, Sequencing, and Transcriptome Assembly

RNA was isolated in triplicate from six tissues (root, leaf, inflorescence stem, flower, early-germinating seed, and seedling) and in duplicate for one tissue (floral bud), for a total of 20 samples, using the Spectrum Plant Total RNA Kit (Sigma-

Aldrich). Twenty RNA-seq libraries were prepared using the TruSeq Stranded mRNA kit (Illumina). Sequencing was performed at the Functional Genomic Center Zurich on one lane of Illumina HiSeq 4000 (paired-end 2×150 bp) for nine samples, and on an Illumina Novaseq 6000 (paired-end 2×150 bp) for the remaining 11 samples ([supplementary table S18, Supplementary Material online](#)). A de novo transcriptome assembly was generated using Trinity v2.8.4 ([Grabherr et al. 2011](#)) using all the 20 RNA-seq samples together, specifying the use of paired and stranded RNA-seq data (`-SS_lib_type RF`) and with the following trimming parameters: "ILLUMINACLIP : 2:30:10 SLIDINGWINDOW : 4:5 LEADING : 5 TRAILING : 5 MINLEN : 25."

Gene Annotation

The gene annotation was performed using a combination of ab initio and homology-based methods. RNA-seq libraries were mapped to the soft-masked *P. veris* genome assemblies using HISAT2 v2.1.0 ([Kim et al. 2015](#)) and a first round of BRAKER2 v2.1.4 ([Hoff et al. 2019](#)) was run with the raw RNA-seq data to train the gene prediction software GeneMark v4.46 ([Borodovsky and Lomsadze 2011](#)) and AUGUSTUS v3.3.2 ([Stanke et al. 2006](#)). Then we aligned the *P. veris* transcriptome against the genome assemblies using GMAP v2019-09-12 ([Wu and Watanabe 2005](#)) and converted the alignment file (originally in .psl format) into a *hints* file with AUGUSTUS script `blat2hints.pl`. We also aligned the proteomes of the 17 angiosperm species listed in [supplementary table S19, Supplementary Material online](#) against the *P. veris* genome assemblies using GenomeThreader v1.7.1 ([Gremme et al. 2005](#)). The GFF files generated by each alignment were sorted, merged together, and converted into two *hints* files with AUGUSTUS script `align2hint.pl`. These two *hints* files (transcriptome and proteome alignments) were used, together with *hints* files containing information on introns and exons (generated in the first BRAKER round) and repetitive elements (generated by converting the .out file output by RepeatMasker), to run BRAKER2 v2.1.4 in ETP-mode (second BRAKER round). Finally, the gene models included in the S locus were manually curated (see [supplementary methods, Supplementary Material online](#)).

BUSCO v4.0.6 ([Manni et al. 2021](#)) was run on the coding sequences of *P. veris* (maternal and paternal haploid assemblies), *Antirrhinum majus*, *Arabidopsis thaliana*, *C. sinensis*, and *D. oleifera*, using the 2,326 single-copy orthologs from the eudicot database (odb v10) to assess the completeness of gene annotations. We also assessed the percentage of each gene model covered by RNA-seq data, by running the ERE and AnnotationEvidence tools of GeMoMa v1.6.2 ([Keilwagen et al. 2016, 2018; supplementary tables S21 and S22, Supplementary Material online](#)).

Synteny Analyses and WGD Identification

MCScan ([Tang et al. 2008](#)) ([https://github.com/tanghaibao/jcvi/wiki/MCscan-\(Python-version\)](https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)); last accessed February 14, 2022; `-min_size = 5`) was used to identify syntenic regions within the maternal haplotype and between the maternal haplotype and other Ericales. For the K_S plots, the

Comparative Genomics Platform (CoGe) SynMap ([Lyons et al. 2008; Lyons and Freeling 2008; Haug-Baltzell et al. 2017](#)) was used to identify syntenic regions and to calculate K_S between collinear gene pairs. Statistically significant peaks representing WGDs were identified in the K_S distribution of *P. veris* using R scripts (https://github.com/gtiley/Ks_plots; last accessed February 14, 2022; [Tiley et al. 2018](#)) which fitted a mixture of 1–5 normal distributions to the K_S histogram. We discarded the detected peak at $K_S = 3.66$ (SD = 0.20) as the method used is prone to overestimate the number of WGDs for K_S values >3.0 ([Tiley et al. 2018](#)).

Estimating a Neutral Substitution Rate for *P. veris*

To obtain an absolute age for the *Pv- α* WGD, we estimated a neutral substitution rate for *P. veris*. We calibrated our substitution rate estimate using the divergence between *P. veris* and the other non-*Primula* Ericales species included in our study (*A. chinensis*, *C. sinensis*, *D. oleifera*, *R. delavayi*, and *V. corymbosum*), which we estimated to be located at $K_S = 1.2$ ([supplementary fig. S18, Supplementary Material online](#)). Previous studies reported a crown age of 97.6 Ma (95% CI = 90.6–107.2 Ma; [Foster et al. 2017; Rose et al. 2018](#)) for Ericales. We used this divergence time to calculate a 95% CI for the neutral substitution rate with the formula $r = K_S/2t$ (where r is the neutral substitution rate and t is the divergence time expressed in years), and obtained a neutral substitution rate of 6.15×10^{-9} (95% CI = 5.60×10^{-9} – 6.62×10^{-9}) substitutions per synonymous site per year. We note that our estimate for the neutral substitution rate is similar to previously reported values for vascular plants ([Lynch and Conery 2000](#)).

Orthologous Gene Sets Identification and Phylogenetic Analyses

We identified orthologous gene sets by analyzing the proteomes of the 13 angiosperm species listed in [supplementary table S23, Supplementary Material online](#) using OrthoFinder v2.3.11 ([Emms and Kelly 2015, 2019](#)). These species were selected as having highly contiguous, well-annotated genome assemblies and/or based on their phylogenetic proximity to *P. veris*; of these species, only the *P. obconica* proteome was derived from a transcriptome assembly, whereas the proteomes of all the other species were obtained from genome assemblies. A list of all orthogroups is presented in [supplementary table S24, Supplementary Material online](#). The phylogeny presented in [figure 3A](#) was generated by OrthoFinder using STAG ([Emms and Kelly 2018](#)): a species tree was inferred from each of the 6,347 orthogroups containing all 13 species; then a greedy consensus tree was created from all the 6,347 species trees inferred from single orthogroups. The STAG support values consist of the fractions of orthogroup trees supporting each bipartition. The consensus species tree was rooted with STRIDE ([Emms and Kelly 2017](#)). The phylogenetic analyses performed to build S-locus gene trees are described in the [supplementary methods, Supplementary Material online](#).

Population Genetic Analyses

To generate variant files for ten *S*-morph *P. veris* individuals resequenced with Illumina short reads, we used the Genome Analysis Toolkit (GATK) v4.1.2.0 (McKenna et al. 2010), following the best practices recommendations (Van der Auwera et al. 2013). Reads were aligned against the maternal haplotype assembly using BWA-MEM v0.7.17 (Li 2013) with default parameters. The alignment was validated and checked for PCR duplicates using PICARD v2.18.4 (<http://broadinstitute.github.io/picard>; last accessed February 14, 2022). Then GATK HaplotypeCaller (–standard-min-confidence-threshold-for-calling 30; –min-base-quality-score 20; –ERC GVCF), GenotypeGVCFs (–include-non-variant-sites true), and SelectVariants (–select-type SNP; –select-type NO_VARIATION) were used to create a VCF file which included only SNPs and invariant sites. We applied a hard filter on the VCF file (GATK VariantFiltration; QD < 2.0; FS > 60.0; MQ < 40.0; HaplotypeScore > 13.0; MQRankSum < -12.5; ReadPosRankSum < -8.0) and excluded the filtered sites. For each sample, an alternative reference was generated by incorporating the variants into the maternal haplotype using the GATK FastaAlternateReferenceMaker. In the alternative assemblies, the sites which did not pass the hard filter were hard masked; this way, only sites called with high confidence were included in each assembly.

Evolution of *S*-Locus Genes

The coding sequences of all *S*-locus genes and their paralogs were extracted from the ten *S*-morph alternative *P. veris* assemblies using BEDtools fastaFromBed v2.28.0 (Quinlan and Hall 2010). ParaAT v2.0 (Zhang et al. 2012) was used to align paralogous gene pairs with MUSCLE v3.8.31 (Edgar 2004) and to calculate K_a and K_s between them with KaKs_Calculator v2.0 (Wang et al. 2010) with default parameters (–c 1, standard genetic code; –m MA [Model Averaging on a set of candidate models], through which parameters across several models of nucleotide substitution are averaged in order to reduce biases arising from model selection). ParaAT was also used to calculate d_N/d_S for the *S*-locus genes and their paralogs. Sequences of the *S*-locus genes and their paralogs were retrieved from published studies (Li et al. 2016; Huu et al. 2020): for *CYP^T*, *CYP734A51*, *GLO^T*, and *GLO1*, we downloaded sequences from *P. forbesii*, *P. maximowiczii*, *P. oreodoxa*, *P. veris*, and *P. vulgaris*, whereas for the remaining *S*-locus genes and their closest paralogs, only *P. veris* and *P. vulgaris* sequences were available (supplementary table S8, Supplementary Material online).

For the McDonald–Kreitman test (MKT), sequences of *S*-locus genes and their paralogs obtained from ten *S*-morph individuals were obtained as described in the previous paragraph. The MKT was carried out on DnaSp v6.12.03 (Rozas et al. 2017) using *P. vulgaris* sequences as outgroup.

To search for accelerated d_N/d_S in *S*-locus genes compared with their paralogs, we used the clade model of EasyCodeML (Gao et al. 2019), a wrapper of CodeML (PAML; Yang 2007). The model C (CmC), which estimates a separate d_N/d_S for each clade, was compared against the null model 2a_rel

(M2a_rel), which assumes a fixed d_N/d_S among clades. A likelihood-ratio test was performed between the two models.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation program—Marie Skłodowska-Curie (grant number 722338—PlantHUB), the Swiss National Science Foundation (grant numbers 31003A_175556/1; 160004; 131726; and 184826), the Deutsche Forschungsgemeinschaft (German Research Foundation; grant numbers SPP 2237; 440370263; and HI 2076/1-1), the Georges and Antoine Claraz Foundation, the Forschungskredit, and the University Research Priority Program “Evolution in Action” of the University of Zurich.

Author Contributions

G.P., P.S., and E.C. designed the study. P.S. and E.C. coordinated and contributed to all phases of the project. G.P. and B.K. cultivated and harvested the plants. G.P., B.K., and D.D. isolated DNA and RNA. G.P., D.D., and W.P. performed the DNA and RNA sequencing. G.P., E.L.-B., N.Y., R.R.C., and S.I.D. assembled the genomes and contributed to gene annotation and most genomic analyses. R.R.C. annotated repetitive elements. E.L.-B. performed the phylogenetic analyses. G.P., P.S., and E.C. wrote the manuscript, with inputs from all authors.

Data Availability

The sequencing data underlying this article are available in the European Nucleotide Archive (ENA) under the study accession PRJEB44353. The maternal and paternal genome assemblies and respective annotations are available in the CoGe comparative genomics platform (<https://genomeevolution.org/coge/GenomeInfo.pl?gid=61149>; <https://genomeevolution.org/coge/GenomeInfo.pl?gid=61151>) and on FigShare (<https://doi.org/10.6084/m9.figshare.14556075>).

References

- Barrett SCH. 2002. The evolution of plant sexual diversity. *Nat Rev Genet.* 3(4):274–284.
- Barrett SCH. 2019. ‘A most complex marriage arrangement’: recent advances on heterostyly and unresolved questions. *New Phytol.* 224(3):1051–1067.
- Becher H, Jackson BC, Charlesworth B. 2020. Patterns of genetic variability in genomic regions with low rates of recombination. *Curr Biol.* 30(1):94–100.e3.
- Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. 2012. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58(3):268–276.
- Borodovsky M, Lomsadze A. 2011. Eukaryotic gene prediction using GeneMark-hmm-E and GeneMark-ES. *Curr Protoc Bioinforma.* 35:4.6.1–4.6.10.
- Branco S, Carpentier F, Rodríguez de la Vega RC, Badouin H, Snirc A, Le Prieur S, Coelho MA, de Vienne DM, Hartmann FE, Begerow D, et al. 2018. Multiple convergent supergene evolution events in mating-type chromosomes. *Nat Commun.* 9(1):2000.

- Burrows BA, McCubbin AG. 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(1):53–67.
- Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci.* 355(1403):1563–1572.
- Charlesworth D. 2016. The status of supergenes in the 21st century: recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. *Evol Appl.* 9(1):74–90.
- Cocker JM, Wright J, Li J, Swarbreck D, Dyer S, Caccamo M, Gilmartin PM. 2018. *Primula vulgaris* (primrose) genome assembly, annotation and gene expression, with comparative genomics on the heterostyly supergene. *Sci Rep.* 8(1):17942.
- Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol.* 13(4):e1002112.
- Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet.* 14(4):262–274.
- Darlington CD, Mather K. 1949. The elements of genetics. London: George Allen & Unwin.
- Darwin C. 1877. The different forms of flowers on plants of the same species. London: Murray.
- de Vos JM, Hughes CE, Schneeweiss GM, Moore BR, Conti E. 2014. Heterostyly accelerates diversification via reduced extinction in primroses. *Proc Biol Sci.* 281(1784):20140075.
- Dolezel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R. 1998. Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann Bot.* 82:17–26.
- Dolezel J, Sgorbati S, Lucretti S. 1992. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol Plant.* 85(4):625–631.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5):1792–1797.
- Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics.* 9:18.
- Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157.
- Emms DM, Kelly S. 2017. STRIDE: species tree root inference from gene duplication events. *Mol Biol Evol.* 34(12):3267–3278.
- Emms DM, Kelly S. 2018. STAG: species tree inference from all genes. *bioRxiv.* 267914.doi:10.1101/267914.[TQ6]
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20(1):238.
- Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 117(17):9451–9457.
- Foster CSP, Sauquet H, van der Merwe M, McPherson H, Rossetto M, Ho SYW. 2017. Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Syst Biol.* 66(3):338–351.
- Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW. 2019. EasyCodeML: a visual tool for analysis of selection using CodeML. *Ecol Evol.* 9(7):3891–3898.
- Gilmartin PM. 2015. On the origins of observations of heterostyly in *Primula*. *New Phytol.* 208(1):39–51.
- Goel M, Sun H, Jiao WB, Schneeburger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 20(1):277.
- Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective population size within a genome. *Genetics* 189(4):1389–1402.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, 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(7):644–652.
- Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. *Inform Soft Technol.* 47(15):965–978.
- Gutiérrez-Valencia J, Hughes PW, Berdan EL, Slotte T. 2021. The genomic architecture and evolutionary fates of supergenes. *Genome Biol. Evol.* 13:evab057.
- Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E. 2017. SynMap2 and SynMap3D: web-based whole-genome synteny browsers Hancock. *Bioinformatics* 33(14):2197–2198.
- Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER. *Methods Mol Biol.* 1962:65–95.
- Huu CN, Kappel C, Keller B, Sicard A, Takebayashi Y, Breuninger H, Nowak MD, Bäurle I, Himmelbach A, Burkart M, et al. 2016. Presence versus absence of CYP734A50 underlies the style-length dimorphism in primroses. *Elife* 5:1–15.
- Huu CN, Keller B, Conti E, Kappel C, Lenhard M. 2020. Supergene evolution via stepwise duplications and neofunctionalization of a floral-organ identity gene. *Proc Natl Acad Sci U S A.* 117(37):23148–23157.
- Jay P, Whibley A, Frézal L, Rodríguez de Cara MÁ, Nowell RW, Mallet J, Dasmahapatra KK, Joron M. 2018. Supergene evolution triggered by the introgression of a chromosomal inversion. *Curr Biol.* 28(11):1839–1845.e3.
- Jiao WB, Schneeburger K. 2020. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun.* 11:1–10.
- Kapitonov VV, Jurka J. 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet.* 9(5):411–412.
- Kappel C, Huu CN, Lenhard M. 2017. A short story gets longer: recent insights into the molecular basis of heterostyly. *J Exp Bot.* 68(21–22):5719–5730.
- Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. 2018. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics.* 19(1):189.
- Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. 2016. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* 44(9):e89.
- Kejnovsky E, Hawkins JS, Feschotte C. 2012. Plant transposable elements: biology and evolution. In: Wendel JF, editor. Plant genome diversity volume 1: plant genomes, their residents, and their evolutionary dynamics. Vienna: Springer-Verlag. p. 17–34.
- Keller B, Thomson JD, Conti E. 2014. Heterostyly promotes disassortative pollination and reduces sexual interference in Darwin's primroses: evidence from experimental studies. *Funct Ecol.* 28(6):1413–1425.
- Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 12(4):357–360.
- Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM, et al. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol.* 36(12):1174–1182.
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27(5):722–736.
- Kotani Y, Henderson ST, Suzuki G, Johnson SD, Okada T, Siddons H, Mukai Y, Koltunow AMG. 2014. The LOSS OF APOMEIOSIS (LOA) locus in *Hieracium praealtum* can function independently of the associated large-scale repetitive chromosomal structure. *New Phytol.* 201(3):973–981.
- Lamichhaney S, Fan G, Widemo F, Gunnarsson U, Thalmann DS, Hoepfner MP, Kerje S, Gustafson U, Shi C, Zhang H, et al. 2016. Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat Genet.* 48(1):84–88.
- Larson DA, Walker JF, Vargas OM, Smith SA. 2020. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. *Am J Bot.* 107(5):773–789.
- Li B, Bickel RD, Parker BJ, Saleh Ziabari O, Liu F, Vellichirammal NN, Simon J-C, Stern DL, Brisson JA. 2020. A large genomic insertion containing a duplicated follistatin gene is linked to the pea aphid male wing dimorphism. *Elife* 9:e50608.

- Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [cited 2019 Dec 2]. Available from: <http://arxiv.org/abs/1303.3997>.
- Li J, Cocker JM, Wright J, Webster MA, McMullan M, Dyer S, Swarbreck D, Caccamo M, Oosterhout C. V, Gilmartin PM, et al. 2016. Genetic architecture and evolution of the S locus supergene in *Primula vulgaris*. *Nat Plants*. 2(12):16188.
- Li J, Webster MA, Wright J, Cocker JM, Smith MC, Badakshi F, Heslop-Harrison P, Gilmartin PM. 2015. Integration of genetic and physical maps of the *Primula vulgaris* S locus and localization by chromosome in situ hybridization. *New Phytol*. 208(1):137–148.
- Llaurens V, Whibley A, Joron M. 2017. Genetic architecture and balancing selection: the life and death of differentiated variants. *Mol Ecol*. 26(9):2430–2448.
- Lloyd DG, Webb CJ. 1992. The Selection of heterostyly. In: Barrett SCH, editor. Evolution and function of heterostyly. Monographs on theoretical and applied genetics. Berlin, Heidelberg: Springer. p. 179–207.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.
- Lyons E, Freeling M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J*. 53(4):661–673.
- Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, et al. 2008. Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol*. 148(4):1772–1781.
- Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 38(10):4647–4654.
- Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 33(4):574–576.
- Mast AR, Kelso S, Conti E. 2006. Are any primroses (*Primula*) primitively monomorphic? *New Phytol*. 171(3):605–616.
- McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351(6328):652–654.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297–1303.
- Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol Evol*. 35(7):561–572.
- Nowak MD, Russo G, Schlapbach R, Huu CN, Lenhard M, Conti E. 2015. The draft genome of *Primula veris* yields insights into the molecular basis of heterostyly. *Genome Biol*. 16:12.
- Okada T, Ito K, Johnson SD, Oelkers K, Suzuki G, Houben A, Mukai Y, Koltunow AM. 2011. Chromosomes carrying meiotic avoidance loci in three apomictic eudicot *Hieracium* subgenus *Pilosella* species share structural features with two monocot apomicts. *Plant Physiol*. 157(3):1327–1341.
- Ou S, Jiang N. 2018. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol*. 176(2):1410–1422.
- Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 20(1):275.
- Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of gene duplication in plants. *Plant Physiol*. 171(4):2294–2316.
- Purcell J, Brelsford A, Wurm Y, Perrin N, Chapuisat M. 2014. Convergent genetic architecture underlies social organization in ants. *Curr Biol*. 24(22):2728–2732.
- Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res*. 26(3):342–350.
- Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol*. 20(1):38.
- Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841–842.
- Ren R, Wang H, Guo C, Zhang N, Zeng L, Chen Y, Ma H, Qi J. 2018. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol Plant*. 11(3):414–428.
- Rose JP, Kleist TJ, Löfstrand SD, Drew BT, Schönenberger J, Sytsma KJ. 2018. Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. *Mol Phylogenet Evol*. 122:59–79.
- Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol*. 34(12):3299–3302.
- Schwander T, Libbrecht R, Keller L. 2014. Supergenes and complex phenotypes. *Curr Biol*. 24:288–294.
- Shi T, Huang H, Barker MS. 2010. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Ann Bot*. 106(3):497–504.
- Shivanna KR, Heslop-Harrison J, Heslop-Harrison Y. 1981. Heterostyly in *Primula*. 2. Sites of pollen inhibition, and effects of pistil constituents on compatible and incompatible pollen-tube growth. *Protoplasma* 107(3–4):319–337.
- Stanke M, Tzvetkova A, Morgenstern B. 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol*. 7 (Suppl 1):S11.
- Su W, Gu X, Peterson T. 2019. TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Mol Plant*. 12(3):447–460.
- Sun Y, Svedberg J, Hiltunen M, Corcoran P, Johannesson H. 2017. Large-scale suppression of recombination predates genomic rearrangements in *Neurospora tetrasperma*. *Nat Commun*. 8(1):1140.
- Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science* 320(5875):486–488.
- Temsch EM, Temsch W, Ehrendorfer-Schratt L, Greilhuber J. 2010. Heavy metal pollution, selection, and genome size: the species of the Zerjav study revisited with flow cytometry. *J. Bot*. 2010:1.
- Thompson MJ, Jiggins CD. 2014. Supergenes and their role in evolution. *Heredity* 113(1):1–8.
- Tiley GP, Barker MS, Gordon Burleigh J. 2018. Assessing the performance of KS plots for detecting ancient whole genome duplications. *Genome Biol Evol*. 10(11):2882–2898.
- Turner JRC. 1967. On supergenes. I. The evolution of supergenes. *Am Nat*. 101:195–221.
- Tuttle EM, Bergland AO, Korody ML, Brewer MS, Newhouse DJ, Minx P, Stager M, Betuel A, Cheviron ZA, Warren WC, et al. 2016. Divergence and functional degradation of a sex chromosome-like supergene. *Curr Biol*. 26(3):344–350.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 43:11.10.1–11.10.33.
- Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics*. 8(1):77–80.
- Wang J, Wurm Y, Nipitwattanaphon M, Riba-Grognuz O, Huang Y-C, Shoemaker D, Keller L. 2013. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* 493(7434):664–668.
- Wang Y, Nie F, Shahid MQ, Baloch FS. 2020. Molecular footprints of selection effects and whole genome duplication (WGD) events in three blueberry species: detected by transcriptome dataset. *BMC Plant Biol*. 20(1):14.

- Wellenreuther M, Bernatchez L. 2018. Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol Evol.* 33(6):427–440.
- Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859–1875.
- Xiong W, He L, Lai J, Dooner HK, Du C. 2014. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc Natl Acad Sci U S A.* 111(28):10263–10268.
- Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol.* 24(8):1586–1591.
- Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc Natl Acad Sci U S A.* 110(19):E1743–E1751.
- Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L. 2012. ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem Biophys Res Commun.* 419(4):779–781.
- Zhang RG, Wang ZX, Ou S, Li GY. 2019. TEsorter: lineage-level classification of transposable elements using conserved protein domains. *bioRxiv.* 800177. doi: 10.1101/800177.