Title:

Evolution of flower color pattern through selection on regulatory small RNAs

Authors:

Desmond Bradley¹, Ping Xu², Irina-Ioana Mohorianu³, Annabel Whibley¹, David Field⁴,⁵, Hugo Tavares¹,⁹, Matthew Couchman¹, Lucy Copsey¹, Rosemary Carpenter¹, Miaomiao Li⁶,⁷, Qun Li⁶, Yongbiao Xue⁶,⁷,⁸, Tamas Dalmay²*, Enrico Coen¹*

Affiliations:

¹ Department of Cell and Developmental Biology, John Innes Center, Colney Lane, Norwich NR4 7UH, UK.

² School of Biological Sciences, University of East Anglia, Norwich, NR4 7TJ.
3 School of Biological Sciences and School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ.

4 Department of Botany and Biodiversity Research, University of Vienna, Faculty of Life Sciences, Rennweg 14, A-1030 Vienna, Austria.

5 IST Austria, Klosterneuburg, Austria.

6 State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences and National Center for Plant Gene Research, Beijing 100101, China.

7 University of Chinese Academy of Sciences, Beijing 100190, China.

8 Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.

9 Current address: Sainsbury Laboratory, University of Cambridge, Cambridge CB2 1LR, UK.

*Corresponding authors

Abstract:
Small RNAs regulate genes in plants and animals. Here we show that population-wide differences in color patterns in snapdragon flowers are caused by an inverted duplication that generates small RNAs. The complexity and size of the transcripts indicate the duplication represents an intermediate on the pathway to microRNA evolution. The small RNAs repress a pigment biosynthesis gene, creating a yellow highlight at the site of pollinator entry. The inverted duplication exhibits steep clines in allele frequency in a natural hybrid zone, showing that the allele is under selection. Thus, regulatory interactions of evolutionarily recent small RNAs can be acted upon by selection and contribute to the evolution of phenotypic diversity.

One Sentence Summary:

Selection acting on an inverted duplication that generates small RNAs leads to evolution of regulatory interactions and phenotypic change.

LENGTH

Main Text:

A convenient system for studying selection in natural populations is afforded by hybrid zones, where closely-related species or populations come into contact (1). Such a hybrid zone has been described for two subspecies of Antirrhinum majus (snapdragon), that
differ in flower color (2), a trait involved in pollinator attraction (3-7). Both subspecies are pollinated by bees but have alternate patterns for guiding flower entry:

*A.m.pseudomajus* flowers are magenta, with a patch of yellow highlighting the bee entry point (Fig. 1A), whereas *A.m.striatum* flowers are yellow with magenta veins at the entry point (Fig. 1B). The magenta and yellow flower color intensities show sharp clines at a hybrid zone (2) where the subspecies come into contact. Production of magenta is regulated by *ROSEA (ROS)* and *ELUTA (EL)* (8-10). *ROS* encodes a MYB-like transcription factor that promotes anthocyanin biosynthetic gene expression in *A.m.pseudomajus* and exhibits a steep cline in allele frequencies at the hybrid zone (2, 9). Distribution of yellow pigment is regulated by *SULF* (Fig. 1B, C), which represses production of the yellow flavonoid aurone in *A.m.pseudomajus* (Fig. 1D) (2, 9, 10). Here we study the molecular nature of *SULF*.

To isolate *SULF* we first mapped it to an interval of ~3Mb on chromosome 4 by sequencing pools of *sulf* and *SULF* phenotypes from a segregating population (fig. S1). In parallel, we carried out a transposon mutagenesis experiment in *A. majus* (*SULF*) and isolated a mutant, *sulf-660*, that was both somatically and genetically unstable (fig. S2A and Methods). Comparing the genome sequence of *sulf-660* and its revertants revealed a single insertion site, within the mapped region of *SULF*, specific to *sulf-660*. Three independent revertants had different excision footprints at this site, confirming that the transposon was responsible for the *sulf* phenotype (fig. S2B).
BLAST searches of the sequence flanking the transposon insertion site revealed regions of 74-88% nucleotide sequence identity to *A.majus* chalcone 4’-O-gluosyltransferase (*Am4’CGT*), which encodes an enzyme involved in synthesis of the yellow pigment aurone (Fig. 2A and table S1) (11). The regions of *Am4’CGT* homology were organized as an inverted duplication in the *A. majus* SULF genome. Both the left and right arms of the duplication carried deletions relative to intact *Am4’CGT*, suggesting they had independently degenerated from a more complete precursor. A contiguous region of inverted homology between the left and right arms spanned a ~590 bp region (red arrows, Fig. 2A), separated by a ~600 bp spacer region, which contained the transposon insertion site of *sulf-660*. Phylogenetic analysis indicated that the SULF inverted repeats were likely generated from *Am4’CGT* recently in the evolution of the *Antirrhinum* lineage (Fig. 2B and fig. S3).

To determine whether the inverted duplication at SULF might be under selection, we compared *A.m.pseudomajus* and *A.m.striatus* populations sampled either side of a hybrid zone. PCR using oligos flanking the inverted repeats gave bands in the range 1.5-2.5 kb for all individuals from the *A.m.pseudomajus* (n=96) but not the *A.m.striatus* populations (n=95), suggesting that the inverted duplication was present at higher frequency in *A.m.pseudomajus* (fig. S4). Sequencing pools of ~50 individuals from each population revealed reduced depth of sequence for *A.m.striatus* compared to *A.m.pseudomajus* over a ~145 kb region around SULF, suggesting that *A.m.striatus* carried deletions relative to *A.m.pseudomajus* in this chromosome region (Fig. 2C).
This conclusion was supported by PCR amplification assays using a range of oligos. Deletion alleles were also observed in resequenced individuals, including a 1.3kb deletion that removed the left arm of the inverted repeat and part of the spacer sequence in *A.m. striatum* pools. Thus, the inverted duplication present in *SULF* of *A.m.pseudomajus* is absent or at low frequency in *A.m.striatum* populations, further demonstrating the requirement for the inverted duplication for *SULF* function.

SNPs in a ~300kb interval containing *SULF*, showed steep clines in allele frequency (Fig. 2D and fig. S5), centered at the same geographic location as clines for *ROS* and flower color (2). SNPs sampled from other positions along chromosome 4 either showed no clines, or clines centered at different geographic locations (Fig. 2D and fig. S5). The significance of the clines at *SULF* was confirmed by comparing DNA sequences from pools of individuals sampled from a transect covering ~20km either side of the hybrid zone. Of the ~7x10^5 polymorphic SNPs on the *SULF* chromosome, 99% showed no allele frequency differences across the transect, and of those that did, more than 99% did not give steep clines aligned with *ROS*. Thus, there is likely to be strong selection acting on *SULF*.

The coincidence of the *SULF* and *ROS* clines suggests that these loci interact. In *A.m. pseudomajus*, where *ROS* confers magenta color, *SULF* could be favored because it restricts yellow to create a contrasting highlight at the bee entry point (Fig. 1A). In *A.m.striatum*, where *ros* confers reduced magenta intensity for much of the flower, *sulf* could be favored because it confers both a striking yellow color and a contrasting...
background to the magenta veins (Fig. 1B). Thus, selection acting on different allele combinations at $SULF$ and $ROS$ allows alternate floral guides to be maintained either side of a hybrid zone. The situation is comparable to selection acting on loci controlling yellow and red coloration of mimetic patterns in *Heliconius* butterflies (12, 13).

Given the structure of the inverted duplication at $SULF$ and its homology to $Am4^{'CGT}$, we hypothesized that $SULF$ represses $Am4^{'CGT}$, and thus restricts yellow flower color, via regulatory small RNAs. To determine whether $SULF$ generated small RNAs, small RNA libraries were prepared from petals of *A.majus* $SULF$ and *sulf*-660. The biggest differences in small RNA abundance mapped to the $SULF$ inverted repeats and corresponded to predominantly 21-mers (Fig. 3A, B). RNA blots probed with $SULF$ confirmed small RNAs from the inverted repeat were present in $SULF$ and absent in *sulf* genotypes, including *A.m.striatum* (Fig. 3C and fig. S6). The small RNAs likely derive from processed transcripts predicted to generate long foldback hairpin RNAs (fig. S7).

If the small RNAs generated by $SULF$ restrict yellow pigmentation by targeting $Am4^{'CGT}$, then $SULF$ and $Am4^{'CGT}$ should exhibit complementary expression patterns. Analysis of RNA extracted from yellow and non-yellow regions of the petals of *A.majus* showed that $SULF$ was preferentially expressed in the non-yellow region, whereas $Am4^{'CGT}$ was mainly expressed in the yellow region (Fig. 3D). The spatial restriction of $Am4^{'CGT}$ was confirmed by RNA in situ hybridization (Fig. 3E).

Overall expression of $Am4^{'CGT}$ was lower in petals of *SULF* compared to *sulf*-660 (Fig. 3F). 5’ RACE on $SULF$ genotypes revealed products for $Am4^{'CGT}$ terminating at a
range of positions, suggesting cleavage at multiple sites (fig. S8). *No cleavage products were found in sulf.* The lack of a single cleavage site in *SULF* genotypes is consistent with the *SULF* inverted duplication generating multiple small RNAs targeting *Am4’CGT* (Fig. 3B). To determine whether *SULF* alleles from the subspecies also varied in their ability to repress *Am4’CGT*, we introgressed *SULF* from *A.m.pseudomajus* (*SULF*<sup>p</sup>) or *A.m.striatum* (*sulfs*) into an *A.majus* background with the same *Am4’CGT* target allele. *Am4’CGT* expression was reduced in both dorsal and ventral petals of *SULF*<sup>p</sup> compared to *sulfs* (Fig. 3F). Thus, *SULF* acts by repressing transcript levels of the target *Am4’CGT* gene in *A.m.pseudomajus* but not in *A.m.striatum*.

If selection on inverted duplications is a common mechanism for establishing regulatory interactions, we might expect the genome to contain a large number of inverted duplications similar to *SULF*. Scanning the *A.majus* genome for inverted duplications with a similar adjusted folding energy to *SULF* revealed many such regions, some of which generated small RNAs (Fig. 4A). However, most of these small RNAs were >21nt long, unlike those generated by *SULF* (circled, Fig. 4B), which were ~21nt. Moreover, the small RNA population generated by *SULF* was of relatively low complexity because of the high abundance of a subset of small RNAs. Based on size and complexity, the profile of small RNAs generated by *SULF* was similar to that of conserved microRNAs (orange spots, Fig. 4B). Given the *SULF* hairpin is about five times longer than a typical conserved microRNA hairpin, these findings suggest that *SULF* generates a functioning long regulatory hairpin RNA.
If only a subset of small RNAs generated by *SULF* are required to inhibit target gene activity, selection would not be able to maintain homology with the target gene *Am4'CGT* over the extended length observed (590bp). This argument implies that *SULF* is of recent evolutionary origin, consistent with the phylogenetic analysis (Fig. 2B). With respect to its young age, *SULF* is similar to other inverted duplications with extended similarity to protein coding regions that encode small RNAs (14-17). Over evolutionary time, functional inverted duplications such as *SULF* might be lost, maintained, or become shorter microRNA hairpins (14, 15, 18-21). The deletions observed in both the left and right arms of the inverted repeat at *SULF*, relative to *Am4'CGT* (Fig. 2A), suggest the process of size reduction may have already occurred to some extent.

Among the many documented cases of loci contributing to natural variation (22), several examples of small regulatory RNAs have been described (23-26). However, these examples involve changes in expression pattern of pre-existing micro-RNAs or creation of new target sites, rather than *de novo* generation of a small regulatory RNA, as observed with *SULF*. The unusual nature of *SULF* may be a matter of chance or may reflect constraints on regulatory mechanisms (27). For example, the biosynthetic pathway to yellow aurone pigment synthesis has fewer steps and has a more limited taxonomic distribution than the magenta anthocyanin pigment synthesis pathway (11, 28). Variation in transcription factors, such as *ROS*, may therefore not be available specifically to modulate yellow patterning. Inverted duplications that generate regulatory RNAs may thus provide a flexible mechanism, complementing that based on transcription factor or *cis*-regulatory variation (22), for modulating or creating novel
expression patterns upon which natural selection may act to generate evolutionary change.

References and Notes


**Acknowledgments**

The sRNA-seq data presented in this study is publicly available on Gene Expression Omnibus56, under accession number GSE91378. Datasets for genomic DNAs are available at the European Nucleotide Archive accession number PRJEB22668 and scripts at linked sites. The authors have no competing interests. We thank Maria-Elena Mannarelli for technical support, Nick Barton for suggestions on the manuscript and Alexandra Rebocho for helpful discussions. This work was supported by BBSRC grant BB/G009325/1 awarded to E.C. Supplement contains additional data.

**Supplementary Materials**

Materials and Methods

Figs. S1 to S8

Table S1

References (30-58)
Fig. 1. Flower color pattern phenotypes

Flower face (left) and side (right) views of *A.majus (A.m.)* species, showing lower ventral (V), lateral (L), and upper dorsal (D) lobes. Bee vision is sensitive to both yellow and the blue component of magenta reflectance. (A) *A.m.pseudomajus*. Magenta with yellow highlight. White arrows indicate bee entry point. (B) *A.m.striatum*. Yellow with magenta highlights. (C) Flowers from plants with *ros EL* from *A.m.striatum (rosS ELs)* and *SULF* from *A.m.pseudomajus (SULFP)*. (D) Schematic showing the pathways to anthocyanin and aurone pigments; chalcone synthase (CHS), chalcone isomerase (CHI), *A.m.* chalcone 4’-O-glucosyltransferase (Am4’CGT) and *A.m.* aureusidin synthase (AmAS1).

Fig. 2. *SULF* locus shows homology to *Am4’CGT* and signatures of selection

(A) *SULF* inverted duplication. Organisation of *Am4’CGT* is shown twice (grey arrows) to indicate regions of homology with *SULF* (CDS = coding sequence). The left and right inverted repeats at *SULF* (red arrows) flank the transposon insertion site of *sulf-660* (black triangle).

(B) Maximum likelihood phylogeny of CGT-related DNA sequences from *Antirrhinum majus* (red), *Mimulus guttatus* (black) and *Linaria vulgaris* (blue). Bootstrap support for nodes with >85% support (red circles, scaled by strength). For extended clade see fig. S2.

(C) Plot of *A.m.striatum* sequence coverage normalized against *A.m.pseudomajus* for pools located at either end of the hybrid zone. Bars indicate genes, with *SULF* locus in red. Double-headed arrow shows region under-represented in *A.m.striatum*. Positions of KASP SNPs used for cline analysis (blue dots).
(D) Clines for KASP markers across the hybrid zone transect. SNP index and chromosome position is indicated above each plot. Markers from SULF show steep clines at the hybrid zone, aligned with clines for ROSI (right). Markers further away from SULF either show no clines (two examples shown), or clines centered at other geographic locations (fig. S4).

Fig. 3. SULF locus makes small RNAs targeting Am4’CGT.

(A) Comparison of total read abundance for small RNAs isolated from libraries of sulf-660 and SULF-661. Small RNAs mapping to the SULF locus in red.

(B) Abundance of small RNAs mapping to SULF from the SULF-661 libraries. Reads with potential to target Am4’CGT (red) and those unable to target (too many mismatches) (grey).

(C) Blot of petal RNA probed with an oligo matching one of the abundant 21-mers, showing signal in ventral and lateral (VL) or dorsal (D) petals in SULF-661 but not sulf-660. U6 = ubiquitin control.

(D) Complementary expression pattern of SULF small RNAs and Am4’CGT expression. Petals (left panel) were dissected into a central (C) yellow region, and peripheral (P) non-yellow region. For SULF expression, small RNA blots were probed with SULF, revealing stronger expression in the peripheral compared to central region (middle panel). For Am4’CGT, RNA was subject to qRT-PCR showing lower expression in the peripheral region (right panel).

(E) Floral bud of A.majus was sectioned to reveal the pigments (top panel), and similar sections probed to show in situ expression of Am4’CGT (purple stain, bottom panel).
(F) q-RT-PCR on petal RNA (total, or dissected into upper and lower regions).

Expression of \textit{Am4'CGT} is reduced in genotypes carrying \textit{SULF} from \textit{A.majus} (\textit{SULF}^M) or \textit{A.m.pseudomajus} (\textit{SULF}^P) compared to those carrying \textit{sulf} from \textit{A.majus} (\textit{sulf}^M) or \textit{A.m.striatum} (\textit{sulf}^S). Standard errors calculated from the means of 3 independent biological samples, each analyzed in triplicate.

**Fig. 4. Expression and frequency distribution of inverted repeats and microRNA genes in \textit{Antirrhinum majus}**

\textbf{(A)} Frequency and expression levels of inverted repeats with folding energies similar to \textit{SULF}, as a function of length of predicted hairpin RNA (including spacer). An inverted repeat is considered expressed if the maximum overall abundance of incident sRNAs, in any library is above a noise threshold (20). Boxed region shows class to which \textit{SULF} belongs.

\textbf{(B)} Average complexity and mean length of small RNAs mapping to inverted repeats (as in \textbf{A}) and microRNA hairpins. Each point corresponds to a predicted transcript with a hairpin-like structure. \textit{SULF} hairpin circled in red. Only sRNAs in the 21-24nt range are considered. Average complexity is the number of different reads (non-redundant) divided by the total number of reads mapping to the hairpin (29). Although \textit{SULF} generates small RNAs throughout the inverted repeats, the high abundance of some leads to a low overall complexity. For inverted repeats, transcript abundance is color coded on a log scale and varies from blue (low abundance, 20) to red (high abundance,160,000). Orange indicates microRNA hairpins.
A. *A. m. pseudomajus* 
ROS^{p} el^{p} SULF^{p}

B. *A. m. striatum* 
ros^{s} EL^{s} sulf^{s}

C. *ros^{s} EL^{s} SULF^{p}*

D. 1 p-coumaryl-CoA + 3 malonyl-CoA

\[
\begin{align*}
\text{CHS} & \quad \rightarrow & \quad \text{CHI} \\
\text{Am 4'CGT} & \quad \rightarrow & \quad \text{Anthocyanins} \\
\text{Am AS1} & \quad \rightarrow & \quad \text{Aurones}
\end{align*}
\]
Supplementary Materials for

Evolution of flower color pattern through selection on regulatory small RNAs

Desmond Bradley, Ping Xu, Irina-Ioana Mohorianu, Annabel Whibley, David Field, Hugo Tavares, Matthew Couchman, Lucy Copsey, Rosemary Carpenter, Miaomiao Li, Qun Li, Yongbiao Xue, Tamas Dalmay*, Enrico Coen*

*Corresponding authors. Email: enrico.coen@jic.ac.uk ; t.dalmay@uea.ac.uk

This PDF file includes:

Materials and Methods
Figs. S1 to S8
Tables S1
Materials and Methods

Plant Material

Populations flanking a hybrid zone in the Pyrenees (2) provided seed for maintaining lines of *A.m.pseudomajus ROS*<sup>P</sup> *SULFP* (from Ventola) and *A.m.striatum ros*<sup>S</sup> *sulf*<sup>S</sup> (from La Molina). Greenhouse plants were grown as described (1), with supplemental lights in winter to give 12-16h days or grown outside in summer upon benches.

*Antirrhinum majus* stock lines are highly inbred and maintained at the John Innes (JI) Centre. Stocks JI7 and JI75 are *INC SULFP*, while JI57 is *inc sulf*. The *inc* mutation reduces magenta (anthocyanin) pigments in the flower (2), while *sulf* results in spread yellow through the petal lobes (3, 4).

JI7 has been sequenced to give a chromosome-build reference genome (BGI, Beijing; http://bioinfo.sibs.ac.cn/am). JI75 has highly active endogenous transposons to generate mutants (5-7). JI660 *sulf* arose from a large-scale mutagenesis experiment using JI75, color and JI661 *SULFP* revertant arose from *sulf-660* due to transposon instability.

*A.majus* crosses to species

*A.majus* JI7 was crossed with *A.m.pseudomajus* J1428 progeny; seed of J1428 was collected at 4,737.5m east of the HZ center (2). F1 plants were self-pollinated due to self-compatible *A.majus* alleles to give family J108. KASP genotyping (see below) of the J108 family of 245 plants identified 15 plants with genotype *SULFP Am4'CGTM*. *A.majus* JI7 was crossed with *A.m. striatum* J1160 progeny; seed of J1160 was collected at 12,364.2m west of the HZ center. F1 plants were self-pollinated to give family J106. KASP genotyping of the J106 family of 249 plants identified 23 plants with genotype *sulf Am4'CGTM*.

Genotyping Crosses

The KASP method (LGC Genomics) was used to identify *SULFP* and *Am4'CGTM* for introgression in F2 families J106 and J108 using the following oligo sets:

*Am4'CGT* oligo set: do.238- GTTGAGATGCCCGGGTTCCCATTG (to detect P or S allele on Chr2:71834391-71834413); do.239- GTTGAGATGCCGGGTCCCGTTCCATTA (to detect M allele); do.240- ACTCATTGGTAAATCAGAGGAGTG (the common reverse oligo on Chr2:71834418-71834441).

*SULFP* oligo set: do.256- CCACTCCGCGACCCATTGAGCT (to detect M allele on Chr4:34930351-34930372); do.257- CCACTCCGCGACCCATGGAGCT (to detect P or S allele); do.258- TCGATGGGATGATGATGTTAATGG (the common reverse oligo on Chr4:34930324-34930347).

KASP Genotyping was performed as described (LGC) and the fluorescence signals discriminating the two alleles was detected in a BioRad CFX96 lightcycler and data processed with the BioRad CFX Manager software v3.1.

Leaf genomic DNA Isolation.

3-6 small young leaves (~1cm long; total 100-200mg) were collected in eppendorf tubes and frozen in liquid nitrogen. Large numbers were collected in 96 tube format on
dry ice or placed in glassine seed bags and dried in silica for long term storage at room temperature. Wild samples were collected in bags in silica gel.

Genomic DNA minipreps used the DNeasy Plant Mini Kit or DNeasy 96 Plant Kit according to instructions (Qiagen). For WGS, 2-5g of leaves were collected on dry ice and isolated by CTAB preps as described (8).

Illumina resequencing

Full details of the analysis pipelines and associated scripts are available here (https://github.com/JIC-CSB/antirrhinum-hz-pipeline). In short, paired end illumina reads were filtered and trimmed for quality using fastq-mcf (ea-utils v.1.1.2-484) or Trimmomatic(v0.32) and mapped to the JI7 reference genome (vIGBDV1) using Stampy (v.1.0.28) All default settings were used except the setting of an explicit substitution rate to account for expected divergence from the reference (--substitutionrate 0.02 in A. majus and A. majus subspecies pools, 0.05 in A. majus x A. sempervirens cross). Alignment file manipulations used SAMtools v1.3.0. After mapping, duplicate reads were excluded using the MarkDuplicates tool in Picard (v1.134; http://broadinstitute.github.io/picard) and local indel realignment using IndelRealigner was performed with GATK(v3.5.0).

SULF Isolation

Mapping SULF

We mapped the sulf locus using a sulf allele from A.sempervirens crossed to A.majus JI7. This population provided clear segregating phenotypes and a large number of useful SNPs for mapping. A.sempervirens Accession AC 1170 was backcrossed to A.majus JI7 three times before self-pollinating and generating a population of sulf/sulf and SULF/- plants. Genomic DNA from pools of 35 sulf/sulf and 35 SULF/- individuals were resequenced and reads mapped to the reference genome. For analysis of homozygous SNP density, allele counts for reference (JI7) and alternate (A. sempervirens) alleles were exported using SAMtools mpileup with the following settings: -q 40 -Q 30 -BA -t AD. BCFtools (v1.3.1) was used to export a vcf of all variable sites, indels excluded, which was converted to table form using GATK VariantsToTable. Filtering thresholds were set following review of empirical distributions. We removed sites with depth <10 or >100 in either pool, sites that had a frequency of <10% for the A. sempervirens allele in either pool and sites that were likely fixed homozygous for the alternate (A. sempervirens) allele. A 100kb window analysis with a 50kb step was applied genome-wide. The Homozygosity Index was calculated as the number of homozygous A. sempervirens sites (i.e. reference allele count equal to zero) divided by the total number of variable sites per window. Windows with fewer than 350 SNPs/100kb were excluded as these regions were deemed likely homozygous JI7. We identified a single genomic region with a high density of homozygous SNPs linked to yellow that mapped to ~3.2Mb chromosome 4 (fig. S1).

Transposon-tagging SULF

We used A.majus JI75, which has highly active endogenous transposons (5-7), to isolate a sulf mutant JI660 in a large-scale mutagenesis experiment. The sulf-660 phenotype had spread yellow, giving rise to an orange color through overlap with magenta (7) and crossing proved it was allelic to sulf JI57. The JI661 SULF revertant
arose from \textit{sulf-660} due to transposon instability. Further independent \textit{SULF} revertant plants were isolated; J216-3 and L142-11.

We used WGS to identify differences between \textit{sulf-660} and \textit{SULF-661} individuals which could be indicative of a transposon insertion specifically in the \textit{sulf} genome using signatures from the distribution of reads with unmapped pairs. These criteria identified 187 loci which differed between \textit{sulf-660} and \textit{SULF-661} and which could reflect a transposon insertion in the \textit{sulf-660} line. Only one of these candidates, Chr4:34879920-34879943, was also located in the region of high sequence differentiation between \textit{sulf/sulf} and \textit{SULF/-} pools.

To confirm that the \textit{sulf-660} line carries a transposon in this region we used long PCR (Phusion Taq as described by the manufacturer, New England Biolabs) with flanking oligos (do.99 and do.104). do.99- TCTATCATGGCTTGATTTACAGCC (Chr4:34879576-34879599); do.104- TTTGCTTAGTGACTTTAACCACC (Chr4:34880053-34880075). The PCR products were cloned into pGEM-T (Promega) as described by manufacturers and Sanger sequenced. \textit{sulf-660} gave a 5kb product with homology to a CACTA transposon, consistent with insertion of a transposable element. These primers amplified a ~500bp fragment in \textit{SULF-661}, other independent \textit{SULF} revertant individuals derived from \textit{sulf-660} and other \textit{A. majus} stock lines. Three different excision footprints were found in the three independent revertants, all of which had restricted yellow phenotypes. The \textit{SULF} region mapped to positions 34,877,442 - 34,880,992 on chromosome 4 (Fig. 2a), which lies within the 3.2Mb interval defined by genetic mapping described above.

Genomic organization

Annotated genes with homology to \textit{Am4'CGT} were identified via BLASTp searches of the \textit{A. majus} reference genome and the \textit{Mimulus guttatus} reference v2.0 (via Phytozome v12.1) with an e-value threshold of <1e^{-60}. A single \textit{Linaria vulgaris} 4'CGT accession (BAE48240.1) was also included (fig. S3). In addition to blastp hits to annotated proteins, regions of \textit{A.majus} Chr4 with >500bp homology to \textit{Am4'CGT} coding sequence were included in DNA alignments (Fig. 2B). Peptide alignments were generated using MUSCLE and DNA alignments using MAFFT both implemented using default settings via the EMBL-EBI multiple sequence alignment website (http://www.ebi.ac.uk/Tools/msa/). Substitution model evaluation and phylogeny construction with Maximum likelihood were performed using MEGA 6. In both cases, alignments were analyzed with partial deletion (site coverage cut-off 75%) and support evaluated with 1000 bootstrap replicates. Resultant trees were visualised using iTOL (https://itol.embl.de/). Gene abbreviations used in Figure 2B are: AmC5 = AnMG04-8860500; MgC6 = Migut.F00274; MgC1 = Migut.F00273; MgC2 = Migut.H00709; MgC5 = Migut.F01071; MgC4 = Migut.F01069; MgC3 = Migut.F01068; AmC3 =AnMG0208611300; AmC2 = AnMG0308611900; AmC1 = AnMG0308612500. Non-annotated but homologous regions were Am H1 = Chr4:34868392-34867516; Am H2 = Chr4:34898959-34898266.

Deletion analysis
Copy number across Chromosome 4 was investigated by comparing coverage depth in pooled sequence datasets from two pools located at either end of the hybrid zone (YP4 and MP11). These pools comprised 50 individuals each. Median coverage values were extracted using GATK (v3.5.0) depth of coverage for windows of 5kb with 1kb step size for a 1Mb region of Chromosome 4. The median coverage of each sample for the entire chromosome was used to correct for differences in sequencing effort for samples and the adjusted median coverage of YP4 was then divided by MP11, with results shown in Fig. 2c.

**Clines in SNPs**

KASP genotyping platform (LGC Genomic) was used to genotype SNP loci adjacent to SULF (and other regions of chromosome 4) to quantify the steepness of clines in allele frequencies and their coincidence with flower color phenotype. We sampled 1722 plants over 3 years (2013, 2014 and 2015) from a natural hybrid zone between *A.m.pseudomajus* and *A.m. striatum* in the Spanish Pyrenees. Individuals were located to within 2 meters with a GPS (Trimble GeoXT datalogger), leaf tissue collected for DNA extraction and one flower taken for phenotyping (see below). To design SNP markers, potential divergent SNP loci were first identified from six WGS Illumina poolSeq analyses, each of 50-52 individuals. Pools YP4, YP2 and YP1 were harvested ~12.8, 1.6 and 1.8 km west respectively, of the hybrid zone center (Latitude 42.32270, Longitude 2.07442). Pools MP11, MP4 and MP2 were from 8, 1.4 or 0.7 km east, respectively, of the center. SNPs were then selected on the basis of excess allele frequency differences in the outermost pools ($\Delta p_{1,6} > 0.6$), a minimum depth of 20 reads in all six pools for 100bp either side of the focal SNP. We used these criteria to design 35 KASP oligos (7 within 185kb of SULF and 28 across chromosome 4) with custom R and Python scripts (SNPextract.py (https://github.com/dfield007/snapdragon). This identified SNPs positions suitable for KASP genotyping platform (LGC Genomics) and extracted the 100bp sequence surrounding each candidate SNP. This included selecting sites with: (i) 30 <depth < 300 in both outer pools at the focal SNP (to reduce the probability of false positives and paralogs), (ii) 30 <depth < 300 for sequences 50bp either direction of focal SNP, (iii) <3 other SNPs within 50bp (to ensure primer efficiency), and (iv) biallelism (a KASPar requirement). DNA extractions and genotyping were carried out by LGC Genomics. Replicate DNA extractions and genotyping of n = 500 individuals confirmed relatively low error rates (mean ~0.15%). Plants were grouped into discrete demes of 200 x 200 meters, and their position collapsed along 25km one-dimensional transect. We fitted a five parameter, symmetric sigmoid cline to the observed genotype counts with the expected allele frequencies $\hat{p}$ along the one-dimensional transect as,

$$\hat{p} = p_0 + \frac{(p_1 - p_0)}{1 + \exp \left(-4 \left(\frac{x - c}{w}\right)\right)}$$

where $x$= spatial coordinates in meters along the transect, $c$ = cline center, $w$ = cline width (1/gradient), $p_0$ = allele frequency at the asymptote in the west (*A.m. striatum* parental allele frequency) and $p_1$ = allele frequency at the asymptote in the east (*A.m. pseudomajus* parental allele frequency). In addition, we fitted a beta-binomial error term
to account for the variance in allele frequencies across demes and to control for population structure along the cline $F_{ST} = \frac{\text{var}(p)}{\bar{p}(1 - \bar{p})}$.

To fit clines to the data, we used a metropolis-hastings (simulated annealing) algorithm to sample the likelihood surface following (9) with a few modifications. Briefly, we begin with a random set of parameters which are changed randomly and the log likelihood $lnL$ is computed at each iteration. When the next iteration $lnL''$ has a greater log likelihood than the previous likelihood $lnL'$ (i.e. $lnL'' > lnL'$), the new parameters are accepted. If the next iteration is lower ($lnL'' < lnL'$), we accept with a probability $lnL''/lnL'$. To ensure ample exploration of the likelihood surface, the jump size for the next set of parameters are adjusted by a factor of 1.05 when accepted (accept scale) and by $(1/1.05)$ when parameters are rejected (reject scale). After some tests of different accept and rejection scales, we found these values achieved efficient mixing and exploration of the likelihood surface with an acceptance rate ~0.5. This algorithm was run for 50,000 iterations with a burnin = 2000. From this we found the best fitting combination of parameters from the maximum log likelihood (max $lnL$). To find the 95% credible regions (95% CI) we assumed the likelihood surface follows a chi-square distribution (max $lnL - 2$). For each locus, we visually inspected the likelihood surface to ensure they were well mixed for each parameter. The algorithm was run on each locus three times with randomly chosen starting parameters to ensure independent runs displayed similar parameters estimates with overlapping 95% CI.

center

PCR Marker Genotyping
A PCR genotyping assay distinguished *A.m.pseudomajus* and *A.m.striatum* flanking populations using the SULF PCR Marker do172-do156 that spanning the inverted repeats region: do.172- AAGTTCATCGCTCTTCAATCTCC (Chr4:34878913-34878935); do.156- TGAGGTGGCTAAATAGTGACCAC (Chr4:34881137-34881159) using standard PCR conditions with annealing at 55°C and extension of 2min. Control oligos to Am4’CGT were used to confirm gDNA presence; do.245- ATATCACCAACCACCATGC (Chr2:71833617-71833637) and do.259-TGTATACGTTGCGACTCAGAC (Chr2:71835432-71835456).

Populations in fig.S3 were A-D, *A.m.striatum*, at ~13, 11, 7 and 3 km west of the Hybrid Zone center, and E-G, *A.m.pseudomajus*, at ~5, 8 and 8 km east of center.

RNA analysis
We compared tissues from JI7 SULF, sulf-660, SULF-661, *A.m.pseudomajus* and *A.m.striatum*. Total RNA was isolated from late developmental stages of petals from young buds to just before opening (stages 3 to 6 of Davies et al.) (29). Each sample contained petals from 5-6 buds (~200-300 mg tissue). Petals of 5-6 flower buds just before opening (stage 6) were cut at the tube lobes boundary, and separated into the upper dorsal lobes separate from the lower ventral/laterals. yellow foci) from the peripheral tissues (edge of lateral and ventral lobes) from 20-24 newly opened flowers.Total RNA was isolated from pooled tissue (~100-200 mg) using RNeasy Plant mini Kit (Qiagen).

To purify small RNAs for library construction we first isolated total RNA from petals sulf-660, SULF-661, *A.m.pseudomajus* or *A.m.striatum*. RNA was extracted using
TriReagent (Ambion) following the manufacturer’s protocol with minor modifications as follows. After Tri-reagent and chloroform extraction, the aqueous phase was mixed with an equal volume of isopropanol and stored at -80°C overnight. The total precipitated RNA, was washed with 80% ethanol twice and re-suspended in Rnase free H2O followed by a second round of phenol-chloroform extraction. The sRNA fraction was enriched using mirVana miRNA isolation kitTM (Ambion) according to the manufacturer’s instructions.

One μg of the enriched sRNA fraction were used for sRNA library construction based on a previously published protocol (10). The libraries were sequenced on Hiseq2500 at Earlham Institute, Norwich, UK.

SULF-derived sRNA molecules were detected through RNA blot analysis as described (11). Five to eight microgram of total RNA were denatured and loaded onto 16% urea-polyacrylamide gels and the RNA was transferred to Hybond NX (Amersham) membrane through semi-dry electrophoresis. Chemical cross-linking used 1-ethyl-3 (dimethylaminopropyl) carbodiimide. The SULF inverted repeats (IR) were cloned in pGEM-T vector (Promega). The linearized and blunted templates were used to generate complementary RNA strands for both repeats. The RNA probes were labelled with [α-32P] UTP using T7 or Sp6 RNA in vitro transcriptase (NEB) at 37°C for one hour. The reaction included: 1xRibomix, 1mM A, C, GTP and 4 μM UTP, 200ng of DNA template, 20units of RNA polymerase, and 0.2-1 μM [α-32P] UTP. To detect specific sRNAs, the complementary oligo DNA was end labelled with [γ-P32] ATP using T4 DNA kinase (NEB).

To map cleavage sites of Am4’CGT mRNA in sulf-660 or SULF-661, 80μg of total RNA was isolated from each and used for mRNA isolation with Dynabeads mRNA direct kit (Ambion) following the manufacturer’s protocol. The cleavage sites of Am4’CGT mRNA were determined by 5’RACE using the Generacer kit (Invitrogen) as described (12). The Am4’CGT primary oligo used was SNAP-3-GSP 5’-CTGGTTGTGAGAGCTCCTCCTTCTCC (Chr2:71834715-71834744); followed by nested primer SNAP-3-targnest 5’-CTCTTCTTCCGAAACAAAGGAAATCACGCT (Chr2:71834694-71834724). As above, total RNA was isolated from A.m.striatum and A.m.pseudomajus to determine any cleavage sites. The first oligo was GSP-0505 5’-CCTCTGCTCTCTCGTACAACGGCCAACC (Chr2:71834997-71835023); followed by nested primer GSP-9003 5’-GACTCAACACCTCTTTCTGCGGCACCC (Chr2:71834890-71834916).

Expression analysis

Based on the genomic sequence at the SULF locus, two primers near the inverted repeats were designed for RT-PCR analysis. About 250ng of mRNA isolated from ventral petals of SULF-661 and sulf-660 was used for reverse transcription reactions where polydT(20) was used as the reverse primer. Sequence specific primers: PX1 5’-GTTCATCGCTCTTCAATCTCCATCATTTTCTAC-3’ (Chr4:34878905-34878947) and PX2 5’-GAACATACCCACTTACTTGGACGTAGTG-3’ (Chr4:34880836-34880864) were used for sequence specific amplification of cDNA expressed from this locus. The cDNA fragments specific to SULF-661 were gel extracted and cloned into pGEM-T vector (Promega) for sequence analysis.
For RNA *in situ* of *Am4′CGT*, a 0.51 kb insert of *Am4′CGT* was generated with oligos do.203 5′-AACATACCCACTTACTCGACG (Chr2:71834277-71834298) and do.204 5′-TCGACATCCACTCTTCTCCAAACC (Chr2:71834764-71834786) and cloned into pGEM-T (Promega) as described by manufacturer. The clone pD569 was used to generate antisense RNA probes for *in situ* as described (13).

For qRT-PCR, RNA was isolated and DNase treated as described (Qiagen). First strand synthesis on 0.5 μg of each genotype with SuperScript III was performed as described (Invitrogen). qPCR *Am4′CGT*M oligos used were AW6 and AW8:

AW6 5′-TCGATTTCTTTGGTTGGCCC (Chr2:71834783-71834802) and AW8 5′-ATTGATCCTCTGCTCTGCGT (Chr2:71835010-71835029). GAPDH reference gene oligos were: GAPDH_1762 5′- CACGAGACGAGCTTCACAAA (Chr4:12237378-12237359) and GAPDH_1781 5′- CTGCCATTAAGGAGGAATCG (Chr4:12238005-12238015; Chr4:12238141-12238150). PCR conditions were as described (14). These conditions gave linear amplification at high efficiency for each set of oligos. We used 3 experimental and 3 biological replicates and relative values (log2 scale) were all compared to the same sample of *sulf-660* made in each analysis.

**RNAseq Bioinformatics**

**mRNA-seq analysis**

50bp single-end libraries generated reads that were quality-filtered as detailed above and mapped to the *A. majus* reference genome using tophat v. 2.0.4. We calculated normalized expression values (RPKM) for transcripts using the cuffdiff tool from the cufflinks package. Statistical analyses were performed using the R package and custom R scripts (15).

**smallRNAs**

The sequencing fastq files were converted to fasta format and reads without Ns were retained for further analysis. The evaluation of quality scores was conducted as in the FastQC suite [https://www.bioinformatics.babraham.ac.uk/projects/fastqc/]. The 3′ adapter was trimmed using perfect string matching on the first 7 nucleotides of the adapter (TGGAATT); the HD signatures, 4 assigned nucleotides at the 3′ and 5′ end of the insert (10, 16) were also trimmed. Next, the files were converted from redundant to non-redundant format (17, 18) and the results were summarized into redundant and non-redundant size class distributions.

In non-redundant format, the reads were mapped to the JI7 reference genome allowing 0, 1 or 2 mis-matches and 0 gaps using PatMaN (19, 20). The reads were also mapped to plant mature miRNAs and miRNA hairpins, retrieved from miRbase, release 21 (21). The sequencing depth of the small RNA libraries was ~25M reads, with ~14M reads matching to the reference genome, full length, with no mis-matches or gaps allowed. The abundances of the reads were normalized using the reads per total method (22-24). The replicate to replicate differential expression was called on offset fold change, offset = 20 (empirically determined) (18, 25) calculated on the proximal ends of maximal confidence intervals built on the available replicates. The differential expression on loci (regions on the genome) was conducted using a simplified form of CoLId (26).
applicable on 2 samples, with 2 replicates each. The sRNA analysis was conducted using custom-made Perl and R scripts. The presence plots were created in R, v 3.4.0. The secondary structures were obtained using RNAfold, part of the Vienna RNA package (27).

A scatter plot was generated to show comparative distribution of abundances between sulf-660 and SULF-661 (Fig.3A). Sequencing reads were obtained from two samples with two biological replicates for each genotype (as above). Control replicate vs replicate plots showed reads mapped along the diagonal indicating similarity. The reads were mapped fill length with no mis-matches or gaps allowed using PaTMaN (19). The abundances were normalized using the per total approach. Average expression between bio reps was calculated as log2((replicate1 + replicate2)/2+1). These were normalized and mapped to the genome using full lengths, with no mis-matches or gaps allowed. The small RNAs derived from the SULF locus were marked in red. Reads mapping to the SULF locus were plotted (Fig.3B). The small RNAs which may target Am4'CGT, were predicted in line with the Allen rules (maximum 4 mis-matches between the sRNA and the target, with no mis-matches on the 10th-11th positions) and plotted in red (15). Small RNAs not fitting this criteria for targetting are plotted in grey.

**Genome scan for inverted repeats**

The scan of the A.majus genome for inverted repeats similar to the SULF was first conducted using existing software [Emboss, palindrome application (28), detectIR (29) and Lirex (30)]. Given the spacer region separating the inverted repeats in SULF, its inverted repeat structure was not detected with existing software. We therefore developed a new method based on a palindrome search coupled with the prediction of secondary structures for the proposed inverted repeats that allowed for spacer regions. Briefly, the genome was scanned using windows of variable length (from 100nt to 1600nt, in increments of 50nt). Consecutive windows had a 50nt overlap. Each window was first scanned for the presence of perfect palindromic motifs (5nt), and regions with 70% of palindromic hits, were then folded using RNAfold and the adjusted minimum free energy calculated. The adjusted minimum free energy is the minimum (optimal) free energy resulting from the RNAfold prediction, normalized per 100nt. This adjustment is performed to ensure comparability between the stability of structures of different lengths (e.g. using the amfe, a haipin-like structure of 100nt can be compared to a structure of 1000nt). Without this adjustment, longer structures will always have lower mfe's because of the higher number of AT and CG pairs which contribute to the lower mfe. Windows with adjusted minimum free energy less than -40 were scanned for the presence of one mature stem (windows with multiple stems were discarded). This minimum energy was chosen based on the properties of SULF. By comparison, conserved microRNAs have a free energy ~ -20 and are therefore largely excluded from the search. Because of computational time limitations, searches with minimum free energies of -20 could only be performed for lengths up to 500nt.

The analysis was conducted on inverted repeats with spacers (for the latter, the window was split into 3 equal regions and the secondary structure was predicted on the concatenated sequence of the first and third fragments). In this search, SULF was returned as encoding a potential ~1450nt hairpin, which excluded the first 100 nt of the
region of contiguous 590 inverted homology. This exclusion arose because this region folded upon itself and such foldback regions were trimmed in the analysis. The distribution of abundances of sRNAs (from the samples described above) incident with these inverted repeats is shown in Fig. 4a. Inverted repeats with an overall/total sRNA abundance less than 20 were considered to be expressed below the noise range, and thus classified as non-expressed. Only *SULF* showed strong differential expression between *sulf-660* and its revertant. The microRNA hairpins used in Fig. 4b were identified by screening the small RNA libraries for matches to conserved plant microRNAs and then mapping them to the genome. The miRNA hairpins were determined using a similar approach as described previously (31).
Fig. S1. SNP homozygosity density plot

(A) A genome-wide scan revealed a single major peak of local homozygosity on chromosome 4. 100kb sliding window analysis (50kb step size) of *A. sempervirens*-derived homozygous SNP density in a pool of 35 individuals with the *sulf* phenotype selected from a segregating population. Regions with no datapoints are likely fixed for either J17 or *A. sempervirens* alleles. (B) The chromosome 4 signal block was 3.2Mb wide (pink shading) and contained the *SULF* locus (dashed red line).
Fig. S2. Isolation of the unstable sulf-660 allele in *A. majus*.

(A) Mutagenesis of the magenta stock *A. majus* JI75 ROS SULF line produced the orange-red sulf-660 mutant. Instability gave rise to the revertant SULF-661, again magenta. To allow the change in yellow pattern to be more easily seen, sulf-660 was crossed with an *inc* mutant (which blocks magenta pigmentation) and an *inc sulf* plant isolated. Growing this plant and its progeny revealed instability in the sulf phenotype, both somatic and germinal, which was due to transposon excision from the sulf-660 locus. Scale bar 1cm.

(B) Revertant SULF plants derived from the unstable sulf-660 were analysed by PCR at the site of the transposon. Three revertants analyzed all showed loss of the transposon at the same site on Chromosome 4. and had three different sequence alterations (footprints) confirming that this locus is SULF. Note SULF-661 has revertant 2 sequence.
Fig. S3. Extended peptide phylogeny for CGT-like genes.
We compared CGT-like genes from *Antirrhinum* (red), *Linaria* (grey) and *Mimulus* (black). AnMG0001316900 corresponds to Am4’CGT. Clade iv is plotted in Fig 2B. Details given in Methods.
A

B

C

<table>
<thead>
<tr>
<th>Regions A – D</th>
<th>Regions E – G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alleles lanes 1-3</td>
<td>88</td>
</tr>
<tr>
<td>Allele lane 4</td>
<td>7</td>
</tr>
<tr>
<td>Alleles lanes 6-12</td>
<td>0</td>
</tr>
</tbody>
</table>
Fig. S4. A PCR marker identifies *A.m.pseudomajus* and *A.m.striatum* SULF alleles forming a sharp cline at a Hybrid Zone.

(A) 95-96 individuals were analyzed from regions flanking a hybrid zone (central colored markers); regions A-D (*A.m.striatum* flank) and E-G (*A.m.pseudomajus* flank). (B) Individual gDNAs were compared to *A.majus* J17 using the PCR Marker do172-do156 flanking the inverted repeats of the SULF locus. A range of PCR allele patterns were found, falling into 3 classes. The first allele class includes lanes 1-3 where no strong bands were found. The second allele class had only 1 major band, lane 4, and represents a deletion of 1.3kb of the inverted repeats region. The last allele class was only seen for the *A.m.pseudomajus* flank and has a range of higher kb bands (lanes 6-14) similar to *A.majus* (lane 14). Size markers are given on left-hand side in kilobases (kb). (C) Frequencies of the 3 allele classes for each flank of the hybrid zone, regions A-D (west of center) and regions E-G (east of center).
Fig. S5. Geographic cline fits for 35 SNP loci across Chromosome 4.
(A) Cline width and center (with 95% confidence intervals) where filled circles indicate SNP loci with a cline fit ($\Delta p_{1,6} > 0.6$) and open circles indicate SNP loci where a cline fit was not performed due to low allele frequency differences ($\Delta p_{1,6} < 0.6$). SNP loci with no cline fits are also indicated with respect to their position along the y-axis (n.f. = no fit). Values for a $ROS$ SNP, highlighted with a dotted line, shown for comparison. SNPs 19-26 derive from the $SULF$ region. (B) Individual SNP loci, with allele frequencies in 200 m demes in relation to distance along the transect. Best fitting cline model indicated with blue curve. SNP loci sorted in order across the chromosome from top left to bottom right.
Fig. S6. Am4’CGT mRNA cleavage products in SULF petals.
Small RNA Blots for A.m.pseudomajus and A.m.striatum. Small RNAs were isolated from (A) A.m.pseudomajus (SULF^p) and A.m.striatum (sulf^s) species petals, or from (B) A.m.pseudomajus (SULF^p) and A.m.striatum (sulf^s) introgressed into A.majus carrying Am4’CGTM. Blots were probed for SULF small RNAs and found ~21 nucleotides (nt). Samples are from dorsal petals of 3-5 pooled independent plants. The same results were found for 2 further independent biological replicates in each case, or similarly from ventral petals. Blots also probed with Ubiquitinn U6 as a loading control.
Fig. S7. RNA mapping to the SULF locus

(A, B) IgV profiles showing depths of mapped RNA reads at a, the SULF locus and b, the NIV locus as a control. RNAseq was made on total petals from small 10 mm long flower buds for A.majus J17 SULF\textsuperscript{M}, A.m.pseudomajus SULF\textsuperscript{P} and A.m.striatum sulf\textsuperscript{S} individuals. Reads are shown as grey peaks (with species SNPs colored). Depths are shown on the right-hand side. Both loci are on Chromosome 4 at the positions indicated. The inverted repeats at SULF are shown as large black arrows, with the site of the transposon insertion in the sulf-660 allele indicated by a triangle (not to scale). (C) RT-PCR identified a 1695 nt transcript (arrow) in petals of SULF-661 (lane 2) but not in sulf-
Sequencing showed that this transcript mapped to the SULF above (red line in A). The transcript includes the inverted repeats, but splicing has removed part of the intervening region. Smaller transcripts were found that included only the first or second repeat. The sulf mutant showed only transcripts containing the first or second repeat alone, but with some transposon sequence.
Fig. S8. **Am4’CGT mRNA cleavage products in SULF petals.**

5’RACE on total RNA from *sulf* and *SULF* petals. (A) Gel regions were isolated (green boxes), cloned and sequenced. *sulf*-660 gave few colonies and 3 clones sequenced did not contain **Am4’CGT** fragments. (B) *SULF*-661 gave many clones, 14 were sequenced; 1 had no insert, while 13 had fragments of **Am4’CGT** that showed 12 different start site sequences (black bars), indicating a range of **Am4’CGT** mRNA fragments in *SULF* petals. Numbers are nucleotides relative to start number 1 of ATG. All start sites on left are as found. All sequences stopped at the same position (at the end oligo) but only sequence to nt 669 is shown.
(C) Gel of 5’RACE products for *A.m striatum* and *A.m pseudomajus*. The clear band in *A.m psuedomajus* (green box) was isolated and 35 clones sequenced; 27 had fragments of *Am4’CGT*, with 24 unique start sites shown.
<table>
<thead>
<tr>
<th>Reference genome match</th>
<th>Query (Am 4'CGT) match</th>
<th>Percent identity</th>
<th>Aligned length</th>
<th>Mismatch Count</th>
<th>Gap Count</th>
<th>e-value</th>
<th>Match orientation</th>
<th>Query region overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr4:27714674-27714521</td>
<td>2706-2857</td>
<td>72.78</td>
<td>158</td>
<td>33</td>
<td>4</td>
<td>2.00E-14 minus</td>
<td>downstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:30856882-30856986</td>
<td>2042-2143</td>
<td>74.53</td>
<td>106</td>
<td>22</td>
<td>2</td>
<td>2.00E-09 plus</td>
<td>CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34781696-34783036</td>
<td>1017-2352</td>
<td>74.43</td>
<td>1369</td>
<td>289</td>
<td>9</td>
<td>0</td>
<td>plus CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34832879-34833033</td>
<td>608-755</td>
<td>77.07</td>
<td>157</td>
<td>25</td>
<td>3</td>
<td>3.00E-25 plus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34833584-34833086</td>
<td>134-365</td>
<td>72.5</td>
<td>240</td>
<td>41</td>
<td>11</td>
<td>2.00E-22 plus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34834885-34834994</td>
<td>653-753</td>
<td>79.28</td>
<td>111</td>
<td>12</td>
<td>4</td>
<td>1.00E-16 plus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34836385-34838049</td>
<td>789-2383</td>
<td>81.58</td>
<td>1683</td>
<td>204</td>
<td>21</td>
<td>0</td>
<td>plus promoter, CDS, downstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34845928-34846236</td>
<td>2415-2735</td>
<td>73.93</td>
<td>326</td>
<td>63</td>
<td>6</td>
<td>4.00E-49 plus</td>
<td>downstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34856740-34857066</td>
<td>533-815</td>
<td>64.63</td>
<td>328</td>
<td>61</td>
<td>11</td>
<td>1.00E-10 plus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34861039-34860898</td>
<td>2153-2291</td>
<td>80.28</td>
<td>142</td>
<td>25</td>
<td>1</td>
<td>2.00E-28 minus</td>
<td>CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34868392-34867516</td>
<td>1278-2165</td>
<td>86.26</td>
<td>888</td>
<td>111</td>
<td>3</td>
<td>0</td>
<td>minus CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34873993-34873478</td>
<td>789-1285</td>
<td>81.24</td>
<td>533</td>
<td>47</td>
<td>11</td>
<td>5.00E-143 minus</td>
<td>upstream, CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34875836-34875294</td>
<td>128-682</td>
<td>73.37</td>
<td>597</td>
<td>63</td>
<td>18</td>
<td>1.00E-99 minus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34877442-34877628</td>
<td>171-383</td>
<td>74.07</td>
<td>216</td>
<td>24</td>
<td>6</td>
<td>9.00E-32 plus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34878446-34878726</td>
<td>620-888</td>
<td>72.11</td>
<td>294</td>
<td>44</td>
<td>13</td>
<td>1.00E-30 plus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34878794-34878849</td>
<td>904-960</td>
<td>88.14</td>
<td>59</td>
<td>2</td>
<td>3</td>
<td>2.00E-09 plus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34878900-34880045</td>
<td>1069-2235</td>
<td>84.15</td>
<td>1186</td>
<td>129</td>
<td>12</td>
<td>0</td>
<td>plus CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34880273-34880045</td>
<td>2585-2811</td>
<td>78.39</td>
<td>236</td>
<td>35</td>
<td>5</td>
<td>2.00E-46 minus</td>
<td>downstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34880865-34880320</td>
<td>1389-1929</td>
<td>88.46</td>
<td>546</td>
<td>58</td>
<td>2</td>
<td>0</td>
<td>minus CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34880992-34880907</td>
<td>882-973</td>
<td>86.96</td>
<td>92</td>
<td>6</td>
<td>3</td>
<td>7.00E-21 minus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34884145-34884056</td>
<td>1778-1868</td>
<td>87.91</td>
<td>91</td>
<td>10</td>
<td>1</td>
<td>5.00E-23 minus</td>
<td>CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34885632-34885537</td>
<td>1443-1691</td>
<td>83.13</td>
<td>249</td>
<td>39</td>
<td>2</td>
<td>6.00E-66 minus</td>
<td>CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34887433-34887098</td>
<td>1093-1456</td>
<td>71.47</td>
<td>368</td>
<td>69</td>
<td>10</td>
<td>4.00E-43 minus</td>
<td>CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34889097-34889921</td>
<td>155-1131</td>
<td>70.77</td>
<td>1098</td>
<td>155</td>
<td>35</td>
<td>6.00E-148 minus</td>
<td>upstream, CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34898959-34898266</td>
<td>1862-2543</td>
<td>80.51</td>
<td>708</td>
<td>98</td>
<td>14</td>
<td>7.00E-179 minus</td>
<td>CDS, downstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34901729-34901618</td>
<td>2358-2462</td>
<td>76.79</td>
<td>112</td>
<td>19</td>
<td>3</td>
<td>3.00E-13 minus</td>
<td>CDS, downstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:34917689-34917086</td>
<td>1297-1924</td>
<td>74.8</td>
<td>639</td>
<td>115</td>
<td>6</td>
<td>2.00E-122 minus</td>
<td>CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34919018-34918765</td>
<td>1017-1294</td>
<td>71.99</td>
<td>282</td>
<td>47</td>
<td>6</td>
<td>4.00E-36 minus</td>
<td>CDS</td>
<td></td>
</tr>
<tr>
<td>Chr4:34921202-34920975</td>
<td>167-415</td>
<td>68.5</td>
<td>254</td>
<td>49</td>
<td>11</td>
<td>9.00E-13 minus</td>
<td>upstream</td>
<td></td>
</tr>
<tr>
<td>Chr4:35191842-35193170</td>
<td>1017-2352</td>
<td>74.74</td>
<td>1366</td>
<td>287</td>
<td>9</td>
<td>0</td>
<td>plus CDS</td>
<td></td>
</tr>
</tbody>
</table>
Table S1.
Homology between *Antirrhinum majus* Chromosome 4 and *Am4 CGT* (AnMG0001316900) and flanking regions as revealed by Blastn. Results were filtered to exclude alignments under 50bp in length or with an e-value of greater than 1.0e⁻⁹. The query sequence comprised the *Am4 CGT* single exon coding sequence (positions 1001-2374 with upstream and downstream flanks (positions 1-1000 and 2375-2874, respectively). Content shows defined loci as Fig.2c, with a (AnMG0307025300.01), b (AnMG0308611900.01), c (AnMG0308612500.01), d (AnMG0308613000.01), e (AnMG0308613000.01) and f (AnMG0208611300.01).