Genetics and Molecular Research

Characterization of ESTs from black locust for gene discovery and marker development

J.X. Wang[1], C. Lu[2], C.Q. Yuan[1], B.B. Cui[3], Q.D. Qiu[1], P. Sun[4], R.Y. Hu[1], D.C. Wu[1], Y.H. Sun[1] and Y. Li[1]

[1]National Engineering Laboratory for Tree Breeding,
Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants,
Ministry of Education, College of Biological Sciences and Technology,
Beijing Forestry University, Beijing, China
[2]Beijing Daxing Fruit and Forestry Institute, Beijing, China
[3]Department of Biology and Chemistry, Baoding University, Hehei, China
[4]Non-Timber Forestry Research and Development Center,
Chinese Academy of Forestry, Zhengzhou, China

Corresponding authors: Y.H. Sun / Y. Li
E-mail: sunyuhan@bjfu.edu.cn / yunli@bjfu.edu.cn

Genet. Mol. Res. 14 (4): 12684-12691 (2015)
Received March 17, 2015
Accepted July 23, 2015
Published October 19, 2015
DOI http://dx.doi.org/10.4238/2015.October.19.12

ABSTRACT. Black locust (*Robinia pseudoacacia* L.) is an ecologically and economically important species. However, it has relatively underdeveloped genomic resources, and this limits gene discovery and marker-assisted selective breeding. In the present study, we obtained large-scale transcriptome data using a next-generation sequencing platform to compensate for the lack of black locust genomic information. Increasing the amount of transcriptome data for black locust will provide a valuable resource for multi-gene phylogenetic analyses and will facilitate research on the mechanisms whereby conserved genes and functions are maintained in the face of species divergence. We sequenced the black locust transcriptome from a cDNA library of multiple tissues and individuals on an Illumina platform, and this produced 108,229,352 clean sequence reads. The high-quality overlapping expressed sequence tags (ESTs)

were assembled into 36,533 unigenes, and 4781 simple sequence repeats were characterized. A large collection of high-quality ESTs was obtained, *de novo* assembled, and characterized. Our results markedly expand the previous transcript catalogues of black locust and can gradually be applied to black locust breeding programs. Furthermore, our data will facilitate future research on the comparative genomics of black locust and related species.

Key words: *Robinia pseudoacacia*; RNA-Seq; Transcriptome; SSRs

INTRODUCTION

Black locust (*Robinia pseudoacacia* L.) is a tree belonging to the subfamily Faboideae. This species is native to the southeastern United States and is widely planted and naturalized elsewhere North America, Europe, southern Africa, and Asia. Black locust was first introduced to China during 1877-1878, and it is currently extensively cultivated in many parts of the country (Li, 1983). This species is well adapted for growth in a wide variety of soils and environmental conditions such as salt, drought, and cold, and it is used for fuel, wood fiber, feedstock, lumber, forage, and beekeeping (Barrett et al., 1990; Sun et al., 2013; Yuan et al., 2013). Consequently, black locust is an ecologically and economically valuable species; however, data regarding its genomics and genome resources are lacking. To date, only 189 nucleotide sequences, 3095 expressed sequence tags (ESTs), and 79 protein sequences have been deposited in the National Center for Biotechnology Information (NCBI) GenBank database.

Information regarding the genetic control of many important traits and fine-scale genetic variation is extremely limited in plants (Niu et al., 2013). Several representative genera of angiosperm trees have genome sizes of 540-2000 Mb. For woody plants, especially those of high heterozygosity such as black locust, whole-genome sequencing requires long-term and expensive investment, and is therefore currently limited to a few species. However, it has been used to obtain information on unigenes through transcriptome sequencing (Park et al., 2006; Crowhurst et al., 2008; Feng et al., 2012).

In many species, the main limitation to understanding and characterizing important traits is the lack of sufficient genetic sequences for the development of high-density genetic maps and marker-assisted breeding studies (Russell et al., 2011). ESTs have been identified as useful sources of simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs), both of which are useful tools for mapping and marker-assisted breeding in plants (Morgante et al., 2002).

RNA-Seq is a powerful tool for transcriptome analysis and it enables investigations of species to be conducted without the use of corresponding sequenced genome information as a reference. It has been widely used in model and non-model species (Wilhelm and Landry, 2009; Schilmiller et al., 2010; Bleeker et al., 2011) to obtain mass sequence data for gene discovery, annotation, and molecular marker development (Pavy et al., 2005). In comparison with traditional laboratory methods, RNA-Seq is a high-throughput technology. It offers considerable advantages for detecting allele-specific expression, splice junction variation, novel transcribed regions, SSRs, and SNPs (Malone and Oliver, 2011; Feng et al., 2012). RNA-Seq is expected to revolutionize the way in which eukaryotic transcriptomes are analyzed (Wilhelm and Landry, 2009). Genome sequencing technology has become progressively more efficient over the past decade; however,

the sequencing of complex genomes remains expensive (Emrich et al., 2007).

Here, we provide the first comprehensive characterization of the transcriptome of *R. pseudoacacia* L., using the Illumina sequencing platform. During our analysis, we identified thousands of molecular markers. Our results provide a valuable new insight into the transcriptome of *R. pseudoacacia* L., and they will facilitate marker-assisted selective breeding.

MATERIAL AND METHODS

Plant materials, RNA extraction, cDNA library, and sequencing

In July 2013, materials were collected from a healthy 10-year-old black locust tree at MiJiaBao Forestry Station, YanQin, Beijing, China. We collected samples from each of eight different tissues: leaf, petiole, stem, flower, root, young pod, bark, and annual shoot. All samples were immediately frozen in liquid nitrogen and transported to the laboratory, where they were stored at -80°C until use in analyses. Total RNA from the samples was extracted using a Takara Mini BEST Plant RNA Extraction Kit following the manufacturer protocol. RNA quality was verified using formaldehyde agarose gel electrophoresis. RNA quantity was determined using a NanoDrop ND-1000 spectrophotometer. For each cDNA library, we used 1000 ng RNA. The mixed cDNA library was sequenced using the Illumina Genome Analyzer IIx platform.

Sequence assembly

Reads were assembled using Trinity. The longest assembled sequences were referred to as contigs. The reads were then mapped back to the contigs with paired-end reads to detect contigs from the same transcript and determine the distances between these contigs. Finally, we obtained sequences that could not be extended on either end. These sequences were defined as unigenes.

Sequence annotation

Functional annotations of unigenes were analyzed using gene ontology (GO) analysis. All unigene sequences were searched against protein databases (Nr, SwissProt, KEGG, and COG) using BLASTx (E-value < 0.0001). Protein functional information was predicted from annotations of the most similar proteins in the databases. GO functional annotations were obtained from SwissProt. GO annotation comprises three ontologies - molecular functions, cellular components, and biological processes. The basic GO unit is a GO term, and every GO term belongs to a type of ontology.

Identification of SSRs

SSR refers to a basic unit of a DNA genome in which a 2-6 nucleotide sequence repeat is widely distributed in different locations of the genome, usually for lengths of ≤200 bp. SSR detection was based on the transcriptome, using assembled unigenes as the reference sequences with the MISA software to locate all the SSRs (Grabherr et al., 2011). Similar criteria for SSR

screening were used in previous studies (Thiel et al., 2003).

RESULTS

Transcriptome sequencing and *de novo* assembly

After RNA extraction, we constructed a cDNA library using pooled RNA from mixed samples of black locusts. After removing adaptors, primer sequences, and poly-A tails, as well as short, long, and low-quality sequences, we obtained 108,229,352 clean reads with an average length of 100 bp. The clean sequence data were submitted to the NCBI Short Read Archive under Accession No. SRR1563113. Using the Trinity software, we obtained 68,053 transcript sequences with an average length of 912 bp (Table 1). All the sequences were assembled, and this resulted in 36,533 non-redundant unigenes with an average length of 878 bp. The N50 of the unigenes was 1317 bp. Among the non-redundant unigenes, 26,031 (71.3%) ranged in length from 200 to 1000 bp; 7694 (20.5%) ranged in length from 1001 to 2000 bp; and 3006 (8.2%) were >2000 bp in length. The length distribution of these unigenes is shown in Figure 1.

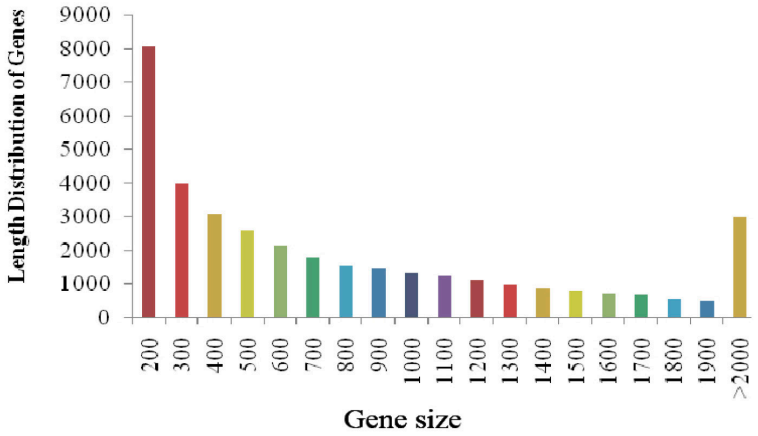

Figure 1. Length distribution of black locust unigenes.

Table 1. Summary of the sequencing and assembly results.

	Total	Minimum length	Median length	Mean length	N50	Maximum length	Total length
Transcripts	68,053	201	671	912	1,335	8,028	62,134,169
Unigenes	36,533	201	622	878	1,317	8,028	32,098,396

Functional annotation of the transcriptome

To determine the possible functions of tagged genes, we annotated the unigenes with GO terms, based on BLASTx searches of sequence comparisons between the descriptors of their closest homolog and the NCBI non-redundant protein database. The functions of the identified genes covered various biological processes, molecular functions, and cellular components. We found that 15,995 of the 36,533 unigenes were assigned to the biological processes category (Figure 2). After metabolic processes (96.8%) and cellular processes (81.2%), the largest proportion of

functionally assigned ESTs fell into single-organism processes (44.75%) and biological regulation (29.7%). In addition, 15,810 unigenes were assigned to the molecular function category. After binding (76.3%) and catalytic activity (62.7%), the largest proportion of functionally assigned ESTs fell into transporter activity (8.2%) and receptor activity (5.6%). Finally, 14,747 unigenes were assigned to the cellular components category. The largest proportion of functionally assigned ESTs fell into cells (88.9%) and cell parts (88.9%).

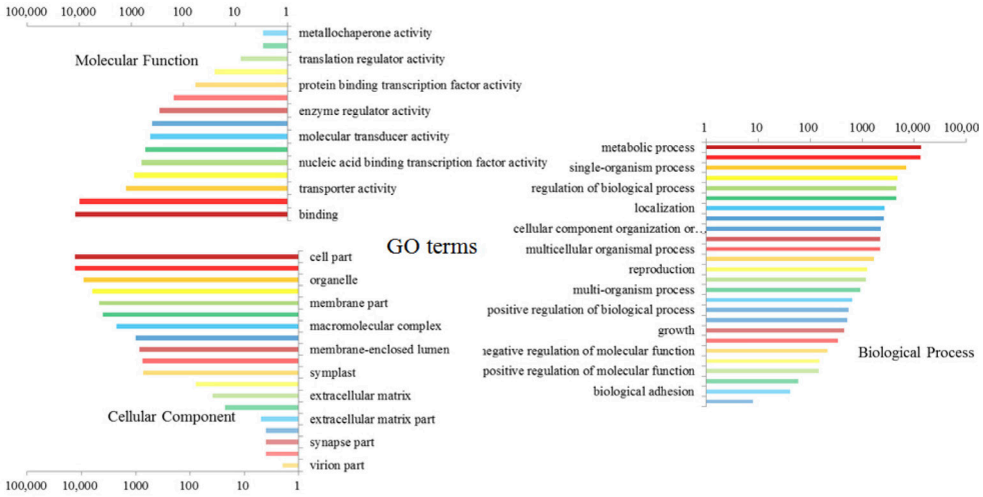

Figure 2. Gene ontology (GO) classification of black locust unigenes.

Identification of SSRs and SNPs

We used the MISA program (Thiel et al., 2003) to identify 4781 SSR loci (Table 2). Of these loci, 547 sequences contained more than one EST-SSR and 36 EST-SSRs were present in compound form. In general, one EST-SSR was found for every 6.71 kb in the unigenes. The most common type of repeat was the trimer, which constituted 48.5% of all SSRs detected. This was followed by dimers (39.5%), tetramers (5.6%), pentamers (3.2%), and hexamers (3.1%) (Figure 3).

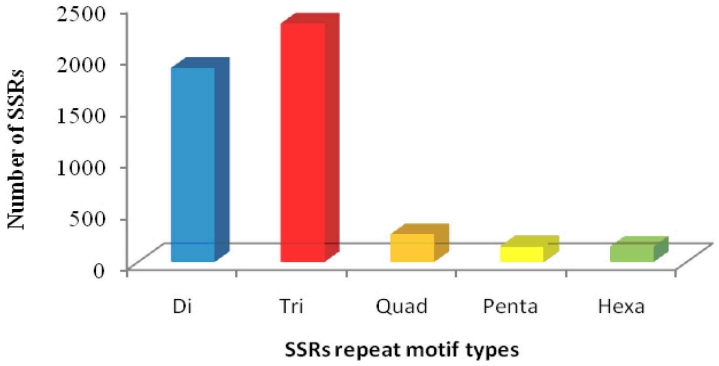

Figure 3. Distribution of simple sequence repeats (SSRs) in black locust ESTs.

We calculated the frequencies of EST-SSRs with different numbers of repeat units. We found that the most common type of repeat unit was 5 repeat motifs, followed by 6, 7, 4, 8, 9, 10, 11, and 12 repeat motifs. Within the detected cSSRs, we identified 147 motif sequence types, of which the most common dimer, trimer, tetramer, pentamer, and hexamer repeats were AG/CT, AAG/CTT, AAAG/CTTT, AAGAG/CTCTT, and AAGGAG/CTCCTT, respectively.

Table 2. Summary of EST-SSR mining results.

Searched item	Result
Total number of sequences examined	36,533
Total length of the examined sequences (bp)	32,098,396
Total number of identified SSRs	4,781
Number of SSR-containing sequences	4,132
Number of sequences containing multiple SSRs	547
Number of SSRs present in compound formation	36
Di-nucleotide repeats	1,889
Tri-nucleotide repeats	2,321
Tetra-nucleotide repeats	270
Penta-nucleotide repeats	152
Hexa-nucleotide repeats	149

DISCUSSION

The read length obtained using 454 pyrosequencing is longer than that obtained using Illumina sequencing; however, the latter has been successfully and increasingly used for many species (Collins et al., 2008; Hegedus et al., 2009; Trick et al., 2009; Wang et al., 2010a,b,c; Wei et al., 2011). A key advantage is the considerably low cost of Illumina sequencing compared with 454 pyrosequencing methods (Wang et al., 2012).

A large number of ESTs for black locust have been obtained from cDNA libraries based on Illumina sequencing, which is a more efficient method than traditional sequencing. In the present study, we obtained 36,533 unigenes with an average length of 878 bp. In most previous studies, cDNA libraries generated from a single tissue sample were used; in contrast, in the present study, we used a normalized cDNA library comprising multiple tissue samples. The large-scale ESTs obtained provide more comprehensive black locust transcriptome information and will facilitate the assembly of black locust ESTs in the future.

A major advantage of next-generation sequencing technology is the capacity to deliver large numbers of gene-based markers from transcriptome sequencing projects. Based on cost and throughput requirements, conventional markers such as restriction fragment length polymorphisms and random amplified polymorphic DNA (RAPD) are being replaced with SSRs and SNPs (Powell et al., 1996). SSRs are considered ideal genetic markers for population and evolutionary studies. In comparison with other genetic markers such as amplified fragment length polymorphisms and RAPD, SSRs have many advantages, including high-throughput capabilities, biallelic loci, and data portability (Vignal et al., 2002; Li et al., 2009). The genome-wide and abundant EST-based SSR markers obtained using next-generation sequencing constitute an effective approach for marker discovery in many plant species, because these markers facilitate the generation of dense genetic maps and have the advantage of higher cross-species transferability.

Previous studies have shown that the main type of repeat in plant EST-SSRs is the trimer, followed by dimers and quadmers; however, the main type of repeat may vary among different

species (Bassam et al., 1991). In the present study, we found that the dominant type of black locust EST-SSR repeat unit was trimer repeats, which constituted 48.5% of all SSRs detected. The most common motif was GAA.

We identified 4781 SSR loci in 36,533 black locust unigenes. The frequency of EST-SSRs was 13.1%. Previous studies have shown that the EST-SSR frequencies of ginkgo, jatropha, poplar, and eucalyptus are 5.97, 6.51, 17.73, and 17.73%, respectively (Nicot et al., 2004; Yadav et al., 2011). The results of our present study differ markedly from these previous findings, possibly because of the different SSR search programs used and variations in the software parameter settings. However, the discrepancies may be derived from the specific characteristics of black locust EST-SSRs.

CONCLUSION

Our present study represents the first large-scale sequencing and analysis of the black locust gene transcriptome and provides the most comprehensive black locust sequence resource to date. To understand the functions and regulatory pathways of different categories, we used GO and KEGG analyses and classified the functional categories based on all the unigenes. The large-scale data obtained using extremely high numbers of SSRs will facilitate the development of oligo-nucleotide microarrays. In addition, our data constitute a reference transcriptome for future RNA-seq experiments in large-scale gene expression assays. Our findings will provide a valuable insight into genetic variation in populations and will facilitate the genetic control of important traits in black locust. Finally, our data will benefit future research into mapping and marker-assisted breeding of black locust.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Science and Technology Support Program (#2012BAD01B0601), the National Science Foundation of China (#31170629), the Scientific Research of the Forest Public Welfare Industry (#201304116), and the Beijing Municipal Science and Technology Commission Project (#Z121100008512002).

REFERENCES

Barrett RP, Mebrahtu T and Hanover JW (1990). Black locust: A multi-purpose tree species for temperate climates. In: Advances in new crops (Janick J and Simon JE, eds.). Timber Press, Portland.

Bassam BJ, Caetano-Anollés G and Gresshoff PM (1991). Fast and sensitive silver staining of DNA in polyacrylamide gels. *Anal. Biochem.* 196: 80-83.

Bleeker PM, Spyropoulou EA, Diergaarde PJ, Volpin H, et al. (2011). RNA-seq discovery, functional characterization, and comparison of sesquiterpene synthases from *Solanum lycopersicum* and *Solanum habrochaites* trichomes. *Plant Mol. Biol.* 77: 323-336.

Collins LJ, Biggs PJ, Voelckel C and Joly S (2008). An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Inform.* 21: 3-14.

Crowhurst RN, Gleave AP, MacRae EA, Ampomah-Dwamena C, et al. (2008). Analysis of expressed sequence tags from *Actinidia*: applications of a cross species EST database for gene discovery in the areas of flavor, health, color and

ripening. *BMC Genomics* 9: 351.
Emrich SJ, Barbazuk WB, Li L and Schnable PS (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* 17: 69-73.
Feng C, Chen M, Xu CJ, Bai L, et al. (2012). Transcriptomic analysis of Chinese bayberry (*Myrica rubra*) fruit development and ripening using RNA-Seq. *BMC Genomics* 13: 19.
Grabherr MG, Haas BJ, Yassour M, Levin JZ, et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29: 644-652.
Hegedus Z, Zakrzewska A, Agoston VC, Ordas A, et al. (2009). Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Mol. Immunol.* 46: 2918-2930.
Li JH (1983). The introduction and development of *Robinia pseudoacacia* in Shandong. *J. Shandong Forest. Sci. Technol.* 4: 73-75.
Li S, Wan H, Ji H, Zhou K, et al. (2009). SNP discovery based on CATS and genotyping in the finless porpoise (*Neophocaena phocaenoides*). *Conserv. Genet.* 10: 2013-2019.
Malone JH and Oliver B (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* 9: 34.
Morgante M, Hanafey M and Powell W (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30: 194-200.
Nicot N, Chiquet V, Gandon B, Amilhat L, et al. (2004). Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). *Theor. Appl. Genet.* 109: 800-805.
Niu SH, Li ZX, Yuan HW, Chen XY, et al. (2013). Transcriptome characterisation of *Pinus tabuliformis* and evolution of genes in the *Pinus* phylogeny. *BMC Genomics* 14: 263.
Park S, Sugimoto N, Larson MD, Beaudry R, et al. (2006). Identification of genes with potential roles in apple fruit development and biochemistry through large-scale statistical analysis of expressed sequence tags. *Plant Physiol.* 141: 811-824.
Pavy N, Paule C, Parsons L, Crow JA, et al. (2005). Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. *BMC Genomics* 6: 144.
Powell W, Morgante M, Andre C, Hanafey M, et al. (1996). The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Mol. Breed.* 2: 225-238.
Russell JR, Bayer M, Booth C, Cardle L, et al. (2011). Identification, utilisation and mapping of novel transcriptome-based markers from blackcurrant (*Ribes nigrum*). *BMC Plant Biol.* 11: 147.
Schilmiller AL, Miner DP, Larson, M, McDowell E, et al. (2010). Studies of a biochemical factory: tomato trichome deep expressed sequence tag sequencing and proteomics. *Plant Physiol.* 153: 1212-1223.
Sun P, Yuan C, Dai L, Xi Y, et al. (2013). Phytohormone and assimilate profiles in emasculated flowers of the black locust (*Robinia pseudoacaia*) during development. *Acta Biol. Hung.* 64: 364-376.
Thiel T, Michalek W, Varshney RK and Graner A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* 106: 411-422.
Trick M, Long Y, Meng J and Bancroft I (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnol. J.* 7: 334-346.
Vignal A, Milan D, SanCristobal M and Eggen A (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol.* 34: 275-305.
Wang B, Guo G, Wang C, Lin Y, et al. (2010a). Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. *Nucl. Acids Res.* 38: 5075-5087.
Wang S, Wang X, He Q, Liu X, et al. (2012). Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. *Plant Cell Rep.* 31: 1437-1447.
Wang XW, Luan JB, Li JM, Bao YY, et al. (2010b). *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* 11: 400.
Wang Z, Fang B, Chen J, Zhang X, et al. (2010c). *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.
Wei W, Qi X, Wang L, Zhang Y, et al. (2011). Characterization of the sesame (*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers. *BMC Genomics* 12: 451.
Wilhelm BT and Landry JR (2009). RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 48: 249-257.
Yadav HK, Ranjan A, Asif MH, Mantri S, et al. (2011). EST-derived SSR markers in *Jatropha curcas* L.: development, characterization, polymorphism, and transferability across the species/genera. *Tree Genet. Genomes* 7: 207-219.
Yuan CQ, Li YF, Wang L, Zhao KQ, et al. (2013). Evidence for inbreeding depression in the tree *Robinia pseudoacacia* L. (Fabaceae). *Genet. Mol. Res.* 12: 6249-6256.