RESEARCH IN CONTEXT: PART OF A SPECIAL ISSUE ON SEXUAL PLANT REPRODUCTION

Floral heteromorphy in *Primula vulgaris*: progress towards isolation and characterization of the $S$ locus

Jinhong Li, Margaret A. Webster, Matthew C. Smith and Philip M. Gilmartin*

School of Biological and Biomedical Sciences, Durham University, Durham DH1 3LE, UK

* For correspondence. E-mail philip.gilmartin@durham.ac.uk

Received: 22 February 2011 Returned for revision: 28 March 2011 Accepted: 16 May 2011 Published electronically: 28 July 2011

- **Background** The common primrose, *Primula vulgaris*, along with many other species of the Primulaceae, exhibits floral heteromorphy in which different individuals develop one of two possible forms of flower, known as pin and thrum. Both flower types are hermaphrodite and exhibit reciprocal positions of male and female reproductive structures, which together with a sporophytic incompatibility system, prevent self-pollination and promote outcrossing. The development of the two different forms of flower is controlled by a co-adapted linkage group of genes known as the $S$ locus.
- **Scope** Here progress towards identification and characterization of these genes is described to provide a molecular genetic explanation of the different floral characteristics that define heterostyly in *Primula* as observed and described by Charles Darwin. Previous work to identify and characterize developmental mutations linked to the *P. vulgaris* $S$ locus, together with the isolation of $S$ locus-linked genes and polymorphic DNA sequences markers, is summarized. The development of tools are described which will facilitate isolation and characterization of the $S$ locus and its environs, including the creation of two expressed sequence tag libraries from pin and thrum flowers, as well as the construction and screening of two bacterial artificial chromosome (BAC) libraries containing thrum genomic DNA. Screening of these libraries with four $S$ locus-linked sequences has enabled us to assemble four BAC contigs representing over 40 individual overlapping BAC clones which represent over 2-2 Mb of $S$ locus-linked genomic sequence. PCR-based approaches for identification of the allelic origin of these BACs are described as well as identification of an additional 14 $S$ locus-linked genes within BAC-end sequences.
- **Conclusions** On-going work to assemble the four $S$ locus-linked contigs into one contiguous sequence spanning the $S$ locus is outlined in preparation for sequence analysis and characterization of the genes located within this region.

**Key words:** *Primula vulgaris*, $S$ locus, floral heteromorphy, heteromorphic flower development.

INTRODUCTION

Floral heteromorphy

Observations of heteromorphic flower development caught the imagination of Charles Darwin who initially documented his studies on heterostyly in dimorphic (Darwin, 1862, 1863) and trimorphic (Darwin, 1864) species in the early 1860s. He subsequently combined his observations on related plant breeding systems in his landmark book *The Different Forms of Flowers on Plants of the Same Species* (Darwin, 1877). Of the species studied by Darwin, it is the primrose, *P. vulgaris*, which has possibly received the greatest attention during the intervening 150 years. However, other dimorphic species including *Fagopyrum esculentum* (buckwheat) (Ali et al., 1998; Matsui et al., 2004; Yassui et al., 2004, 2008; Ota et al., 2006), which was also studied by Darwin (1877) and *Turnera subulata* (white alder) (Shore et al., 2006; Labonne et al., 2008, 2009, 2010; Labonne and Shore, 2011) have also been established as important model heterostylosus species which have contributed to our understanding of this elegant out-breeding system which prevents self-pollination within hermaphroditic flowers.

Classical analysis of floral heteromorphy in Primula

Dimorphic flower development in *Primula* has become a textbook example of floral heteromorphy (Richards, 1997), which is also referred to as heterostyly. Studies have focused on the Mendelian (Bateson and Gregory, 1905; Ernst, 1925; Crosby, 1949; Lewis and Jones, 1993; Kurian and Richards, 1997), evolutionary (Charlesworth and Charlesworth, 1979c; Piper and Charlesworth, 1986; Conti et al., 2000; Mast and Conti, 2006) and population (Crosby, 1940; Charlesworth and Charlesworth, 1979a; Piper et al., 1984) genetics of the system, as well as developmental (Heslop-Harrison et al., 1981; Shivanna et al., 1981; Webster and Gilmartin, 2003, 2006) and more recently molecular genetic (Manfield et al., 2005; McCubbin et al., 2006; Li et al., 2007, 2008, 2010) aspects of pin and thrum flower development. Darwin’s studies of the two different forms of flower in *Primula* underpin all these studies.

In dimorphic *Primula* species, pin flowers contain a long style which presents the stigma at the opening of the mouth of the flower (Fig. 1B). In thrum flowers, it is the anthers which are located the mouth of the flower (Fig. 1A). Similarly, the height of the short style found in thrum flowers corresponds to the position of anthers in pin flowers (Darwin, 1862). The reciprocal position of male and female
reproductive structures in the two hermaphrodite forms of flower facilitates reciprocal pollination by insects (Darwin, 1862). Other morph-specific differences have been documented including the length of stigmatic papillae (Heslop-Harrison et al., 1981), the size of pollen (Darwin, 1862) and diameter of the mouth of the flowers (Webster and Gilmartin, 2006). Pin flowers possess longer stigmatic papillae than thrum flowers, thrum flowers produce larger pollen grains than pin flowers, and the mouth of the flower is wider in thrums than pins. In addition, the presence of a sporophytic self-incompatibility system which reinforces the architectural constraints on self-fertilization implies the presence of both pollen- and style-specific incompatibility genes associated with the S locus (Crosby, 1949; Shivanna et al., 1981; Richards, 1997; McCubbin, 2008). In addition, two genes, Mg and Ma, associated with flower pigment production (Kurian, 1996; Richards, 1997), as well as developmental mutations affecting organ identity, such as Hose in Hose (Ernst, 1942; Webster, 2005; Li et al., 2010) and sepaloid (Webster, 2005; Li et al., 2008), as well as the leaf shape mutant Oakleaf (Webster, 2005), have also been shown to be inherited with one or other of the two different forms of flower.

Detailed observations of developmental timing and cell size and shape have revealed that long styles in pin flowers contain epidermal cells which are twice as long as those found in thrum styles; differential cell elongation therefore provides an explanation for different style lengths in the two forms of flower (Heslop-Harrison et al., 1981; Webster and Gilmartin, 2006). In contrast, observations of the corolla epidermal cell length above and below the point of anther attachment in the two forms of flower reveals that, notwithstanding the different height of the anthers in pin and thrum flowers, the cell size below the point of anther attachment to the corolla is the same, suggesting that differences in cell division are responsible for the differential elevation of anthers during development (Webster and Gilmartin, 2006). Comparison between different stages of pin and thrum flower development revealed that those genes controlling style elongation in pin flowers operate at an earlier stage of flower development than those which control anther elevation in thrum flowers (Webster and Gilmartin, 2006).

Although Darwin observed that pin–thrum and thrum–pin crosses were required to facilitate effective pollination (Darwin, 1877), and that crosses within morphs did not typically yield seed, others provided the basis for our current understanding of the genetics of heteromorphy in Primula (Bateson and Gregory, 1905; Crosby, 1949; Ernst, 1955). Pin plants are homozygous recessive for the s allele of the S locus. Thrum plants are heterozygous, carrying a dominant S allele and a recessive s allele (Crosby, 1949; Lewis and Jones, 1993). Crosses between pin (ss) and thrum (Ss) plants always yield a one to one ratio of pin to thrum progeny (Darwin, 1862; Crosby, 1949). It is also known from Ernst’s work (Ernst, 1933, 1936b, 1955), and subsequent re-analysis of his data (Lewis and Jones, 1993), that the S locus comprises a co-adapted linkage group of genes which control different aspects of heteromorphy. At least three genes, G, P and A, are clustered at the S locus and control the length of the gynae
cium (G), pollen size (P) and anther position (A). Integration of the genetic observations (Ernst, 1955; Lewis and Jones, 1993) with the developmental biology (Webster and Gilmartin, 2006) of pin and thrum flowers tells us: (a) that the dominant G allele is required for suppression of cell elongation in thrum styles and that in the absence of G, style cells elongate to twice the length in pin flowers; (b) that the dominant A allele promotes increased cell division below the point of anther attachment in the corolla to elevate the anthers to the mouth of the thrum flowers. These interpretations will prove helpful in the identification of candidate genes responsible for these developmental effects.

Specific allelic combinations of GPA and gpa alleles at the S locus in P. vulgaris are possibly maintained by recombination suppression or by the very close chromosomal association of these genes. Interestingly, in T. subulata, there is no evidence for recombination suppression at the S locus (Labonne et al., 2007). From the analysis of rare S locus recombinants which produce long and short homostyle flowers (Crosby, 1940; Ernst, 1955; Dowrick, 1956; Bodmer, 1960; Charlesworth and Charlesworth, 1979b; Lewis and Jones, 1993), the
suggested gene order within the S locus is GPA. In *P. vulgaris*, pin plants are therefore *gpa*/*gpa* and thrum plants *GPa*/*gpa*. However, this gene order may not be the same in all species of *Primula* (Kurian and Richards, 1997; Richards, 1997). Plants producing long and short homostyle flowers result from inter-morph fertilization events following meiotic recombination between *G* and *A*; the size of their pollen is determined by which side of locus *P* recombination occurred. The partial linkage of two anthocyanin pigment genes (Kurian, 1996; Richards, 1997), and the developmental mutants *Hose in Hose* (Ernst, 1936a; Webster and Grant, 1990; Li et al., 2010), *sepaloid* (Webster, 2005; Li et al., 2008) and *Oakleaf* (Webster, 2005) with the S locus provides examples of genes which flank the S locus but are not associated with the control of heteromorphy. Some of these mutations have proved valuable in defining the boundaries of the S locus (Webster, 2005; Li et al., 2008, 2010).

**Molecular genetic analysis of floral heteromorphy in Primula**

Initial studies to identify S locus-linked genes focused on the isolation of polymorphic sequences associated with the pin and thrum alleles of the *P. vulgaris* S locus. An F1 commercial hybrid variety of *P. vulgaris* (‘Blue Jeans’) derived from in-bred parental lines by the Farmen Seed Company, Naples, Italy was selected. It was reasoned that morph-specific sequence differences associated with thrum plants would represent polymorphisms between the dominant (S) and recessive (s) alleles of the S locus. This assumption resulted in the identification and isolation of the first *Primula* S locus-linked sequence using 10-bp random primers by random amplification of polymorphic DNA (RAPD) PCR (Manfield et al., 2005). The original 323-bp RAPD product was cloned and sequenced to enable design of specific 15-bp primers which enabled amplification of this sequence as a sequence characterized amplified region (SCAR) marker (Manfield et al., 2005). Southern hybridization on pin and thrum genomic DNA using this fragment as a probe also revealed a restriction fragment length polymorphism (RFLP) associated with this region (Manfield et al., 2005). Although the sequence does not represent an expressed gene, it shows tight linkage with the S locus with no observed recombination. This marker was subsequently named *P. vulgaris* S locus polymorphism 1 (*PvSLP1*).

Other analyses using the highly reproducible and reliable technique of fluorescent differential display (FDD) (Kreps et al., 2000; Kuno et al., 2000; Scutt et al., 2002) with cDNA from pin and thrum flowers were designed to identify genes that are differentially regulated in response to the S locus (Li et al., 2007). Surprisingly of the 20 transcripts identified as differentially expressed by FDD, none proved to be morph-specific as determined by northern analysis (Li et al., 2007). However, subsequent analysis of selected sequences revealed that the sensitivity of FDD had in fact identified allelic polymorphisms in S locus-linked genes which are expressed in both pin and thrum flowers (Li et al., 2007). The two S locus-linked (SLL) genes identified by this approach were designated *PvSLL1* and *PvSLL2*. Parallel studies by others have been successful in identifying genes that are differentially expressed in response to the S locus (McCubbins et al., 2006). These sequences will be valuable in future studies to understand the cascade of events leading to the development of two forms of flowers orchestrated by genes at the S locus. Although the two S locus-linked genes, *PvSLL1* and *PvSLL2*, are not differentially regulated between flower morphs, they have provided valuable insight into the architecture of the S locus.

Mendelian genetic analysis indicates that there are two S locus alleles, dominant (S) and recessive (s) (Crosby, 1949; Lewis and Jones, 1993; Richards, 1997). However, studies on pin and thrum alleles of *PvSLL1* and *PvSLL2* reveal that there are three alleles for each of these loci in the study population (Li et al., 2007). This apparent discrepancy can be rationalized by the explanation that there are two functional alleles for key genes at the S locus, but that different sequence polymorphisms, which may or may not affect function, have arisen and become fixed due to the limited recombination at the S locus. It is therefore possible to identify polymorphisms in pin-specific and thrum-specific alleles of S-linked genes. The study population carries two different pin alleles of *PvSLL1* and *PvSLL2* which are referred to as P1 and P2, but only one thrum (T) allele (Li et al., 2007). It is therefore possible that multiple neutral sequence variants of pin, and indeed thrum, alleles of S locus-linked genes, both integral to, and flanking, the S locus could exist. This insight will prove useful in subsequent molecular analyses of S locus-associated sequences.

In parallel to the characterization of S locus-linked sequences derived from RAPD and FDD analyses, the availability of three developmental mutations that show linkage to the S locus has also been exploited: *Hose in Hose* (Ernst, 1936a; Webster and Grant, 1990; Li et al., 2010), *sepaloid* (Webster, 2005; Li et al., 2008) and *Oakleaf* (Webster, 2005). Previous studies have shown that *Hose in Hose*, a dominant floral mutation resulting in the conversion of sepals to petals, is linked to the S locus (Ernst, 1942; Webster and Grant, 1990) and other recent studies have revealed that the mutation is within the *P. vulgaris* B function MADs box gene, *Globosa* (*PvGlo*) (Li et al., 2010). The studies on the recessive *sepaloid* mutation, which produces flowers with no petals but only sepals and carpels, appears to be a loss of B function MADs box gene mutation, but it does not map to the *PvGlo* locus, suggesting another S locus-linked gene involved in petal identity (Li et al., 2008). The dominant *Oakleaf* mutation, which produces highly serrated leaves and abnormal petals, is also linked to the S locus (Webster, 2005) but the gene responsible is as yet unknown. The availability of S locus-linked sequences and genetic resources have provided us with the tools to isolate and characterize the *P. vulgaris* S locus.

The present studies on heteromorphic flower development in *Primula* were prompted by a desire to identify and understand the genes which control and co-ordinate the development of two forms of flower in this species. Here, progress towards the identification and characterization of the S locus in *P. vulgaris*, as presented at the 21st International Congress on Sexual Plant Reproduction in Bristol 2010, is summarized.

**MATERIALS AND METHODS**

**cDNA preparation and 454 sequencing**

Plant materials (2- to 11-mm buds and nearly open flowers) from pin and thrum plants were pooled for total RNA extraction using RNAqueous total RNA isolation kits (Ambion Cat.
Bacterial artificial chromosome (BAC) library construction and screening

In-house BAC library construction was performed following Hongbin Zhang’s Manual of BAC Protocols, Laboratory for Plant Genomics & GENEFinder Genomic Resource, Texas A&M University, USA (http://hbz7.tamu.edu/homelinks/tool/bac_content.htm) with modifications. Megabase-sized genomic DNA of *P. vulgaris* ‘Blue Jeans’ was isolated from young leaves pooled from thrum plants. DNA was partially digested with HindIII and cloned into BAC vector pCC-BAC using a CopyControl BAC cloning kit (Epicentre) in a 1.5-mL microcentrifuge tube. Cells were resuspended by vigorous vortexing. EpiBlue solution (10 μL) (Epicentre) was added to each tube and thoroughly mixed. Aliquots (25 μL) were loaded onto a 0.8% (w/v) agarose tube in 1 × TAE followed by electrophoresis at 4.5 V cm⁻¹ for 3 h alongside of the BAC-Tracker Supercoiled DNA ladder (Epicentre). The gel was stained with ethidium bromide and destained in distilled water before size estimation by visualization under UV light. In total, 108 random BAC clones were examined to determine the average insert size.

Pulse-field gel electrophoresis fingerprinting of BAC DNA

BAC DNA (3 μg) was digested with HindIII (Fermentas) in a 30 μL reaction overnight, precipitated in ethanol and resuspended in TE before loading on a 1% agarose gel and fractionated by pulsed field gel electrophoresis in 0.5× TBE using a CHEF-DR II (Bio-Rad) under the following conditions: initial switch time 0.1 s, final switch time 0.1 s, 600 V cm⁻¹, for 4 h at 14°C. After electrophoresis, the gel was stained in 1× SYBR gold (Invitrogen) for 30 min and visualized and photographed under UV light.

Allele-specific PCR

Individual PCR reactions were performed using an MyiQ iCycler PCR machine (Bio-Rad) with 100 ng of genomic DNA from individual pin and thrum plants and 10 ng BAC DNA as positive control, in a 50 μL reaction with GoTaq DNA polymerase (Promega Part# M829B) in 1× Green GoTaq Flexi Buffer (Promega Part# M891A). Pairs of forward (F) and reverse (R) primers were designed to forward (F) and reverse (R) BAC-end sequence and designated FF, FR, RF, RR as follows: BAC-90B2.RF, BAC-90B2.RR, BAC-91J7.RF, BAC-91J7.RR, BAC-47I4.FF, BAC-47I4.FR. PCR conditions and primers are available on request.
RESULTS

Construction and characterization of pin and thrum flower expressed sequence tag (EST) libraries

In parallel to the present studies to identify individual S locus-linked genes, a transcriptomics-based approach has been undertaken to identify genes that are differentially regulated between pin and thrum flowers. 454-based DNA sequencing of cDNA derived from both pin flowers and thrum flowers was used. As starting material, separate pools of pin and thrum flowers ranging in size from 2 mm to 11 mm were used to encompass the key developmental stages during which differential style elongation and differential anther elevation occur between the two forms of flower. Pin and thrum flower cDNA was used for 454 sequencing library construction, sequence acquisition and sequence assembly at the Liverpool Centre for Genomics Research to create two EST libraries representing sequences independently derived from pin and thrum flowers.

Table 1A provides summary details of these analyses. Two independent sequencing runs were undertaken for each library and the data pooled. The combined number of sequencing reads from pin and thrum was 116 463 and 181 555, respectively, each representing over 26 million and 41 million nucleotide bases. Individual sequencing reads ranged from just over 30 bases to approx. 600 bases. Assembly of these individual sequencing reads (Table 1B) yielded 13 990 contigs from the pin cDNA and 18 242 contigs from the thrum cDNA with lengths ranging from 90 to 3108 nucleotides. The majority of contigs are represented in both pin and thrum libraries. Those contigs that are present in only pin or thrum libraries, or show different representation in the two libraries will be further characterized to determine whether they represent morph-specific genes. Comparison of the pooled assembled contigs to the arabidopsis genome sequence identified 9352 open reading frames showing similarity to known arabidopsis genes. A full analysis of differential expression and sequence comparisons will form the basis of future work. Availability of these pin- and thrum-specific EST collections will not only facilitate the analysis of differential expression profiles in the two forms of flower, but will also support and facilitate gene discovery and annotation within the chromosomal sequence obtained from a BAC contig containing the S locus and its environs.

**Table 1A Analysis of pin and thrum flower EST datasets**

(A) Comparison of read number, cumulative read length and read length characteristics following 454 DNA sequence data obtained from pin and thrum flower cDNA

<table>
<thead>
<tr>
<th></th>
<th>Sequence pin #1</th>
<th>Sequence pin #2</th>
<th>Sequence pin total</th>
<th>Sequence thrum #1</th>
<th>Sequence thrum #2</th>
<th>Sequence thrum total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of reads</td>
<td>21 582</td>
<td>94 881</td>
<td>116 463</td>
<td>33 653</td>
<td>147 902</td>
<td>181 555</td>
</tr>
<tr>
<td>Sum of read (nt)</td>
<td>5053 385</td>
<td>21 794 249</td>
<td>26 847 634</td>
<td>7763 162</td>
<td>33 712 330</td>
<td>41 475 492</td>
</tr>
<tr>
<td>Min. length (nt)</td>
<td>38</td>
<td>38</td>
<td>38</td>
<td>40</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>Max. length (nt)</td>
<td>668</td>
<td>597</td>
<td>668</td>
<td>571</td>
<td>518</td>
<td>571</td>
</tr>
<tr>
<td>Mean length (nt)</td>
<td>234</td>
<td>230</td>
<td>231</td>
<td>231</td>
<td>228</td>
<td>228</td>
</tr>
<tr>
<td>s.d.</td>
<td>51</td>
<td>52</td>
<td>51</td>
<td>51</td>
<td>51</td>
<td>51</td>
</tr>
</tbody>
</table>

(B) Summary data of contig assembly from 454 sequence analysis for pin and thrum flower cDNA showing the number of assembled contigs and characteristics of the contig assemblies

<table>
<thead>
<tr>
<th></th>
<th>Pin EST sequences</th>
<th>Thrum EST sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of contigs</td>
<td>13 990</td>
<td>18 242</td>
</tr>
<tr>
<td>Sum of contigs (nt)</td>
<td>4182 953</td>
<td>5887 200</td>
</tr>
<tr>
<td>Min. length (nt)</td>
<td>94</td>
<td>90</td>
</tr>
<tr>
<td>Max. length (nt)</td>
<td>2642</td>
<td>3108</td>
</tr>
<tr>
<td>Mean length (nt)</td>
<td>299</td>
<td>323</td>
</tr>
<tr>
<td>s.d.</td>
<td>223</td>
<td>240</td>
</tr>
</tbody>
</table>
The first library was constructed in-house using partial HindIII digestion of thrum genomic DNA which was ligated into the HindIII cloning site of BAC vector pCC1-BAC. Initial characterization of this library revealed 14,976 clones, with an average insert size of 130 kb (see Materials and methods) which represented 4 × coverage of the 489 Mb *P. vulgaris* genome. A second commercial BAC library from thrum DNA was then commissioned from Lucigen (Middleton, WI, USA) using pSMART-BAC as a vector. Randomly sheared genomic DNA was ligated via BstXI linkers into the vector to yield a library of 35,712 clones with an average insert size of 70 kb (see Materials and methods) and representing over 5 × coverage of the genome. In combination, the two libraries contain 50,688 clones representing over 9 × genome coverage. The two libraries were independently gridded into 384-well microtitre plates and clone cultures arrayed onto nylon membranes in 4 × 4 features containing duplicates for each clone. Each clone was given a unique identifier as defined by its microtitre plate location.

BACs corresponding to *PvSLP1* (BAC-81B15), *PvSLL1* (BAC-6K4), *PvSLL2* (BAC-56H19) and *PvGlo* (BAC-pCCGlo1 and BAC-pCCGlo2), were isolated by hybridization screening of the arrayed BAC clone filters. These founding BACs were then used to initiate assembly of four contigs as shown in Fig. 2. BAC DNA was prepared which, following restriction endonuclease digestion, was characterized by pulse-field gel electrophoresis to determine insert size. The vector pSMART-BAC contains SP6 and SR4 primer sequences; these were designated, respectively, as forward and reverse to aid in insert-end identification. Similarly with pCC1-BAC, the T7 and pCC1R primer sequences were, respectively, designated as forward and reverse and these were used for BAC-end sequencing. PCR analysis using primers designed to BAC-end sequences, together with forward and reverse vector primers was used to orientate each insert within the BAC vector and thus designate one end of the insert as forward and one as reverse as shown in Fig. 2. BAC-end sequence data were also used to design PCR primers for amplification of BAC-end probes for subsequent rounds of BAC library screening. Initial screens for *PvGlo* used the pCC1-BAC:HindIII library. All other and subsequent screens have so far used the pSMART-BAC:random shear library. Reiterative rounds of hybridization, BAC isolation, BAC-end sequencing and PCR analysis will confirm whether such BACs represent different alleles of the locus or are clones which have been misidentified by the BAC walk stepping into another chromosomal region due to cross-hybridization of the end probe to a noncontiguous BAC. To date over 40 overlapping BACs have been identified by this reiterative screening process.

As additional BAC clones were identified by hybridization screening, they were each characterized to confirm their overlapping relationship with the previous clone to ensure that the walk was contiguous and had not jumped following cross-hybridization to a non-S locus-linked region of the genome. This was done in three ways. (1) Overlap of each new BAC with the previous BAC was confirmed by PCR analysis. For each new BAC, the BAC-end primers were not only used to prepare probes for the next round of screening from the new BAC, but were also used for PCR analysis on the previous BACs. In the majority of cases, primers derived from one end of each new BAC were shown, as expected, to amplify the correct size product from the previous BAC. However, in some cases, this backwards amplification did not yield the expected PCR product. (2) Every new end probe derived from each new BAC was used as a probe to confirm hybridization to the previous BAC. Only those new BACs which had been identified by hybridization using an end probe from an existing BAC which also showed reciprocal back-hybridization to the original clone were chosen for further analysis. (3) BACs which fulfilled the reciprocal PCR and reciprocal hybridization screens for each step of the walk were subsequently characterized by BAC DNA fingerprinting to demonstrate overlap.

Restriction digestion of BAC DNA with HindIII followed by pulse field gel electrophoresis of BACs (Fig. 3) in the same order as they appear in the contig (Fig. 2) provided further evidence of BAC overlap. For example BACs 6011B and 86G11 from Contig 1 (Figs 2A and 3A) share common restriction fragment bands demonstrating overlap between these two clones. In some cases where the BACs are particularly large such as pCCGlo1 and pCCGlo2 (Figs 2D and 3D) the banding pattern is much more complex, but still reveals common bands with neighbouring BACs. Other similar examples are clearly visible in Fig. 3. However, in some cases such as BAC 90B2, there is much less obvious commonality of restriction fragments with BAC 87M14, its neighbour as identified by hybridization and PCR screening. The combined BAC DNA fingerprinting data presented in Fig. 3 reveal that the majority of BACs share a common overlapping restriction profile with their neighbours. Some of the BACs shown in Fig. 2, which either carry small inserts or overlap almost completely with another BAC are not included in the data presented in Fig. 3.

In those cases where PCR amplification using primers derived from a new BAC failed to amplify the same-size fragment from the original BAC, or where reciprocal cross-hybridization of BAC-end sequences did not confirm overlap of the new clone with the original, the BAC was not included in assembling the contig. However, if a new BAC met both these criteria, but failed to show a common overlapping finger print pattern with its neighbours, it was still included in the contig as a possible allelic variant. On-going BAC sequencing will confirm whether such BACs represent different alleles of the locus or are clones which have been misidentified by the BAC walk stepping into another chromosomal region due to cross-hybridization of the end probe to a non-contiguous BAC. To date over 40 overlapping BACs have been identified by this reiterative screening process.

The availability of BAC-end sequence data made it possible to undertake an initial screen for additional S locus-linked genes. All BAC-end sequences were used to interrogate GenBank by BLAST sequence analysis. This analysis identified a number of sequence similarities between our BAC-end sequences and known genes. Selected examples of these genes are presented in Table 2. For simplicity, only the corresponding arabidopsis genes that were identified by BLAST sequence analysis with a cut off E-value of 1e-20 are presented. Sequence similarities with genes from other plant species, and those obtained at a lower E-value threshold are not shown. It is interesting to note that arabidopsis sequence similarities obtained from both forward ends of overlapping BACs, BAC-51M17 and
Fig. 2. Assembly of BAC contigs spanning S locus-linked sequences. S locus associated BACs were identified by hybridization using four previously defined S locus-linked sequences: PvSLL1 (A), PvSLL2 (B), PvGlo (C) and PvSLP1 (D). The four founding BACs (identity codes in bold), were used to initiate walks in both directions to identify overlapping BAC contigs through reiterative end-sequencing, probe preparation and hybridization. BAC identity codes and sizes are shown. BAC vector pSMART-BAC contains SP6 forward (blue) and SR4 reverse (red) primer sequences at either end of the insert site. BAC vector pCCBAC contains T7 forward (green) and pCC1R reverse (yellow) primer sequences at either end of the insert site. Left (L) and right (R) ends of each contig were assigned arbitrarily. The number of BACs in each contig is indicated, together with the cumulative sequence length of BACs contributing to each contig. The cumulative non-overlapping length of all BAC clones is 2.23 Mb.
BAC-46L18 (contig 1), reveal neighbouring genes located on arabidopsis chromosome 1 (Table 2). However, there appears to be no conserved synteny in this region with any one arabidopsis chromosome as other BAC-end sequences reveal homologies to arabidopsis genes located on chromosomes 2, 4 and 5.

Definition of pin and thrum alleles in S locus-derived BACs

The availability of sequences for the S locus-linked genes, PvSLL1, PvSLL2 and PvGlo, made it possible to distinguish pin- and thrum-specific alleles for these sequences by PCR or by direct DNA sequence comparisons. However, the different alleles are not sufficiently different to prevent cross-hybridization. Similarly for PvSLP1, although this sequence represents a thrum-specific PCR amplification product, DNA sequences contained within this fragment are not unique to the thrum genome and cross-hybridization to pin DNA is observed when this sequence is used as a hybridization probe (Manfield et al., 2005). Consequently, these four probes could potentially identify BACs containing genomic DNA originating from either the pin or thrum alleles of the S locus. Having isolated the founding BACs, these were screened using allele-specific PCR primers derived from each of the sequences. Amplification of BAC-81B15 using the SCAR marker primers for PvSLP1 reveals that this BAC originates from the thrum allele. Similar analysis of BAC-6K4, BAC-56H19, BAC-pCCGlo.1 and BAC-pCCGlo.2 revealed that they all originated from pin alleles.

During subsequent reiterative BAC filter hybridization screening using BAC-end sequences we did not seek initially to define allelic origin of each additional BAC clone. However, once a tiling path of overlapping BACs within each contig had been assembled, analyses were initiated to determine the allelic origin of each of the BACs in the contigs. Using available BAC-end specific primers we initially screened for allele-specific PCR amplification profiles to assign allelic origin to individual BACs. Examples of these analyses are shown in Fig. 4. A panel of pin and thrum plants of known genotype, with respect to the P1, P2 and T alleles of PvSLL1, which shows complete linkage to the S locus, was used. In some cases, as exemplified by PCR

**Fig. 3.** RFLP analysis of overlapping BAC clones. Pulse field agarose gel analysis of overlapping BACs, as indentified by their individual code, following digestion with HindIII is shown. The length (in kilobases) of DNA size markers (M) is shown. (A) BACs contributing to contig 1 (PvSLL1); (B) BACs contributing to contig 2 (PvSLL2); (C) BACs contributing to contig 3 (PvGlo); (D) BACs contributing to contig 4 (PvSLP1) as shown schematically in Fig. 2.
and thrum plants of known genotype, with respect to (Fig. 4), it was not possible to distinguish between alleles as using primers derived from the reverse end of BAC-90B2 end-specific primers are shown, along with the approximate sizes in base pairs.

**Table 2. Analysis of BAC-end sequence to identify open reading frames**

<table>
<thead>
<tr>
<th>BAC-end</th>
<th>Contig</th>
<th>Similar to arabidopsis</th>
<th>E-value</th>
<th>Protein/function in Arabidopsis thaliana</th>
</tr>
</thead>
<tbody>
<tr>
<td>51M17-F</td>
<td>1</td>
<td>AT1G21740</td>
<td>1e-20</td>
<td>Unknown function</td>
</tr>
<tr>
<td>46L18-F</td>
<td>1</td>
<td>AT1G21750</td>
<td>9e-24</td>
<td>ATPDIL1-1, protein disulfide isomerase-like</td>
</tr>
<tr>
<td>20E12-R</td>
<td>1</td>
<td>AT1G44414</td>
<td>2e-25</td>
<td>Unknown function</td>
</tr>
<tr>
<td>20L16-R</td>
<td>1</td>
<td>AT1G77390</td>
<td>9e-24</td>
<td>TAM, CYCA1, meiosis cell cycle gene NADK2</td>
</tr>
<tr>
<td>82I1-R</td>
<td>2</td>
<td>AT1G21640</td>
<td>1e-35</td>
<td>STP11, sugar transporter 11</td>
</tr>
<tr>
<td>28H15-F</td>
<td>2</td>
<td>AT5G23270</td>
<td>3e-37</td>
<td>QKY, putative transmembrane protein</td>
</tr>
<tr>
<td>28H15-R</td>
<td>2</td>
<td>AT1G74720</td>
<td>1e-97</td>
<td>AQP, putative transmembrane protein</td>
</tr>
<tr>
<td>17A21-R</td>
<td>3</td>
<td>AT2G04280</td>
<td>4e-60</td>
<td>Unknown function</td>
</tr>
<tr>
<td>8119-F</td>
<td>3</td>
<td>AT4G29090</td>
<td>4e-30</td>
<td>Ribonuclease H-like superfamily protein</td>
</tr>
<tr>
<td>56H19-R</td>
<td>3</td>
<td>AT1G78780</td>
<td>3e-18</td>
<td>Pathogenesis-related family protein</td>
</tr>
<tr>
<td>2808-R</td>
<td>4</td>
<td>AT5G09220</td>
<td>2e-47</td>
<td>AAP2, AMINO ACID PERMEASE 2</td>
</tr>
<tr>
<td>77E4-R</td>
<td>4</td>
<td>AT2G12480</td>
<td>8e-23</td>
<td>Serine carboxypeptidase-like 43 (SCP43)</td>
</tr>
<tr>
<td>pCCGo2-R</td>
<td>4</td>
<td>AT5G42240</td>
<td>4e-40</td>
<td>Serine carboxypeptidase-like 42 (SCP42)</td>
</tr>
<tr>
<td>9D20-F</td>
<td>4</td>
<td>AT5G52560</td>
<td>7e-20</td>
<td>ATUSP, involved in pollen development,</td>
</tr>
</tbody>
</table>

BAC-end sequence was used to search the arabidopsis TAIR10 protein database using BLASTX (http://www.arabidopsis.org/Blast/). The origins of BAC-end sequences are shown as well as their contig location. Arabidopsis genes were identified using a cut-off E-value of 1e-20 are shown as well as their encoded proteins and, where known, their function.

Using primers designed to the right-hand ends of BAC-90B2 (contig 3) and BAC-91J7 (contig 2), and left-hand end of BAC-47I4 (contig 2) as indicted. PCR products of genomic DNA are shown alongside amplification products derived from the original BAC DNA, together with a no-DNA control. The different alleles amplified by the BAC end-specific primers are shown, along with the approximate sizes in base pairs.

![Fig. 4. Definition of allelic origin of BAC clones by PCR. Agarose gel fractionation of PCR amplification products derived from genomic DNA of pin and thrum plants of known genotype, with respect to S locus alleles P1, P2 and T are shown. Primers were designed to the right-hand ends of BAC-90B2 (contig 3) and BAC-91J7 (contig 2), and left-hand end of BAC-47I4 (contig 2) as indicted. PCR products of genomic DNA are shown alongside amplification products derived from the original BAC DNA, together with a no-DNA control. The different alleles amplified by the BAC end-specific primers are shown, along with the approximate sizes in base pairs.](image-url)

Using primers derived from the reverse end of BAC-90B2 (Fig. 4), it was not possible to distinguish between alleles as the PCR primers yielded similar products irrespective of genotype. However, in other cases this approach made it possible to gain insight to the allelic origin of individual BACs. PCR analysis using primers derived from the reverse end of BAC-91J7 did not amplify a product from plants homozygous for P1, but produced products of similar size from plants carrying either the P2 or the T allele (Fig. 4). In both these examples, further work is on-going to provide a definitive assignment.

In contrast, PCR analysis using primers derived from the forward end of BAC-47I4 provided clear confirmation on allelic origin (Fig. 4). No PCR products were amplified from the P1 allele. However, the P2 allele yielded a 310-bp amplification product and the T allele yielded a 630-bp PCR product. A faint 630-bp band can also be seen in the P2P2 plant cation product and the T allele yielded a 630-bp PCR product. However, the presence of the 630-bp product from BAC-47I4, together with the absence of the P2-specific band, reveals that this BAC derives from the thrum allele of the S locus.

In those cases where simple PCR analysis did not define allelic origin, restriction endonuclease digestion of PCR products is being used to reveal allelic variants. An example using BAC-81B15 is shown in Fig. 5. PCR amplification of *Primula* genomic DNA samples using primers derived from the forward end of BAC-81B15 produced faint amplification products from the P1 allele, and robust amplification from both P2 and T alleles. DNA sequence analysis of these PCR products revealed the differential presence of an Hpy CH4V restriction endonuclease cleavage site in different alleles. Subsequent digestion of PCR products with Hpy CH4V revealed the presence of an approx. 500-bp thrum-specific fragment which was also present in the BAC (Fig. 5). The allelic origin of every BAC is currently being characterized within the four contigs ahead of BAC sequence analysis.

**DISCUSSION**

As a prelude to the positional cloning of the *Primula S* locus we have used a series of approaches to identify S locus-linked
sequences using in-bred commercial lines of *P. vulgaris*, as well as the analysis of *S* locus-linked mutant phenotypes arising in horticultural populations. The availability of *F*₂ hybrid plants originating from in-bred commercial lines enabled us to search for thrum-specific polymorphisms between individuals with limited sequence variability outside the *S* locus. By using this probe to screen for BAC clones carrying *PvSLP1*, a 55-kb BAC (BAC-85B15) was identified which has been shown to be derived from the thrum allele of the *S* locus, both by PCR screening using the original thrum-specific SCAR primers (Manfield *et al.*, 2005), and independently by analysis of BAC-end sequences (Fig. 5). Subsequent walks in both directions by using BAC-end probes and reiterated library screens have identified a tiling path of ten BACs amounting to 595 kb. As for all four BAC contigs (Fig. 2), the degree of overlap between the BACs is not yet known to determine the physical length of each contig. Further analyses are also on-going to determine the allelic origin of each BAC. Given the proximity of *PvSLP1* to the *S* locus, this sequence represents a key starting point for a BAC walk even though it is not yet known whether the forward or reverse ends of BAC-85B15 face the *S* locus. However, by walking in both directions from this initial marker, one direction will take us towards the *S* locus.

Similarly, it is known that the *S* locus-linked gene *PvSLL1* is also very tightly linked to the *S* locus (Li *et al.*, 2007). This probe identified BAC-6K4 which contains a 60-kb insert. By using end-probe sequences and reiterated screening 16 BACs have been incorporated into a contig which includes a total of 865 kb. Using *PvSLL2* as a probe, BAC-56H19 which contains a 50 kb insert was isolated; this clone has made it possible to isolate four further overlapping BACs in both directions and representing 250 kb. *PvSLL2* is not completely linked to the *S* locus so represents a potential flanking marker for the locus (Li *et al.*, 2007). Contig 4 comprises eight overlapping BACs encompassing 565 kb, and was initiated with *PvGlo* which also represents a flanking marker for the *S* locus (Li *et al.*, 2010). Currently the potential overlap between these contigs is being characterized and continuing BAC library screens to obtain maximum coverage of the region. In the majority of cases, BAC-end probes have identified one or more hybridising BACs within the library. However, in some cases, end probes have contained highly repetitive sequences which have identified large numbers of hybridising BACs. These individual BACs are now being sequenced to identify low- or single-copy number sequences to use as probes to continue the chromosome walk. Assembly of contigs 1–4 into a single contig will confirm the order of the molecular markers and provide a contiguous span for 454 DNA sequence analysis.

The current focus on defining allelic origin of each BAC will make it possible to pool BACs derived from specific alleles for pooled 454 DNA sequence analysis. Individual BACs will also be sequenced to facilitate assembly of a sequence scaffold across the region. Fourteen new *S* locus-linked genes have already been identified within BAC-end sequence. Analysis of these sequences will not only provide additional allele-specific markers, but will also contribute to our assembly of a genetic map spanning the *S* locus. Analysis of the chromosomal location of each of the 14 arabidopsis sequences most closely related to these BAC-end genes does not provide any insight into syntenic relationships, other than the presence of two neighbouring arabidopsis genes in overlapping BAC clones from contig 1 (Fig. 2 and Table 2). As a result of sequence analysis and assembly of the BAC contigs it will be possible to explore the evolutionary relationships of gene order at, and surrounding, the *Primula S* locus with other species for which whole genome sequence data are available. The functions of the arabidopsis genes identified by *Primula* BAC-end sequences do not provide any insight into *S* locus function and it remains to be determined how far these genes are from the co-adapted linkage group of *G, P* and *A* which comprise the *S* locus. This information will only be revealed once the physical and genetic maps have been integrated and the on-going sequence analysis of the region completed. However, it is interesting to note that one of the gene similarities is involved in meiosis and another is involved in pollen development. It is anticipated that genes involved in different aspects of flower development, and unrelated to the control of heteromorphy will, by chance, be present in regions flanking the *S* locus. Indeed, it has already been shown that both *PvGlo* (Li *et al.*, 2010) and *sepaloid* (Li *et al.*, 2008), which have roles in petal identity, are linked to the *S* locus, as is *PvSLL2*, a gene with similarity to the *Constats* flower-timing genes (Li *et al.*, 2007).

The availability of DNA sequence spanning and flanking the *S* locus in *Primula* will enable us to identify all open reading frames within this region. This analysis will be greatly facilitated by the availability of pin- and thrum-specific flower cDNA libraries which will also aid in gene-structure predictions. From our classical genetic analyses we already have a range of plants carrying recombination events both within and flanking the *S* locus. Analysis of these plants, together with an on-going mutant screen for long and short homostyles will enable us to home in on the *S* locus and genes associated with it to create the opportunity to identify and characterize candidates for *G, P* and *A* which control the *Primula* breeding system as described by Darwin 150 years ago (Darwin, 1862).

Our progress towards an understanding of the *Primula S* locus represents just one avenue of research aimed at understanding the mechanisms of floral heteromorphy. Parallel work on *F. esculentum* and *T. subulata* also offer exciting opportunities to identify genes involved in heterostyly. The availability of *S* locus-linked markers (Ali *et al.*, 1998; Ota *et al.*, 2006), genetic maps (Yasui *et al.*, 2004; Konishi *et al.*, 2006) and BAC libraries (Yasui *et al.*, 2008) in *F. esculentum*, and the significant progress towards identification of the *S* locus in *T. subulata* following characterization of *S* locus-linked genes (Tamari and Shore, 2006; Labonne *et al.*, 2009), development of a high resolution map of this region (Labonne *et al.*, 2008), analysis of *S* locus deletion mutants (Labonne *et al.*, 2010) and positional cloning of a BAC encompassing the *T. subulata S* locus (Labonne and Shore, 2011), all demonstrate that we are close to gaining a molecular understanding of floral heteromorphy in more than one species. Once *S* locus sequences are available for comparison, opportunities will present themselves for a comparative study of these independently evolved systems that control development of different forms of flowers on plants of the same species.
ACKNOWLEDGEMENTS

We thank Neil Hall and other members of the Liverpool Centre for Genomic Research for their help with 454 high-throughput sequencing. We also thank David Westhead and John Whitaker at the University of Leeds for their help in transcriptome sequence assembly and Martin Lappage at Leeds University Experimental Gardens and Mike Hughes at Durham University Botanical Gardens for help with maintenance of plants. This work was supported by the Biotechnology and Biological Sciences Research Council (BB/H019278/1) to P.M.G. and J.L., the Gatsby Charitable Foundation and Durham University.

LITERATURE CITED


Darwin CR. 1877. The different forms of flowers on plants of the same species. London: John Murray.


