EXCLI Journal 2014;13:392-400 – ISSN 1611-2156
Received: September 13, 2013, accepted: September 26, 2013, published: April 07, 2014

Original article:

EST SEQUENCING AND GENE EXPRESSION PROFILING IN *SCUTELLARIA BAICALENSIS*

Nam Il Park[1,†], Ik Young Choi[2,†], Beom-Soon Choi[2], Young Seon Kim[3], Mi Young Lee[3], Sang Un Park[4*]

[1] Department of Plant Science, Gangneung-Wonju National University, 7 Jukheon-gil, Gangneung-si, Gangwon-do, 210-702, Korea
[2] National Instrumentation Center for Environmental Management, Seoul National University, 599 Gwanangno, Daehak-dong, Gwanak-gu, Seoul, 151-921, Korea
[3] KM-Based Herbal Drug Research Group, Korea Institute of Oriental Medicine, Daejeon, 305-811, Korea
[4] Department of Crop Science, Chungnam National University, 99 Daehak-ro, Yuseong-Gu, Daejeon, 305-764, Korea

[†] These authors contributed equally to this work.

***** Corresponding author: Sang Un Park, Department of Crop Science, Chungnam National University, 99 Daehak-ro, Yuseong-Gu, Daejeon, 305-764, Korea. Phone: +82-42-821-5730; Fax: +82-42-822-2631; e-mail: supark@cnu.ac.kr

ABSTRACT

Scutellaria baicalensis is an important medicinal plant, but few genomic resources are available for this species, as well as for other non-model plants. One of the major new directions in genome research is to discover the full spectrum of genes transcribed from the whole genome. Here, we report extensive transcriptome data of the early growth stage of *S. baicalensis*. This transcriptome consensus sequence was constructed by *de novo* assembly of shotgun sequencing data, obtained using multiple next-generation DNA sequencing (NGS) platforms (Roche/454 GS_FLX+ and Illumina/Solexa HiSeq2000). We show that this new approach to obtain extensive mRNA is an efficient strategy for genome-wide transcriptome analysis. We obtained 1,226,938 and 161,417,646 reads using the GS_FLX and the Illumina/Solexa HiSeq2000, respectively. *De novo* assembly of the high-quality GS_FLX and Illumina reads (95 % and 75 %) resulted in more than 82 Mb of mRNA consensus sequence, which we assembled into 51,188 contigs, with at least 500 bp per contig. Of these contigs, 39,581 contained known genes, as determined by BLASTX searches against non-redundant NCBI database. Of these, 20,498 different genes were expressed during the early growth stage of *S. baicalensis*. We have made the expressed sequences available on a public database. Our results demonstrate the utility of combining NGS technologies as a basis for the development of genomic tools in non-model, medicinal plant species. Knowledge of all described genes and quantitation of the expressed genes, including the transcription factors involved, will be useful in studies of the biology of *S. baicalensis* gene regulation.

Keywords: *de novo* assembly, expression profiling, next generation sequencing, *Scutellaria baicalensis*

EXCLI Journal 2014;13:392-400 – ISSN 1611-2156
Received: September 13, 2013, accepted: September 26, 2013, published: April 07, 2014

INTRODUCTION

Microarray hybridization systems have been used traditionally to profile the expression of known genes in a given organism or tissue (Watson et al., 1998). Total RNA sequencing, using a high-throughput DNA sequencing method, is another method that not only efficiently provides RNA sequence data but also profiles gene expression and identifies novel genes in non-model organisms (Parchman et al., 2010; Zeng et al., 2010; Logacheva et al., 2011). The ability to obtain whole genome transcriptome data efficiently, rapidly, and at a low cost has been facilitated by the use of next-generation DNA sequencing (NGS) technologies such as the Roche/454 GS_FLX+ and Illumina/Solexa HiSeq2000.

Although NGS platforms have been used to derive transcriptome data, many geneticists and biologists do not have extensive transcriptome sequence data available to perform the challenging *de novo* assembly of NGS data derived from multiple NGS platforms.

The pyrosequencing approach used by the Roche/454 GS_FLX+ is useful for *de novo* assembly to construct a reference gene expression profile in a novel genome (Ekblom et al., 2010; Kumar and Blaxter, 2010; Hsiao et al., 2011; Sloan et al., 2012). The long sequence reads obtained with the Roche/454 GS_FLX+ (average read-length, 600 bp) allows discovery of unigenes, while reducing incorrect assembly of homologous sequence regions. It is estimated that about 1 million reads (of approximately 600 Mb) per 1 PicoTiterPlate run on the Roche/454 GS_FLX+ is sufficient to provide 15× coverage of the whole transcriptome derived from a 40-Mb genome.

In contrast, the short read sequences of the Illumina/Solexa HiSeq2000 provide a cost-effective means to profile gene expression and allow mapping to known reference genes (Matsumura et al., 2010; Oshlack et al., 2010; Wickramasinghe et al., 2012) but are not suitable for discovering novel genes, because *de novo* assembly of these data is challenging. However, data from the HiSeq2000 should allow efficient extension of a reference RNA data set when using multiplatform sequence data, viz., the sub-data from the HiSeq2000 and the preliminary contig assembly data derived from the Roche/454 GS_FLX+ reads, in a hybrid *de novo* assembly approach. However, such a hybrid total mRNA assembly has not yet been reported for construction of a transcriptome reference sequence data set.

Scutellaria baicalensis, one of 50 fundamental herbs used in traditional Chinese medicine, is a species of flowering plant in the Lamiaceae family. This plant is used in traditional Chinese medicine to treat inflammation, respiratory tract infections, diarrhea, dysentery, liver disorders, hypertension, hemorrhaging, and insomnia (Nishikawa et al., 1999; Li et al., 2000). *S. baicalensis* is rich in flavones, a class of flavonoids produced by plants. Of these, **baicalin**, **baicalein**, and **wogonin** exhibit biological and pharmacological properties, including antioxidative and cancer prevention effects, and have been useful for treating and preventing coronary heart disease (Martens and Mithöfer, 2005). Correlating gene expression with flavone biosynthesis in *S. baicalensis* allowed engineering of a gene for increased production of flavones in a transgenic hairy root culture system (Xu et al., 2010; Park et al., 2011). Insufficient *S. baicalensis* sequencing data are available to study functional genomics in this plant. However, applying NGS technology to *S. baicalensis* may facilitate study of the functional genes important for metabolic engineering in this plant.

In this study, we performed a *de novo* assembly of multiplatform sequencing data, obtained using both the Roche/454 GS_FLX+ and Illumina/Solexa HiSeq2000 sequencing technologies and characterized the expression profiles of these genes to establish a novel transcriptome reference data set for *S. baicalensis*.

EXCLI Journal 2014;13:392-400 – ISSN 1611-2156
Received: September 13, 2013, accepted: September 26, 2013, published: April 07, 2014

MATERIALS AND METHODS

Plant material and RNA isolation

Sterilized *S. baicalensis* seeds were placed on agar-solidified MS culture medium (Murashige and Skoog, 1962). The seeds were germinated and grown at 25 °C in a growth chamber with a 16-h photoperiod for 3 months. The young plants were then used to study normal gene expression from the whole genome, in the early growth stage of this plant. Total RNA was isolated from pooled root, stem, and leaf organs using a total RNA isolation kit (GeneAll, cat. #305-101, Seoul, Korea).

454 GS_FLX and Illumina sequencing

Total mRNA was transcribed and shotgun sequenced using both the Roche/454 GS_FLX+ and Illumina/Solexa HiSeq2000 NGS systems. For the GS_FLX+ platform, mRNA was purified from total RNA using Sera-Mag Magnetic Oligo (dT) beads (Illumina cat. #RS-930-1001, San Diego, CA, USA), while cDNA synthesis, NGS library construction, and DNA sequencing were performed following the manufacturer's instructions (Roche Diagnostics, Mannheim, Germany). mRNA purification from total RNA, cDNA synthesis, library construction, and DNA sequencing were performed following the manufacturer's sequence protocol using the HiSeq2000 (Illumina).

Construction of a transcriptome for the reference sequence

De novo assembly of the pooled sequencing data from both NGS platforms was performed using Trinity software (http://TrinityRNASeq.sourceforge.net).

Prior to assembly, the raw data from both NGS sequencing platforms were filtered for a Phred quality score of at least 20 and for read-lengths of at least 200 bp for the GS_FLX+ data and 75 bp for the HiSeq2000 data. Therefore, only high quality sequencing data with at least QC20, i.e., 1 % of DNA sequencing error ratio on each base were used during assembly.

Gene annotation

A total of 51,188 contigs with a consensus sequence of at least 500 bp was used to predict genes with Augustus software (http://augustus.gobics.de/) and were annotated using BLASTX based on sequence similarity to known proteins in the non-redundant NCBI database. The genes were classified using Gene Ontology (GO) terms. Moreover, the number of different genes with matching gene descriptions was determined.

Gene expression quantification

Total gene expression was quantified by mapping the filtered HiSeq2000 sequence data to the 51,188 reference contigs constructed from the hybrid assembly of GS_FLX+ and HiSeq2000 data. All reads mapped to the reference sequence were normalized by the RPKM method (reads per kilobase of exon model per million mapped reads), a formula reported by Mortazavi et al., which is used to quantify gene expression (Mortazavi et al., 2008).

RESULTS AND DISCUSSION

Sequencing and de novo assembly

We obtained 624,555,295 bp from 1,226,938 reads on the GS_FLX+ and 15,361,260,265 bp from 161,417,646 reads on the HiSeq2000 platform. Of these, 624,555,205 bp (99 %) and 15,361,260,265 bp (71 %) from the raw DNA FASTA sequences derived from the GS_FLX+ and HiSeq2000 platforms, respectively, were verified as having a high quality score of at least QC20 (Table 1). A total of 96,681,919 bp of high-quality DNA sequence data from 96,663 contigs were combined for *de novo* assembly. Thus, we obtained approximately 82 Mb of sequence, comprising 51,188 contigs with a read-length of at least 500 bp and an average contig read-length of 1,616 bp (Table 2).

EXCLI Journal 2014;13:392-400 – ISSN 1611-2156
Received: September 13, 2013, accepted: September 26, 2013, published: April 07, 2014

Table 1: Summary of sequencing data derived from the GS_FLX+ and HiSeq2000 platforms

	Reads		Length (bp)		Coverage depth for an estimated transcriptome length of 40 Mb
GS_FLX+					
Raw FASTA sequence	1,289,901		629,029,941		16
High quality sequence after trimming*	1,226,938	(95 %)	624,555,205	(99 %)	16
HiSeq2000					
Raw FASTA sequence	215,200,340		21,735,234,340		543
High quality sequence after trimming*	161,417,646	(99 %)	15,361,260,265	(71 %)	384
Sum of GS_FLX+ & HiSeq2000					
Raw FASTA sequence	216,490,241		22,364,264,281		559
High quality sequence after trimming*	162,644,584	(75 %)	15,985,815,470	(71 %)	400

*High quality sequence after trimming is the data that remains after filtering for a quality score of at least 20, and removing short read-lengths of less than 200 bp and 75 bp on GS_FLX+ and HiSeq2000 reads, respectively.

Table 2: Summary of contigs and read-length from the *de novo* assembly of high-quality sequencing data using Trinity software

	Contigs	Length (bp)	Average length (bp)/contig
All contigs	96,663	96,681,919	1,000
Contigs with at least 500 bp	51,188	82,633,331	1,614

De novo assembly was performed by pooling high quality sequencing data from GS_FLX+ and HiSeq200 platforms using Trinity software

The number of contigs with lengths of 500–600 bp was 5,927 (11.6 %). A further 5,940 contigs (11.6 %) had a length of at least 3000 bp, whereas 31,050 contigs (60.7 %) had a length of at least 1,000 bp (Figure 1). This result demonstrated the advantage of combining data from multiple platforms to obtain transcriptome information and construct contigs, compared to previous transcriptome studies that used data from a single platform (454 pyrosequencing of Roche/454 GS_FLX+) (Pazos-Navarro et al., 2011; Rowland et al., 2012; Sloan et al., 2012). A single NGS platform has been used previously for transcriptome analysis, generating a low number of contigs from a small amount of pyrosequencing data and successfully identified defense response-related genes in avocado (*Persea americana*) to the root pathogen *P. cinnamon* (Mahomed and van den Berg, 2011). A large number (625,342) of pyrosequencing reads from the Roche/454 GS_FLX+ platform was also used previously to construct 15,284 contigs, with a length >300 bp, using *de novo* assembly in the transcriptome analysis of the tree peony (*Paeonia suffruticosa* Andrews) (Gai et al., 2012), providing a gene discovery resource in this plant.

The large contigs established in our study could be useful for plant breeders or geneticists to understand transcriptome data from *S. baicalensis*. All annotated gene sequences have been uploaded to a publicly accessible database (ftp://ngs.snu.ac.kr/hwanggeum/).

EXCLI Journal 2014;13:392-400 – ISSN 1611-2156
Received: September 13, 2013, accepted: September 26, 2013, published: April 07, 2014

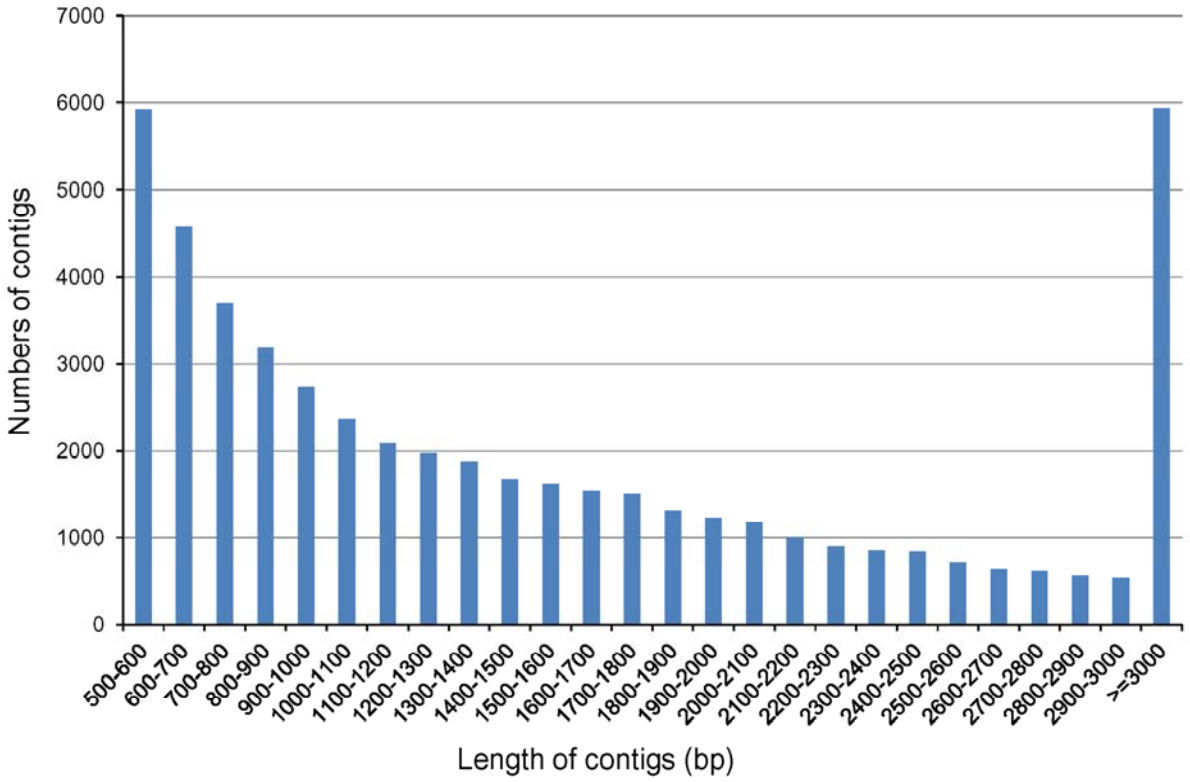

Figure 1: Distribution of the number of contigs based on contig length. The average length of contigs was 1,616 bp in 51,188 contigs of at least 500-bp length. Of these contigs, 5,940 (11.6 %) had a length of at least 3,000 bp, whereas 31,050 (60.7 %) had a length of at least 1,000 bp.

Gene annotation

To analyze the transcriptome, we used BLASTX searches of the 51,818 contigs for gene annotation. We identified 39,581 contigs with similarity (E-value cutoff: e-7) to sequences in the non-redundant NCBI database. Of these, we identified a total of 20,498 different genes expressed in the early growth stage of *S. baicalensis*. We found that an average of two annotated contigs matched to the same gene, based on the descriptions in the NCBI public database. Again, the large number of different genes identified demonstrates the effectiveness of combining NGS data from multiple platforms to discover expressed genes when studying the transcriptome of a novel species.

We summarized the distribution of 20,498 different genes, identified through BLASTX-searches in NCBI, in different taxonomic categories (Table 3). More than 44 % (9,746 genes) of the total 20,498 different genes had matches with genes in *Vitis vinifera*. Moreover, almost 30 % of the genes identified in *S. baicalensis* were expressed in *Ricinus communis* and *Populus trichocarpa*.

All annotated genes were further classified using GO terms analysis, based on the cellular component, function, and biological process categories (Figure 2). A total of 15,401, 13,730, and 14,626 genes were matched to these three categories, respectively. A total of 15,294 and 11,651 genes were classified to the cell and organelle, based on the cellular component. The three highest groupings for the molecular function category were binding (9,384), catalytic activity (8,211), and transporter activity (1,460). In the biological processing group, the categories of cellular process- and metabolic process-related genes contained the highest number of genes (12,324 and

EXCLI Journal 2014;13:392-400 – ISSN 1611-2156
Received: September 13, 2013, accepted: September 26, 2013, published: April 07, 2014

10,553 genes, respectively). Specifically, we identified 1,193 immune system-related genes and 240 viral reproduction genes. Transcription factors (TFs), which are key to plant development and the responses of plants to environmental stressors (Broun, 2004; Mehrtens et al., 2005; Du et al., 2009), are another important element in the regulation of gene expression. We found 126 TFs, including putative TF data; these genes provide useful insights into *S. baicalensis* regulation.

Table 3: Distribution of different genes matched to the NCBI database using BLASTX searches

Taxonomy category	Different genes	Frequency (%)
Vitis vinifera	9746	47.5
Ricinus communis	3114	15.2
Populus trichocarpa	2954	14.4
Glycine max	465	2.3
Nicotiana tabacum	434	2.1
Arabidopsis thaliana	337	1.6
Arabidopsis lyrata subsp. *lyrata*	333	1.6
Solanum lycopersicum	308	1.5
Solanum tuberosum	210	1.0
Medicago truncatula	134	0.7
Oryza sativa Japonica Group	129	0.6
Sorghum bicolor	107	0.5
Capsicum annuum	67	0.3
Other plants	2160	10.5
Total	20498	100 %

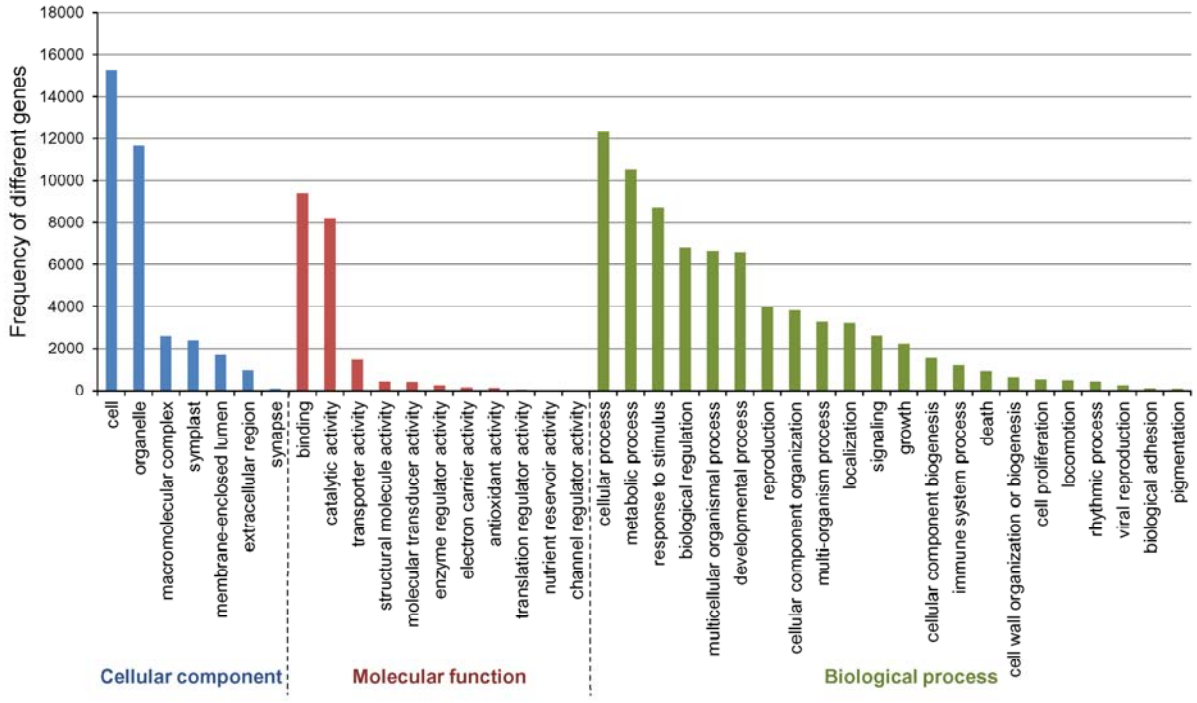

Figure 2: Distribution of annotated genes according to Gene Ontology terms on the basis of cellular localization, molecular function, and biological processes. A total of 1,5401, 1,3730, and 1,4626 genes had matches in the cellular component, molecular function, and biological process sectors, respectively

EXCLI Journal 2014;13:392-400 – ISSN 1611-2156
Received: September 13, 2013, accepted: September 26, 2013, published: April 07, 2014

Gene expression quantification

A total of 152,917,924 (94.7 %) of 161,417,646 reads from the HiSeq2000 platform were mapped to the reference contig sequence. All 51,188 contigs were identified as expressed using the HiSeq2000 RNA sequencing data mapping method. We found that the HiSeq2000 RNA sequencing data comprehensively covered the *S. baicalensis* transcriptome. In the RPKM quantification of gene expression, the gene expression quantifier ranged from 0.007 to 4,554 in 51,188 expressed sequence tags (ESTs), incorporating 39,581 known genes and 11,607 unknown genes, with an average quantifier for gene expression of 12.7. The first 5,119 (10 %) highly expressed genes, with an average quantifier of 88.15, were expressed more than 6.9-fold higher than the average quantifier for all ESTs (12.7; data not shown). A total of 718 ESTs was highly expressed with a quantifier of 127 on RPKM analysis (Figure 3). Of the 718 highly expressed genes, six ESTs were photosynthesis-related genes. The quantitative expression data are available on a publicly accessible database (ftp://ngs.snu.ac.kr/hwanggeum/) and will be useful as a standard expression profile against which other quantitative expression data from other *S. baicalensis* research fields can be compared.

CONCLUSION

This is the first study to identify a comprehensive transcriptome of the early growth stage of *S. baicalensis* using *de novo* assembly of sequence data derived from multiple NGS platforms. The combination of mRNA sequences from the Roche/454 GS_FLX and Illumina/Solexa HiSeq 2000 was effective for identifying a large set of unigenes through *de novo* assembly of whole-genome data from a single species. A total of 51,188 contigs, at least 500 bp long, were obtained in this transcriptome analysis. Of these, at least 20,498 different genes were identified as being expressed in the early growth stage in *S. baicalensis*. Knowledge of all described genes and quantitation of the expressed genes, including the TFs involved, will provide useful insights into gene regulation in *S. baicalensis*.

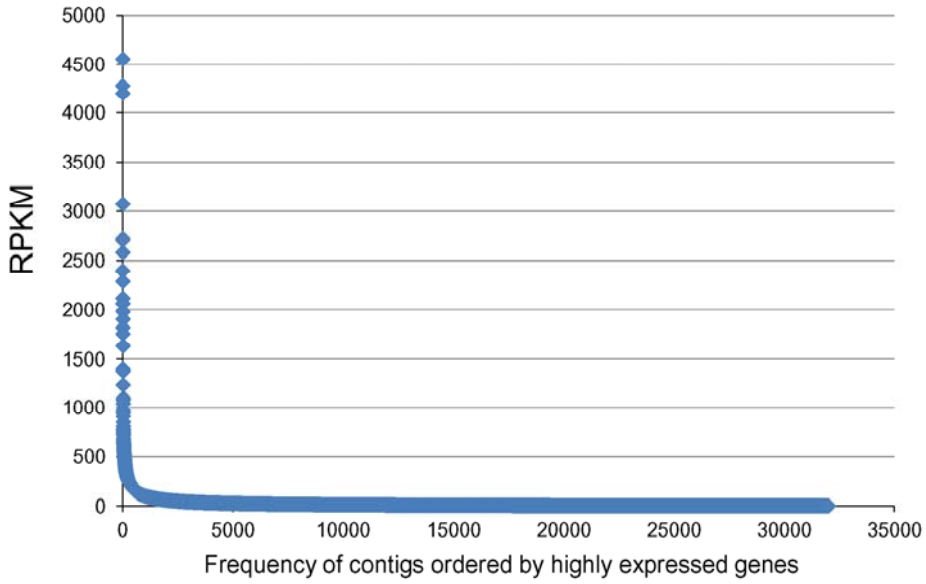

Figure 3: RPKM (reads per kilobase of exon model per million mapped reads) of contigs ordered by highly expressed genes. The first 32,000 genes demonstrating high expression are presented on the chart.

ACKNOWLEDGMENTS

This work was carried out with the support of "Cooperative Research Program for Agriculture Science & Technology Development (Project No. PJ906938)" Rural Development Administration, Republic of Korea.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Broun P. Transcription factors as tools for metabolic engineering in plants. Curr Opin Plant Biol 2004;7: 202-9.

Du H, Zhang L, Liu L, Tang XF, Yang WJ, Wu YM et al. Biochemical and molecular characterization of plant MYB transcription factor family. Biochemistry (Moscow) 2009;74:1-11.

Ekblom R, Balakrishnan C, Burke T, Slate J. Digital gene expression analysis of the zebra finch genome. BMC Genomics 2010;11:219.

Gai S, Zhang Y, Mu P, Liu C, Liu S, Dong L et al. Transcriptome analysis of tree peony during chilling requirement fulfillment: Assembling, annotation and markers discovering. Gene 2012;497:256-62.

Hsiao YY, Chen YW, Huang SC, Pan ZJ, Fu CH, Chen WH et al. Gene discovery using next-generation pyrosequencing to develop ESTs for Phalaenopsis orchids. BMC Genomics 2011;12: 360.

Kumar S, Blaxter M. Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 2010; 11:571.

Li H, Murch S, Saxena P. Thidiazuron-induced *de novo* shoot organogenesis on seedlings, etiolated hypocotyls and stem segments of Huang-qin. Plant Cell Tissue Organ Culture 2000;62:169-73.

Logacheva M, Kasianov A, Vinogradov D, Samigullin T, Gelfand M, Makeev V et al. De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum). BMC Genomics 2011;12: 30.

Mahomed W, van den Berg N. EST sequencing and gene expression profiling of defence-related genes from Persea americana infected with Phytophthora cinnamomi. BMC Plant Biol 2011;11:167.

Martens S, Mithöfer A. Flavones and flavone synthases. Phytochemistry 2005;66:2399-407.

Matsumura H, Yoshida K, Luo S, Kimura E, Fujibe T, Albertyn Z et al. High-throughput superSAGE for digital gene expression analysis of multiple samples using next generation sequencing. PLoS ONE 2010; 5:e12010.

Mehrtens F, Kranz H, Bednarek P, Weisshaar B. The arabidopsis transcription factor MYB12 is a flavonol-specific regulator of phenylpropanoid biosynthesis. Plant Physiol 2005;138:1083-96.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5: 621-8.

Murashige T, Skoog F. A revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiol Plantarum 1962;15:473-97.

Nishikawa K, Furukawa H, Fujioka T, Fujii H, Mihashi K, Shimomura K et al. Flavone production in transformed root cultures of *Scutellaria baicalensis* Georgi. Phytochemistry 1999;52:885-90.

Oshlack A, Robinson M, Young M. From RNA-seq reads to differential expression results. Genome Biol 2010;11:220.

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

Park NI, Xu H, Li X, Kim SJ, Park SU. Enhancement of flavone levels through overexpression of chalcone isomerase in hairy root cultures of *Scutellaria baicalensis.* Funct Integr Genomics 2011;11:491-6.

Pazos-Navarro M, Dabauza M, Correal E, Hanson K, Teakle N, Real D et al. Next generation DNA sequencing technology delivers valuable genetic markers for the genomic orphan legume species, *Bituminaria bituminosa*. BMC Genet 2011;12:104.

Rowland L, Alkharouf N, Darwish O, Ogden E, Polashock J, Bassil N et al. Generation and analysis of blueberry transcriptome sequences from leaves, developing fruit, and flower buds from cold acclimation through deacclimation. BMC Plant Biol 2012;12:46.

Sloan DB, Keller SR, Berardi AE, Sanderson BJ, Karpovich JF, Taylor DR. De novo transcriptome assembly and polymorphism detection in the flowering plant Silene vulgaris (Caryophyllaceae). Mol Ecol Resour 2012;12:333-43.

Watson A, Mazumder A, Stewart M, Balasubramanian S. Technology for microarray analysis of gene expression. Curr Opin Biotech 1998;9:609-14.

Wickramasinghe S, Rincon G, Islas-Trejo A, Medrano J. Transcriptional profiling of bovine milk using RNA sequencing. BMC Genomics 2012;13: 45.

Xu H, Park NI, Li X, Kim YK, Lee SY, Park SU. Molecular cloning and characterization of phenylalanine ammonia-lyase, cinnamate 4-hydroxylase and genes involved in flavone biosynthesis in *Scutellaria baicalensis*. Bioresource Technol 2010; 101:9715-22.

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