MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) doi: 10.1111/j.1755-0998.2011.03079.x

De novo transcriptome assembly and polymorphism detection in the flowering plant *Silene vulgaris* (Caryophyllaceae)

DANIEL B. SLOAN,* STEPHEN R. KELLER,† ANDREA E. BERARDI,* BRIAN J. SANDERSON,* JOHN F. KARPOVICH‡ and DOUGLAS R. TAYLOR*
*Department of Biology, University of Virginia, Charlottesville, VA 22903, USA, †Appalachian Laboratory, University of Maryland Center for Environmental Science, Frostburg, MD 21532, USA, ‡Department of Computer Science, University of Virginia, Charlottesville, VA 22903, USA

Abstract

Members of the angiosperm genus *Silene* are widely used in studies of ecology and evolution, but available genomic and population genetic resources within *Silene* remain limited. Deep transcriptome (i.e. expressed sequence tag or EST) sequencing has proven to be a rapid and cost-effective means to characterize gene content and identify polymorphic markers in non-model organisms. In this study, we report the results of 454 GS-FLX Titanium sequencing of a polyA-selected and normalized cDNA library from *Silene vulgaris*. The library was generated from a single pool of transcripts, combining RNA from leaf, root and floral tissue from three genetically divergent European subpopulations of *S. vulgaris*. A single full-plate 454 run produced 959 520 reads totalling 363.6 Mb of sequence data with an average read length of 379.0 bp after quality trimming and removal of custom library adaptors. We assembled 832 251 (86.7%) of these reads into 40 964 contigs, which have a total length of 25.4 Mb and can be organized into 18 178 graph-based clusters or 'isogroups'. Assembled sequences were annotated based on homology to genes in multiple public databases. Analysis of sequence variants identified 13 432 putative single-nucleotide polymorphisms (SNPs) and 1320 simple sequence repeats (SSRs) that are candidates for microsatellite analysis. Estimates of nucleotide diversity from 1577 contigs were used to generate genome-wide distributions that revealed several outliers with high diversity. All of these resources are publicly available through NCBI and/or our website (http://silenegenomics.biology.virginia.edu) and should provide valuable genomic and population genetic tools for the *Silene* research community.

Keywords: 454 sequencing, microsatellites, nucleotide diversity, *Silene vulgaris*, single-nucleotide polymorphisms, transcriptome

Received 23 June 2011; revision received 1 September 2011; accepted 15 September 2011

Introduction

Silene L. is a large and diverse genus of flowering plants within the Caryophyllaceae, consisting of approximately 700 predominantly herbaceous species (Brach & Song 2006). Many of these species are of great interest in a variety of ecological and evolutionary fields, including breeding system and sex chromosome evolution, host–pathogen dynamics, biological invasions, organelle genome transmission and evolution, plant–pollinator interactions, metapopulation dynamics, and speciation (reviewed in Bernasconi *et al*. 2009).

Silene vulgaris (Moench) Garcke ('bladder campion') is one of the most well-studied species in this genus. It is phenotypically diverse and widely distributed across its native range of Eurasia and North Africa (Marsden-Jones & Turrill 1957; Jalas & Suominen 1987). It has also been introduced to other regions such as North America, South America and Australia where it colonizes disturbed habitats such as roadsides, railroad tracks and agricultural settings (Marsden-Jones & Turrill 1957; Randall 2002). This history of expansion into non-native regions has been the focus of studies investigating the influence of invasive range expansions on population genetic structure and local adaptation (Taylor & Keller 2007; Keller *et al*. 2009; Keller & Taylor 2010). *Silene vulgaris* is also well studied for its gynodioecious breeding system (Clapham *et al*. 1952; Desfeux *et al*. 1996), meaning that populations consist of a mixture of hermaphroditic and female (i.e. male-sterile) individuals. Male sterility in this species is probably determined by a complex interaction between mitochondrially encoded cytoplasmic male

Correspondence: Daniel B. Sloan, Department of Ecology and Evolutionary Biology, West Campus, Yale University, West Haven, CT 06516, USA, Fax: 203-737-3109; E-mail: daniel.sloan@yale.edu

sterility factors and nuclear restorers of fertility (Charlesworth & Laporte 1998; Taylor *et al.* 2001; Olson & McCauley 2002). Evolution of this breeding system has motivated studies on the rate of sequence evolution and the maintenance of genetic diversity in *S. vulgaris* organelle genomes (Ingvarsson & Taylor 2002; Houliston & Olson 2006; Barr *et al.* 2007; Sloan *et al.* 2008a; Touzet & Delph 2009), as well as on the mode of inheritance of these genomes, which is characterized by a low frequency of paternal 'leakage' (McCauley *et al.* 2007; Bentley *et al.* 2010). *Silene vulgaris* has also been of interest because of its complex evolutionary relationship with the specialist pathogen *Microbotryum*, i.e. anther smut fungus (Antonovics *et al.* 2002; Le Gac *et al.* 2007; Refregier *et al.* 2008; Sloan *et al.* 2008b), and because of its history of local adaptation to metal-contaminated soils (Schat & Ten Bookum 1992; Brown *et al.* 1995; Schat *et al.* 1996).

Despite the research interests in *Silene*, none of the species in this genus (including *S. vulgaris*) are genetic model systems. Although genomic and population genetic resources are generally lacking in *Silene*, a set of polymorphic microsatellite markers has been identified through a combination of studies (Juillet *et al.* 2003; Tero & Schlötterer 2005; Teixeira & Bernasconi 2007; Molecular Ecology Resources Primer Development Consortium 2010a,b), including an analysis of ESTs in *S. latifolia* (Moccia *et al.* 2009). This earlier expressed sequence tag (EST) study was based on costly Sanger sequencing and therefore was limited to a total of 3662 cDNA clones, a modest sample compared to the current and ever-improving standards for high-throughput sequencing.

The advent of 'next-generation' DNA sequencing technology in recent years has created unprecedented opportunities for generating genomic information in previously uncharacterized systems. Roche's 454 GS-FLX platform has been widely used for transcriptome sequencing in non-model organisms (Margulies *et al.* 2005; Wheat 2010), with a single run now capable of producing approximately one million sequencing reads with a mean length of ~400 bp and a mode of >500 bp. These read lengths enable effective assembly of sequence data even in the absence of a reference genome sequence and have spurred a recent flurry of transcriptome projects in non-model organisms (e.g. Vera *et al.* 2008; Meyer *et al.* 2009; Schwartz *et al.* 2010), including numerous angiosperms (Novaes *et al.* 2008; Alagna *et al.* 2009; Barakat *et al.* 2009; Bellin *et al.* 2009; Soltis *et al.* 2009; Wang *et al.* 2009; Buggs *et al.* 2010; Peng *et al.* 2010; Sun *et al.* 2010; Zeng *et al.* 2010, 2011; Angeloni *et al.* 2011; Bai *et al.* 2011; Blanca *et al.* 2011; Dutta *et al.* 2011; Franssen *et al.* 2011; Fu *et al.* 2011; Kaur *et al.* 2011; Logacheva *et al.* 2011; Lu *et al.* 2011; McDowell *et al.* 2011; Portnoy *et al.* 2011; Swarbreck *et al.* 2011; Yang *et al.* 2011). In this study, we report the results of using 454 sequencing to analyse a normalized cDNA library derived from a pooled sample of multiple tissue types and populations of *S. vulgaris*, including functional annotation of expressed gene content, marker discovery and an analysis of nucleotide diversity. Coupled with the recent release of 454 sequence data from non-normalized transcriptome libraries divided across multiple *Silene* species (Blavet *et al.* 2011), our study presents a valuable resource for advancing research in this genus.

Materials and methods

Study species, source populations and growing conditions

Like most *Silene* species, *S. vulgaris* is diploid ($2n = 24$). Its (haploid) genome size or *C*-value has been estimated to be 1.13 pg (1.10 Gb) (Siroky *et al.* 2001). The genetic population structure of this species has been the subject of multiple studies in both its native and introduced ranges, using a combination of amplified fragment length polymorphisms and plastid DNA sequences (Taylor & Keller 2007; Keller *et al.* 2009; Keller & Taylor 2010). These studies have found that the genetic diversity of *S. vulgaris* is subdivided at a regional scale among three major demes with strong geographical signatures, approximately corresponding to the southern, eastern and western portions of its native European range. To maximize the sampling of genetic diversity in this study, we chose one family from each of these three European demes (Table 1). Seeds from field collections originally made in July 2004 were sown in Fafard 3B soil mix (Conrad Fafard, Inc.; Agawam, MA) and germinated on a mist bench in July 2009. After germination, seedlings were grown in the University of Virginia greenhouses with regular watering, biweekly treatment with 20-20-20 NPK fertilizer and supplemental lighting provided on a 16-h/8-h light/-dark cycle.

Table 1 Summary of *Silene vulgaris* source populations

Population code (family number)	Location	Deme*	RNA source
ALN1 (v 5)	Alençon, France	West	100% leaf
BOL (v 1)	Bollène, France	South	50% leaf; 50% root
SEE5 (v 2)	Seefeld, Austria	East	50% leaf; 50% flower

*As defined by Bayesian clustering of amplified fragment length polymorphism genotype data (Keller & Taylor 2010).

RNA extraction

Eight weeks after sowing, total cellular RNA was extracted using an RNeasy Plant Mini Kit (Qiagen, Inc., Valencia, CA, USA). RNA was extracted from the youngest cauline leaf of a single individual from each of the three families (ALN1, BOL and SEE5). To improve the representation of genes expressed in a tissue-specific fashion, we also extracted RNA from root and floral tissue. The entire root system was harvested from a maternal sibling of the BOL plant used for leaf extraction. It was then rinsed thoroughly with tap water and stripped of most of its fine, secondary roots to remove soil. To further minimize contamination from external soil microbes, the tissue was washed in a 5% bleach solution for 30 s and then rinsed thoroughly with tap water. Approximately 50 mg of tissue from the upper portion of the root system was used for extraction. For the floral sample, a single RNA extraction was performed on tissue pooled from two individuals from the SEE5 maternal family. An entire male-sterile flower was included as well as two stamens and two petals from a single hermaphroditic flower. The hermaphroditic flower was from the same individual used for leaf extraction. Pooling of tissue from both hermaphrodite and male-sterile flowers was performed to increase the chance of capturing transcripts specific to either of the two flower types.

Extracted RNA was quantified using a NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific, Inc., Wilmington, DE, USA), and RNA quality was verified by running an RNA Pico 6000 chip on a Bioanalyzer 2100 (Agilent Technologies, Inc., Santa Clara, CA, USA). RNA was pooled in equal quantities from all three families. The contributions from the BOL and SEE5 families were split evenly between leaf and root/floral RNA, whereas only leaf RNA was included from the ALN1 family (Table 1).

cDNA library construction and 454 sequencing

A pooled RNA sample (8 μg) was provided to Indiana University's Center for Genomics and Bioinformatics (CGB) for construction of a cDNA library using previously described protocols (Schwartz *et al*. 2010; Carter *et al.* in press). Library construction involved selection for polyadenylated (polyA$^+$) transcripts to enrich for protein-coding mRNAs and normalization to improve the representation of low-copy transcripts and thereby maximize gene discovery. The resulting library was sequenced with a full picotiter plate on a 454 GS-FLX sequencer with Titanium reagents (Roche Applied Science, Indianapolis, IN, USA). Sequencing was performed in the Department of Biology's Genomics Core Facility at the University of Virginia, using standard 454 protocols. Image and signal processing were conducted with gsRunProcessor version 2.3, and the resulting sequence data were processed by the CGB to remove custom adapters involved in cDNA library construction.

Transcriptome assembly

Assembly of transcriptome data sets is complicated by basic biological features of eukaryotic gene expression, including alternative splicing (Modrek & Lee 2002). In this study, use of (non-inbred) diploid organisms and the pooling of RNA from multiple individuals added further complexity. In particular, the assembly process must account for the co-existence of highly related but distinct sequences that could represent alternative splice variants of the same allele, different alleles from the same locus or entirely different loci (i.e. paralogs). To address these complexities, Roche's GS *de Novo* Assembler version 2.3 ('Newbler') employs a network or graph-based approach (Fig. 1) to describe the connectivity between assembled contigs. For example, in the case of two alternative splice variants, a single shared sequence can be connected to two different flanking sequences. In such cases, Newbler introduces breaks into the assembly and splits individual reads into pieces that are placed in different contigs. These split reads can then be used to define the alternative connections between contigs. Newbler uses these data to organize assembled contigs into clusters (termed 'isogroups'). These groups are intended to represent all contigs from a given genetic locus (although grouping of closely related paralogous loci is also possible). Within each isogroup, contigs can be connected in different permutations (termed 'isotigs'), each of which can be loosely thought of as a specific splice variant or allele.

We assembled our cleaned 454 sequencing reads with Newbler using the following command line settings: –cdna, –het, –it 500, –ig 1000, –icc 200. After assembly, output files were processed with a perl script (fix454AllContigsForCdnaAssembly.pl) provided by Roche to correct known errors associated with the length of reported contig sequences from Newbler version 2.3. The reported assembly statistics and subsequent annotation steps excluded contigs that were <100 bp in length.

Transcriptome annotation

Annotation methods were based on the approach described previously by Meyer *et al.* (2009). Assembled sequences (both contigs and isotigs) were searched against the NCBI nr and Uniprot Swiss-Prot protein databases with NCBI-BLASTX version 2.2.22 (Camacho *et al.* 2009) to putatively identify homologous genes in other species. BLAST output was parsed with modified

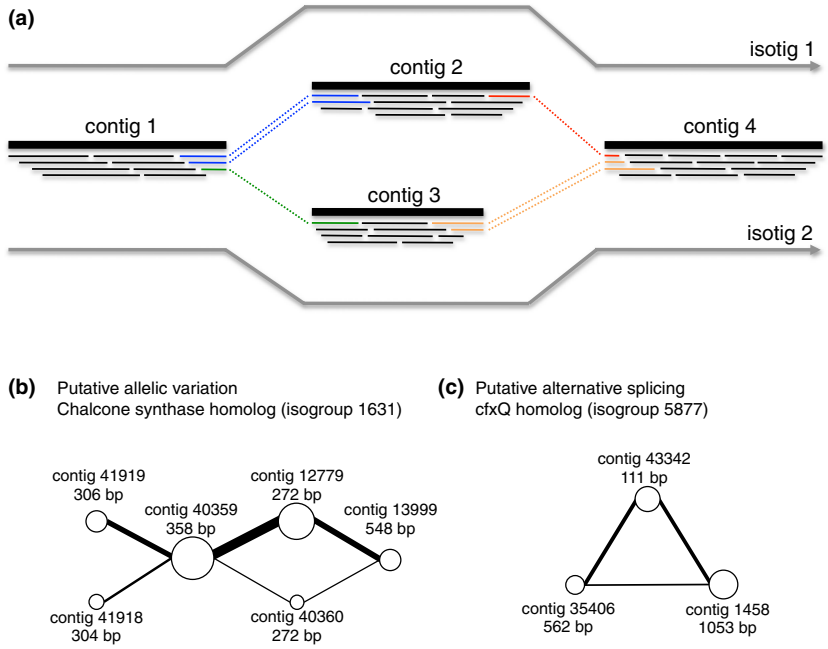

Fig. 1 Graph-based transcriptome assembly with Newbler version 2.3. (a) A stylized example of a single 'isogroup' containing four contigs (thick black lines). The fine lines represent individual sequencing reads, and the dotted lines connect the two ends of reads that have been split between two contigs. The outer grey arrows trace the paths of the different possible 'isotigs', which can represent allelic variants or alternative splice forms. The ball and stick networks depict actual isogroups from the *Silene vulgaris* transcriptome assembly that are consistent with either allelic variants (b) or alternative splice forms (c), but note that allelism and alternative splicing have not been experimentally verified in either of these cases. The circles represent contigs, with circle size reflecting the average read depth for the contig. The thickness of lines connecting contigs is proportional to the number of linking reads.

versions of published scripts (Meyer *et al.* 2009). These scripts use BioPerl modules (Stajich *et al.* 2002) to obtain gene name and taxonomic information from NCBI based on the accession number for each BLAST hit. The results were filtered to return the top hit, excluding genes annotated with terms such as 'hypothetical', 'predicted', 'uncharacterized', 'unknown' and 'unnamed'. Assembled sequences were also searched against the Uniprot TrEMBL database, and the resulting BLAST output was analysed with Blast2GO version 2.4.2 (Conesa *et al.* 2005), using default parameter settings to assign each sequence with gene ontology (GO) terms (biological process, molecular function and cellular component). Finally, assembled sequences were searched against the Pfam database using NCBI-RPSTBLASTN to identify conserved protein domains. All BLAST searches were executed with an *e*-value cut-off of 0.001 and a maximum of 50 hits. BLAST searches were performed locally using either formatted BLAST databases downloaded from NCBI (nr and Pfam) or FASTA files downloaded from Uniprot (Swiss-Prot and TrEMBL) and subsequently converted to local BLAST databases. All sequence databases were downloaded on 25 April 2010. Because of the computationally intensive nature of searching tens of thousands of sequences against multiple large BLAST databases, these searches were executed using hundreds of Linux nodes on the University of Virginia's Cross Campus Grid (XCG).

The taxonomic distribution of BLAST hits was analysed with MEGAN version 4.32. This software package uses the NCBI taxonomy to identify the most recent common ancestor of the set of BLAST hits from each query sequence. Analysis was performed on the BLASTX results from searching the assembled transcriptome contigs against the NCBI nr database, using default parameters in MEGAN (min support 5, min score 35, top per cent 10, win score 0).

Single-nucleotide polymorphism (SNP) detection

To identify sequence variants within our pooled sample, all reads were mapped against the assembled contig sequences using Roche's GS Reference Mapper version 2.3 with the –cdna and –cref command line options. The high-confidence sequence variants reported by GS Reference Mapper were subsequently filtered to identify all biallelic SNPs in regions with high sequence coverage (≥20×) and with a high minor allele frequency (≥33%). Reported SNPs were also limited to positions with at least 200 bp of flanking sequence on each side to permit sufficient sequence for primer development for SNP assays.

Simple sequence repeat detection

Simple sequence repeats (SSRs) (i.e. microsatellites) are often highly polymorphic within species and therefore represent valuable tools for population genetic analysis. The assembled isotig sequences were analysed with msatcommander version 0.8.2 (Faircloth 2008) to identify all cases of 2- to 6-bp motifs present in at least four tandemly repeated copies and to design primers in sequences flanking these SSRs. Because multiple isotigs from the same isogroup can share sequences, the

resulting output was filtered to exclude duplicate SSRs within the same isogroup.

Expressed sequence tag nucleotide diversity

The pooled sequencing of three genetically divergent lines allowed us to characterize the distribution of nucleotide diversity across the transcriptome using recently derived estimators that correct for pooled samples (Futschik & Schlötterer 2010). For this analysis, we trimmed 454 reads to a minimum length of 50 bp and quality score >20. We then created a pseudo-reference genome based on the contigs from our *de novo* (Newbler) assembly for read mapping. Because the Newbler assembly output generates multiple contigs from a given homologous region of the genome (i.e. isogroup), we first clustered contigs based on a 95% similarity threshold using the algorithm CD-HIT-EST version 4.5.4 (Huang *et al.* 2010). This similarity-based clustering reduced the reference transcriptome from 46 953 contigs to 40 184 contigs. Clustered contigs were then used as a pseudo-reference genome to map trimmed 454 reads using the bwasw algorithm for long reads in BWA version 0.8.5a (Li & Durbin 2010). Mapped reads were converted to SAM format, alignments were filtered for MAPQ quality <20, and SNPs were identified using the pileup command in Samtools version 0.1.16 (Li *et al.* 2009).

We used the PoPoolation version 1.017 pipeline (Kofler *et al.* 2011) to calculate nucleotide diversity based on the number of segregating sites (θ_W) and pairwise differences (θ_π), corrected for the biases introduced by pooling of samples. Population genetic summary statistics were computed for each alignment using the PoPoolation script Variance-at-positions.pl with the following input settings: –pool-size 6, –min-count 6, –min-coverage 20, –max-coverage 1000, –min-qual 20. After excluding alignments of <500 bp from the resulting output, we obtained a total of 1577 contigs for further analysis. Analysis of contig nucleotide diversity was performed as a heuristic tool to describe the overall levels of coding region SNP diversity found in our transcriptome data set, and to permit a means for identifying those regions that display exceptionally high or low diversity. These analyses are not appropriate for estimating the nucleotide diversity as a property of the species because of variability in transcript presence and abundance that may introduce bias into estimates of nucleotide diversity from EST sequences, even in normalized libraries like those used here.

Results

454 sequencing

Sequencing of the *S. vulgaris* cDNA library on a full 454 plate produced a total of 968 840 reads with an average read length of 391.9 bp for a total yield of 379.7 Mb of sequence data. After processing to remove custom library adapters, the total read number was reduced to 959 520 with an average length of 379.0 bp and a total yield of 363.6 Mb.

Transcriptome assembly and annotation

Newbler assembled 86.7% of the 959 520 reads into a total of 40 964 contigs of at least 100 bp in length. Of the remaining 127 269 reads, 57 662 were characterized as singletons, while the other 69 607 were excluded from the assembly by Newbler based on criteria such as length, quality, repetitiveness and chimerism. The longest assembled contig is 4929 bp, while the mean contig length is 620 bp (Fig. 2a). Split sequencing reads connect the assembled contigs into a total of 18 178 isogroups, which can be approximately thought of as the number of

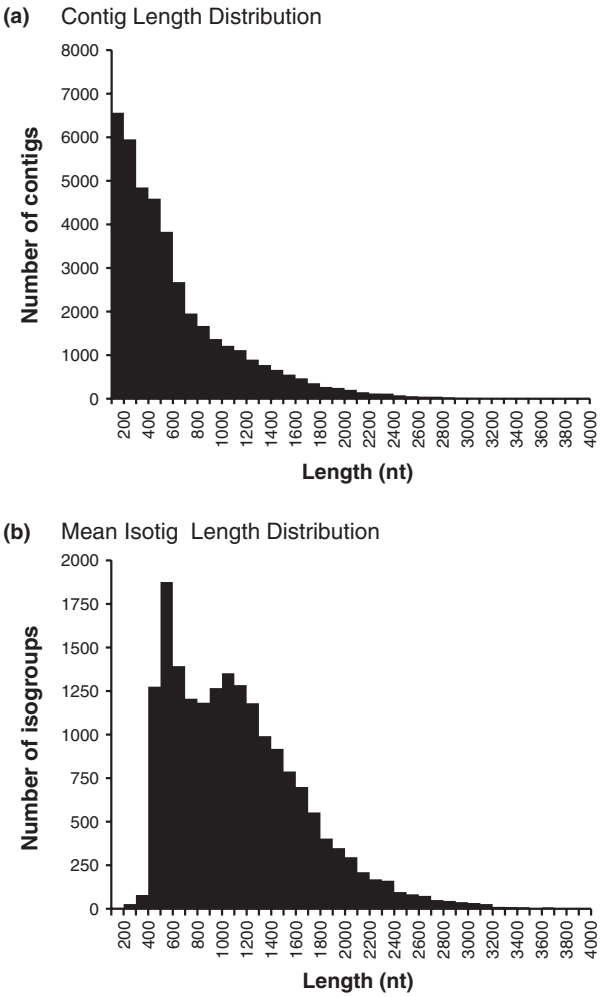

Fig. 2 Length distributions for contigs (a) and isotigs (b). Isotig lengths are reported as the mean for each isogroup.

distinct loci (Fig. 1). These isogroups contain as many as 122 contigs, but most isogroups (66.9%) contain only a single contig and the mean number of contigs per isogroup is only 2.6. Annotation results indicate that many of the largest and most complex isogroups represent highly duplicated gene families (e.g. retrotransposons and NBS-LRR genes, which are known to be involved in pathogen resistance) rather than individual genetic loci. When an isogroup contains multiple contigs, they can be connected in various permutations to form a suite of isotigs for that isogroup (approximately corresponding to the set of alternate alleles or splice variants). The assembly produced a total of 37 976 isotigs (including those composed of only a single contig). Each isogroup is associated with anywhere from 1 to 318 isotigs with a mean of 2.1, and each isotig comprises anywhere from 1 to 22 contigs with a mean of 3.8. The average and maximum isotig lengths are 1333 and 6230 bp, respectively (Fig. 2b). The average sequencing coverage for isotigs was 11.6× with a maximum of 78.2× (Fig. 3). Raw sequencing reads were deposited to NCBI's Short Read Archive (SRA037583). Trimmed reads and assembled sequences are available for download at http://silenegenomics.biology.virginia.edu.

Assembled sequences were annotated based on the results of BLAST comparisons against public sequence databases. Approximately 70% of contigs had a BLAST hit with an *E*-value of 1e-6 or better against one or more public sequence databases, and almost half of all contigs were assigned at least one GO annotation term (Table 2). Annotation data are available in a searchable online database and for download at http://silenegenomics.biology.virginia.edu.

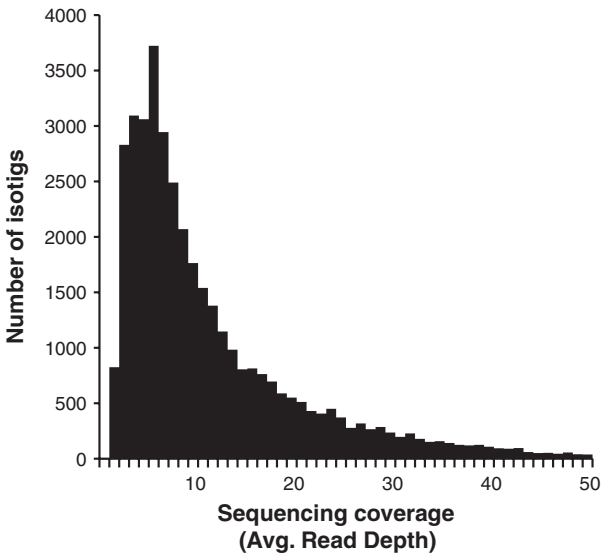

Fig. 3 Distribution of sequencing coverage across isotigs.

Table 2 Annotation summary

	Contigs	Isotigs	Isogroups
NCBI nr	70.1 (62.4)	89.2 (79.7)	88.4 (79.6)
Uniprot Swiss-Prot	50.0	69.7	67.9
Pfam	41.4	64.3	62.4
GO annotation (any)	47.7	59.6	60.7
Biological process	26.8	31.8	33.0
Molecular function	37.0	46.1	46.2
Cellular component	29.4	35.4	39.9

Values indicate the percentage of sequences/groups with one or more significant BLAST hits/annotations based on an *E*-value cut-off of 1e-6. For the NCBI nr database, values in parentheses indicate the results after excluding hits with uninformative annotations (see Materials and methods).

Taxonomic analysis assigned more than 97% of contigs with significant hits to a most recent common ancestor within the land plant lineage. Although *Silene* is unambiguously classified within the Caryophyllales, only 2.5% of contigs with significant hits were assigned specifically to this group, reflecting the limited availability of caryophyllid sequence data in public databases. In contrast, 47.0% were assigned within the rosids, which include well-characterized genomic model systems such as *Arabidopsis thaliana*, *Glycine max*, *Medicago truncatula*, *Populus trichocarpa*, *Ricinus communis* and *Vitis vinifera*. A total of 64 contigs were assigned to taxonomic nodes outside the land plants and off the *Silene* line of descent (Fig. 4). Although such contigs represent potential cases of contamination or horizontal gene transfer in our data set, most of these assignments were based on marginal BLAST hits near the significance threshold and often involved low-complexity sequences, which are prone to producing false-positive hits. Nevertheless, there are some sequences that are clearly of bacterial or viral origin, providing a distinct signature of plant pathogens within the *Silene* transcriptome data set. Most notably, the contigs within isogroup 296 exhibit clear homology to the cucumopine synthase gene encoded within the T_DNA region in the pRi2659 plasmid of *Agrobacterium rhizogenes*, which is transferred to plant host genomes to mediate *Agrobacterium* pathogenesis (Brevet *et al.* 1988; Suzuki *et al.* 2001). In addition, contig 749 appears to code for a protein that is closely related to the RNA-dependent RNA polymerase of the *Vicia* cryptic virus (Blawid *et al.* 2007).

SNP and SSR detection

GS Reference Mapper successfully mapped 69.5% of reads back onto the assembled transcriptome contigs. From these, we identified an initial total of 107 488 high-confidence SNPs or small indels (i.e. '454HCDiffs' from

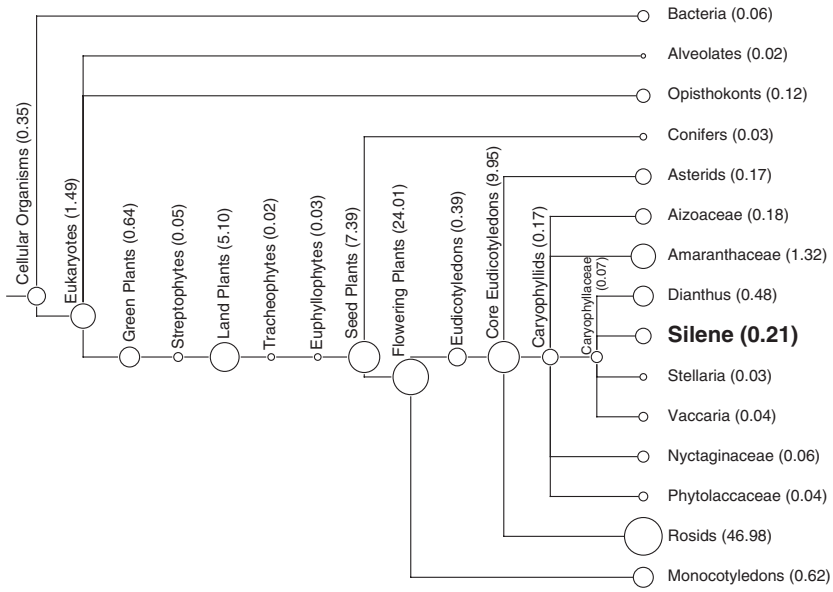

Fig. 4 Taxonomic analysis of assembled contigs. The size of each circle reflects the number of contigs assigned to the corresponding taxonomic node by MEGAN based on BLASTX searches against the NCBI nr database. The percentage of all assigned contigs is indicated in parentheses.

GS Reference Mapper). To increase the reliability of SNP identification, we filtered these results based on multiple criteria including read depth and allele frequency (see Materials and methods). The resulting data set includes a total of 13 432 biallelic SNPs distributed across 3435 contigs (3308 different isogroups). At these rather stringent thresholds, the distribution of SNPs per isogroup varied from 1 to 42, with an average of 3.0 SNPs per kb.

The program msatcommander identified a total of 1320 unique SSRs with up to 14 repeats and sufficient flanking sequence to design primers. This list of candidates represents a potentially large expansion of the modest number of microsatellite markers currently available for *S. vulgaris* (Juillet *et al*. 2003). In addition, a recent study of EST-SSRs in *S. latifolia* (Moccia *et al*. 2009) found a high degree of transferability across the genus, so many of the SSRs identified in our analysis may be useful for studies of other *Silene* species. However, the loci and primer sequences reported in our study have not been screened for variability, so some of these markers may not be polymorphic. In addition, because this analysis was based on cDNA sequences, a subset of these primers may span intron/exon boundaries and fail to amplify genomic DNA templates. An earlier study on microsatellite development in *Silene* tested primer pairs designed from EST sequences and found that 41% (30 of 74) produced successful amplification of polymorphic markers (Moccia *et al*. 2009).

Nucleotide diversity in this transcriptome resource

We characterized the nucleotide diversity of ESTs in this transcriptome resource after correcting for sampling within our pooled library. Among the 1577 alignments that passed quality and length filters, median estimates of EST nucleotide diversity in this transcriptome were similar for segregating sites ($\theta_W = 0.0088$) and pairwise differences ($\pi = 0.0085$). The shape of the distribution was unimodal but with a skewed upper tail, containing many contigs with high EST nucleotide diversity (Fig. 5). These contigs may represent genes with interesting evolutionary histories that have lead to high levels of accumulated polymorphism, but the possibility that some of the identified SNPs actually represent variation between duplicated genes (i.e. paralogs) cannot be ruled out. Functional annotations for contigs in the 99th percentile of the empirical distribution of nucleotide diversity ($\pi \geq 0.0357$ or $\theta_W \geq 0.0347$) are provided in Table S1 (Supporting information).

Discussion

The recent advances in DNA sequencing technology have been heralded as a 'democratizing' force that is bringing genomic tools in reach of individual laboratories and researchers working outside the confines of classic genetic model systems (Rothberg & Leamon 2008). With a single 454 sequencing run, we were able to broadly characterize expressed gene content and generate large sets of candidate markers for population genetic analysis in a species for which there were essentially no pre-existing genomic resources. This data set should provide a valuable tool for expediting research to address the many biological questions that have made the genus *Silene* a preferred object of study. For example, *Silene* mitochondrial genomes have recently been found to have lost numerous protein genes that are present in other angiosperms (Sloan *et al*. 2010). With a simple BLAST search of

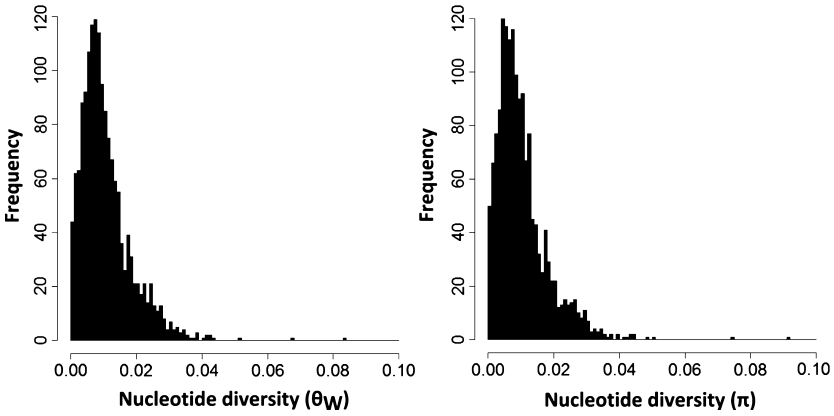

Fig. 5 Empirical distributions of nucleotide diversity (θ_W and π) in this *Silene vulgaris* transcriptome resource, estimated from the pooled sample of normalized cDNA from three families.

our newly developed transcriptome resources, it is now possible to show that *Silene vulgaris* contains homologs for at least nine of these mitochondrial genes, indicating that they have been functionally transferred to the nucleus (DBS, unpublished data). Furthermore, our data provide an opportunity to identify candidate genes involved in traits of ecological and evolutionary interest in *Silene*. For example, hundreds of sequences were annotated as containing NBS-LRR or PPR domains, which are known to be involved in host–pathogen interactions and restoration of cytoplasmic sterility, respectively (Belkhadir *et al.* 2004; Fujii *et al.* 2011).

Additionally, the availability of genome-wide SNP and SSR markers opens the door to population genomic approaches to characterize demographic history and selection (Stinchcombe & Hoekstra 2008; Siol *et al.* 2010). The SNP and SSR resources reported here add considerably to the availability of genomic markers in *Silene*, and our preliminary analysis of EST nucleotide diversity suggests an abundance of polymorphism for future molecular population genetic studies.

While the analysis of nucleotide diversity is useful for summarizing the levels of polymorphism in this transcriptome resource and for identifying those ESTs with exceptionally high diversity, these values should not viewed as directly comparable to estimates from genomic DNA or interpreted as an estimate of the nucleotide diversity in *Silene vulgaris* as a species. In particular, nucleotide diversity from EST sequences may be affected by individual or tissue-specific gene expression. For example, genes that are exclusively expressed in root or floral tissue should be represented by only one family in our pooled sample (Table 1), resulting in a downward bias on diversity estimates. We also expect values of θ_W to be biased slightly downwards owing to the necessity of applying a minimum allele count filter to guard against false-positive SNPs (and this is also why we do not report Tajima's *D*, which is sensitive to the frequency of rare variants). Thus, while additional sampling of genomic sequence will be necessary to infer demographic history and the effects of selection acting on particular gene regions, these initial results demonstrate the promise of next-generation sequencing for generating empirical distributions of nucleotide diversity across the genome of a non-model organism.

The rapid proliferation of DNA sequencing platforms and methods requires researchers to weigh the relative merits of different sequencing technologies and methodological approaches for generating genomic data for non-model organisms (Wall *et al.* 2009). For example, to accomplish our dual goals of broadly characterizing gene content and identifying polymorphic markers, we used a normalized library that pooled RNA from multiple source populations and tissue types. By pooling samples (without any form of multiplex tagging), we were able to minimize our library construction costs and still capture sequence diversity across populations and tissue types, while library normalization allowed us to obtain coverage for a greater number of genes (particularly those expressed at low levels). Combined with newly developed methods for analysing pooled genomic sequences (Futschik & Schlötterer 2010; Kofler *et al.* 2011), this approach led to a highly cost-effective means for generating polymorphism and functional annotation across a broad swath of the *Silene* transcriptome.

However, these choices come with trade-offs. Because our samples are anonymously pooled, we cannot tie sequences to specific individuals or tissue types. Likewise, normalization comes at the expense of (at least partially) losing quantitative information on levels of gene expression. Given that technical variance among replicates can be substantial for non-normalized transcriptome data (McIntyre *et al.* 2011), especially at low coverage, our decision to sequence a normalized cDNA library probably sacrificed little in terms of reliable data on gene expression but came with the benefit of greater exon sampling. Combining our current data set with other sequencing technologies should allow for addressing

additional biological questions. With the current 454-based reference assembly in hand, more cost-effective, short-read technologies (e.g. Illumina and SOLiD) may be used to analyse sequence variants and expression levels in specific individuals, tissue types or experimental treatments.

As read lengths continue to increase across platforms (Illumina sequencing can now produce paired-end runs of a 100 bp or more), *de novo* assembly will become more feasible without relatively expensive technologies such as Sanger or 454 (Li *et al*. 2010). Such efforts will also benefit from improvements in bioinformatic resources. A comparative study of multiple assemblers found that the recently updated version of Newbler (version 2.5) offers significant gains over the version 2.3 assembler used in this study and that the best results are obtained from merging output from multiple assemblers (Kumar & Blaxter 2010). Therefore, reanalysis of our raw data will probably yield improvements in assembly length and quality.

In combination with the results from another recent *Silene* transcriptome study (Blavet *et al*. 2011), our data provide a valuable new tool for ecological and evolutionary research in this genus. These two *Silene* transcriptome studies complement each other well. With longer 'Titanium' reads and deeper sequencing of a normalized library from a single species, our study provides a broad characterization of gene content. In contrast, Blavet *et al*. used non-normalized libraries split across multiple species, generating valuable information on interspecific diversity and gene expression levels. In addition to providing an immediate resource for the *Silene* research community, these data sets should also represent a useful tool for future projects such as whole-genome sequencing and annotation in *Silene*.

Acknowledgements

We thank Keithanne Mockaitis and Indiana University's CGB for cDNA library construction and John Chuckalovcak for operation of the 454 sequencer at the University of Virginia. We are also grateful to Carey Hill and Hamp Carruth for their assistance with website design and hosting and three anonymous reviewers for comments that improved an earlier version of this manuscript. This study was supported by funding from the NSF (MCB-1022128) and a CLU grant from the University of Virginia.

References

Alagna F, D'Agostino N, Torchia L *et al.* (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics*, **10**, 399.

Angeloni F, Wagemaker CAM, Jetten MSM *et al*. (2011) *De novo* transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques. *Molecular Ecology Resources*, **11**, 662–674.

Antonovics J, Hood M, Partain J (2002) The ecology and genetics of a host shift: *Microbotryum* as a model system. *American Naturalist*, **160**, S40–S53.

Bai X, Rivera-Vega L, Mamidala P, Bonello P, Herms DA, Mittapalli O (2011) Transcriptomic signatures of ash (*Fraxinus* spp.) phloem. *PLoS ONE*, **6**, e16368.

Barakat A, DiLoreto DS, Zhang Y *et al.* (2009) Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. *BMC Plant Biology*, **9**, 51.

Barr CM, Keller SR, Ingvarsson PK, Sloan DB, Taylor DR (2007) Variation in mutation rate and polymorphism among mitochondrial genes in *Silene vulgaris*. *Molecular Biology and Evolution*, **24**, 1783–1791.

Belkhadir Y, Subramaniam R, Dangl JL (2004) Plant disease resistance protein signaling: NBS-LRR proteins and their partners. *Current Opinion in Plant Biology*, **7**, 391–399.

Bellin D, Ferrarini A, Chimento A *et al.* (2009) Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species. *BMC Genomics*, **10**, 555.

Bentley KE, Mandel JR, McCauley DE (2010) Paternal leakage and heteroplasmy of mitochondrial genomes in *Silene vulgaris*: evidence from experimental crosses. *Genetics*, **185**, 961–968.

Bernasconi G, Antonovics J, Biere A *et al.* (2009) *Silene* as a model system in ecology and evolution. *Heredity*, **103**, 5–14.

Blanca J, Canizares J, Roig C, Ziarsolo P, Nuez F, Pico B (2011) Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (Cucurbitaceae). *BMC Genomics*, **12**, 104.

Blavet N, Charif D, Oger-Desfeux C, Marais G, Widmer A (2011) Comparative high-throughput transcriptome sequencing and development of SiESTa, the *Silene* EST annotation database. *BMC Genomics*, **12**, 376.

Blawid R, Stephan D, Maiss E (2007) Molecular characterization and detection of Vicia cryptic virus in different *Vicia faba* cultivars. *Archives of Virology*, **152**, 1477–1488.

Brach AR, Song H (2006) eFloras: new directions for online floras exemplified by the Flora of China Project. *Taxon*, **55**, 188.

Brevet J, Borowski D, Tempe J (1988) Identification of the region encoding opine synthesis and of a region involved in hairy root induction on the T-DNA of cucumber-type Ri plasmid. *Molecular Plant-Microbe Interactions*, **1**, 75–79.

Brown SL, Chaney RL, Angle JS, Baker AJM (1995) Zinc and cadmium uptake by hyperaccumulator *Thlaspi caerulescens* and metal tolerant *Silene vulgaris* grown on sludge-amended soils. *Environmental Science & Technology*, **29**, 1581–1585.

Buggs RJA, Chamala S, Wu W *et al*. (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology*, **19**, 132–146.

Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Carter J, Smith Z, Mockaitis K (in press) In: *Methods in Molecular Biology* (ed. Springer P).

Charlesworth D, Laporte V (1998) The male-sterility polymorphism of *Silene vulgaris*: analysis of genetic data from two populations and comparison with *Thymus vulgaris*. *Genetics*, **150**, 1267–1282.

Clapham AR, Tutin TG, Warburg EF (1952) *Flora of the British Isles*. Cambridge University Press, Cambridge.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Desfeux C, Maurice S, Henry JP, Lejeune B, Gouyon PH (1996) Evolution of reproductive systems in the genus *Silene*. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **263**, 409–414.

Dutta S, Kumawat G, Singh BP *et al.* (2011) Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. *BMC Plant Biology*, **11**, 17.

Faircloth BC (2008) msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources*, **8**, 92–94.

Franssen SU, Shrestha RP, Brautigam A, Bornberg-Bauer E, Weber APM (2011) Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. *BMC Genomics*, **12**, 1.

Fu CH, Chen YW, Hsiao YY *et al.* (2011) OrchidBase: a collection of sequences of the transcriptome derived from orchids. *Plant & Cell Physiology*, **52**, 238–243.

Fujii S, Bond CS, Small ID (2011) Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 1723–1728.

Futschik A, Schlötterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, **186**, 207–218.

Houliston GJ, Olson MS (2006) Nonneutral evolution of organelle genes in *Silene vulgaris*. *Genetics*, **174**, 1983–1994.

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Ingvarsson PK, Taylor DR (2002) Genealogical evidence for epidemics of selfish genes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11265–11269.

Jalas J & Suominen J (eds) (1987) *Atlas Florae Europaeae. Distribution of Vascular Plants in Europe. III. Caryophyllaceae*. Cambridge University Press, Cambridge.

Juillet N, Freymond HÉ, Degen L, Goudet JÉÔ (2003) Isolation and characterization of highly polymorphic microsatellite loci in the bladder campion, *Silene vulgaris* (Caryophyllaceae). *Molecular Ecology Notes*, **3**, 358–359.

Kaur S, Cogan NOI, Pembleton LW *et al.* (2011) Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genomics*, **12**, 265.

Keller SR, Taylor DR (2010) Genomic admixture increases fitness during a biological invasion. *Journal of Evolutionary Biology*, **23**, 1720–1731.

Keller SR, Sowell DR, Neiman M, Wolfe LM, Taylor DR (2009) Adaptation and colonization history affect the evolution of clines in two introduced species. *The New Phytologist*, **183**, 678–690.

Kofler R, Orozco-terWengel P, De Maio N *et al.* (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE*, **6**, e15925.

Kumar S, Blaxter ML (2010) Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics*, **11**, 571.

Le Gac M, Hood ME, Fournier E, Giraud T (2007) Phylogenetic evidence of host-specific cryptic species in the anther smut fungus. *Evolution*, **61**, 15–26.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li R, Fan W, Tian G *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Logacheva MD, Kasianov AS, Vinogradov DV *et al.* (2011) *De novo* sequencing and characterization of floral transcriptome in two species of buckwheat (*Fagopyrum*). *BMC Genomics*, **12**, 30.

Lu FH, Yoon MY, Cho YI *et al.* (2011) Transcriptome analysis and SNP/SSR marker information of red pepper variety YCM334 and Taean. *Scientia Horticulturae*, **129**, 38–45.

Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Marsden-Jones EM, Turrill WB (1957) *The Bladder Campions:* Silene maritima *and* S. vulgaris. Ray Society, London.

McCauley DE, Sundby AK, Bailey MF, Welch ME (2007) Inheritance of chloroplast DNA is not strictly maternal in *Silene vulgaris* (Caryophyllaceae): evidence from experimental crosses and natural populations. *American Journal of Botany*, **94**, 1333.

McDowell ET, Kapteyn J, Schmidt A *et al.* (2011) Comparative functional genomic analysis of *Solanum* glandular trichome types. *Plant Physiology*, **155**, 524–539.

McIntyre LM, Lopiano KK, Morse AM *et al.* (2011) RNA-seq: technical variability and sampling. *BMC Genomics*, **12**, 293.

Meyer E, Aglyamova GV, Wang S *et al.* (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, **10**, 219.

Moccia MD, Oger-Desfeux C, Marais GA, Widmer A (2009) A White Campion (*Silene latifolia*) floral expressed sequence tag (EST) library: annotation, EST-SSR characterization, transferability, and utility for comparative mapping. *BMC Genomics*, **10**, 243.

Modrek B, Lee C (2002) A genomic view of alternative splicing. *Nature Genetics*, **30**, 13–19.

Molecular Ecology Resources Primer Development Consortium (2010a) Permanent genetic resources added to Molecular Ecology Resources database 1 August 2009–30 September 2009. *Molecular Ecology Resources*, **10**, 232–236.

Molecular Ecology Resources Primer Development Consortium (2010b) Permanent genetic resources added to Molecular Ecology Resources database 1 October 2009–30 November 2009. *Molecular Ecology Resources*, **10**, 404–408.

Novaes E, Drost DR, Farmerie WG *et al.* (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics*, **9**, 312.

Olson MS, McCauley DE (2002) Mitochondrial DNA diversity, population structure, and gender association in the gynodioecious plant *Silene vulgaris*. *Evolution*, **56**, 253–262.

Peng Y, Abercrombie LL, Yuan JS *et al.* (2010) Characterization of the horseweed (*Conyza canadensis*) transcriptome using GS-FLX 454 pyrosequencing and its application for expression analysis of candidate non-target herbicide resistance genes. *Pest Management Science*, **66**, 1053–1062.

Portnoy VD, Pollock A, Karchi S *et al.* (2011) Use of non-normalized, non-amplified cDNA for 454-based RNA sequencing of fleshy melon fruit. *The Plant Genome*, **4**, 36–46.

Randall RP (2002) *A Global Compendium of Weeds*. R.G. and F.J. Richardson, Melbourne.

Refregier G, Le Gac M, Jabbour F *et al.* (2008) Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation. *BMC Evolutionary Biology*, **8**, 100.

Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. *Nature Biotechnology*, **26**, 1117–1124.

Schat H, Ten Bookum WM (1992) Genetic control of copper tolerance in *Silene vulgaris*. *Heredity*, **68**, 219–229.

Schat H, Vooijs R, Kuiper E (1996) Identical major gene loci for heavy metal tolerances that have independently evolved in different local populations and subspecies of *Silene vulgaris*. *Evolution*, **50**, 1888–1895.

Schwartz TS, Tae H, Yang Y *et al.* (2010) A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences. *BMC Genomics*, **11**, 694.

Siol M, Wright SI, Barrett SC (2010) The population genomics of plant adaptation. *The New Phytologist*, **188**, 313–332.

Siroky J, Lysak MA, Dolezel J, Kejnovsky E, Vyskot B (2001) Heterogeneity of rDNA distribution and genome size in *Silene* spp. *Chromosome Research*, **9**, 387–393.

Sloan DB, Barr CM, Olson MS, Keller SR, Taylor DR (2008a) Evolutionary rate variation at multiple levels of biological organization in plant mitochondrial DNA. *Molecular Biology and Evolution*, **25**, 243–246.

Sloan DB, Giraud T, Hood ME (2008b) Maximized virulence in a sterilizing pathogen: the anther-smut fungus and its co-evolved hosts. *Journal of Evolutionary Biology*, **21**, 1544–1554.

Sloan DB, Alverson AJ, Storchova H, Palmer JD, Taylor DR (2010) Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. *BMC Evolutionary Biology*, **10**, 274.

Soltis DE, Albert VA, Leebens-Mack J *et al*. (2009) Polyploidy and angiosperm diversification. *American Journal of Botany*, **96**, 336–348.

Stajich JE, Block D, Boulez K *et al*. (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Research*, **12**, 1611–1618.

Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity*, **100**, 158–170.

Sun C, Li Y, Wu Q *et al*. (2010) De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genomics*, **11**, 262.

Suzuki K, Tanaka N, Kamada H, Yamashita I (2001) Mikimopine synthase (mis) gene on pRi1724. *Gene*, **263**, 49–58.

Swarbreck SM, Lindquist EA, Ackerly DD, Andersen GL (2011) Analysis of leaf and root transcriptomes of soil-grown *Avena barbata* plants. *Plant & Cell Physiology*, **52**, 317–332.

Taylor DR, Keller SR (2007) Historical range expansion determines the phylogenetic diversity introduced during contemporary species invasion. *Evolution*, **61**, 334–345.

Taylor DR, Olson MS, McCauley DE (2001) A quantitative genetic analysis of nuclear-cytoplasmic male sterility in structured populations of *Silene vulgaris*. *Genetics*, **158**, 833–841.

Teixeira S, Bernasconi G (2007) High prevalence of multiple paternity within fruits in natural populations of *Silene latifolia*, as revealed by microsatellite DNA analysis. *Molecular Ecology*, **16**, 4370–4379.

Tero N, Schlötterer C (2005) Isolation and characterization of microsatellite loci from *Silene tatarica*. *Molecular Ecology Notes*, **5**, 517–518.

Touzet P, Delph LF (2009) The effect of breeding system on polymorphism in mitochondrial genes of *Silene*. *Genetics*, **181**, 631–644.

Vera JC, Wheat CW, Fescemyer HW *et al*. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, **17**, 1636–1647.

Wall PK, Leebens-Mack J, Chanderbali AS *et al*. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**, 347.

Wang W, Wang Y, Zhang Q, Qi Y, Guo D (2009) Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics*, **10**, 465.

Wheat CW (2010) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica*, **138**, 433–451.

Yang SS, Tu ZJ, Cheung F *et al*. (2011) Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. *BMC Genomics*, **12**, 199.

Zeng S, Xiao G, Guo J *et al*. (2010) Development of a EST dataset and characterization of EST-SSRs in a traditional Chinese medicinal plant, *Epimedium sagittatum* (Sieb. Et Zucc.) Maxim. *BMC Genomics*, **11**, 94.

Zeng Y, Conner JA, Ozias-Akins P (2011) Identification of ovule transcripts from the Apospory-Specific Genomic Region (ASGR)-carrier chromosome. *BMC Genomics*, **12**, 206.

Data Accessibility

Raw sequence data are available from the NCBI Sequence Read Archive (SRA037583). Trimmed sequences, assembly data and additional resources are available at our website (http://silenegenomics.biology.virginia.edu).

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Summary and annotation data for contigs distributed in the 99th quantile of nucleotide diversity estimated by π or θ_W.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.