American Journal of Botany
Celebrating 100 years 1914–2014

American Journal of Botany 101(2): 327–337. 2014.

GENOME ASSEMBLY AND ANNOTATION FOR RED CLOVER (*TRIFOLIUM PRATENSE*; FABACEAE)[1]

JAN IŠTVÁNEK[2], MICHAL JAROŠ[3], ALEŠ KŘENEK[3], AND JANA ŘEPKOVÁ[2,4]

[2]Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic; and [3]CERIT-SC, Institute of Computer Science, Masaryk University, Brno, Czech Republic

- *Premise of the study:* Red clover (*Trifolium pratense*) is an important forage plant from the legume family with great importance in agronomy and livestock nourishment. Nevertheless, assembling its medium-sized genome presents a challenge, given current hardware and software possibilities. Next-generation sequencing technologies enable us to generate large amounts of sequence data at low cost. In this study, the genome assembly and red clover genome features are presented.
- *Methods:* First, assembly software was assessed using data sets from a closely related species to find the best possible combination of assembler plus error correction program to assemble the red clover genome. The newly sequenced genome was characterized by repetitive content, number of protein-coding and nonprotein-coding genes, and gene families and functions. Genome features were also compared with those of other sequenced plant species.
- *Key results:* Abyss with Echo correction was used for de novo assembly of the red clover genome. The presented assembly comprises ~314.6 Mbp. In contrast to leguminous species with comparable genome sizes, the genome of *T. pratense* contains a larger repetitive portion and more abundant retrotransposons and DNA transposons. Overall, 47 398 protein-coding genes were annotated from 64 761 predicted genes. Comparative analysis revealed several gene families that are characteristic for *T. pratense*. Resistance genes, leghemoglobins, and nodule-specific cystein-rich peptides were identified and compared with other sequenced species.
- *Conclusions:* The presented red clover genomic data constitute a resource for improvement through molecular breeding and for comparison to other sequenced plant species.

Key words: assessment of assembly software; *de novo* assembly; Fabaceae; genome annotation; red clover; *Trifolium pratense.*

Red clover (*Trifolium pratense* L.; Fabaceae) is a very important forage plant in many countries around the world. It serves also as a green manure crop and temporary cover crop. It belongs to the tribe Trifoliae, which comprises approximately 240 species of annual and perennial herbs, both wild and cultivated. It is an outcrossing species with a gametophytic self-incompatibility system. Its high level of heterozygosity has hampered intensive genetic and genomic analyses. Red clover is difficult to self, inbreeding by selfing usually cannot be continued beyond two or three generations because of loss of vigor. Neither inbred lines nor doubled haploids are available. Red clover (x = 7), with its genome size estimated to be 418 Mbp (1C = 0.43 pg; Vižintin et al., 2006), is a leguminous plant and is capable of fixing atmospheric nitrogen.

[1]Manuscript received 23 September 2013; revision accepted 27 November 2013.

The authors thank the Ministry of Agriculture of the Czech Republic (grant no. QI111A019) and the Ministry of Education, Youth and Sports of the Czech Republic (grant no. CZ.1.07/2.4.00/31.0155) for financial support. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the program "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005), is appreciated. In particular, access to the CERIT-SC computing and storage facilities provided under the program Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ. 1.05/3.2.00/08.0144 is acknowledged.

[4]Author for correspondence (e-mail: repkova@sci.muni.cz)

doi:10.3732/ajb.1300340

The leguminous order is among the most examined orders in the plant kingdom. Five plants from this group have been sequenced: the model plants *Medicago truncatula* Gaertn. (Young et al., 2011) and *Lotus japonicus* L. (Sato et al., 2008) and crop plants soybean [*Glycine max* (L.) Merrill.; Schmutz et al., 2010], pigeon pea [*Cajanus cajan* (L.) Millsp; Varshney et al., 2012], and chickpea (*Cicer arietinum* L.; Varshney et al., 2013). The genome structure of red clover has been investigated using fluorescence in situ hybridization (FISH; Sato et al., 2006; Kataoka et al., 2012). Recently, the first consensus high-density linkage map was constructed; it is 836.6 cM long with 1414 simple sequence repeats (SSRs), 181 amplified fragment length polymorphisms (AFLPs), and 204 restriction fragment length polymorphisms (RFLPs) (Isobe et al., 2009).

Currently, 7262 SSRs and 228 RFLPs are included in the Clover Garden database (http://clovergarden.jp/). The genome structures of red clover, white clover, *M. truncatula*, and *L. japonicus* have been compared using SSR markers (Isobe et al., 2012). While macrosynteny has been confirmed across the four legume species, the genomic structure between white clover and *M. truncatula* was shown to have a higher degree of conservation than that of the two clover species (Isobe et al., 2012). Red clover has not previously been sequenced, and it is the first sequenced species from the leguminous plants that is not a food crop or model plant. Moreover, its medium-sized genome poses a challenge for contemporary de novo assembly programs because these systems were, until recently, focused mainly on bacterial genome assembly (Chaisson and Pevzner, 2008).

Next-generation sequencing (NGS) is developing rapidly as various sequencing technologies are being introduced for practical work (Mardis, 2013). The data produced can differ in

many features: runtime, price, read length, and error rate. Nevertheless, Illumina sequencing technology is still the standard for high-throughput massively parallel sequencing (Quail et al., 2012). In comparison with Sanger sequencing, Illumina reads have shorter read lengths and higher error rates, but can provide greater genome coverage at lower cost (Lindblad-Toh et al., 2005). Large quantities of short reads, however, present a challenge for many bioinformatic tools and computer equipment. Suitable assembly of genomes remains a daunting problem, although considerable progress has been made in the last few years (Salzberg et al., 2012). Reducing error rate of base-calls also appears to have important practical implications for assembly, because it simplifies the resolution of imperfect matches in read overlaps or de Bruijn graphs (Langmead et al., 2009; Simpson et al., 2009). Error correction is mainly based on statistical evaluation of probability of potential error at each position utilizing high coverage.

In this study, de novo assembly of the red clover genome is presented, and its genome is described, including a comparison with the genome of other sequenced plant species. We present here ~314.6 Mbp of draft nucleotide sequences for red clover. First, however, we tested assemblers and error correction programs on shorter data sets created from the closely related species *M. truncatula* to test for the best combination. This study therefore offers also practical insight into program settings because they significantly influence the quality of results.

MATERIALS AND METHODS

***Simulated and real sequencing data*—**Test data sets were created from chromosome 2 (33 Mbp) of *M. truncatula* (assembly release Mt 3.0, http://medicagohapmap.org/), a close relative of *T. pratense* (both from tribe Trifoliae), specifically because the accuracy of testing assemblies can be measured only with well-annotated references. Approximately 16.3 million paired-end reads 101 bp long were simulated in each data set using the script written by Zhang et al. (2011). Parameters were set according to the characteristics of real data sets, with fragment size 430 bp, and average genome coverage 50×. Very similar *k*-mer frequencies between simulated reads and *T. pratense* reads had been counted by the program Tallymer (Kurtz et al., 2009; Appendix S1, see Supplemental Data with the online version of this article) while showing comparable repeat content between data sets. The data sets differed in their error rates when simulating the sequencing errors. The first data set contained 0.1% error bases (DS0.1), which is the error rate of the read after quality filtering and removal of wrongly called bases at the 3′-end of the read (Minoche et al., 2011). The second data set had 1% error bases (DS1) and served to indicate the capability of the software to handle more sequencing errors. Capability of compensating sequencing errors is especially important in the case of new sequencing technologies that are being released and which generate higher error rates (e.g., PacBio).

Real data were obtained from the Tatra variety of *T. pratense*. Seeds were procured from GeneBank of the Crop Research Institute, Prague-Ruzyně, Czech Republic, accession 13T0200327. Leaves were collected from 30-d-old, greenhouse-grown plants. Genomic DNA was extracted from nuclei isolated from ~10 g of young leaves using the method developed by Zhang et al. (1995). DNA was isolated from 16 pooled plants.

A paired-end genomic DNA library was constructed by IGA Technology Services (Udine, Italy) with a TruSeq DNA-seq kit. Clusters were generated in a flow cell by the cBot system (IGA Technology Services S.R.L., Udine, Italy), and the library was run on a HiSeq2000 using a standard Illumina sequencing workflow.

Trifolium pratense paired-end reads 101 bp long were obtained from a single genomic library. The average fragment size was 430 bp and genome coverage of ~58.8× was achieved. Sequence reads are available at the Sequence Read Archive of NCBI under accession SRP022158, and the project has been deposited in the DDBJ/EMBL/GenBank under accession ASHM00000000. The version described in this paper is version ASHM01000000.

***Bioinformatic tools evaluation*—**Eight assemblers (Appendices S2 and S3, with the online supplemental data, summarize different testing configurations) were tested on simulated data to evaluate their success in the de novo assembly: Edena v3 (Hernandez et al., 2008), Velvet v1.2 (Zerbino and Birney, 2008), SOAPdenovo v1.05 (Li et al., 2010b), CABOG v7.0 (Myers et al., 2000), Abyss v1.3.3 (Simpson et al., 2009), Pasha v1.0.3 (Liu et al., 2011), SGA (Simpson and Durbin, 2012), and Gossamer v1.2.2 (Conway et al., 2012). The assemblers were assessed in terms of contiguity and accuracy. Contiguity refers to the fragmentation of assembly. Number of contigs, largest contig, average contig length, N50 (which is the size of the smallest contig or scaffold such that 50% of the genome is contained in contigs of size N50 or larger), and N90 were compared. Accuracy was measured as every contig was aligned to the reference by the Nucmer program (Delcher et al., 2002) and counted as correct if an identity of 95% was achieved and at least 80% of the contig length was properly aligned (for comparison with 99% identity and 99% coverage, see online Appendices S2 and S3).

To evaluate assemblers for their overall performance (in testing of both data sets), the following assembly statistics were chosen: number of contigs, average contig length, largest contig, N50, and N90 from contiguity statistics, as well as the percentage of correct contigs and percentage of correct assembly from accuracy statistics. Each assembler (more precisely, the individual run settings of the assemblers) was graded based on the results (the best result obtaining the most points). Assemblers were then compared based on the total number of points acquired in both data sets. The best performing assembler was picked to assemble the real data sets.

For error correction, the following software systems were used: Coral v1.4 (Salmela and Schröder, 2011), Echo v1.11 (Kao et al., 2011), Reptile v1.1 (Yang et al., 2010), and Quake v0.3.4 (Kelley et al., 2010). All read error correction programs were set according to their authors' instructions. To assess the effect of read error correction on the assembly characteristics, the contiguity and accuracy were measured in three assemblers (Edena, Abyss, and CABOG) for each correction, as described above. The performance of each software algorithm was also tested for the successful recognition of sequencing errors. We proposed a simple assessment of the error correction software based on single nucleotide polymorphism (SNP) calling. Because we introduced errors imitating sequencing errors into the simulated data, the variants/errors can be detected as SNPs. The correction programs were evaluated based on how many SNPs were found (the fewer the better because the original data were errorless). Two alignment software systems were used for uncorrected and corrected read alignment: bwa v0.6.1 (Li and Durbin, 2009) and bowtie v0.12.7 (Langmead et al., 2009). Samtools v0.1.18 (Li et al., 2009) was used for SNP calling. Correction programs were also assessed using the Error Correction Evaluation Toolkit (ECR) (Yang et al., 2012) to confirm our findings independently. For the results of error correction software, see online Appendices S4–S7.

***Characterization of the newly sequenced genome*—**The repeat content of the genome of *T. pratense* was characterized by RepeatExplorer (Novák et al., 2013), which is implemented in the Galaxy platform (http://galaxyproject.org). A total of 2 136 920 (~0.5×) Illumina reads were used. Repetitive sequences were identified using similarity-based clustering analysis. Clusters containing more than 0.1% of used reads were inspected more closely. For purposes of annotation, the graphical representations of repeats were examined (Novák et al., 2010). BLAST (http://blast.ncbi.nlm.nih.gov; Altschul et al., 1990) searches of contigs assembled with CAP3 (implemented in RepeatExplorer) were inspected for hits with available repeat sequences in GenBank databases. The program Dotter (Sonnhammer and Durbin, 1995) was used to inspect structural features such as tandem subrepeats in contigs, and clview (http://compbio.dfci.harvard.edu/tgi/software) was used to identify insertion sites in potential transposable elements.

The application SSR Locator (da Maia et al., 2008) was used to mine SSRs in the red clover genome, and also for primer design. An SSR site was defined as a monomer occurring at least 12×, a dimer at least 6× times, tri- and tetramers at least 4×, and penta-, and hexamers occurring at least 3×. The number of PCR products was predicted for each primer pair.

Using the gene predictor program AUGUSTUS (Stanke et al., 2004) with an Arabidopsis-trained matrix, ab initio prediction of complete and partial genes was performed in those assembled genomic contigs of *T. pratense* longer than 200 bp. Gene function determination and annotation were done by Blast2GO v2.6.6 (Conesa et al., 2005). Gene functions were assigned using BLASTP against the refseq protein database while linking BLAST hits (e-value: $1e^{-6}$) to the Gene Ontology (GO) database v1.2 (http://www.geneontology.org). InterProScan v4.8 (http://www.ebi.ac.uk/Tools/pfa/iprscan/; Zdobnov and Apweiler, 2001) determined motifs and domains of genes against protein databases ProDom, PRINTS, Pfam, Gene3D, PANTHER, SuperFamily, SignalP, TMHMM, PIR, SMART, TIGR, PROFILE, and PROSITE. Gene Ontology IDs were obtained for each

gene from the corresponding InterPro entry. The protein-encoding genes were classified into functional categories according to GO, and the results were summarized in plant GOslim functional categories. Every gene was compared with the KEGG database (release 67.1; http://www.genome.jp/kegg) protein entries, and the pathway in which the gene might be involved was determined.

Nonprotein coding (tRNA, rRNA, miRNA, and snRNA) genes were predicted by INFERNAL v1.1 (Nawrocki et al., 2009) against the Rfam database (release 11.0; rfam.sanger.ac.uk/).

Analysis of orthologous genes was performed by the program OrthoMCL (Li et al., 2003). All predicted protein sequences (proteomes) from the sequenced legume plant clades Millettioid (*Cajanus cajan*, and *Glycine max*) and Galegoid (*Trifolium pratense*, *Medicago truncatula*, *Lotus japonicus*, and *Cicer arietinum*), with one outgroup species (*Arabidopsis thaliana*; TAG Initiative, 2000), were analyzed to circumscribe sets of orthologous genes. First, BLASTP (e-value: $1e^{-5}$) was used to identify best-hit pairs for species-by-species and within-species comparisons. This hit matrix served as the basis for ortholog definition using OrthoMCL (inflation parameter I = 1.5). Orthologous gene groups were then organized into species-specific and higher taxonomic level groups. The red clover-specific genes were identified. The red clover genes having significant hits (cut-off ≤ $1e^{-5}$) with protein sequences of Fabaceae species were identified as legume-specific candidates, and those lacking any significant similarity with plant species were identified as candidate *Trifolium*-specific genes. GO terms in these genes were examined.

The red clover gene transcripts encoding for transcription factors (TFs) were identified using the hidden Markov model (HMM) profiles available in the PFAM database v27.0 (www.pfam.sanger.ac.uk) and those generated from the domain alignments available at the Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de/v3.0/; Pérez-Rodríguez et al., 2010) for 84 families using a HMMER search. We used similar criteria to identify the TFs belonging to various families as the Plant Transcription Factor Database. Gene distribution in different transcription factor families was compared with other sequenced legume plants and *Arabidopsis*.

Resistance genes were predicted based on gene annotation, the HMMER search against HMM profiles available in the PFAM database and the presence of R-gene domains using the InterProScan and InterPro databases (Hunter et al., 2009). Genes for leghemoglobins were identified with BLASTP searches (cut-off ≤ $1e^{-15}$) against known those for leghemoglobins from *M. truncatula* (Young et al., 2011), *G. max* (Schmutz et al., 2010), and *C. arietinum* (Jain et al., 2013). Nodule-specific cystein-rich peptides were predicted with TBLASTX search against known nodule-specific cystein-rich peptides in *M. truncatula* (cut-off ≤ $1e^{-5}$; Young et al., 2011).

RESULTS

***Testing of assembly tools*—**High-quality de novo assembly is a very delicate task because each species has its own unique genome structure; e.g., the repetitive content in plants is highly variable (Thompson et al., 1996; Haberer et al., 2005). For this reason, before red clover assembly, the assemblers were first tested to choose the best-performing assembler and error correction tool.

It is clear even at first glance that the assembler algorithms differ greatly, because they provided diverse results, and that even the proportion of erroneous bases plays a significant role in some of the assemblers (see Appendix 1). In some instances, the higher proportion of incorrect bases resulted in severe fragmentation of the assembly (Edena, CABOG); in other cases, the effect on fragmentation was negligible (Abyss, Pasha). Importantly, assemblers showed significant differences in retaining accuracy of the assembly in the contig creation or in the following scaffolding step (Pasha, SOAP). An overall comparison of assemblers based on their performance for each data set is summarized in Appendix 2. Upon considering the accuracy of assembly, we found that Abyss attained the best results among the compared assemblers and chose it to assemble the red clover genome because it best balanced the contiguity and accuracy of the assembly.

Error correction tools were evaluated in terms of their practical impact on the assembly (measured by assembly statistics and SNP number found after correction) and from a statistical viewpoint (measured by ECR). For both, Echo provided the best results (Appendix 3). That outcome was achieved even in spite of its having the least sensitivity and the most false negatives (FNs) in DS0.1. Echo was calculated to have the best gain primarily because of its very low number of false positives (FPs). In addition, the improvement of assembly characteristics (less fragmented) is apparent when compared with the uncorrected data (shown in Appendices 1 and 3).

Error correction tools were also tested on Edena and CABOG, which gave results comparable to those of Abyss. In the case of CABOG, the error correction very significantly reduced the assembly fragmentation. For these results, see Appendices S4 and S5.

***Genome assembly*—**In total, 243.6 million paired-end reads were obtained from the red clover genome, comprising 24 605 851 492 sequenced nucleotides. On the basis of the results using the simulated data sets, Echo (run parameters: –k 15 –b 10000000 –nh 8192) was chosen to correct the reads and Abyss (k64) to assemble data (Table 1). The resulting assembly is slightly more contiguous than in the testing data sets, which possibly is a consequence of the data being produced from one genomic library with a fixed fragment size. Despite greater variability in the real data compared with the simulation, 3′-end trimming and Echo correction reduced its effect, resulting in a less-fragmented assembly (Appendix S8). While studies using several genomic libraries with different fragment sizes have gotten better results, creating additional libraries is technically more difficult and more expensive (Salzberg et al., 2012).

As in other studies (Salzberg et al., 2012), we encountered troubles with software tools in handling such large amounts of

TABLE 1. *Trifolium pratense* genome assembly, gene annotation and nonprotein coding genes.

	Value
Assembly features	
Number of scaffolds	176 760
Total span (Mbp)	314.6
Average scaffold length (bp)	1780
N50 (scaffolds) (bp)	4750
N90 (scaffolds) (bp)	765
Longest scaffold (bp)	58 296
Number of contigs	236 989
Average contig length (bp)	1305
N50 (contigs) (bp)	2937
N90 (contigs) (bp)	516
Longest contig (bp)	42 067
GC content (%)	32.8
Protein coding genes	
Number of annotated genes	47 398
Mean gene length (bp)	1784.8
Mean number of exons per gene	3.6
Mean number of introns per gene	3.0
Mean exon length (bp)	248.6
Mean intron length (bp)	295.2
Single-exon genes	7844
Nonprotein coding genes	
Number of miRNA genes	4719
Mean length of miRNA genes (bp)	130.6
Number of rRNA genes	109
Mean length of rRNA genes (bp)	224.9
Number of tRNA genes	874
Mean length of tRNA genes (bp)	73
Number of snRNA genes	584
Mean length of snRNA genes (bp)	105.9

data. Although Abyss was used to assemble the real sequencing data, we also tried other assemblers. In addition to Abyss, SOAPdenovo and Velvet also completed assembly of the *T. pratense* data. Their results conformed with those from using the test data sets, thus showing the same trends independently of the amount of data provided (Appendices S9 and S10). However, to confirm our way of testing assembler performances on closely related, well-described species and to verify the results obtained in test data sets and the real data set, we have also tested assemblers on the publicly available *M. truncatula* reads (SRR965418, SRR965430) from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra). These reads have the closest characteristics to our red clover data set. Abyss, SOAP, and Velvet were used; nevertheless, Velvet crashed from insufficient operating memory. The trends were confirmed for the other two assemblers; however, a drop in accuracy was observed, possibly as a result of greater complexity of the whole genome compared with a single chromosome (Appendix S11). Therefore, a lower accuracy is also expected in red clover genome assembly. The other assemblers failed to generate the output. They crashed after a few hundred hours of running time (run of assembler was limited to max. 2 mo for a specific run), mainly from insufficient operating memory (available 500 GB RAM) or for unknown reasons.

***Genome characterization*—**The red clover genome was evaluated to describe and characterize a newly sequenced genome. Repetitive content of the genome along with protein coding and nonprotein coding genes were predicted. Gene functions were examined and related to their pathways. Shared gene families were identified among the chosen plants.

In the clustering-based approach to repeat characterization, clusters contained 63% of all analyzed reads, with 20% of the reads being assigned to the 12 largest clusters representing the most abundant repetitive elements in the genome (Fig. 1). Based on detailed inspection, the clusters were divided into repeat classes representing more than 45% of the genome (Table 2). The most prevalent repetitive elements belong to Ty-1/Copia (12.22%), which are slightly more abundant than Ty-3/Gypsy elements (9.76%). All main plant lineages of retrotransposons are present in the red clover genome, although their abundances differ substantially. Each retrotransposon subclass is mainly represented by just one lineage: the Chromovirus and Maximus lineage for Ty-3/Gypsy and Ty-1/Copia, respectively. Present DNA transposons belong to all main groups, with PIF/Harbinger and Mutator being the most prevalent.

Possible SSR loci were predicted in red clover, and primers were designed. A total of 86 434 SSR sites were found (online Appendix S12).

A total of 64 761 complete and partial genes were predicted in the *T. pratense* genome, with 48 130 predicted proteins being equal to or longer than 100 amino acids. The total number of predicted genes is possibly increased by pseudogenes and partial genes, and this number can be improved by future RNA sequencing. Nevertheless, 47 398 (73.2%) genes were fully annotated and their functions were predicted. For every gene, the participation in a particular biological process as well as molecular function were predicted (Fig. 2).

Comparison at higher taxonomic levels (Fig. 3) revealed 12 192 orthologous groups (gene families) conserved between legumes and *Arabidopsis thaliana* as the outgroup species and only 202 and 181 orthologous groups specific for galegoid (*T. pratense*, *M. truncatula*, *L. japonicus*, and *C. arietinum*) or millettioid (*C. cajan*, and *G. max*) species, respectively. There were 16 773 orthologous groups conserved between galegoid and millettioid species. Examining the orthologous gene groups provides an important foundation for comparative biology and functional inference in red clover, because genes with simple orthologous relationships often have conserved functions whereas genes duplicated more recently relative to speciation often underlie functional diversification.

Genes specific to *Trifolium pratense* and to legumes were identified. Using BLAST searches, we identified 9926 (15.3%) genes as candidate red-clover-specific genes, which did not show similarity to any sequence analyzed (Appendix S13). The legume-specific and *Trifolium*-specific genes are listed in Appendix S14. This proportion is higher than in chickpea (10%; Garg et al., 2011; Jain et al., 2013) and *Arabidopsis* (4.9%; Lin et al., 2010), but is closer to rice (17.4%; Campbell et al., 2007). The other 5212 genes were identified as legume-specific. The analysis of GO terms revealed reproduction among the biological processes and binding function among the molecular functions

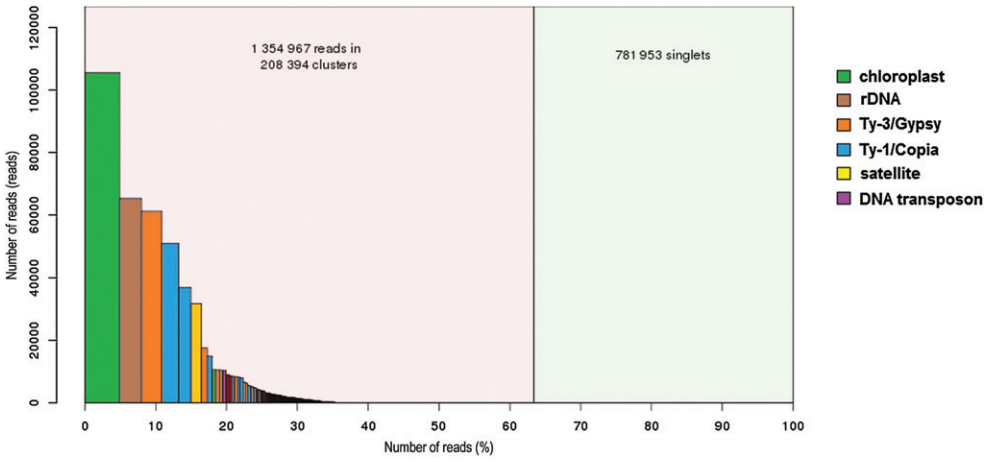

Fig. 1. Size distribution and repeat composition of clusters generated by similarity-based partitioning of *Trifolium pratense* reads. The cumulative proportion of clusters in the genome is shown along the *x*-axis. Bars on the histogram represent individual clusters; bar height corresponds to numbers of reads in the clusters and colors to the types of repetitive elements.

TABLE 2. Repeat composition of the *Trifolium pratense* genome estimated from the Illumina sequencing data.

Repeat type	Lineage	Genome proportion [%]	
Retroelements		23.81	
Ty-3/Gypsy		9.76	
	Chromovirus		4.94
	Tat/Ogre		2.41
	Athila		1.99
	other		0.42
Ty-1/Copia		12.22	
	Maximus		7.29
	Angela		1.18
	Bianca		0.69
	Tork		0.63
	Ivana/Oryco		0.59
	AleII		0.55
	TAR		0.42
	AleI/Retrofit		0.08
	other		0.79
LINE		1.30	
SINE		0.36	
other		0.17	
DNA transposons		6.07	
	PIF/Harbinger		1.23
	Mutator		1.14
	RC/Helitron		0.79
	hAT		0.57
	Mariner		0.29
	CACTA		0.16
	MITE		1.52
	other		0.37
Satellite repeats		2.58	
rDNA		5.63	
unclassified		7.05	
Total		45.14	

as the most abundant for *Trifolium*-specific gene families (Appendix S15).

On the basis of the HMM profile search results, we identified 4939 (7.6%) red clover genes belonging to 84 transcription factor families (Appendix S16). FAR1-, Tify-, NAC-, AP2-EREBP-, and Trihelix-domain-containing proteins were the most prevalent with totals of 943, 439, 256, 228, and 174, respectively. The comparison of selected transcription factor families with other sequenced legume plants and *Arabidopsis* is available in Appendix S17.

Gene annotation and presence of specific R-protein domains led to the identification of 687 R-genes candidates: 406 NBS-LRR genes, and 281 receptor-like kinases (Appendix S18). Moreover, BLAST searches led to the identification of 11 leghemoglobin genes and 542 potential nodule-specific cystein-rich peptides in the red clover genome, a number comparable with the number of nodule-specific cystein-rich peptides found previously in *M. truncatula* (Young et al., 2011). Identified peptides were assigned into proper families (Appendix S19).

DISCUSSION

In our study, testing of suitable software preceded assembly of the genome of *T. pratense* because these systems had been evaluated mainly by their authors and very often only in terms of contiguity (Li et al., 2010a; Liu et al., 2011). Recently, three extensive studies have been published concerning the performance of assemblers on several organisms with different genome lengths (Earl et al., 2011; Salzberg et al., 2012; Bradnam et al., 2013), but testing data from the plant kingdom had not been included. Testing of assemblers for the use on a particular species was more recently suggested by Bradnam et al. (2013) because results of one assembler may vary among the species.

Technology for sequencing is a rapidly evolving field, but it is limited by its methodology. With the standard Illumina protocol, biases can be transferred into NGS data. For example, amplification of genomic fragments by PCR can produce biases in GC content (Aird et al., 2011). Also PCR error can result in a false SNP site. However, a PCR-free library preparation can be the solution (Quail et al., 2008; Kozarewa et al., 2009).

On the basis of our results, using contiguity as the sole criterion for assembly evaluation was clearly insufficient and biased even with the small error rate. Despite the use of relatively relaxed accuracy parameters, substantial differences were observed among the assemblies. The accuracy of an assembly has proven to be as important as contiguity because a coherent assembly often arises from misassembled long contigs and wrongly formed scaffolds, thus artificially increasing the assembly statistics. This type of error is the most prevalent and is also the most significant (Salzberg et al., 2012). Because assembly with the longest contigs or scaffolds is usually preferred by researchers, there is a high probability of such errors (SOAPdenovo and scaffolding step of Pasha). On the other hand, even accuracy should not be the only criterion for assembly assessment as it may result in an accurate but very fragmented assembly (as in the case of SGA). What is needed, therefore, is a balance between contiguity and accuracy.

We demonstrated that error correction can improve assemblies, in some instances, very significantly (Kao et al., 2011; Yang et al., 2012). But the benefit of this correction cannot be predicted because of the specific genomic structure of different species. However in our case, the improvement was very significant (average contig length was improved by 22%, N50 by 37). We therefore believe that it, too, should be included in genome assembly studies. The different approaches vary in their success; however, those adopting *k*-mer frequencies (Echo) are currently more effective than others. Our findings, which provide an evaluation from a practical viewpoint, were also independently verified by the ECR evaluation. This toolkit provides insight into the type and quantities of errors by sorting them into categories based on error-correction success, namely, true positives (TPs), false positives (FPs), and false negatives (FNs). It also calculates the sensitivity (sensitivity = TPs / total errors) and gain [gain = (TPs − FPs) / (TPs + FNs)] of the error-correction process.

We compared the genome of *T. pratense* with sequenced genomes of the legume family and other plant species to support our findings. In contrast to the other sequenced leguminous plants (Sato et al., 2008; Schmutz et al., 2010; Young et al., 2011; Varshney et al., 2012, 2013), the most prevalent repetitive elements belong to the Ty-1/Copia family, which is followed by Ty-3/Gypsy elements. In the leguminous plants, *G. max* (Schmutz et al., 2010), with its genome size of 1.1 Gbp, has the largest content of repetitive DNA (61%). While the genome of *T. pratense* contains almost as many repeats as in chickpea (58.14%; Varshney et al., 2013) and pigeon pea (51.67%; Varshney et al., 2012), its genome is only half their size (738.09 Mbp for chickpea, 833.07 Mbp for pigeon pea). *M. truncatula* (Young et al., 2011) and *L. japonicus* (Sato et al., 2008), whose genomes are of comparable size, have only 30.50%

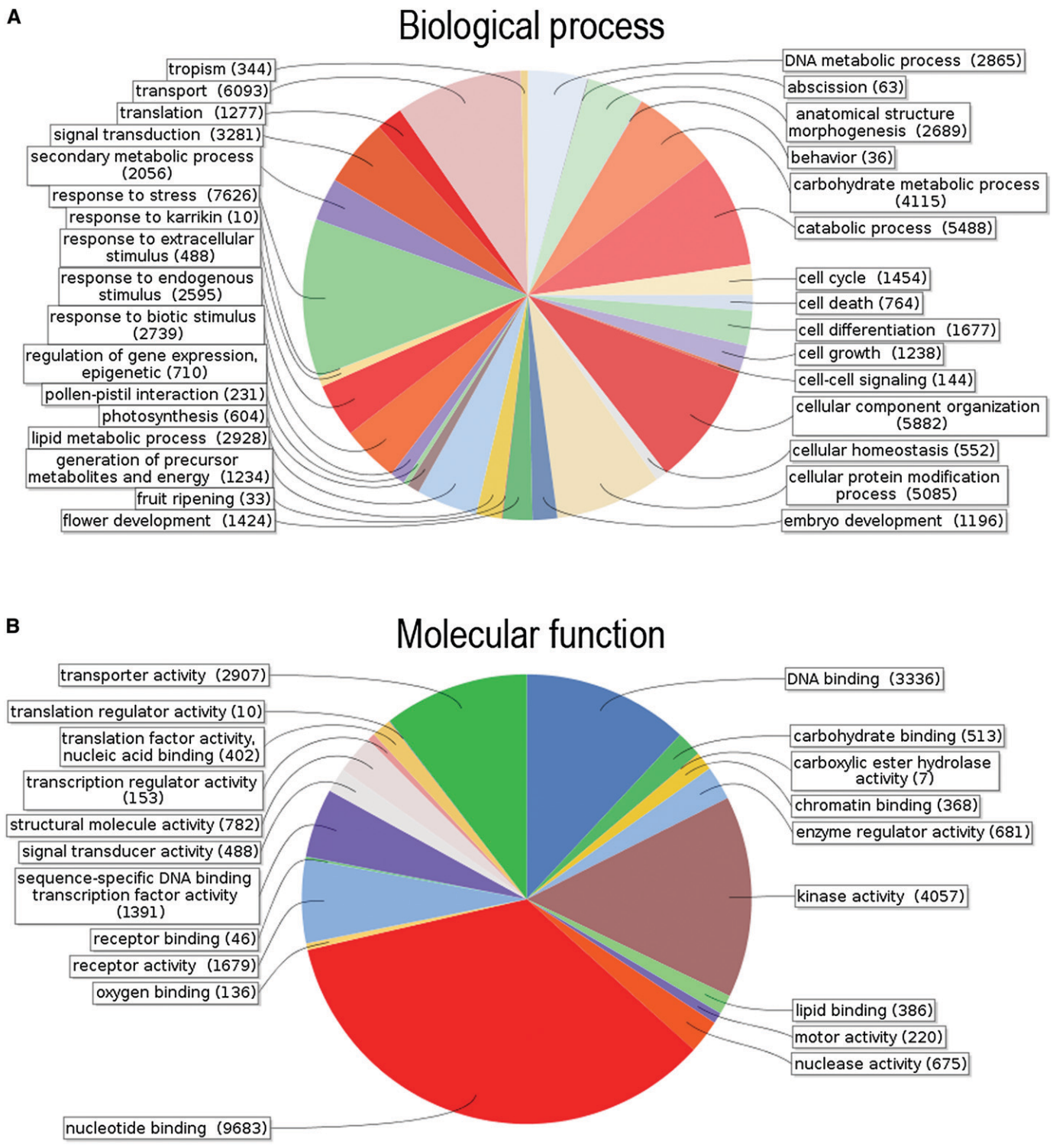

Fig. 2. Classification of red clover genes into plant GOslim categories. (A) Results for GOslim classes of biological process. (B) Results for GOslim classes of molecular function.

and 37.58% repetitive content, respectively. DNA transposons in the red clover genome are more abundant than in *M. truncatula* (3.40%), *L. japonicus* (3.31%), and pigeon pea (4.53%), but are less than half as abundant as in *G. max* (16.50%). DNA transposons can be divided into several groups, with PIF/Harbinger the most prevalent and CACTA the least. When compared with the other legumes, the abundance of DNA transposon classes is clearly not conserved among these plant species, in contrast to the retrotransposons. CACTA is the most abundant transposon class in *G. max*, Mutator in *M. truncatula*, and PIF/Harbinger in

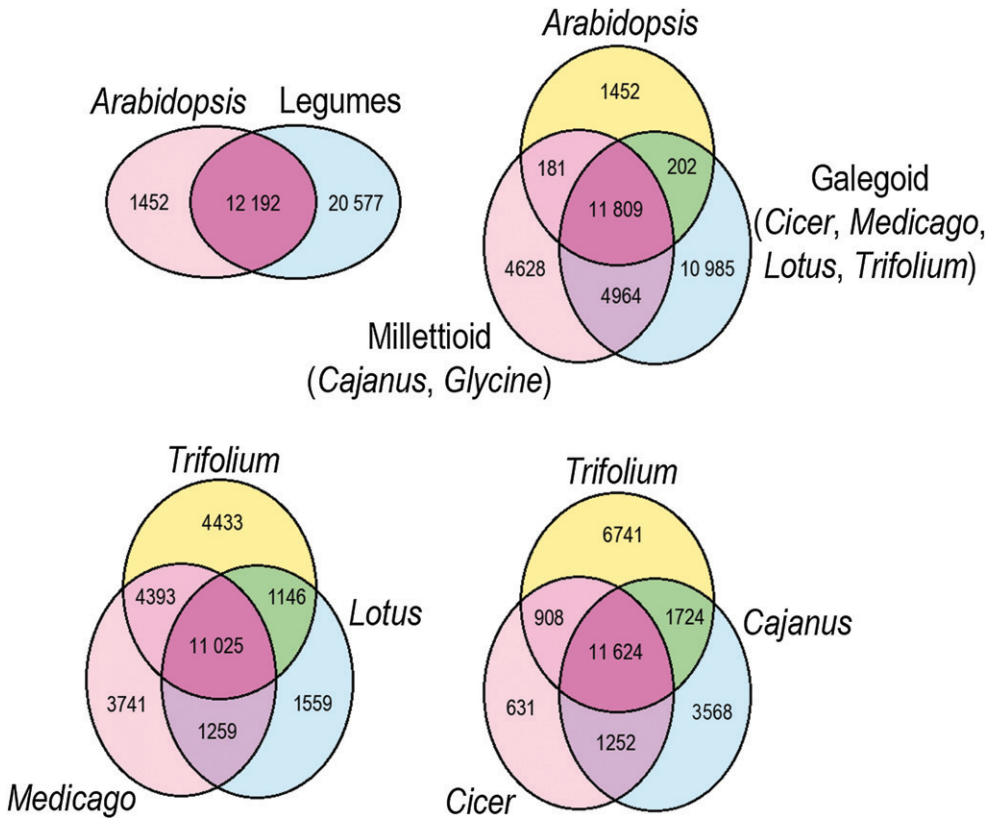

Fig. 3. Shared and unique gene families in legume species from Galegoid (*Trifolium pratense*, *Medicago truncatula*, *Lotus japonicus*, and *Cicer arietinum*) or Millettioid clade (*Cajanus cajan*, and *Glycine max*) and *Arabidopsis thaliana*.

T. pratense. The genome of red clover also contains a substantial proportion of simple sequence repeats, which can be used for genetic mapping of qualitative trait loci (QTLs) or for molecular breeding approaches such as marker-assisted selection. In these approaches, the detected SSR loci can be used for genetic fine-mapping, genotyping of specific genes or regions, or identifying variability in new genetic resources.

A total of 64 761 complete and partial genes were predicted for the genome of *T. pratense*, and 47 398 were fully annotated. The number of annotated genes (Table 3) is very close to the total number of genes in *M. truncatula* (48 066 excluding transposable elements [TEs]; Young et al., 2011) and *Populus trichocarpa* (45 555; Tuskan et al., 2006), but the number differs substantially from those for *Oryza sativa* (39 045 excluding TEs; RAGP v7; http://rice.plantbiology.msu.edu/; Kawahara et al., 2013) and especially *Arabidopsis thaliana* (28 775 excluding TEs; TAIR10; http://www.arabidopsis.org; Initiative TAG, 2000), *Zea mays* (32 540; Schnable et al., 2009) and *Vitis vinifera* (30 434; Jaillon et al., 2007). Although, 14 322 of 62 388 proteins in *M. truncatula* are related to TEs, in red clover only 212 genes were identified based on the gene annotation. However, this number will very likely increase with improvements in genome annotation. On the basis of the closer relationships between these two leguminous plants, we can presume the number of transcripts belonging to transposable elements to be very similar to that in *M. truncatula*. RNA sequencing can improve this prediction in the future. The fact that *T. pratense* has the shortest genes among the compared plant species is likely due to natural variability, because the average gene length differs substantially throughout a given sequenced plant species. Red clover and *M. truncatula* nevertheless have comparable numbers of exons per gene, in contrast to the other plants, and may be an effect of their close relationship. The red clover gene contains three introns on average. The intron length is a major difference in the compared plants. These values are likely related to genome size, as was previously postulated for *Zea mays* (Wei et al., 2009). In contrast to their many diverse features, exon length tends to be rather constant among plants, despite

TABLE 3. Overall statistics of predicted genes in the *Trifolium pratense* genome and comparison with other sequenced plant species.

Statistic	*Trifolium pratense*	*Medicago truncatula*	*Arabidopsis thaliana*	*Oryza sativa*	*Zea mays*	*Populus trichocarpa*	*Vitis vinifera*
No. of genes [TEs]	47 398 [212] [a]	62 388 [14 322]	33 602 [4827]	55 986 [16 941]	32 540	45 555	30 434
Gene length (bp)	1785	1953	1857	2853	3757	2300	3399
Exons per gene	3.6	3.7	5.3	4.9	5.3	4.3	5
Exon length (bp)	248.6	253	280	254	304	254	130
Intron length (bp)	295.2	411	155	413	516	379	213

[a] Number of genes annotated to be related to transposable elements (TEs).

the different sizes of the genome in the compared plants, since coding regions are less likely to be affected by genome size differences.

A total of 5212 legume-specific genes were predicted in the presented red clover genome assembly. However, this number is expected to increase with the availability of new legume sequences. Simultaneously, sequences found to be species-specific will decrease because more sequences would show homology to these newly sequenced genes from other legumes. Features of legume-specific and *Trifolium*-specific genes differ from those of the other genes. Shorter and fewer introns and an unusual GC content in lineage-specific genes have also been observed in studies of Campbell et al. (2007), Lin et al. (2010), and Jain et al. (2013). Even though the origin of lineage-specific genes is not exactly clear, several causes may contribute to the prediction of lineage-specific genes, such as lateral gene transfer and/or gene duplication followed by rapid sequence divergence, and de novo emergence from non-genic sequences (Jain et al., 2013). Some of the lineage-specific genes might also arise from genome assembly and annotation artifacts as well.

Red clover has more genes containing a transcription factor compared with other legumes and *Arabidopsis* (Libault et al., 2009; Schmutz et al., 2010). However, different abundances of transcription factor families are common among plants (http://planttfdb.cbi.pku.edu.cn/). For example, AP2-EREBP transcription factor is more than twice as abundant as it is in legumes and *Arabidopsis*, except for soybean in which it is much more prevalent. On the other hand, bZIP-, LOB-, and WRKY-domain-containing proteins occur in similar numbers.

In the presented assembly, 687 R-genes were identified and assigned to an appropriate R-gene class. The comparison with other sequenced species is available in Appendix S20. Red clover has about a half of the resistance genes that have been found in *M. truncatula* and *G. max* (Young et al., 2011; Jain et al., 2013). The number of resistance genes is closer to *C. arietinum*, although the receptor-like kinases in *C. arietinum* are nearly twice as abundant. Also, the quantity of receptor-like kinases in red clover is smaller than in other compared species. Nevertheless, great differences in the abundance of other R-gene classes are also observed. For example, red clover's most abundant R-genes belong to the CC-NBS-LRR class, even more than in *M. truncatula*. However in total, *M. truncatula* has significantly more genes in all other R-gene classes compared with red clover. Because of the complexity of R-genes, it is also possible that the current assembly might not capture all of the resistance genes. Further analysis can improve our findings.

The homology hits led to the identification of 11 leghemoglobin proteins, which play a significant role in the fixation of air nitrogen (Ott et al., 2005). The number of leghemoglobins found is similar to that found in *M. truncatula*. Phylogenetic analysis of the leghemoglobin genes of *T. pratense*, *M. truncatula*, *G. max*, and *C. arietinum* showed that red clover genes clustered with their homologues from other species, and only three clustered distinctly (Appendix S21).

Red clover is a very important intercrop plant that is useful in sustaining arable land. The genome and description presented herein should provide relevant additional information for plant breeders and researchers in creating new varieties based on molecular breeding and selection. Moreover, the description of specific gene families in the red clover genome and its comparison with other legumes should help with the utilization of presented genomic information into practice.

LITERATURE CITED

AIRD, D., M. G. ROSS, W. S. CHEN, M. DANIELSSON, T. FENNELL, C. RUSS, D. B. JAFFE, ET AL. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* 12: R18.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, AND D. J. LIPMAN. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

BRADNAM, K. R., J. N. FASS, A. ALEXANDROV, P. BARANAY, M. BECHNER, I. BIROL, S. BOISVERT, ET AL. 2013. Assemblathon 2: Evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* 2: 10.

CAMPBELL, M. A., W. ZHU, N. JIANG, H. LIN, S. OUYANG, K. L. CHILDS, K. L. HAAS, ET AL. 2007. Identification and characterization of lineage-specific genes within the *Poaceae. Plant Physiology* 145: 1311–1322.

CHAISSON, M. J., AND P. A. PEVZNER. 2008. Short read fragment assembly of bacterial genomes. *Genome Research* 18: 324–330.

CONESA, A., S. GÖTZ, J. M. GARCIA-GOMEZ, J. TEROL, M. TALON, AND M. ROBLES. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.

CONWAY, T., J. WAZNY, A. BROMAGE, J. ZOBEL, AND B. BERESFORD-SMITH. 2012. Gossamer—A resource-efficient *de novo* assembler. *Bioinformatics* 28: 1937–1938.

DA MAIA, L. C., D. A. PALMIERI, V. Q. DE SOUZA, M. M. KOPP, F. I. F. DE CARVALHO, AND A. C. DE OLIVEIRA. 2008. SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International Journal of Plant Genomics* 2008: 412696.

DELCHER, A. L., A. PHILLIPPY, J. CARLTON, AND S. L. SALZBERG. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* 30: 2478–2483.

EARL, D. A., K. BRADNAM, J. ST. JOHN, A. DARLING, D. LIN, J. FAAS, H. YU, ET AL. 2011. Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Research* 21: 2224–2241.

GARG, R., R. K. PATEL, S. JHANWAR, P. PRIYA, A. BHATTACHARIEE, G. YADAV, S. BHATIA, ET AL. 2011. Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. *Plant Physiology* 156: 1661–1678.

HABERER, G., S. YOUNG, A. K. BHARTI, H. GUNDLACH, C. RAYMOND, G. FUKS, E. BUTLER, ET AL. 2005. Structure and architecture of the maize genome. *Plant Physiology* 139: 1612–1624.

HERNANDEZ, D., P. FRANCXOIS, L. FARINELLI, M. OSTERAS, AND J. SCHRENZEL. 2008. *De novo* bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Research* 18: 802–809.

HUNTER, S., R. APWEILER, T. K. ATTWOOD, A. BAIROCH, A. BATEMAN, D. BINNS, P. BORK, ET AL. 2009. InterPro: The integrative protein signature database. *Nucleic Acids Research* 37: D211–D215.

ISOBE, S., H. HISANO, S. SATO, H. HIRAKAWA, K. OKUMURA, K. SHIRASAWA, S. SASAMOTO, ET AL. 2012. Comparative genetic mapping and discovery of linkage disequilibrium across linkage groups in white clover *Trifolium repens* (L.). *Genes Genomes Genetics* 2: 607–617.

ISOBE, S., R. KÖLLIKER, H. HISANO, S. SASAMOTO, T. WADA, I. KLIMENKO, K. OKUMURA, AND S. TABATA. 2009. Construction of a consensus linkage map for red clover *Trifolium pratense* (L.). *BMC Plant Biology* 9: 57.

JAILLON, O., J. M. AURY, B. NOEL, A. POLICRITI, C. CLEPET, A. CASAGRANDE, N. CHOISNE, ET AL. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449: 463–467.

JAIN, M., G. MISRA, R. K. PATEL, P. PRIYA, S. JHANWAR, A. W. KHAN, N. SHAH, ET AL. 2013. A draft genome of the pulse crop chickpea (*Cicer arietinum* L.). *Plant Journal* 74: 715–729.

KAO, W. C., A. H. CHAN, AND Y. S. SONG. 2011. Echo: A reference-free short-read error correction algorithm. *Genome Research* 21: 1181–1192.

KATAOKA, R., M. HARA, S. KATO, S. ISOBE, S. SATO, S. TABATA, AND N. OHMIDO. 2012. Integration of linkage and chromosome maps of red clover *Trifolium pratense* (L.). *Cytogenetic and Genome Research* 137: 60–69.

KAWAHARA, Y., M. DE LA BASTIDE, J. P. HAMILTON, H. KANAMORI, W. R. MCCOMBIE, S. OUYANG, D. C. SCHWARTZ, ET AL. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6: 4.

KELLEY, D. R., M. C. SCHATZ, AND S. L. SALZBERG. 2010. Quake: Quality-aware detection and correction of sequencing errors. *Genome Biology* 11: R116.

KOZAREWA, I., Z. NING, M. A. QUAIL, M. J. SANDERS, M. BERRIMAN, AND D. J. TURNER. 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods* 6: 291–295.

KURTZ, S., A. NARECHANIA, J. C. STEIN, AND D. WARE. 2009. A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* 9: 517.

LANGMEAD, B., C. TRAPNELL, M. POP, AND S. L. SALZBERG. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10: R25.

LI, H., AND R. DURBIN. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760.

LI, R., W. FAN, G. TIAN, H. ZHU, L. HE, J. CAI, Q. HUANG, ET AL. 2010a. The sequence and *de novo* assembly of the giant panda genome. *Nature* 463: 311–317.

LI, H., B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN, N. HOMER, G. MARTH, ET AL. 2009. The sequence alignment/map SAM, format and SAM tools. *Bioinformatics* 25: 2078–2079.

LI, L., C. J. STOECKERT, AND D. S. ROOS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* 13: 2178–2189.

LI, R., H. ZHU, J. RUAN, W. QIAN, X. FANG, Z. SHI, Y. LI, ET AL. 2010b. *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20: 265–272.

LIBAULT, M., T. JOSHI, V. A. BENEDITO, D. XU, M. K. UDVARDI, AND G. STACEY. 2009. Legume transcription factor genes: What makes legumes so special? *Plant Physiology* 151: 991–1001.

LIN, H., G. MOGHE, S. OUYANG, A. IEZZONI, S. H. SHIU, X. GU, AND C. R. BUELL. 2010. Comparative analyses reveal distinct sets of lineage-specific genes within *Arabidopsis thaliana. BMC Evolutionary Biology* 10: 41.

LINDBLAD-TOH, K., C. M. WADE, T. S. MIKKELSEN, E. K. KARLSSON, D. B. JAFFE, M. KAMAL, M. CLAMP, ET AL. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438: 803–819.

LIU, Y., B. SCHMIDT, AND D. L. MASKELL. 2011. Parallelized short read assembly of large genomes using de Bruijn graphs. *BMC Bioinformatics* 12: 354.

MARDIS, E. R. 2013. Next-generation sequencing platforms. *Annual Review of Analytical Chemistry* 6: 287–303.

MINOCHE, A. E., J. C. DOHM, AND H. HIMMELBAUER. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biology* 12: R112.

MYERS, E. W., G. G. SUTTON, A. L. DELCHER, I. M. DEW, D. P. FASULO, M. J. FLANIGAN, S. A. KRAVITZ, ET AL. 2000. A whole-genome assembly of drosophila. *Science* 287: 2196–2204.

NAWROCKI, E. P., D. L. KOLBE, AND S. R. EDDY. 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics* 25: 1335–1337.

NOVÁK, P., P. NEUMANN, AND J. MACAS. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* 11: 378.

NOVÁK, P., P. NEUMANN, J. PECH, J. STEINHAISL, AND J. MACAS. 2013. RepeatExplorer: A galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics* 29: 792–793.

PÉREZ-RODRÍGUEZ, P., D. M. RIAÑO-PACHÓN, L. G. CORRÊA, S. A. RENSING, B. KERSTEN, AND B. MUELLER-ROEBER. 2010. PlnTFDB: Updated content and new features of the plant transcription factor database. *Nucleic Acids Research* 38 (Database issue): D822–D827.

OTT, T., J. T. VAN DONGEN, C. GUNTHER, L. KRUSELL, G. DESBROSSES, H. VIGEOLAS, V. BOCK, ET AL. 2005. Symbiotic leghemoglobins are crucial for nitrogen fixation in legume root nodules but not for general plant growth and development. *Current Biology* 15: 531–535.

QUAIL, M. A., I. KOZAREWA, F. SMITH, A. SCALLY, P. J. STEPHENS, R. DURBIN, H. SWERDLOW, AND D. J. TURNER. 2008. A large genome center's improvements to the Illumina sequencing system. *Nature Methods* 5: 1005–1010.

QUAIL, M. A., M. SMITH, P. COUPLAND, T. D. OTTO, S. R. HARRIS, T. R. CONNOR, A. BERTONI, ET AL. 2012. A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341.

SALMELA, L., AND J. SCHRÖDER. 2011. Correcting errors in short reads by multiple alignments. *Bioinformatics* 27: 1455–1461.

SALZBERG, S. L., A. M. PHILLIPPY, A. ZIMIN, D. PUJU, T. MAGOC, S. KOREN, T. J. TREANGEN, ET AL. 2012. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* 22: 557–567.

SATO, S., S. ISOBE, E. ASAMIZU, N. OHMIDO, R. KATAOKA, Y. NAKAMURA, T. KANEKO, ET AL. 2006. Comprehensive structural analysis of the genome of red clover *Trifolium pratense* L. *DNA Research* 12: 301–364.

SATO, S., Y. NAKAMURA, T. KANEKO, E. ASAMIZU, T. KATO, M. NAKAO, S. SASAMOTO, ET AL. 2008. Genome structure of the legume, *Lotus japonicus. DNA Research* 15: 227–239.

SCHMUTZ, J., S. B. CANNON, J. SCHLUETER, J. MA, T. MITROS, W. NELSON, AND D. L. HYTEN. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 463: 178–183.

SCHNABLE, P. S., D. WARE, R. S. FULTON, J. C. STEIN, F. WEI, S. PASTERNAK, C. LIANG, ET AL. 2009. The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326: 1112–1115.

SIMPSON, J., K. WONG, S. JACKMAN, J. SCHEIN, S. JONES, AND I. BIROL. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research* 19: 1117–1123.

SIMPSON, J. T., AND R. DURBIN. 2012. Efficient *de novo* assembly of large genomes using compressed data structures. *Genome Research* 22: 549–556.

SONNHAMMER, E. L. L., AND R. DURBIN. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167: GC1–GC10.

STANKE, M., R. STEINKAMP, S. WAACK, AND B. MORGENSTERN. 2004. Augustus: A web server for gene finding in eukaryotes. *Nucleic Acids Research* 32: W309–W312.

TAG [THE ARABIDOPSIS GENOME] INITIATIVE. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* 408: 796–815.

THOMPSON, H. L., R. SCHMIDT, AND C. DEAN. 1996. Identification and distribution of seven classes of middle repetitive DNA in the *Arabidopsis thaliana* genome. *Nucleic Acids Research* 24: 3017–3022.

TUSKAN, G. A., S. DIFAZIO, S. JANSSON, J. BOHLMANN, I. GRIGORIEV, U. HELLSTEN, N. PUTNAM, ET AL. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596–1604.

VARSHNEY, R. K., W. CHEN, Y. LI, A. K. BHARTI, R. K. SAXENA, J. A. SCHLUETER, M. T. A. DONOGHUE, ET AL. 2012. Draft genome sequence of pigeonpea *Cajanus cajan*, an orphan legume crop of resource-poor farmers. *Nature Biotechnology* 30: 83–89.

VARSHNEY, R. K., C. SONG, R. K. SAXENA, S. AZAM, S. YU, A. G. SHARPE, S. CANNON, ET AL. 2013. Draft genome sequence of chickpea *Cicer arietinum*, provides a resource for trait improvement. *Nature Biotechnology* 31: 240–246.

VIŽINTIN, L., B. JAVORNIK, AND B. BOHANEC. 2006. Genetic characterization of selected *Trifolium* species as revealed by nuclear DNA content and its rDNA region analysis. *Plant Science* 170: 859–866.

WEI, F., J. ZHANG, S. ZHOU, R. HE, M. SCHAEFFER, K. COLLURA, D. KUDRNA, ET AL. 2009. The physical and genetic framework of the maize B73 genome. *PLOS Genetics* 5: e1000715.

YANG, X., S. P. CHOCKALINGAM, AND S. ALURU. 2012. A survey of error correction methods for next generation sequencing. *Briefings in Bioinformatics* 14: 56–66.

YANG, X., K. DORMAN, AND S. ALURU. 2010. Reptile representative tiling for short read error correction. *Bioinformatics* 26: 2526–2533.

YOUNG, N. D., F. DEBELLE, G. E. D. OLDROYD, R. GEURTS, S. B. CANNON, M. K. UDVARDI, V. A. BENEDITO, ET AL. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* 480: 520–524.

ZDOBNOV, E. M., AND R. APWEILER. 2001. InterProScan—An integration platform for the signature recognition methods in InterPro. *Bioinformatics* 17: 847–848.

ZERBINO, D., AND E. BIRNEY. 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research* 18: 821–829.

ZHANG, H. B., Z. ZHAO, X. DING, A. H. PATERSON, AND R. A. WING. 1995. Preparation of megabase-size DNA from plant nuclei. *Plant Journal* 7: 175–184.

ZHANG, W., J. CHEN, Y. YANG, Y. TANG, J. SHANG, AND B. SHEN. 2011. A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLoS ONE* 6: e17915.

APPENDIX 1. Comparison of eight assemblers on testing data sets (*Medicago truncatula*) with 0.1% and 1% error rates.

Assembler (insert size set to 430 bp)	No. of contigs	Average contig length	Largest contig	N50	N90	Correct of contigs (%) [a]	Correct assembly (%) [a]
Error rate 0.1% (DS0.1)							
Abyss (k64)	10 140	2716.3	238 755	37 706	4271	71.43	86.22
CABOG	1115	21 417.51	159 483	41 340	9084	85.65	77.21
Edena	3752	7138.02	544 127	98 109	9493	63.33	66.82
Gossamer (k62)	13 135	1970.06	61 231	10 958	935	73.38	84.9
Pasha contigs (k31)	49 785	499.97	25 957	2229	134	88.59	96.62
Pasha scaffolds (k31)	2307	10 806.11	176 725	43 896	9316	41.7	5.96
SGA	158 970	261.46	33 139	1104	103	88.48	94.18
SOAP contigs (63mer)	256 123	118.17	12 172	691	32	21.99	75.69
SOAP scaffolds (63mer)	3395	7281.25	144 220	37 893	6997	42.97	9.51
Velvet (k61)[b]	18 089	1441.18	61 075	8068	455	90.49	97.37
Velvet (k61;430)[b]	14 404	1801.69	61 439	8135	586	89.14	96.69
Error rate 1% (DS1)							
Abyss (k64)	12 820	2149.43	192 245	34 435	3529	69.7	84.31
CABOG	150 511	266.49	27 734	410	110	87.97	93.58
Edena	41 563	686.96	39 115	4188	150	78.64	94.62
Gossamer (k62)	20 687	1263.3	40 953	5592	514	78.64	91.53
Pasha contigs (k31)	49 781	499.43	25 955	2186	135	88.65	96.61
Pasha scaffolds (k31)	2360	10 525.78	181 184	39 502	9088	42.58	6.44
SGA	139 480	281.35	35 770	1799	101	80.63	91.62
SOAP contigs (63mer)	1 304 176	50.98	846	49	27	11.59	23.8
SOAP scaffolds (63mer)	6590	3555.95	127 804	22 152	4033	38.01	1.46
Velvet (k61)[b]	26 906	983.29	36 097	4734	277	81.57	96.12
Velvet (k61;430)[b]	22 553	1166.02	36 097	4810	415	79.4	95.59

[a] Assembly accuracy with parameters set to 95% identity and 80% contig alignment.
[b] Comparison of automatically derived and user-defined insert size.

APPENDIX 2. Evaluation of the performance of assemblers based on their results in both testing data sets of *Medicago truncatula*.

Assembler (insert size set to 430 bp)	Points from DS0.1 (max = 77)	Points from DS1 (max = 77)	Sum of points (max = 154)	Overall assembler performance
Abyss (k64)	50	56	106	0.69
CABOG	60	29	89	0.58
Edena	56	40	96	0.62
Gossamer (k62)	41	51	92	0.60
Pasha contigs (k31)	32	40	72	0.47
Pasha scaffolds (k31)	52	59	111	0.72 [b]
SGA	27	29	56	0.36
SOAP contigs (63mer)	10	9	19	0.12
SOAP scaffolds (63mer)	46	51	97	0.63 [b]
Velvet (k61) [a]	42	48	90	0.58
Velvet (k61; 430) [a]	46	50	96	0.62

[a] Comparison of automatically derived and user-defined insert size.
[b] Very low accuracy of the assembly (see Appendix 1).

APPENDIX 3. Impact of error correction on assembly statistics and statistical evaluation.

Assembler (k64; insert size set to 430 bp)	No. of contigs	Mean contig length	N50	Correct contigs (%)	Correct assembly (%)	SNP-bwa	SNP-bowtie	Total errors	TP	FP	FN	Sensitivity (%)	Gain
Error rate 0.1% (DS0.1)													
Abyss-Coral	19644	1433.91	9896	47.52	55.81	6776	4096	1265373	1251635	6472888	13738	98.91	–4.13
Abyss-Echo (k15)	9211	2982.73	35378	56.95	59.94	784	253	1265501	1129832	69241	135669	89.28	0.84
Abyss-Reptile (k13)	10034	2745.98	37918	55.1	59.9	4172	3559	1265598	1185172	169020	80426	93.65	0.8
Abyss-Quake (k16)	10039	2742.25	38366	55.25	59.96	4213	3532	1108846	1089340	138224	19506	98.24	0.86
Error rate 1% (DS1)													
Abyss-Coral	19794	1416.98	9832	65.15	84.49	7129	4087	12380705	12309833	7772824	70872	99.43	0.37
Abyss-Echo (k15)	10301	2681.45	35125	69.94	85.14	1655	1062	12380616	11370956	317439	1009660	91.85	0.89
Abyss-Reptile (k13)	11107	2483.69	37498	69.89	84.05	4586	3233	12380854	9718854	1225085	2662000	78.5	0.69
Abyss-Quake (k16)	12383	2219.72	34209	71.2	85.37	4668	2950	3344669	3257616	411882	87053	97.4	0.85