TÜBİTAK

Turkish Journal of Agriculture and Forestry

http://journals.tubitak.gov.tr/agriculture/

Research Article

Turk J Agric For
(2015) 39: 470-488

doi:10.3906/tar-1409-70

Single nucleotide polymorphism discovery through Illumina-based transcriptome sequencing and mapping in lentil

Hülya YILMAZ TEMEL[1], Deniz GÖL[2], Hilal Betül KAYA AKKALE[1], Abdullah KAHRİMAN[3], Muhammed Bahattin TANYOLAÇ[1,*]

[1]Department of Bioengineering, Ege University, İzmir, Turkey
[2]Ministry of Agriculture, Food Analysis Laboratory, İzmir, Turkey
[3]Department of Field Crops, Harran University, Şanlıurfa, Turkey

Received: 15.09.2014 • **Accepted:** 14.10.2014 • **Published Online:** 12.06.2015 • **Printed:** 30.06.2015

Abstract: Lentil, which belongs to the family Leguminosae (Fabaceae), is a diploid (2n = 2x = 14 chromosomes) self-pollinating crop with a genome size of 4063 Mbp. Because of its nutritional importance and role in the fixation of nitrogen from the atmosphere, lentil is a widely used crop species in molecular genetic studies. By using DNA markers, to date, a limited number of polymorphic bands have been generated. Therefore, it is necessary to develop additional markers to saturate the genome at high density. Single nucleotide polymorphism (SNP) markers are promising for this purpose because of their abundance, stability, and heredity; they can be used to generate a large number of markers over a short distance that are distributed in both intragenic and intergenic regions. Transcriptome sequencing technology was applied to 2 lentil genotypes, and cDNAs were sequenced using the Illumina platform. A total of 111,105,153 sequence reads were generated after trimming. The high-quality reads were assembled, producing 97,528 contigs with an N50 of 1996 bp. The Genome Analysis Tool Kit Unified Genotyper algorithm detected 50,960 putative SNP primers. A genetic linkage map was constructed by using JoinMap4.0 and the map consists of 7 major linkage groups that could be represented as 7 chromosomes of lentil. The extensive sequence information and large number of SNPs obtained in this study could potentially be used for future high-density linkage map construction and association mapping. The large number of contigs obtained in this study could be used for the identification of orthologous transcripts from cDNA data on other organisms.

Key words: *Lens culinaris* Medik., transcriptome, linkage map, single nucleotide polymorphism, simple sequence repeat, inter-simple sequence repeat

1. Introduction

Lentil, which belongs to the family Leguminosae (Fabaceae), is an important food source for people around the world (Fikiru et al., 2007). Lentil represents the greatest source of protein after soybeans and hemp (Callaway, 2004). In addition to its nutritional importance, this crop plays a role in the fixation of nitrogen from the atmosphere and the formation of nitrogen in the soil, which replenishes nutrients and maintains soil productivity (Wong, 1980). Lentils are drought-tolerant (Karim Mojein et al., 2003) and are grown in many areas around the world. Geographically, this crop is widely cultivated in West Asia and the Indian subcontinent, North Africa, South Europe, South and North America, and Australia (Erskine, 1997). The major lentil-producing regions of the world are Asia and the West Asia/North Africa region (Erskine et al., 1998), and lentil is currently under cultivation in more than 35 countries (Yadav et al., 2007). Yadav et al. (2007) also reported that 99% of the world's lentil production is provided by 20 countries, with the most important lentil-producing countries being Australia, Canada, the United States, Bangladesh, China, India, Iran, Nepal, Syria, and Turkey. It is thought that lentil originated in and has been consumed since prehistoric times; it was one of the first crops to be cultivated, exhibiting a history dating back 8000 years, which is why it is referred to as an 'ancient orphan crop' (Yadav et al., 2007).

Lentil exhibits a genome size of approximately 4063 Mbp (Arumuganathan and Earle, 1991) and a 2n = 2x = 14 chromosome number. To understand the genetic structure of large genomes such as that of lentil, it is necessary to discover many markers to characterize the genome. Different types of markers, such as morphological, isozyme, restriction fragment length polymorphism (RFLP) (Havey and Muehlbauer, 1989), random amplified polymorphic DNA (RAPD) (Eujayl et al., 1998), amplified fragment length polymorphism (AFLP) (Eujayl et al., 1998), inter-simple sequence repeat (ISSR) (Rubeena and

* Correspondence: bahattin.tanyolac@ege.edu.tr

Taylor, 2003), simple sequence repeat (SSR) (Hamwieh et al., 2005), and intron-targeted amplified polymorphic gene-based markers (Phan et al., 2007), have been used to construct genetic maps for lentil. Tanyolac et al. (2010) constructed a molecular linkage map for lentil using AFLP, ISSR, RAPD, and morphological markers. However, these types of markers generate a limited number of polymorphic bands in the lentil genome because the variation among germplasms is narrow. It is therefore necessary to develop DNA markers to generate a robust map and saturate the genome with high-density markers. Single nucleotide polymorphism (SNP) markers appear to be promising regarding the generation of a large number of markers within a short distance along chromosomes that are evenly distributed throughout the genome. SNP markers are used in many studies because of their abundance in the genome and the availability of techniques for multiplex SNP genotyping (Hyten et al., 2010a; Shirasawa et al., 2010). SNP markers can be assayed and exploited as high-throughput molecular markers (Trick et al., 2009). In recent years, with new developments in sequencing technology, SNP discovery and SNP genotyping platforms have become important tools for performing high-throughput analyses in many crops such as tomato (Shirasawa et al., 2010), bean (Hyten et al., 2010b; Cortés et al., 2011), barley (Close et al., 2009) and *Brassica* (Trick et al., 2009). There are 3 strategies for performing genome-wide SNP discovery in nonmodel organisms: reduction of genome complexity and sequencing methods such as reduced-representation library sequencing, restriction site-associated DNA sequencing, and whole-genome sequencing and cDNA sequencing (Helyar et al., 2012).

Next-generation sequencing (NGS) technology is currently preferred over traditional sequencing methods because traditional methods are expensive, low-throughput, and time-consuming. While high-throughput methods for performing SNP assays are now reducing the cost of genotyping, SNP discovery is still expensive in crops that have not been sequenced (Hyten et al., 2010b). NGS technologies such as the Roche-454 Genome Sequencer, Illumina Genome Analyzer, and ABI SOLID System platforms are reliable, cost-effective tools for conducting genome-wide analyses of genetic variations between populations (Wang et al., 2010; Helyar et al., 2012). The superiority of the 454 pyrosequencing system is the longer read length obtained compared to the other 2 NGS platforms (Wang et al., 2010). However, many studies have demonstrated that the Illumina System is a rapid, cost-effective platform for SNP genotyping, molecular marker development, and gene discovery (Croucher et al., 2009; Anithakumari et al., 2010; Wang et al., 2010). Recently, Sharpe et al. (2013) developed 3'-cDNA reads derived from 9 *L. culinaris* and 2 *L. ervoides* genotypes using 454 pyrosequencing technology, identified SNPs, and constructed the first comprehensive SNP-based genetic map for *L. culinaris*. In another study, Verma et al. (2013) developed a high-quality expressed gene catalogue and SSR primer pairs by de novo assembly of short sequence reads of lentil (*Lens culinaris* Medik.) transcriptome.

A number of molecular marker linkage maps have been developed for lentil. In earlier studies, researchers used isozymes (Zamir and Ladizinsky, 1984; Tadmor et al., 1987) and other morphological markers (Tadmor et al., 1987) to develop a map of the *Lens* genome. The first genetic map of lentil was constructed by Havey and Muehlbauer (1989) using RFLP markers for a *L. culinaris* × *L. orientalis* cross. The map included a small number of markers and covered a small part of the genome. After this study, different genetic linkage maps were published in lentil by using different molecular markers such as RFLP, RAPD, AFLP, ISSR, and SSR markers (Tahir et al., 1993; Eujayl et al., 1998; Duran et al., 2003; Rubeena and Taylor, 2003; Hamwieh et al., 2005; Tullu et al., 2008; Tanyolac et al., 2010; Gupta et al., 2012). Recently, Sharpe et al. (2013) developed a genetic linkage map for a recombinant inbred line (RIL) population developed from the parents CDC Robin × 964a-46. The map consisted of 543 markers (6 SSRs, 537 contigs) and 7 linkage groups with 834.7 cM total map distance, and average marker distance was 1.53 cM between 2 markers.

Here we present a study in which SNP discovery was conducted in the parents of the Precoz and WA8649041 lentil cultivars through sequencing of whole cDNA strands via Illumina platform sequencing. We also selected a subset of SNPs for amplifying in the RIL population and constructed a genetic linkage map by using SNP, SSR, and ISSR markers.

2. Materials and methods

2.1. Plant material

Two parents (Precoz and WA8649041) and 101 RILs were used as plant material. Lentil seeds were obtained from Washington State University, Pullman, WA, USA. The population was developed from a Precoz × WA8649041 cross, with single-seed descent until the F7 generation. A total of 101 seeds from the RILs were grown in an experimental field of the Ministry of Agriculture in Ankara, Turkey.

2.2. Isolation of total RNA, cDNA library construction, and transcriptome sequencing

Tissue samples (roots, shoots, leaves, branches, and flowers) were harvested from Precoz and WA8649041 plants, placed in aluminum foil, labeled, and finally stored in liquid nitrogen. Total RNA was isolated using the RNeasy Plant Mini Kit (QIAGEN, Valencia, CA, USA, Cat. Number: 74903). The obtained RNA concentration and quality

were checked using a NanoDrop spectrophotometer, and the quality was checked by running on 0.8% agarose gel. mRNA was purified from total RNA (1 μg) and fragmented into pieces of 200–500 bp using poly-T oligo-attached magnetic beads through 2 rounds of purification. Cleaved RNA fragments primed with random hexamers were reverse-transcribed into first-strand cDNA using SuperScript II reverse transcriptase (Life Technologies Inc., Grand Island, NY, USA) and random primers. The RNA template was then removed, and a replacement strand was synthesized to generate double-stranded (ds) cDNA. The fragments were subsequently end-repaired and A-tailed, and adapters were ligated. The cDNA template was finally purified and enriched via PCR, and the quality of the amplified libraries was verified through capillary electrophoresis (Bioanalyzer, Agilent Technologies Inc., Santa Clara, CA, USA). Following qPCR using SYBR Green PCR Master Mix (Life Technologies Inc.), the libraries were combined with index tags in equimolar amounts in the pool. Cluster generation was carried out in the flow cell of the cBot automated cluster generation system (Illumina Inc., San Diego, CA, USA). The flow cell was then loaded into the Illumina Hiseq 2000 sequencing system (Illumina Inc.) at DNA Link Inc. in Seoul, South Korea, and paired-end sequencing was performed with a 2 × 100 bp read length.

2.3. Sequence data analysis, de novo assembly, and SNP detection

The Illumina CASAVA (v.1.8.2) pipeline was used for initial sequence processing and base-calling. The results were received in FASTQ files, which contain read sequences and associated quality scores. All samples' raw data passed the initial quality control using FastQC (v.0.10.1) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Next, we used cutadapt (v.0.9.5) (Martin, 2011) to remove any reads that were contaminated with Illumina adapters. Prior to assembly, all the reads were cleaned and trimmed using Sickle (v.1.1) (https://github.com/najoshi/sickle) with default settings. For the simulated genomic data, the de novo assembly (Robertson et al., 2010) was performed by Velvet (v.1.2.03)/Oases (v.0.2.08) (Zerbino and Birney, 2008; Schulz et al., 2012) and ABySS-PE (v. 1.3.4)/Trans-ABySS (v.1.4.4) (Simpson et al., 2009) on total reads, clean reads, and control reads, respectively. Velvet/Oases (Zerbino and Birney, 2008; Schulz et al., 2012) as run using the k-mer lengths of 25 to 75 along with other default parameters. ABySS-PE/Trans-ABySS was run using lengths of 27 to 63 followed by merging the results. We assembled each dataset using almost identical assembly parameters to compare their performance. The Burrows–Wheeler aligner (BWA; v. 0.5.9rc1) (Li and Durbin, 2009) was used to align the sequencing reads, with default parameters. We used the Genome Analysis Tool Kit (GATK) Unified Genotyper (McKenna et al., 2010) for improvement of alignments, genotype calling, and refining with recommended parameters. Forward and reverse SNP primers for each variant position were designed using Primer3 (v. 2.3.5) (Rozen and Skaletsky, 1999). The assembled transcript sequences were scanned against the nr protein sequence database to identify homologous sequences using BLASTx (2.2.27+) with an E-value threshold of 1e-05. Gene Ontology (GO) annotation was performed using BLAST2GO (v. 2.5.0) (Conesa et al., 2005) to obtain cellular component, molecular function, and biological process terms. A simplified flowchart overview of the steps followed in the assembly process is outlined in Figure 1. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to identify potential pathways represented in the transcriptome (Kanehisa and Goto, 2000).

2.4. DNA isolation

Young leaf samples were collected from the parents and lines and stored at –80 °C until isolation. Genomic DNA was extracted from a total of 103 lentil genotypes using the CTAB-PVP protocol (Doyle and Doyle, 1990). Leaf samples placed in Eppendorf tubes were ground using TissueLyser (Teknogen Co., İzmir, Turkey), and the obtained DNA was resuspended in 100 μL of TE buffer. RNase A (Thermo Scientific Co., Lafayette, CO, USA) and proteinase K (Thermo Scientific Co.) were added to each sample to remove RNA and protein contamination, followed by storage at –20 °C. The concentration and quality of the obtained DNA were checked using a NanoDrop spectrophotometer, and the quality of the isolated DNA was monitored in 0.8% agarose gels.

2.5. Marker validation

The first 500 SNP primers were selected among detected SNPs. PCR amplifications with these primers were carried out using the MJ Research PTC200 Tetrad model thermal cycler (MJ Research, Incline Village, NV, USA). The primers showing polymorphisms were applied to 101 individuals resulting from a Precoz × WA8649041 cross (total of 103 DNA samples).

SSR markers that were identified by Hamwieh et al. (2005) and Rajesh et al. (2008) were screened in the parents to determine the polymorphic ones. Polymorphic markers were used to genotype the population individuals. A total of 25 ISSR primers 15 to 23 nucleotides in length were purchased from the Biotechnology Laboratory of the University of British Columbia, Vancouver, BC, Canada, and were also used to screen the parents for polymorphism.

The forward SSR, ISSR, and SNP primer sequences were modified by adding an M13 tail (CACGACGTTGTAAAACGAC) to the 5' end and the M13 primers were labelled with 2 different fluorescent dyes, IRD 700 and IRD 800, to universally label the PCR

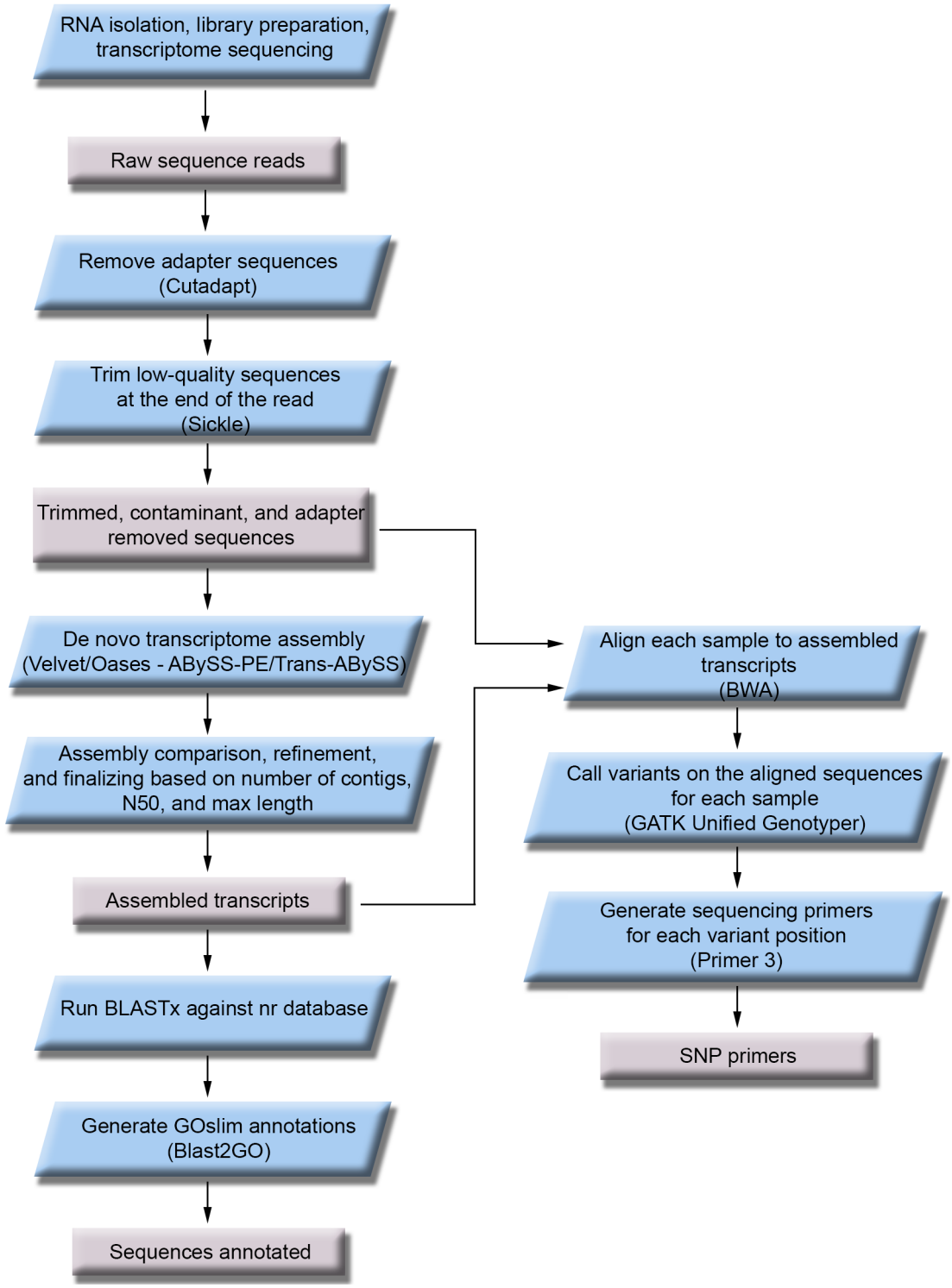

Figure 1. Overview of the different steps for sequencing, de novo assembly of the Illumina reads, and subsequent annotation of the resulting transcriptome and SNP discovery.

products. PCR amplification mixtures were prepared for each sample by mixing 4 µL of 5X Go Taq Flexi Buffer (Promega Corporation, Fitchburg, WI, USA), 0.4 µL of 10 mM dNTPs, 1.6 µL of 10 mM $MgCl_2$, 0.2 µL of forward and 0.8 µL of reverse primer (at 10 µM), 0.8 µL of M13 primer (IRD 700/800), 5 U of Go Taq DNA polymerase

(Promega Corporation), 0.8 µL of M13 primer, 5 µL of diluted genomic DNA (20 ng/µL), and 7.16 µL of water. The SSR, ISSR, and SNP PCR amplification experiments were conducted in accordance with the procedures described by Maccaferri et al. (2008). The PCR products (SSR, SNP, ISSR) were loaded in a 8% denaturing polyacrylamide gel in 1X Tris-borate-EDTA buffer; 1500 V and 40 mA were used. To further identify polymorphisms, the PCR products were analyzed using a LiCor 4300 s DNA Analyzer. Image processing for the SSR, SNP and ISSR fragments was performed using SAGA software (LiCor Biosciences, Lincoln, NE, USA). Each polymorphic band was scored visually as present in mother (a) or present in father (b) across all 101 genotypes for each primer pair.

2.6. Construction of linkage map

A set of 420 SNP, 15 SSR, and 29 ISSR markers were used to construct a linkage map by using JoinMap4.0 (Ooijen, 2006). For linkage analysis a LOD score of 3.00 and recombination fraction 0.40 were used and the Kosambi mapping function was applied to calculate the distances between markers (Kosambi, 1944). The linkage groups were numbered according to the linkage groups previously mapped by Hamwieh et al. (2005). The linkage groups presented in this study were constructed using MAPCHART 2.2 for Windows (Vorrips, 2002).

3. Results

3.1. Illumina transcriptome sequencing, de novo assembly, and SNP discovery

To define nucleotide diversity in the genic regions of the lentil genome, the parents of the RIL population (Precoz and WA8649041) were selected for transcript profiling using Illumina short-read sequencing technology. A total of 113,126,056 raw sequence reads were generated, corresponding to a cumulative 4058 Mbp of sequences for Precoz and 7467 Mbp for WA8649041. Trimming of low-quality sequences at the end of each read resulted in the removal of 2,020,903 sequences. A total of 111,105,153 high-quality reads were obtained, ranging in size from 10 bp to 101 bp, with an average length of 83 bp. A summary of the acquired sequencing data is presented in Table 1. High-quality reads were assembled using the BWA (v. 0.5.9rc1) (Li and Durbin, 2009), producing 97,528 contigs. The size of the contigs ranged from 100 to 19,077 bp, with an N50 of 1996 bp. Following alignment of the sequences, a total of 27,893,323 reads had been aligned for Precoz and 52,046,936 for WA8649041, corresponding to 70.78% and 72.59%, respectively. An overview of the sequencing and assembly statistics for the lentil transcriptome is presented in Table 2, and the assembled transcript length histogram is shown in Figure 2.

Table 1. Summary of sequencing data and trimming of 2 lentil genotypes.

Sample	Raw sequence reads	Reads after trimming	Average cleaned reads (bp)
Precoz	40,182,396	39,410,060	97
WA8649041	72,943,660	71,695,093	97.21
Total	113,126,056	111,105,153	

Table 2. Overview of the sequencing and assembly for lentil transcriptome sequencing statistics.

Total number of raw reads	113,126,056
Total number of reads after trimming	111,105,153
Average length of high-quality reads	97 bp
Total number of trimmed sequences	2,020,903
Sequence length for assembly	10,791,989,752 bp
Total number of contigs	97,528
Total number of reads aligned	79,940,259
Total number of bases aligned	7,672,999,121
Minimum contig length (bp)	100
Maximum contig length (bp)	19,077
Average contig length (bp)	1433
Total number of isotigs	23,398

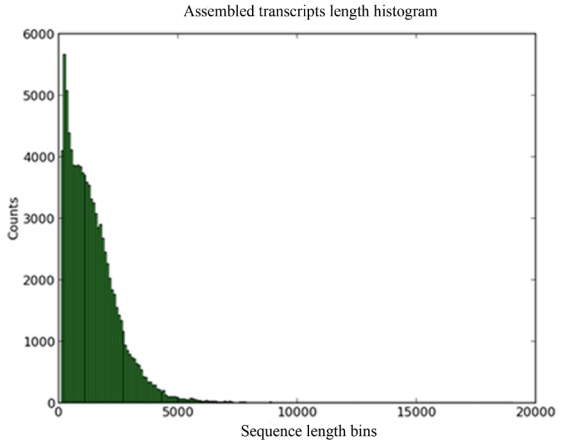

Figure 2. Histogram of the lengths of assembled transcripts.

All of the obtained sequencing reads were deposited into the Short Read Archive of the National Center for Biotechnology Information (NCBI) and can be accessed under accession number NCBI SRP026548.

3.2. SNP detection and marker polymorphisms

Performing SNP discovery with the GATK Unified Genotyper algorithm detected 50,960 putative SNP primers among 97,528 separate contigs. Since the number of SNPs is 50,960, detailed information on the developed SNPs is provided only representatively in Table 3.

SNP validation was conducted using a subset of 500 randomly chosen primers developed via transcriptome sequencing. Among the set of SNP primers that were genotyped, 420 were polymorphic (84%) and 48 were monomorphic. In the remaining 32 (6%) assays, amplification failed in 101 individuals from the RIL population.

3.3. Functional annotation of SNPs

The GO annotation results for the lentil consensus sequences for the cellular component, molecular function, and biological process categories were assigned using Blast2GO, which is a universal analysis tool for functional genomics research (Conesa et al., 2005). GO terms were assigned corresponding to a total of 88,251 sequences. An overview of the GO results for the assembled data is presented in Table 4 and the detailed results are provided representatively in Tables 5–7. Among the 35,774 sequences in the molecular function class, binding (44%) and catalytic activity (43%) constituted the major categories, followed by transporter activity (5%), molecular transducer activity (3%), structural molecule activity (2%), enzyme regulator activity (1%), transcription regulator activity (1%), and electron carrier activity (1%). The metabolic process category under the biological process classification, which included 45,099 sequences, contributed the largest proportion of annotations (35%), followed by the cellular

Table 3. Detailed sequence information and forward and reverse primer sequences of the designed SNP primer pairs (representative table).

Sequence ID	Primer left sequence	Primer right sequence
Locus_10001_Transcript_1/1_Confidence_0.000_Length_1440	CGAACCTTGTGAACCTTAGCAC	ATAGACTCCCCGAGCATGGT
Locus_10001_Transcript_1/1_Confidence_0.000_Length_1440	CCTTCCAAGAGATCGAGCACA	TCCTAGAATTTGGACCCAATTGG
Locus_10004_Transcript_2/5_Confidence_0.158_Length_218	AGGTGATGTTCCATCTCATGTGA	GTGCAGGTCACATGTTCTAGT
Locus_10007_Transcript_2/2_Confidence_0.333_Length_808	AATGGTTTTTGGTTCGGCGG	TCGATCTACCGTAATTTTAGGGTCA
Locus_10007_Transcript_2/2_Confidence_0.333_Length_808	AATGGTTTTTGGTTCGGCGG	TCGATCTACCGTAATTTTAGGGTCA
Locus_10007_Transcript_2/2_Confidence_0.333_Length_808	AGTCTGATTCAACAGAGCGAGA	TCGATCTACCGTAATTTTAGGGTCA
Locus_1000_Transcript_1/2_Confidence_0.750_Length_1377	TTCCTTCTTCCACAACCCCT	GAGTGACGGGTGGAAAGGAG
Locus_1000_Transcript_2/2_Confidence_0.000_Length_1381	TGTGGCCAAGACAGAAACACA	CTTGAATAATCACGCGCCGC
Locus_10010_Transcript_1/1_Confidence_0.800_Length_238	TGTGGACATAAAAGCTGCACT	TGAGTGATGACTTGAGATCCCT
Locus_10012_Transcript_1/1_Confidence_0.000_Length_903	GTTTATCCCGAGGGCATGGT	TGGAGGAGAAAGAAAGAGGTCT
Locus_10013_Transcript_1/1_Confidence_0.000_Length_343	CGAAGCGTTGAGTATACCGGA	TCCCACATTGCTCCATCTGAG
Locus_10015_Transcript_1/4_Confidence_0.333_Length_2091	TCCCCGTTGTTGAAAACACA	TGCAGCCTTACAGACAGTCA
Locus_10015_Transcript_1/4_Confidence_0.333_Length_2091	CCATGTCTTCGTGGCTGAGA	GCTCACAAGCTAATCGACACTG

Table 3. (Continued).

Locus_10015_Transcript_1/4_Confidence_0.333_Length_2091	TCAGCTGAAAGGTGCTTCCA	TTGCATGCAAATAAGTGCTCA
Locus_10015_Transcript_3/4_Confidence_0.333_Length_2550	TCAATGTTTTGGGGTTGAGGC	AACGAAGGGGGTGGATTTCC
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	TTGGATTATAAGGACAACCGGT	ACATAAGTTCCCTAACTCTCAGCC
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	TTGGATTATAAGGACAACCGGT	CCAGTAGACTACACCACTTGGG
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	ACCGGTATAATCTCAACTCCGA	GCACCTTCCAGTAGACTACACC
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	AGCCAAAAACAAACATGCCAGT	AACACACATGAGATACACCAAAAA
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	AGATCTTTGATAGAACATTATTGCGG	TGAGTTAATTAACACACATGAGATACA
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	TCTTTGATAGAACATTATTGCGGAA	TCCCATCTCAAAAGAGAAATGAAAA
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	TGCGGAATTTGATCCATGTGT	TCCAAATCCCATCTCAAAAGAGA
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	TGCGGAATTTGATCCATGTGT	ACCGGTTGTCCTTATAATCCAA
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	AGAGAAAGCATTAATAAAGATGCTTCT	TGTTGTGCTCTTGATATCCGA
Locus_10015_Transcript_4/4_Confidence_0.111_Length_3362	AGAGAAAGCATTAATAAAGATGCTTCT	TGTTGTGCTCTTGATATCCGA
Locus_10018_Transcript_1/1_Confidence_0.750_Length_738	ACCCTTGAATAAGATATTCTACCAGT	AGAGAGTAGAGAGTAGTAACTAGTGT
Locus_10018_Transcript_1/1_Confidence_0.750_Length_738	ACCCTTGAATAAGATATTCTACCAGT	AGAGAGTAGAGAGTAGTAACTAGTGT
Locus_10018_Transcript_1/1_Confidence_0.750_Length_738	ACGTCTTGTTTGCTTCATTTTTAGT	GAGAAACAAAAGAAAGTGAGAAATTGA
Locus_10018_Transcript_1/1_Confidence_0.750_Length_738	ACGTCTTGTTTGCTTCATTTTTAGT	GAGAAACAAAAGAAAGTGAGAAATTGA
Locus_1001_Transcript_1/10_Confidence_0.190_Length_814	TATTGCCATGGATGAGGGGG	GCACAGCTAGGTTTCTCGGT
Locus_1001_Transcript_2/10_Confidence_0.119_Length_851	TGGAAGATTGGTGATGAAAGTGA	GCACAGCTAGGTTTCTCGGT
Locus_1001_Transcript_2/10_Confidence_0.119_Length_851	TCTCACACTTTCCTTCTTCCTCT	CCGTGATGGTTTTCAGGACC
Locus_10024_Transcript_1/1_Confidence_0.000_Length_104	GAGGGCGTTAGGGTTCTGAG	CGGAAACTTGCGCGTGATTA
Locus_10026_Transcript_1/2_Confidence_0.632_Length_296	AGGTGGAAGCTTTTTATCTTTTGAGA	GCATTGAATTTCTGGGTTTTGCA
Locus_10026_Transcript_1/2_Confidence_0.632_Length_296	AGGTGGAAGCTTTTTATCTTTTGAGA	GCATTGAATTTCTGGGTTTTGCA
Locus_10026_Transcript_1/2_Confidence_0.632_Length_296	AGGTGGAAGCTTTTTATCTTTTGAGA	GCATTGAATTTCTGGGTTTTGCA
Locus_10026_Transcript_1/2_Confidence_0.632_Length_296	AGGTGGAAGCTTTTTATCTTTTGAGA	GCATTGAATTTCTGGGTTTTGCA
Locus_10026_Transcript_1/2_Confidence_0.632_Length_296	TTGGAGTTGCATGTGCGAGA	GCATTGAATTTCTGGGTTTTGCA
Locus_10026_Transcript_1/2_Confidence_0.632_Length_296	AGGTGGAAGCTTTTTATCTTTTGAGA	TGCAGTACTCTTAACCTGCACC
Locus_10026_Transcript_1/2_Confidence_0.632_Length_296	AGGTGGAAGCTTTTTATCTTTTGAGA	TGCAGTACTCTTAACCTGCACC
Locus_10026_Transcript_2/2_Confidence_0.895_Length_340	AGGTGGAAGCTTTTTATCTTTTGAGA	GCATTGAATTTCTGGGTTTTGCA
Locus_10026_Transcript_2/2_Confidence_0.895_Length_340	CCATTATGGAAATGTTTTGTTGCGT	GCATTGAATTTCTGGGTTTTGCA

Table 4. Gene ontology results of assembled data.

Total number of sequences for GO annotation	88,251
Total number of sequences in molecular function class	35,774
Total number of sequences in biological process class	45,099
Total number of sequences in cellular component class	7378

Table 5. Detailed GO results for the lentil consensus sequences for the molecular function classification (representative table).

GO ID	Term	#Seqs
GO:0003674	molecular_function	22,844
GO:0005488	binding	15,749
GO:0003824	catalytic activity	15,399
GO:0005215	transporter activity	1805
GO:0060089	molecular transducer activity	992
GO:0005198	structural molecule activity	695
GO:0009055	electron carrier activity	484
GO:0030234	enzyme regulator activity	283
GO:0030528	transcription regulator activity	190
GO:0016209	antioxidant activity	177
GO:0000166	nucleotide binding	5774
GO:0016740	transferase activity	5705
GO:0016787	hydrolase activity	5663
GO:0043167	ion binding	4602
GO:0003676	nucleic acid binding	3809
GO:0016491	oxidoreductase activity	3573
GO:0005515	protein binding	2023
GO:0022857	transmembrane transporter activity	1344
GO:0022892	substrate-specific transporter activity	1238
GO:0004871	signal transducer activity	992
GO:0048037	cofactor binding	983
GO:0016874	ligase activity	952
GO:0046906	tetrapyrrole binding	838
GO:0016829	lyase activity	699
GO:0003735	structural constituent of ribosome	525
GO:0016853	isomerase activity	490
GO:0003700	sequence-specific DNA binding transcription factor activity	381
GO:0030246	carbohydrate binding	314
GO:0019842	vitamin binding	292
GO:0051540	metal cluster binding	241
GO:0003682	chromatin binding	138
GO:0060589	nucleoside-triphosphatase regulator activity	137
GO:0031406	carboxylic acid binding	135
GO:0004601	peroxidase activity	129
GO:0042910	xenobiotic transporter activity	94
GO:0004857	enzyme inhibitor activity	88
GO:0008289	lipid binding	84
GO:0043176	amine binding	84
GO:0000156	two-component response regulator activity	75
GO:0008047	enzyme activator activity	72
GO:0003712	transcription cofactor activity	59
GO:0051184	cofactor transporter activity	56
GO:0043169	cation binding	4599
GO:0017076	purine nucleotide binding	4492
GO:0032553	ribonucleotide binding	4483
GO:0016772	transferase activity, transferring phosphorus-containing groups	3089
GO:0016817	hydrolase activity, acting on acid anhydrides	1905
GO:0003677	DNA binding	1721
GO:0016788	hydrolase activity, acting on ester bonds	1395

Table 6. Detailed GO results for the lentil consensus sequences for the biological process classification (representative table).

GO ID	Term	#Seqs
GO:0008150	biological_process	19,215
GO:0008152	metabolic process	15,613
GO:0009987	cellular process	14,700
GO:0051179	localization	3052
GO:0065007	biological regulation	2830
GO:0050896	response to stimulus	2685
GO:0016043	cellular component organization	1435
GO:0023052	signaling	1049
GO:0032501	multicellular organismal process	762
GO:0032502	developmental process	718
GO:0044085	cellular component biogenesis	669
GO:0016265	death	297
GO:0000003	reproduction	279
GO:0071554	cell wall organization or biogenesis	264
GO:0051704	multiorganism process	173
GO:0002376	immune system process	165
GO:0040007	growth	138
GO:0008283	cell proliferation	119
GO:0040011	locomotion	79
GO:0016032	viral reproduction	72
GO:0044237	cellular metabolic process	11,555
GO:0044238	primary metabolic process	11,268
GO:0043170	macromolecule metabolic process	7879
GO:0006807	nitrogen compound metabolic process	4996
GO:0009058	biosynthetic process	4943
GO:0044281	small molecule metabolic process	3502
GO:0055114	oxidation-reduction process	3034
GO:0051234	establishment of localization	2968
GO:0050789	regulation of biological process	2671
GO:0009056	catabolic process	1733
GO:0051716	cellular response to stimulus	1509
GO:0006950	response to stress	1388
GO:0006996	organelle organization	921
GO:0033036	macromolecule localization	783
GO:0042221	response to chemical stimulus	691
GO:0051641	cellular localization	638
GO:0007275	multicellular organismal development	625
GO:0065008	regulation of biological quality	516
GO:0048856	anatomical structure development	512
GO:0043933	macromolecular complex subunit organization	476
GO:0022607	cellular component assembly	459
GO:0034621	cellular macromolecular complex subunit organization	423
GO:0007049	cell cycle	421
GO:0065009	regulation of molecular function	321
GO:0009719	response to endogenous stimulus	308
GO:0008219	cell death	296
GO:0007017	microtubule-based process	284
GO:0048869	cellular developmental process	266
GO:0022613	ribonucleoprotein complex biogenesis	256

Table 7. Detailed GO results for the lentil consensus sequences for the cellular component classification (representative table).

GO ID	Term	#Seqs
GO:0005575	cellular_component	19,240
GO:0005623	cell	19,181
GO:0044422	organelle part	3374
GO:0043234	protein complex	2317
GO:0043228	nonmembrane-bounded organelle	1687
GO:0005622	intracellular	16,585
GO:0044425	membrane part	3395
GO:0008287	protein serine/threonine phosphatase complex	131
GO:0005737	cytoplasm	13,756
GO:0030529	ribonucleoprotein complex	908
GO:0044436	thylakoid part	683
GO:0034357	photosynthetic membrane	673
GO:0009521	photosystem	565
GO:0019866	organelle inner membrane	261
GO:0005789	endoplasmic reticulum membrane	217
GO:0030117	membrane coat	138
GO:0000151	ubiquitin ligase complex	104
GO:0016469	proton-transporting two-sector ATPase complex	83
GO:0031968	organelle outer membrane	55
GO:0030119	AP-type membrane coat adaptor complex	52
GO:0043231	intracellular membrane-bounded organelle	14,002
GO:0044428	nuclear part	906
GO:0005740	mitochondrial envelope	302
GO:0031966	mitochondrial membrane	296
GO:0044451	nucleoplasm part	245
GO:0005743	mitochondrial inner membrane	238
GO:0044454	nuclear chromosome part	86
GO:0030118	clathrin coat	79
GO:0031981	nuclear lumen	668
GO:0009535	chloroplast thylakoid membrane	629
GO:0009523	photosystem II	550
GO:0009522	photosystem I	505
GO:0005777	peroxisome	127
GO:0042579	microbody	127
GO:0000228	nuclear chromosome	105
GO:0071013	catalytic step 2 spliceosome	97
GO:0031969	chloroplast membrane	96
GO:0009941	chloroplast envelope	96
GO:0005773	vacuole	94
GO:0005798	Golgi-associated vesicle	76
GO:0015934	large ribosomal subunit	71
GO:0005819	spindle	63
GO:0009570	chloroplast stroma	103
GO:0015629	actin cytoskeleton	60

process (33%), localization (7%), biological regulation (6%), response to stimulus (6%), and cellular component organization (3%) categories. In the cellular component classification, which included 7378 sequences, the observed categories were as follows: organelle parts (46%), protein complexes (31%) and nonmembrane-bounded organelles (23%). Graphs of the functional classifications are provided in Figures 3–5.

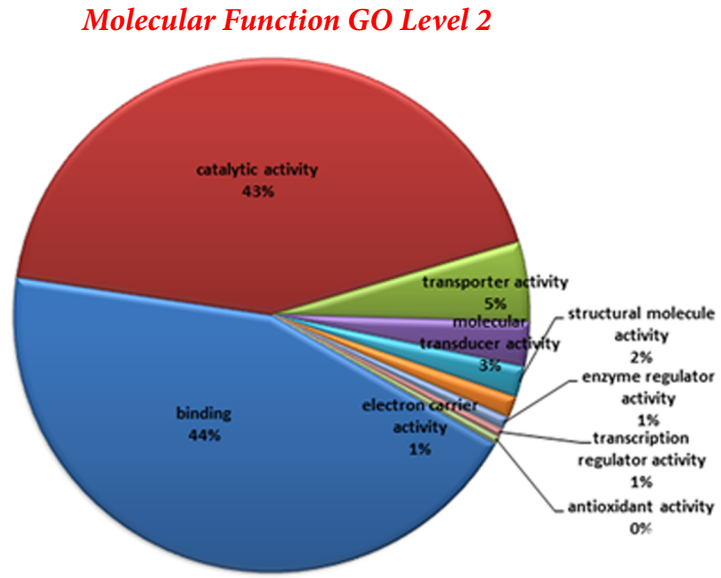

Figure 3. Pie-chart representation of the GO annotation results for the lentil consensus sequences for the molecular function classification. Total number of sequences is 35,774.

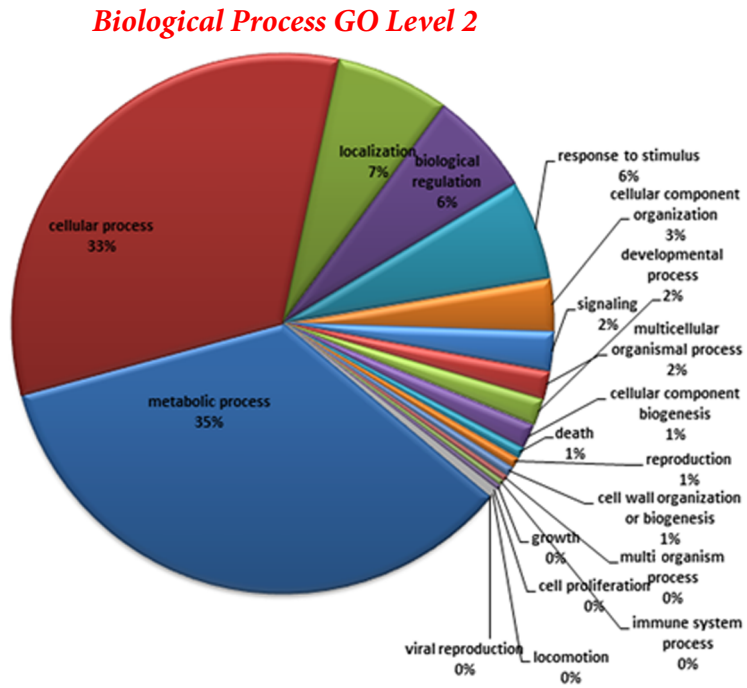

Figure 4. Pie-chart representation of the GO annotation results for the lentil consensus sequences for the biological process classification. The total number of sequences is 45,099.

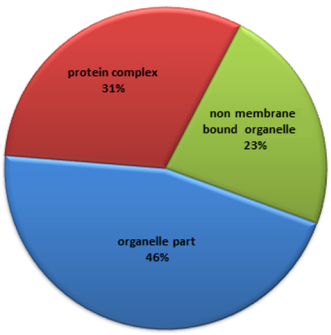

Figure 5. Pie-chart representation of the GO annotation results for the lentil consensus sequences for the cellular component classification. The total number of sequences is 7378.

3.4. Metabolic pathway analysis

A pathway-based analysis run using the KEGG database allowed us to understand the biological functions and interactions of the identified genes (Liu et al., 2013). According to the KEGG results, 2797 sequences were assigned to 144 KEGG pathways associated with metabolic processes (e.g., D-alanine metabolism, lysine degradation, and carotenoid biosynthesis). Detailed information is provided representatively in Table 8.

3.5. Genetic map construction

A genetic map comprising SNP, SSR, and ISSR markers was constructed using a LOD score of 3.00 and Kosambi mapping function with maximum recombination value of 0.45. Totally, 388 markers cover the genome of 432.8 cM with average marker density of 1 marker per 1.11 cM. The length of linkage groups (LGs) varied from 15.7 cM (LG3) to 106.1 cM (LG7) (Figure 6) LG3 is the shortest linkage group with 5 markers and LG7 is the longest with 106 markers. Totals of 376 SNP, 3 SSR (SSR19, SSR33, SSR562), and 9 ISSR (UBC318_1, UBC721, UBC79_2, UBC98, UBC840_2, UBC502_1, UBC808_1, UBC807_10, UBC807_6) markers could be mapped on the map. The LG group characteristics are presented in Table 9.

Table 8. Detailed KEGG results, enzymes represented in the transcriptome (representative table).

Enzyme	Ezyme Id	Enzyme	Ezyme Id
isopenicillin-N epimerase	ec:5.1.1.17	zeaxanthin epoxidase	ec:1.14.13.90
D-alanine---D-alanine ligase	ec:6.3.2.4	9-cis-epoxycarotenoid dioxygenase	ec:1.13.11.51
D-amino-acid transaminase	ec:2.6.1.21	phytoene synthase	ec:2.5.1.32
saccharopine dehydrogenase (NADP+, L-glutamate-forming)	ec:1.5.1.10	ent-copalyl diphosphate synthase	ec:5.5.1.13
enoyl-CoA hydratase	ec:4.2.1.17	10-deacetylbaccatin III 10-O-acetyltransferase	ec:2.3.1.167
histone-lysine N-methyltransferase	ec:2.1.1.43	taxadien-5alpha-ol O-acetyltransferase	ec:2.3.1.162
dihydrolipoyllysine-residue succinyltransferase	ec:2.3.1.61	gibberellin 3beta-dioxygenase	ec:1.14.11.15
aldehyde dehydrogenase (NAD+)	ec:1.2.1.3	gibberellin 2beta-dioxygenase	ec:1.14.11.13
acetyl-CoA C-acetyltransferase	ec:2.3.1.9	ent-kaurene oxidase	ec:1.14.13.78
3-hydroxyacyl-CoA dehydrogenase	ec:1.1.1.35	taxane 13alpha-hydroxylase	ec:1.14.13.77
saccharopine dehydrogenase (NADP+, L-lysine-forming)	ec:1.5.1.8	taxane 10beta-hydroxylase	ec:1.14.13.76
oxoglutarate dehydrogenase (succinyl-transferring)	ec:1.2.4.2	enoyl-CoA hydratase	ec:4.2.1.17
L-aminoadipate-semialdehyde dehydrogenase	ec:1.2.1.31	aldehyde dehydrogenase (NAD+)	ec:1.2.1.3
glutaryl-CoA dehydrogenase	ec:1.3.99.7	(S)-limonene 6-monooxygenase	ec:1.14.13.48
D-amino-acid transaminase	ec:2.6.1.21	(-)-endo-fenchol synthase	ec:4.2.3.10
D-amino-acid transaminase	ec:2.6.1.21	(S)-limonene 6-monooxygenase	ec:1.14.13.48
glutamate dehydrogenase [NAD(P)+]	ec:1.4.1.3	secologanin synthase	ec:1.3.3.9
cytokinin dehydrogenase	ec:1.5.99.12	(+)-neomenthol dehydrogenase	ec:1.1.1.208
capsanthin/capsorubin synthase	ec:5.3.99.8	adenosylmethionine decarboxylase	ec:4.1.1.50
abscisic-aldehyde oxidase	ec:1.2.3.14	S-methyl-5-thioribose-1-phosphate isomerase	ec:5.3.1.23
violaxanthin de-epoxidase	ec:1.10.99.3	homoserine dehydrogenase	ec:1.1.1.3
carotene 7,8-desaturase	ec:1.14.99.30	acireductone synthase	ec:3.1.3.77

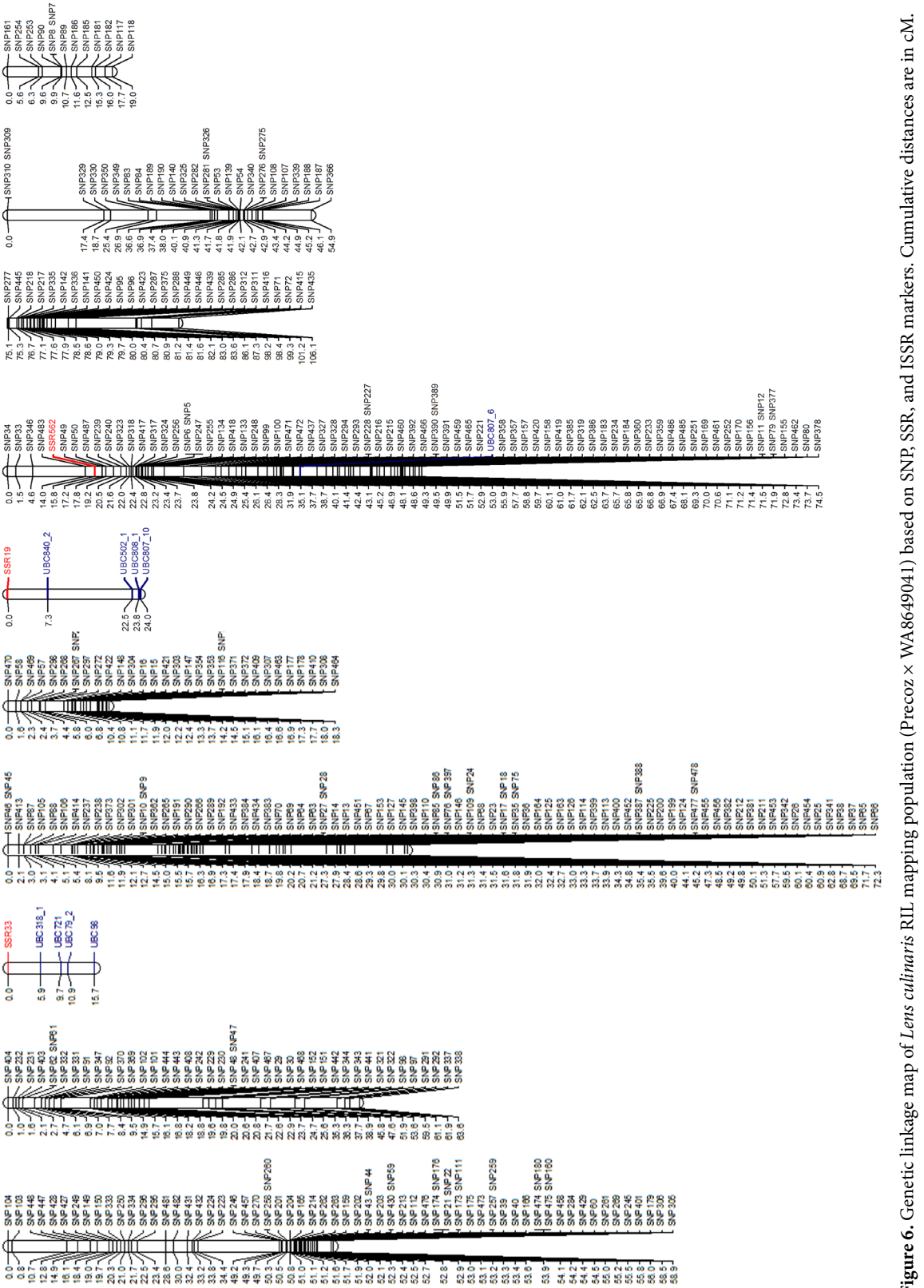

Figure 6. Genetic linkage map of *Lens culinaris* RIL mapping population (Precoz × WA8649041) based on SNP, SSR, and ISSR markers. Cumulative distances are in cM.

Table 9. Characteristics of the genetic linkage map of lentil.

Linkage group (LG)	Length of LG (cM)	Number of markers	Average distance between markers (cM)
LG1	58.9	69 (SNP)	0.85
LG2	63.6	43 (SNP)	1.47
LG3	15.7	5 (1 SSR, 4 ISSR)	3.14
LG4	72.3	88 (SNP)	0.82
LG5	18.3	32 (SNP)	0.57
LG6	24	5 (1 SSR, 4 ISSR)	4.80
LG7	106.1	106 (104 SNP, 1 SSR, 1 ISSR)	1.00
LG8	54.9	27	2.03
LG9	19	13	1.46
Total	432.8	388	1.11

4. Discussion

In the current study, de novo discovery of 50,960 SNPs based on Illumina transcriptome sequencing of 2 cultivars (Precoz and WA8649041) was performed using the Illumina platform. The Illumina platform was chosen from the commercially available NGS platforms because of its ease of use and the superior data quality, high throughput, and appropriate read lengths that can be generated for de novo transcriptome assembly (Varshney et al., 2009; Kaya et al., 2013).

SNPs were chosen because they represent natural sequence variations in genomes (Xu et al., 2009) in both intragenic and intergenic regions (Shirasawa et al., 2010) and can be used as genetic markers to construct high-density genetic maps (Brookes, 1999). With the availability of high-throughput analysis technologies, the abundance, stability, and heredity of SNPs have led them to be used as important markers, replacing traditional molecular markers such as AFLPs, RFLPs, and SSRs for association and genome mapping studies (Xu et al., 2009). While the frequency of SNPs is 1 SNP per 1000 bp of contiguous sequence for cultivated soybean (Zhu et al., 2003; Choi et al., 2007), it is 1 SNP per 425 bp for *Glycine soja*, which is a wild ancestor of soybean (Hyten et al., 2006).

It is possible to acquire a wealth of sequence information in nonmodel organisms using NGS technologies (Vera et al., 2008; Hale et al., 2009; Wheat, 2010; Der et al., 2011; Seeb et al., 2011). Marker development is affected by contig length, and the obtained contig length depends on the sequencing platform (Helyar et al., 2012). An increased contig length increases the coverage depth and the number of reads assembled (Lai et al., 2012). According to the selected sequencing platform, longer or shorter reads are obtained. Without a reference genome, estimating the number of genes sequenced and assessing the precision of the contig assembly are challenging (Parchman et al., 2010; Helyar et al., 2012). If misassemblies of sequences occur because of homologous or paralogous genes, it cannot be verified (Helyar et al., 2012). To obtain deep assemblies of redundant contigs, which is necessary for SNP discovery, a genome reduction step is required for nonmodel organisms (Slate et al., 2009). Transcriptome sequencing is one way to achieve genome reduction for nonmodel organisms, despite the challenges due to differential gene expression among individuals (Seeb et al., 2011).

Genome sequencing and high-throughput methods produce large amounts of data, which can be used to identify gene modulatory networks (Li et al., 2005). Combining complex trait analysis with transcriptome analysis is an important step in molecular genetic studies (Li et al., 2005). Transcriptome SNPs are associated with genes or functional regions of the genome (Xu et al., 2012; Kaya et al., 2013), which is why a number of researchers have focused on combining high-throughput transcriptome data and quantitative trait loci (QTL) detection to understand biological pathways related to complex traits (Mootha et al., 2003; Kirst et al., 2004; Schadt et al., 2005). This new approach is referred to as 'genetical genomics' or 'integrative genomics' and involves the use of gene and genetic marker expression levels to define genomic regions referred to as expression quantitative trait loci (eQTL) (Mignon et al., 2009).

4.1. Sequence assembly and SNP detection

Transcriptome sequencing is useful for generating abundant sequence information, such as identifying SNPs, and understanding the biological processes of cells

(Birol et al., 2009). For some species for which a reference genome is not available, NGS technology is used for draft sequencing (Varshney et al., 2009). In this study, due to the absence of available reference sequences, de novo assembly was performed for cDNA from lentil cultivars (Feldmeyer et al., 2011; Lai et al., 2012). The assembly software ABySS was used for this purpose, which can assemble billions of short reads. The superiority of this software lies in a distributed representation of the de Bruijn graph, which allows parallel computation of the assembly algorithm across a computer network (Simpson et al., 2009).

Sequencing of cDNA libraries generated a total of 113,126,056 raw sequence reads, which is a greater number than that obtained in a previous study by Sharpe et al. (2013) (1.03×10^6 reads for a genotype), who examined 9 *L. culinaris* and 2 *L. ervoides* genotypes to discover SNPs. In another study, 6 genotypes were examined to discover SSRs in lentil and 1.38×10^6 reads were generated (Kaur et al., 2011). In olive, Kaya et al. (2013) performed transcriptome sequencing of cDNA from 5 distinct olive genotypes. The resulting 126,542,413 sequencing reads in this study were assembled into 22,052 contigs and identified 2987 SNP markers. Similar to our study, transcriptome sequencing using the Illumina platform in olive was successfully performed and the authors obtained high-quality reads for SNP discovery. Different numbers of reads have been obtained through transcriptome sequencing in peanut (Wu et al., 2013), coconut (Fan et al., 2013), black pepper (Joy et al., 2013), sweet potato (Wang et al., 2010), ramie (Liu et al., 2013), and olive (Kaya et al., 2013). This finding could be explained by the different expression levels present in the examined tissues collected from different growing stages (Wu et al., 2013). Verma et al. (2013) examined only 1 genotype (Precoz) and generated 119,855,798 raw sequence reads and 91,282,242 reads after trimming.

The complete read dataset acquired in the present study was assembled into 97,528 contigs. Sharpe et al. (2013) identified 27,921 contigs, which is fewer than were obtained in the present study. While the length of the contigs in the present study varied from 100 to 19,077 bp, with an N50 of 1996 bp, the corresponding values reported by Kaur et al. (2011) were 114 to 6479 bp, with an average of 717 bp. We obtained 50,960 putative SNPs from the 2 parents, which is greater than the maximum number of SNPs acquired by Sharpe et al. (2013), who identified fewer SNPs for the 9 *L. culinaris* and 2 *L. ervoides* genotypes they examined than were found in the current study, and also greater than the 38,587 SNPs detected for melon with Sanger and 454 sequencers (Blanca et al., 2011). This higher number of SNPs in our study is mainly due to the genomic variation between parent genotypes, as more diversity between genotypes sequenced may lead to more variation to discover SNPs.

Comparative mapping analyses have demonstrated a direct relationship between the chromosomes of *Medicago truncatula* and *L. culinaris* based on defining complete homology (Phan et al., 2007). Following the functional annotation of the obtained reads, it was observed that many transcripts were identical to proteins found in *M. truncatula*. Some of these proteins, such as subtilis-like serine protease, protein abci7, lipoxygenase, somatic embryogenesis receptor kinase, hypothetical protein MTR_7g117150, and spermatogenesis-associated protein, were found to have orthologs in the *M. truncatula* genome. Similar results were obtained by Kaur et al. (2011) and Sharpe et al. (2013).

4.2. SNP validation

A total of 500 SNP primers were selected for validation, of which 468 produced amplification products; among those, 420 primers produced polymorphic markers and finally 377 SNPs were be able to be mapped in the lentil genome, which means that 75% percent of total SNPs joined into the linkage map. Thirty-two primers (6%) failed to result in amplification. About 93% of the SNPs showed amplification, and 48 primers were monomorphic (9%). Genotyping failures can be caused by the presence of undetected introns or false positives (Wang et al., 2008). There are 2 potential reasons why the SNP primers might have shown monomorphism: either false positive predictions were generated, or the SNPs were not present in the tested parent genotypes (Hubert et al., 2010; Helyar et al., 2012). Different percentages of polymorphisms have been obtained among the SNPs discovered in many studies (Hyten et al., 2010a, 2010b; Shirasawa et al., 2010; Helyar et al., 2012; Kaya et al., 2013; Loridon et al., 2013; Sharpe et al., 2013). Hyten et al. (2010a) reported that, among 3072 SNPs used in genotyping, 9% failed to produce a good assay on RIL mapping populations. In another study (Shirasawa et al., 2010), out of the 1536 SNPs genotyped in tomato, 13% failed to be genotyped. In a SNP discovery study on Atlantic herring (Helyar et al., 2012), from a panel of 1536 SNPs that were genotyped, 19% failed to amplify. Sharpe et al. (2013) reported that of the 1052 SNP assays, 154 (14%) completely failed. This situation shows that this is a common problem in generating SNPs from transcriptomes. For SNP primers that do not show amplification, the reason may be errors occurring in massive sequencing technologies, such as sequencing errors, PCR artifacts, and errors in the mapping of short reads to the reference sequence (Blanca et al., 2012).

4.3. Functional annotation

By searching the annotated sequences and the associated GO classifications, we evaluated the completeness of our transcriptome results. To generate the maximum amount of information from the lentil transcriptome sequences, tissue samples from different parts of lentil plants (roots,

shoots, leaves, branches, and flowers) were harvested for RNA isolation due to the different expression levels among tissues (Sharpe et al., 2013). A large number of contigs matched the sequences of known proteins (especially *M. truncatula* proteins). The contigs without BLAST hits corresponded to 3' or 5' untranslated regions, noncoding RNAs, and short sequences that did not contain known protein domains and might have been lentil-specific genes (Wang et al., 2010). A large number of sequences and wide coverage can generate satisfactory transcriptome sequence information (David et al., 2010). The present results support the earlier reports indicating that Illumina sequencing is an inexpensive, efficient, and reliable tool for transcriptome characterization in nonmodel species (Wang et al., 2010).

4.4. Genetic linkage map

The constructed map consists of 7 major groups, as Sharpe et al. (2013) found. Two additional minor linkage groups with 5 markers each were also constructed. These 7 linkage groups could be represented as 7 chromosomes of lentil. It was found that the average distance between adjacent markers was 1.11 cM, which is close to the number that Sharpe et al. (2013) found (1.53 cM). Even though the length of the constructed map was shorter than the map of Sharpe et al. (2013), if the distance between 2 markers is compared, it can be concluded that the markers in the current study are closer to each other than in the map constructed by Sharpe et al. (2013). The current map covers 432.8 cM of the lentil genome, which was approximately half of the length of Sharpe et al. (2013). Some gaps were detected at LG1, LG2, LG6, LG7, and LG8. This could be due to low polymorphism in that region and it also could be possible that the markers were generated from the genic region; therefore, the gap regions could be intergenic regions (Folta et al., 2009; Sun et al., 2013). SSR19 and SSR33 are mapped respectively on LG6 and LG3, as they were mapped by Hamwieh et al. (2005). The SSR562 marker (Rajesh et al., 2008) was mapped on LG7, which was previously mapped on LG10 (Varlı, 2009).

In conclusion, in this study, the lentil transcriptome was characterized via de novo sequencing using the Illumina platform, without the presence of a reference genome. The comprehensive sequence information and large number of SNPs obtained in this study can potentially be used for genetic characterization, high-density linkage map analyses, map-based cloning, comparative genomics research, detection of genetic variation among landraces and individuals in a population, genome-wide analyses of molecular variation, and genome-based QTL analysis. They can also be employed for association mapping studies in natural lentil populations. The availability of annotated transcriptome sequence information will help to accelerate the isolation and characterization of genes in different pathways and might be useful in molecular genetic approaches for lentil breeding. The great amount of data generated in this study will be useful for performing genetic analysis in *Lens culinaris* and provides an additional resource to Sharpe et al.'s (2013) data. The SNP information generated in this study could also be used for designing SNP arrays for high-throughput genome-wide association studies (Xu et al., 2012). The large number of contigs obtained in the present study can further be used for the identification of transcripts from cDNA data for other organisms generated in an assembly step. The genetic map is an intraspecific gene-based map that consists of SNP, SSR, and ISSR markers. This SNP-based linkage map will be useful for marker-assisted selection in lentil breeding and for future mapping studies of other populations in lentil.

Acknowledgments

We acknowledge Jinsung Kim and Hanwool Lee from DNA Link Inc. (Seoul, South Korea) for their kind help in transcriptome sequencing and depositing the cDNA sequences into the NCBI. We also acknowledge Abdulkadir Aydoğan, who carried out the field experiment at the Ministry of Agriculture in Ankara. This manuscript was funded by the Technological and Scientific Research Council of Turkey with the project number of TÜBİTAK-110O361.

References

Anithakumari AM, Tang J, van Eck HJ, Visser RG, Leunissen JA, Vosman B, van der Linden CG (2010). A pipeline for high throughput detection and mapping of SNPs from EST databases. Mol Breeding 26: 65–75.

Arumuganathan K, Earle ED (1991). Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9: 208–218.

Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R (2009). *De novo* transcriptome assembly with ABySS. Bioinformatics 25: 2872–2877.

Blanca JM, Cañizares J, Ziarsolo P, Esteras C, Mir G, Nuez F, Garcia-Mas J, Pico MB (2011). Melon transcriptome characterization. SSRs and SNPs discovery for high throughput genotyping across the species. Plant Genome 4: 118–131.

Blanca J, Esteras C, Ziarsolo P, Perez D, Fernandez-Pedrosa V, Collado C, de Pablos RR, Ballester A, Roig C, Cnizares J et al. (2012). Transcriptome sequencing for SNP discovery across *Cucumis melo*. BMC Genomics 13: 280.

Brookes AJ (1999). The essence of SNPs. Gene 234: 177–186.

Callaway JC (2004). Hempseed as a nutritional resource: an overview. Euphytica 140: 65–72.

Choi IY, Hyten DL, Matukumalli LK, Song Q, Chaky JM, Quigley CV, Chase K, Lark KG, Reiter RS, Yoon MS (2007). A soybean transcript map: gene distribution, haplotype and single nucleotide polymorphism analysis. Genetics 176: 685–696.

Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S et al. (2009). Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10: 582.

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

Cortés AJ, Chavarro MC, Blair MW (2011). SNP marker diversity in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 123: 827–845.

Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, Quail MA, He M, Assefa S, Bahler J et al. (2009). A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Res 37: e148.

David JP, Coissac E, Melodelima C, Poupardin R, Riaz MA, Proust AC, Reynaud S (2010). Transcriptome response to pollutants and insecticides in the dengue vector *Aedes aegypti* using next-generation sequencing technology. BMC Genomics 11: 216.

Der JP, Barker MS, Wickett NJ, dePamphilis CW, Wolf PG (2011). *De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. BMC Genomics 12: 99.

Doyle JJ, Doyle JL (1990). Isolation of plant DNA from fresh tissue. Focus 12: 13–15.

Duran Y, Fratini R, Garcia P, Perez de la Vega M (2003). An intersubspecific genetic map of *Lens*. Theor Appl Genet 108: 1265–1273.

Erskine W (1997). Lessons for breeders from land races of lentil. Euphytica 93: 107–112.

Erskine W, Chandra S, Chaudhry M, Malik IA, Sarker A, Sharma B, Tufail M, Tyagi MC (1998). A bottleneck in lentil: widening its genetic base in South Asia. Euphytica 101: 207–211.

Eujayl I, Baum M, Powell W, Erskine W, Pehu E (1998). A genetic linkage map of lentil (*Lens* sp.) based on RAPD and AFLP markers using recombinant inbred lines sp.) based on RAPD and AFLP markers. Theor Appl Genet 97: 83–89.

Fan H, Xiao Y, Yang Y, Xia W, Mason AS, Xia Z, Qiao F, Zhao S, Tang H (2013). RNA-Seq analysis of *Cocos nucifera*: transcriptome sequencing and *de novo* assembly for subsequent functional genomics approaches. PLoS One 8: e59997.

Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M (2011). Short read Illumina data for the *de novo* assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 12: 317.

Fikiru E, Tesfaye K, Bekele E (2007). Genetic diversity and population structure of Ethiopian lentil (*Lens culinaris* Medikus) landraces as revealed by ISSR marker. Afr J Biotechnol 6: 1460–1468.

Folta KM, Gardiner SE (2009). Genetics and Genomics of Rosaceae. Vol. 6. New York, NY, USA: Springer Science and Business Media.

Gupta M, Verma B, Kumar N, Chahota RK, Rathour R, Sharma SK, Bhatia S, Sharma TR (2012). Construction of intersubspecific molecular genetic map of lentil based on ISSR, RAPD and SSR markers. Journal of Genetics 91: 279–287.

Hale MC, McCormick CR, Jackson JR, DeWoody JA (2009). Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. BMC Genomics 10: 203.

Hamwieh A, Udupa SM, Choumane W, Sarker A, Dreyer F, Jung C, Baum M (2005). A genetic linkage map of *Lens* sp. based on microsatellite and AFLP markers and the localization of fusarium vascular wilt resistance. Theor Appl Genet 110: 669–677.

Havey MH, Muehlbauer FJ (1989). Linkages between restriction fragment length, isozyme, and morphological markers in lentil. Theor Appl Genet 77: 395–401.

Helyar SJ, Limborg MT, Bekkevold D, Babbucci M, Van-Houdt JKJ, Maes GE, Bargelloni L, Nielsen RO, Bendixen C, Taylor MI et al. (2012). SNP discovery using next generation transcriptomic sequencing in Atlantic herring (*Clupea harengus*). PLoS One 7: e42089.

Hubert S, Higgins B, Borza T, Bowman S (2010). Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). BMC Genomics 11: 191.

Hyten DL, Choi IY, Song Q, Specht JE, Carter TE, Shoemaker RC, Hwang EY, Matukumalli LK, Cregan PB (2010a). A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci 50: 960–968.

Hyten DL, Song Q, Fickus EW, Quigley CV, Lim JS, Choi IY, Hwang EY, Pastor-Corrales M, Cregan PB (2010b). High-throughput SNP discovery and assay development in common bean. BMC Genomics 11: 475.

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB (2006). Impacts of genetic bottlenecks on soybean genome diversity. P Natl Acad Sci USA 103: 16666–16671.

Joy N, Asha S, Mallika V, Soniya EV (2013). *De novo* transcriptome sequencing reveals a considerable bias in the incidence of simple sequence repeats towards the downstream of 'Pre-miRNAs' of black pepper. PLoS One 8: e56694.

Kanehisa M, Goto S (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

Karim Mojein H, Alizadeh HM, Majnoon Hoseini N, Payghmbari SA (2003). Effect of herbicides and hand weeding in control of weed in winter and spring sown lentil (*Lens culinaris* L.). Iranian Journal of Crop Sciences 6: 68–79.

Kaur S, Cogan NO, Pembleton LW, Shinozuka M, Savin KW, Materne M, Forster JW (2011). Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. BMC Genomics 12: 265.

Kaya HB, Cetin O, Kaya H, Sahin M, Sefer F, Kahraman A, Tanyolac B (2013). SNP Discovery by Illumina-based transcriptome sequencing of the olive and the genetic characterization of Turkish olive genotypes revealed by AFLP, SSR and SNP markers. PLoS One 8: e73674.

Kirst M, Myburg AA, De Leon JP, Kirst ME, Scott J, Sederoff R (2004). Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiol 135: 2368–2378.

Kosambi DD (1944). The estimation of map distance from recombination values. Ann Eugen 12: 172–175.

Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U et al. (2012). Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol J 10: 743–749.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754–1760.

Li H, Lu L, Manly KF, Chesler EJ, Bao L, Wang J, Zhou M, Williams RW, Cui Y (2005). Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Genet 14: 1119–1125.

Liu T, Zhu S, Tang Q, Chen P, Yu Y, Tang S (2013). *De novo* assembly and characterization of transcriptome using Illumina paired-end sequencing and identification of CesA gene in ramie (*Boehmeria nivea L. Gaud*). BMC Genomics 14: 125.

Loridon K, Burgarella C, Chantret N, Martins F, Gouzy J, Prosperi JM, Ronfort J (2013). Single-nucleotide polymorphism discovery and diversity in the model legume *Medicago truncatula*. Mol Ecol Resourc 13: 84–95.

Maccaferri M, Sanguineti MC, Corneti S, Ortega JLA, Salem MB, Bort J, DeAmbrogio E, del Moral LFG, Demontis A, El-Ahmed A et al. (2008). Quantitative trait loci for grain yield and adaptation of durum wheat (*Triticum durum* Desf.) across a wide range of water availability. Genetics 178: 489–511.

Martin M (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10–12.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.

Mignon GL, Désert C, Pitel F, Leroux S, Demeure O, Guernec G, Abasht B, Douaire M, Roy PL, Lagarrigue S (2009). Using transcriptome profiling to characterize QTL regions on chicken chromosome 5. BMC Genomics 10: 575.

Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F et al. (2003). Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. P Natl Acad Sci USA 100: 605–610.

Ooijen V (2006). JoinMap4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations. Wageningen, the Netherlands: Kyazma B.V.

Parchman T, Geist K, Grahnen J, Benkman C, Buerkle C (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11: 180.

Phan HTT, Ellwood SR, Hane JK, Ford R, Materne M, Oliver RP (2007). Extensive macrosynteny between *Medicago truncatula* and *Lens culinaris* ssp. *culinaris*. Theor Appl Genet 114: 549–558.

Rajesh PN, White D, Saha G, Chen W, Muehlbauer F (2008). Development and genetic analysis of SSR markers in lentil. In: Plant & Animal Genomes XVI Conference, 12-16 January 2008, San Diego, CA, USA.

Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ et al. (2010). *De novo* assembly and analysis of RNA-seq data. Nat Methods 7: 909–912.

Rozen S, Skaletsky H (1999). Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology 132: 365–386.

Rubeena FR, Taylor PWJ (2003). Construction of an intraspecific linkage map of lentil (*Lens culinaris* ssp. *culinaris*). Theor Appl Genet 107: 910–916.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37: 710–717.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012). Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–1092.

Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, Seeb LW (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol Ecol Resourc 11 (Suppl. 1): 1-8.

Sharpe AG, Ramsay L, Sanderson LA, Fedoruk MJ, Clarke WE, Li R, Kagale S, Vijayan P, Vandenberg A, Bett KE (2013). Ancient orphan crop joins modern era: gene-based SNP discovery and mapping in lentil. BMC Genomics 14: 192.

Shirasawa K, Isobe S, Hirakawa H, Asamizu E, Fukuoka H, Just D, Rothan C, Sasamoto S, Fujishiro T, Kishida Y et al. (2010). SNP discovery and linkage map construction in cultivated tomato. DNA Res 17: 381–391.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009). ABySS: A parallel assembler for short read sequence data. Genome Res 19: 1117–1123.

Slate J, Gratten J, Beraldi D (2009). Gene mapping in the wild with SNPs: guidelines and future directions. Genetica 136: 97–107.

Sun L, Yang W, Zhang Q, Cheng T, Pan H, Xu Z, Zhang J, Chen C (2013). Genome-wide characterization and linkage mapping of simple sequence repeats in mei (*Prunus mume* Sieb. et Zucc.). PLoS One 8: e59562.

Tadmor Y, Zamir D, Ladizinsky G (1987). Genetic mapping of an ancient translocation in the genus *Lens*. Theor Appl Genet 73: 883–892.

Tahir M, Simon CJ, Muehlbauer FJ (1993). Gene map of lentil: a review. LENS Newsletter 2: 3–10.

Tanyolac B, Ozatay S, Kahraman A, Muehlbauer F (2010). Linkage mapping of lentil (*Lens culinaris* L.) genome using recombinant inbred lines revealed by AFLP, ISSR, RAPD and some morphologic markers. Agricultural Biotechnology and Sustainable Development 2: 1–6.

Trick M, Long Y, Meng J, Bancroft I (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnol J 7: 334–346.

Tullu A, Tar'an B, Warkentin T, Vandenberg A (2008). Construction of an intraspecific linkage map and QTL analysis for earliness and plant height in lentil. Crop Sci 48: 2254–2264.

Varlı İ (2009). Use of SSR markers in construction of genetic linkage map of lentil. MSc, Harran University, Şanlıurfa, Turkey (in Turkish with English abstract).

Varshney RK, Nayak SN, May GD, Jackson SA (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27: 522–530.

Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17: 1636–1647.

Verma P, Shah N, Bhatia S (2013). Development of an expressed gene catalogue and molecular markers from the *de novo* assembly of short sequence reads of the lentil (*Lens culinaris* Medik.) transcriptome. Plant Biotechnol J 11: 894–905.

Vorrips RE (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. Journal Hered 93: 77–78.

Wang SL, Sha ZX, Sonstegard TS, Liu H, Xu P (2008). Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics 9: 45.

Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y (2010). *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). BMC Genomics 11: 726.

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

Wong PP (1980). Nitrate and carbohydrate effects on nodulation and nitrogen fixation (acetylene reduction) activity of lentil (*Lens esculenta* Moench). Plant Physiol 66: 78–81.

Wu N, Matand K, Wu H, Li B, Li Y, Zhang X, He Z, Qian J, Liu X, Conley S et al. (2013). *De novo* next-generation sequencing, assembling and annotation of *Arachis hypogaea* L. Spanish botanical type whole plant transcriptome. Theor Appl Genet 126: 1145–1149.

Xu J, Ji P, Zhao Z, Zhang Y, Feng J, Wang J, Li J, Zhang X, Zhao L, Liu G (2012). Genome-wide SNP discovery from transcriptome of four common carp strains. PLoS One 7: e48140.

Xu JY, Xu GB, Chen SL (2009). A new method for SNP discovery. Biotechniques 46: 201–208.

Yadav SS, McNeil DL, Stevenson PC (2007). Lentil: An Ancient Crop for Modern Times. Dordrecht, the Netherlands: Springer.

Zamir D, Ladizinsky G (1984). Genetics of allozyme variants and linkage groups in lentil. Euphytica 33: 329–336.

Zerbino DR, Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003). Single-nucleotide polymorphisms in soybean. Genetics 163: 1123–1134.