De novo Transcriptome Assembly and Dynamic Spatial Gene Expression Analysis in Red Clover

Manohar Chakrabarti, Randy D. Dinkins,* and Arthur G. Hunt*

Abstract
Red clover (Trifolium pratense L.) is a cool-season forage legume grown throughout the northeastern United States and is the most widely planted forage legume after alfalfa (Medicago sativa L.). Red clover provides high-value feed to the livestock because of high protein content and easy digestibility. To date, genomic resources for red clover are scarce. In the current study, a de novo transcriptome assembly of red clover was constructed representing different tissue types. The draft assembly consists of 37,565 contigs with N50 and average contig length of 1707 and 1262 bp, respectively. A comparative study with three other legume species displayed a high degree of sequence conservation between red clover and other legumes. The assembled transcriptome was annotated to allow identification of desirable genes. In particular, a genome-wide identification of red clover transcripts encoding putative transcription factors was performed. A comparative gene expression analysis between different tissue types was performed using the assembled transcriptome as the reference, which revealed dynamic gene expression patterns across different tissue types and also identified spatially dynamic gene coexpression clusters. Genes representing tissue-enriched clusters were subjected to gene ontology (GO) enrichment analysis to identify over-represented functional groups. Identification of these tissue-enriched gene coexpression clusters can help in future research focusing on developmental studies across tissues or in biotechnological improvement of red clover.

Core Ideas
- De novo transcriptome assembly of red clover (Trifolium pratense L.)
- Functional annotation of the assembled transcriptome
- Global identification of red clover transcripts encoding putative transcription factors
- Identification of tissue-enriched gene coexpression clusters

Red clover is a diploid (2n = 14), naturally cross-pollinated, cool-season forage legume (Smith et al., 1985). In general, clover has three primary centers of diversity: Eurasia, America, and Africa. Red clover probably originated in southeastern Europe and Asia Minor and inhabits temperate regions of the world (Taylor, 1985). In the United States, it is the most widely grown forage legume after alfalfa and is grown throughout the northeastern United States. If the climate is favorable, red clover can adapt to a wide range of soils. It can fix atmospheric N through symbiosis with Rhizobium species and is estimated to fix approximately 125 to 220 kg...
of N ha\(^{-1}\) yr\(^{-1}\) (LeRue and Patterson, 1981). It is suitable for hay, silage, pasture, and soil improvement and can be grown either as monocrop or with other grasses (Smith et al., 1985). It can be easily established, it is fast-growing, and can also be intercropped with silage corn (\textit{Zea mays} L.) to control soil erosion (Wall et al., 1991). Digestibility, protein content, and voluntary intake make red clover a high-value feed for livestock (Frame et al., 1998). It also increases micronutrient content in forage mixture (Lindstrom et al., 2014). In the last several decades, cultivation of clover has declined mostly because of low price of the nitrogenous fertilizers; however, with the burgeoning interest in organic farming and sustainable agriculture, N-fixing forage legumes, like red clover, are attracting new attention (Taylor, 2008).

Genomic resources for red clover are very scarce compared with other model plant species. The genomes of several legume species have been sequenced including medicago (\textit{Medicago truncatula} Gaertn.), chickpea (\textit{Cicer arietinum} L.), and pigeonpea (\textit{Cajanus cajan} L. Huth) (Varshney et al., 2012, 2013; Young et al., 2011). Based on the flow cytometry analysis, red clover was estimated to have a genome size of 440 Mb as compared with 135 Mb for \textit{Arabidopsis thaliana} (L.) Heynh. and 257 Mb for Medicago (Sato et al., 2005). Because red clover is self-incompatible, it is predicted to have a highly polymorphic genome. Genomic heterozygosity within a population was also found to be higher than the interpopulation genomic heterozygosity (Campos-de-Quiroz and Ortega-Klose, 2001; Kongkiatngam et al., 1995; Milligan, 1991). No high-quality genome sequence for red clover has been reported yet. Being a nonmodel crop, it also has fewer expressed sequence tags (ESTs) than model legume species. However, with the dwindling cost of sequencing, next-generation sequencing has become the method of choice to build up genomic resources for the nonmodel crops (Bhardwaj et al., 2013; Pradhan et al., 2014; Wu et al., 2014). Rapid progress in the next generation sequencing technology has allowed researchers to underpin molecular regulation of developmental processes and responses to various stimuli including biotic and abiotic stresses on a genome-wide scale. De novo transcriptome assemblies based on next-generation sequencing have been developed for several nonmodel plants, such as sweet potato (\textit{Ipomoea batatas} (L.) Lam.), safflower (\textit{Carthamus tinctorius} L.), sesame (\textit{Sesamum indicum} L.), peanut (\textit{Arachis hypogaea} L.), pigeonpea, lentil (\textit{Lens culinaris} Medik.), pennisycress (\textit{Thlaspi arvense} L.), cacao (\textit{Theobroma cacao} L.), and many more (Dorn et al., 2013; Kudapa et al., 2012; Lulín et al., 2012; Tao et al., 2012; Teixeira et al., 2014; Verma et al., 2013; Wei et al., 2011; Zhang et al., 2012). Several different next-generation sequencing platforms are now available in the market, and of these, Illumina is most widely used (Metzker, 2010).

A previous study has reported the transcriptome assembly of red clover using leaf tissues from drought-sensitive and drought-tolerant genotypes (Yates et al., 2014). However, to study the transcriptional impact of different physiological and developmental processes and also to identify candidate genes involved in such processes, it will be important to study transcriptional networks in different tissue types. Under this backdrop, we have generated a de novo transcriptome assembly of red clover using RNA obtained from different tissue samples. This assembly was subsequently compared with other legumes, used for genome-wide identification of transcripts encoding different plant transcription factor families, and adapted for a gene expression analysis to assess transcriptional dynamics across different tissues and have identified dynamic spatial gene coexpression networks.

**Experimental Procedures**

**Plant Material and RNA Isolation**

Red clover plants (cultivar Kenland) were grown under greenhouse conditions with a 16:8 h light–dark cycle. Three individual lines and three clones for each line were used for the experiment. Clones for individual lines were separated and transplanted at seven to eight leaf stage and were grown in 5.1 by 5.1 by 5.1 cm square pots. Leaf, root, and flower tissues were collected in triplicate, where each replicate was comprised of three clones of one line. All samples were collected at the same time of the day (between 10 AM and noon) to eliminate any diurnal variations. Leaf and root tissues were collected at 2 wk after transplantation, whereas flower samples were collected after a plant develop several flowers. For leaf and flower, whole organs were sampled, whereas for the root tissue, parts of the root were sampled. RNA extraction was done using Trizol reagent (Life Technologies), per manufacturer’s instructions, followed by purification with RNA minispin columns (Enzymax LLC).

**RNA Sequencing Library Preparation and Illumina Sequencing**

The RNA sequencing (RNA-seq) libraries were made with 1 µg of total RNA as described previously (Hunt, 2015). Briefly, total RNA was poly (A) enriched with oligo dT beads followed by RNA fragmentation. Fragmented RNA was reverse transcribed with primer consisting of unique barcodes, a sequencing adaptor for the Illumina platform, and a random hexamer at the very 3’ end. Reverse transcription was followed by strand switching using the so-called SMART technology (Clontech Laboratories, Inc.). Next, complementary DNA were subjected to two rounds of AMPure (Agencourt AMPure XP beads, Beckman Coulters, Inc.) selection, followed by polymerase chain reaction amplification and one subsequent round of AMPure purification. The quality of the prepared libraries was assessed with Agilent high-sensitivity DNA chips (Agilent Technologies). Concentrations of libraries were measured using Qubit fluorometer with Qubit dsDNA HS assay kit (Life Technologies). Libraries were sequenced using Illumina MiSeq and HiSeq.
platforms. Raw sequences were demultiplexed, adapter, and barcode sequences were trimmed, and ribosomal RNA sequences were removed using *A. thaliana* ribosomal RNA sequences. For the Illumina MiSeq run, we could not distinguish between barcodes for four samples because of some base calling issues; hence, reads from these four MiSeq libraries were used for the de novo assembly but not for the gene expression analysis.

**De Novo Assembly of the Transcriptome**

De novo assembly was performed using the CLC Genomics Workbench version 8.0 (CLC bio, 2015) using k-mer of 64 (maximum k-mer allowed by CLC) and with default bubble size. For the de novo assembly, different approaches were tried. In one approach, all processed reads from this experiment were used to create an assembly. In a second approach, all processed reads from this experiment, along with reads from a previously reported study (NCBI BioProject accession PRJNA219226) (Yates et al., 2014), were used for another round of de novo assembly. For the final assembly, all the sequences from this experiment along with the contigs from the previously published assembly were used (Yates et al., 2014). To check the completeness of the assembled transcriptome core eukaryotic genes mapping approach (CEGMA) software housed in iPlant was used (Parra et al., 2007).

**Comparative Analysis of Assembled Transcriptome and Functional Annotation**

Contigs from the assembled red clover transcriptome were compared with *A. thaliana*, Medicago (Mtcatula_285_Mt4.0.v1.protein), soybean [Glycine max (L.) Merr.] (Gmax_275_Wm82.a2.v1.protein), and common bean (Phaseolus vulgaris L.) (Pvulgaris_218_v1.0.protein) peptide sequences using BLASTx. Sequences for *A. thaliana* were downloaded from RefSeq (http://www.ncbi.nlm.nih.gov/refseq), whereas all the legume proteome sequences were downloaded from Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html). BLAST results against *A. thaliana* were imported into Blast2GO and GO annotations and Enzyme Commission (EC) classifications were performed in Blast2GO (Conesa et al., 2005; Götzt et al., 2008). Further functional analysis of the transcriptome was performed in the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (KAAS) (http://www.genome.jp/tools/kaas/) to assign KEGG Orthology (KO) terms to the transcripts and also to map them to KEGG pathways (Moriya et al., 2007). Red clover transcripts were mapped to the flavonoid and isoflavonoid biosynthesis pathway in the KEGG database using KO terms.

**Genome-Wide Identification of Transcription Factors**

The protein sequences of all plant transcription factors and assignment of these sequences to different transcription factor families were downloaded from Plant Transcription Factor Database (PlnTFDB) (http://plntfdb.bio.uni-potsdam.de/v3.0/downloads.php) (Pérez-Rodríguez et al., 2010). Red clover transcripts were subjected to BLASTx analysis against the PlnTFDB peptide sequences with an *E*-value cutoff <1 × 10^-5.

**Differential Gene Expression and Gene Coexpression Cluster Analysis**

Gene expression analysis was performed with the Transcriptomics Analysis suite in CLC Genomics Workbench using the assembled transcriptome as the reference for the read mapping. A principal component analysis was done using the normalized expression values of different samples using CLC Genomics Workbench. To identify differentially expressed genes the ‘empirical analysis of DEG’ option in CLC Genomics Workbench was used. To filter differentially expressed genes between tissue types, a fold change of >5, >50 reads in at least one sample, and *p*-value (Bonferroni corrected) <0.01 were used. A list of differentially expressed genes that passed through this filter was prepared and used further for the coexpression clustering analysis. A soft clustering approach using fuzzy *c*-means algorithm was performed in R using Mfuzz package (Futschik and Carlisle, 2005). For the clustering analysis, gene expression values were standardized such that the mean is zero and standard deviation is one. Genes represented by tissue-enriched clusters were subjected to a GO enrichment analysis using agriGO and REVIGO (Du et al., 2010; Supek et al., 2011).

**Results and Discussion**

**Illumina Sequencing and De Novo Assembly of the Red Clover Transcriptome**

In this project, RNA-seq libraries were prepared with RNA isolated from leaf, root, and flower tissues of red clover. These libraries were sequenced on Illumina MiSeq and HiSeq2000 instruments. The initial processing of these data involved removal of ribosomal RNA-related reads and subsequent demultiplexing and trimming to remove the Illumina adapter sequences. This process altogether generated ~98 million processed reads. The returns from the sequencing runs are summarized in the Supplemental Table S1. Several approaches were tested for the de novo assembly of the red clover transcriptome. De novo assembly with all the processed reads from the current project using the assembler provided in the CLC Genomics Workbench package generated 30,566 contigs with N50 and average contig length of 623 and 598 bp, respectively. When the RNA-seq reads from a previous study (NCBI Short Read Archive Bioproject PRJNA219226) (Yates et al., 2014) were included with the reads generated for the current study and assembly performed using CLC Genomics Workbench, the result was 64,423 contigs with N50 and average contig length of 515 and 522 bp, respectively. Alternatively, when the reads from the current project were assembled onto the contigs from PRJNA219226, 37,565 contigs with N50 and average contig length of 1707 and 1262 bp,
respectively, were obtained as compared with 45,181 contigs with N50 and average contig length of 622 and 933 bp, respectively, for the assembly reported in PRJNA219226 (Yates et al., 2014). Since the last approach resulted in better N50 and average contig length, further analyses were performed using this transcriptome assembly. Different statistics for the final transcriptome assembly are represented in Table 1 and Fig. 1.

Table 1. Red clover de novo assembly statistics.

<table>
<thead>
<tr>
<th>Assembly parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N75</td>
<td>1,032</td>
</tr>
<tr>
<td>N50</td>
<td>1,707</td>
</tr>
<tr>
<td>N25</td>
<td>2,628</td>
</tr>
<tr>
<td>Min.</td>
<td>67</td>
</tr>
<tr>
<td>Max.</td>
<td>13,660</td>
</tr>
<tr>
<td>Average</td>
<td>1,262</td>
</tr>
<tr>
<td>Count</td>
<td>37,565</td>
</tr>
<tr>
<td>Total assembly length (bp)</td>
<td>47,388,793</td>
</tr>
</tbody>
</table>

To assess the completeness of the assembled transcriptome, CEGMA, which searches a set of 248 most highly conserved core eukaryotic genes (CEGs) in the assembled dataset and determine if they are complete or partial (Parra et al, 2007), was used. Completeness analysis in CEGMA with the chosen assembly resulted in 214 (86.29%) complete and 239 (96.37%) partial CEGs. This compares favorably with the prior study that yielded 184 complete (74.19%) and 235 (94.76%) partial CEGs (CEGs are also included in the partial CEGs list). These results indicate that the current assembly has an excellent collection of full-length sequences and encodes a representative set of proteins.

Comparison of the Assembled Transcriptome with Arabidopsis and Other Legumes

To further assess the assembled transcriptome, a comparison with A. thaliana protein sequences using BLASTx was performed; for this, an E-value cut off of 1 × 10^-6 was chosen as defining relevant or authentic matches. This resulted in 30,174 hits and involved 80.3% of the red clover contigs (Fig. 2). Comparative transcriptome analyses between red clover and other legume species, including Medicago, soybean, and common bean pairwise BLASTx searches (E-value cut off 1 × 10^-6) were also performed. There were 33,026 matches with Medicago (involving 87.9% of the red clover contigs), 32,289 with soybean (86% of the red clover contigs), and 32,058 with common bean (85.3% of the red clover contigs). There were 31,558 transcripts shared among red clover and three other legume transcriptomes (Fig. 2D; BLAST hits against Arabidopsis and legume species are represented in Supplemental Dataset S1). This high level of similarity between red clover and other legume species indicates that the assembled red clover transcriptome provides a good coverage of the homologous legume sequences.

Functional Annotation of Red Clover Transcripts

The red clover transcriptome was functionally annotated using Blast2GO (Conesa et al., 2005; Götz et al., 2008). The output of a BLASTx comparison of the red clover contigs against the Arabidopsis protein database was imported in Blast2GO to carry out GO mapping and functional characterization. Transcripts were grouped into three main GO categories: biological processes, molecular function, and cellular component. In total, 30,145 (80.3%) of the red clover transcripts were assigned GO terms (Fig. 3; Supplemental Fig. S1). In the molecular function category, catalytic activity and binding are the most highly represented groups. Further classification of molecular function category revealed kinase activity, DNA binding, nucleotide binding, and transporter activity as widely represented classes. In the biological process category, cellular process, metabolic process, and single-organism process are the most abundant groups. More detailed classification of biological process GO category showed response to stress, cellular component organization, and stress as highly represented groups. Further classification of biological process GO category revealed kinase activity, DNA binding, nucleotide binding, and transporter activity as widely represented classes. In the cellular component category, cellular, organelle, and membrane are the most highly represented groups, and further categorization of this GO classification revealed plastid, plasma membrane, cytosol, and mitochondria as highly represented groups.

To further improve the functional annotation of the assembled transcriptome, EC classifications for 10,095 transcripts (26.87% of the contigs) were obtained; the combined GO and EC annotations of red clover transcriptome is given in Supplemental Dataset S2. Additionally, red clover transcripts were assigned to various pathways in the KEGG database using KAAS (Moriya).
This analysis distributed 6340 transcripts (16.88%) into 330 different KEGG pathways (Fig. 4). The KEGG pathway mapping of red clover transcriptome for the flavonoid and isoflavonoid biosynthesis pathway using KO terms revealed that our transcriptome assembly captures significant portions of these two important secondary metabolite biosynthesis pathways, which were implicated in the nodulation process during legume–rhizobium symbiosis and also to enrich nutritional quality (Supplemental Fig. S2, S3) (Dixon and Pasinetti, 2010; Ferreyra et al., 2012; Subramanian et al., 2007).

**Genome-Wide Identification of Transcription Factors in Red Clover**

Transcription factors are key regulators of gene expression; hence, identification of transcription factors in a particular species is worthwhile. To identify red clover transcripts encoding various transcription factors, a
Fig. 3. Multilevel gene ontology (GO) annotation of red clover transcriptome. Gene Ontology annotation was performed using Blast2GO. (A), (B), and (C) represent biological process, molecular function, and cellular component, respectively. The BLASTx result against Arabidopsis was imported to Blast2GO to perform GO annotation.
BLASTx search was conducted using as queries all the transcription factors sequences available in the PlnTFDB database (Pérez-Rodríguez et al., 2010). This analysis yielded 8373 red clover transcripts (accounting for 22.29% of the assembled transcripts). This set of transcripts encodes putative members of 81 transcription factor families (Fig. 5; Supplemental Dataset S3). The most highly represented transcription factor families were MADS, C3H, FAR1, NAC, PHD, bHLH, WRKY, and Myb-related.

**Dynamic Spatial Gene Expression in Red Clover**

The tissues sampled in the library preparations allow for the determination of dynamic spatial gene expression, such that assembled contigs with identifiable expression patterns may be determined and characterized. Accordingly, a gene expression analysis was conducted using the assembled red clover transcriptome as the reference for mapping the reads. The mapping statistics for the different samples and libraries are given in Supplemental Table S1. A principal component analysis was performed using the normalized expression values of transcripts of all tissue samples, and this is represented as a correlation scatter plot in Supplemental Figure S4. The global gene expression comparison of different tissue samples indicates a distinct gene expression program for the three tissues, except for one replication of leaf samples, which was removed from the subsequent gene expression analysis. To explore the tissue-wise gene expression pattern further, differentially expressed genes were identified using the following criteria: number of reads per sample is >50 in at least one sample, fold change is more than ±fivefold for a given pair-wise comparison, and p-value (Bonferroni corrected)
is <0.01 in any pairwise comparison between the tissue types. Altogether, this analysis yielded 2630 differentially expressed genes in all pairwise tissue type comparisons. These differentially expressed genes were used further to identify clusters of coexpressed genes, which represent spatially enriched expression patterns. Coexpression clusters were identified by the Mfuzz package (Futschik and Carlisle, 2005) using soft clustering approach with fuzzy c-means algorithm. This analysis generated 10 clusters, which revealed dynamic gene expression patterns across the three tissue types (Fig. 6; Supplemental Dataset S4).

Cluster 1 includes genes with low expression in leaf and root and high expression in flower and thus represents genes with enriched expression in floral tissues. Among the genes in this cluster several are of interest. The flower-enriched cluster includes a R2R3 MYB transcription factor with homology to ATMYB21 (encoded by AT3g27810). MYB21 along with MYB24 plays vital roles in jasmonate-triggered stamen and pollen maturation in Arabidopsis. Arabidopsis myb21 mutants display reduced anther filament, delayed anther dehiscence, and decreased male fertility (Mandaokar et al., 2006). This flower-enriched cluster also contains a transcript encoding MADS domain transcription factor PISTILLATA, encoded by AT5g20240 in Arabidopsis. PI (PISTILLATA) together with AP3 (APETALA 3) constitutes the class B organ identity genes in the classical ABC model of floral organ development. AP3 and PI are necessary to provide petal and stamen identity in the Arabidopsis flower. Also, AP3 and PI are sufficient to bestow petal and stamen identity in conjunction with class A and C genes, respectively (Krizek and Meyerowitz, 1996; Wuest et al., 2012).

Clusters 3, 6, and 10 include genes with low expression in root and flower and high expression in leaf and, hence, represent leaf-enriched genes. Leaf-enriched clusters include transcripts encoding enzymes involved in photosynthesis and also different components of photosynthetic apparatus. This includes rubisco activase, a chloroplast protein, which catalyzes the activation of rubisco (ribulose-bisphosphate carboxylase/oxygenase) (Salvucci et al., 1985). Additionally, leaf-enriched gene clusters include transcripts encoding different constituents of photosystem (PS) I and PSII reaction centers, such as PSI subunit D2, PSI subunit F, PSI subunit R, and PSII subunit O2 (Haldrup et al., 2000, 2003; Liu et al., 2009; Murakami et al., 2002). Leaf-enriched clusters also contain of transcripts encoding different light-harvesting complexes (LHCs) associated with PSI and PSII, such as LHCA2, LHCA3, LHCA4, LHCb2.1, LHCb3, LHCb5, and several others. These LHCs act as light-harvesting antenna, which capture light and transfer energy via resonance energy transfer to their respective reaction centers of photosystems (Ruban, 2015).

Cluster 5 represents genes with low expression in leaf and flower and high expression in root thereby representative of root-enriched genes. The root-enriched cluster includes transcripts encoding enzymes involved in the phenylpropanoid biosynthetic pathway. One such example is a cytochrome P450, CYP73A5, encoding CINNAMATE 4-HROXYLASE (encoded by AT2g30490 in Arabidopsis) involved in formation of p-coumarate from cinnamate during phenylpropanoid metabolism. Mutants for this gene in Arabidopsis (termed as reduced epidermal fluorescence 3, ref3) displayed reduced accumulation of different phenylpropanoids and decreased lignin deposition and changed lignin monomer composition (Schilmiller et al., 2009). Lignin deposition and composition might play an important role in modifying the digestibility of forage crops such as red clover. Additionally, p-coumarate can leads to formation of p-coumaroyl CoA, a precursor for isoflavonoid biosynthesis. During legume–rhizobium symbiosis, some flavones and isoflavones were shown to induce rhizobial nodulation genes (Subramanian et al., 2007).
Other clusters also showed dynamic gene expression patterns across the three tissues. Genes in Cluster 2 show high expression in roots and leaves and low expression in flower. This cluster includes transcripts encoding enzymes involved in phenylpropanoid metabolism and flavonoid biosynthetic pathway. Cluster 2 includes transcript with homology to Arabidopsis Phenyl Ammonia Lyase 1 (PALL, encoded by At2g37040), which catalyzes deamination of phenylalanine to trans-cinnamic acid and ammonia (Fraser and Chapple, 2011). Also, this cluster harbors transcripts homologous to Chalcone Synthase (Transparent Testa 4, TT4, encoded by At15g13930) and Chalcone Isomerase (Transparent Testa 5, TT5, encoded by At13g55120) involved in the flavonoid biosynthesis (Ferreya et al., 2012). The Chalcone Synthase (CHS) genes are expanded in legume species; Medicago and soybean harbor CHS genes as clusters of five to seven coregulated genes, which indicates toward evolutionary significance of flavonoids in legumes (Dixon and Pasinetti, 2010). Another product from the flavonoid biosynthetic pathway proanthocyanidins were also reported to reduce bloat in ruminant animals and thus can be of importance to forage species like red clover (Dixon and Pasinetti, 2010). Additionally, this cluster includes transcript homologous to Caffeoyl CoA 3-O-methyltransferase (CCoAOMT, encoded by At4g34050) involved in the conversion of caffeoyl CoA to feruloyl CoA, which can go into lignin biosynthesis (Ferreyra et al., 2012). Lignin content was implicated with the digestibility of the forage legume alfalfa (Reddy et al., 2005). The association of these genes with this cluster suggests important roles for the associated metabolic pathways in both roots and leaves.

Genes represented by Clusters 4 and 7 show higher expressions in leaf and flower than roots. These clusters include a gene orthologous to Arabidopsis Carotenoid Cleavage Dioxygenase 4 (CCD4, encoded by At4g19170), which is a major regulator of carotenoid turnover in senescing leaves and seed tissues (Gonzalez-Jorge et al., 2013). The expression of this gene in red clover was highest in leaves, followed by moderate expression in flower, and low expression in root. Thus this gene can be a target for biofortification of provitamin A in red clover foliage to nutritionally enrich the forage quality. These two clusters also harbor a gene similar to Arabidopsis Xyloglucan ENDOTRANSGLUCOSYLASE 6 (XTH6, encoded by At15g65730). Expressions patterns of the red clover gene follow the similar pattern as that of Arabidopsis gene as visualized in the AtGenExpress visualization tool (Schmid et al., 2005). Xyloglucan is an important constituent of primary cell wall of the dicot plants. It can bind with cellulose and forms cross-link between neighboring cellulose microfibrils and thereby contributes to the cell wall strength and extensibility (McCann et al., 1990; Rose et al., 2002). Xyloglucan-metabolizing enzymes, such as XTH, can be important for cell wall loosening, which can be a vital determinant for the digestibility of forage crops. Clusters 4 and 7 also contain a transcript with homology to 3-ketoacyl-CoA synthase 6 (KCS6 or CER6 encoded by At1g68530), involved in the biosynthesis of very long chain fatty acids (Smirnova et al., 2013). Similar to the expression pattern in red clover, Arabidopsis ortholog was also found to be expressed at higher levels in leaf and flower than the root. Additionally, overexpression of CER6 increased surface wax accumulation in Arabidopsis (Hooker et al., 2002). Such deposition of surface wax can be implicated in nonstomatal water loss and pathogen response (Joubes et al., 2008). CER6 in tomato (Solanum lycopersicum L.) was shown to be involved in floral development, sexual reproduction, and fertility (Smirnova et al., 2013). Thus, the red clover ortholog can possibly be involved both in reproductive processes and stress responses.

Genes belonging to these clusters can also be useful to isolate promoters with high expression in above-ground tissues, which can be used to drive expression of target genes with significance to forage nutrition (such as genes belonging to flavonoid or isoflavonoid biosynthesis or carotenoid turnover) and digestibility (such as genes involved in processes like cell wall loosening and lignin biosynthesis). Alternative to this approach, recently developed genome-editing tools, such as CRISPR-cas9 (Brooks et al., 2014; Feng et al., 2014; Jiang et al., 2013; Li et al., 2013; Nekrasov et al., 2013) and TALEN (Joung and Sander, 2013; Lor et al., 2014) can also be used to specifically modify target genes mentioned in the preceding sections to improve nutritional quality and digestibility of the forage legume red clover.

Functional Classification of the Tissue-Enriched Gene Expression Clusters

To further characterize the tissue-enriched gene coexpression clusters, a GO enrichment analysis was conducted using the singular enrichment analysis tool in agriGO (Du et al., 2010). The list of GO terms from this analysis was exported to the REVIGO web server to reduce redundant GO terms and graphically represent the results (Supek et al., 2011). A treemap depicting overrepresented GO terms in the tissue-enriched gene coexpression clusters is shown in Supplemental Fig. S5 through S7. Reflecting the examples described in the preceding paragraphs, the leaf enriched clusters (Cluster 3, 6, and 10 were combined for the analysis) displayed a higher representation of genes associated with photosynthesis, pigment biosynthesis, and the generation of precursor metabolites and metabolism, whereas the root enriched cluster (Cluster 5) had a higher representation of genes associated with responses to stress, phenylpropanoid biosynthesis, responses to stimulus, and posttranslational protein modification. The flower enriched cluster (Cluster 1) showed a tendency toward genes associated with multicellular organismal development, reproduction, and transport. Thus, the GO enrichment analyses of tissue-enriched clusters show overrepresentation of predicted classes of genes in different tissue types and thereby validate our tissue-enriched gene coexpression clusters.

Cluster 2 (representing transcripts with highest expression in root followed by leaves and low expression
in flowers) showed higher representation of genes related to cellular modified amino acid biosynthesis, phenylpropanoid biosynthesis, and response to stimulus. These classes suggest important roles for the associated processes in both leaf and root function. As discussed above, some of these are consistent with studies in other plants. However, the broader scope illustrated here indicates a significant extent of possible multifunctionality. Also as discussed above, genes in this cluster may well provide biochemical functionalities and gene expression tools that will be useful in the improvement of forage quality in red clover and other forages.

Clusters 4 and 7 (representing transcripts with high expression in leaf and flower and low expression in roots) had a preponderance of genes related to cellular lipid metabolism, cellular modified amino acid biosynthesis, organic acid metabolism, aromatic compound biosynthesis, and secondary metabolism. These associations reflect important contributions of secondary metabolism to both leaf and flower function.

Conclusion
In the present study, we have described the construction and functional annotation of a de novo transcriptome assembly for red clover, a nonmodel forage legume. We have used this assembly to study tissue-enriched gene expression and have identified novel sets of genes with functions in different tissues. We have also identified a large number of genes that encode putative transcription factors; this will benefit future studies on regulation of gene expression in red clover. Altogether, our study expands the existing genomic resources for red clover and provides a platform for future genetic and biotechnological improvement of red clover.

Supplemental Information Available
Supplemental information is available with the online version of this article.
Supplemental Table S1. Sequencing and mapping statistics of red clover transcriptome.
Supplemental Fig. S1. Level 2 GO annotation of red clover transcriptome.
Supplemental Fig. S2. KEGG pathway mapping of red clover transcriptome for flavonoid biosynthesis.
Supplemental Fig. S3. KEGG pathway mapping of red clover transcriptome for isoflavonoid biosynthesis.
Supplemental Fig. S4. Principal component analysis of all red clover tissue samples.
Supplemental Fig. S5. Over-represented GO categories in leaf-enriched gene coexpression clusters.
Supplemental Fig. S6. Over-represented GO categories in root-enriched gene coexpression clusters.
Supplemental Fig. S7. Over-represented GO categories in flower-enriched gene coexpression clusters.
Supplemental Dataset S1. Blast output of red clover transcriptome against Arabidopsis and legumes.
Supplemental Dataset S2. Functional annotation of red clover transcriptome assembly.

Supplemental Dataset S3. Red clover contigs encoding putative transcription factors.

Sequence Deposition
All Illumina sequence reads generated in this project were submitted to the NCBI Short Read Archive (SRA) under the BioProject PRJNA287846 (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA287846).

Acknowledgments
We would like to thank Troy Bass for his excellent support in maintaining red clover plants in the greenhouse and Carol Von Lanken for excellent technical assistance. This work was funded through a specific cooperative agreement with USDA–ARS–Forage-Animal Production Research Unit (58-6440-3-005).

References


