Plant Omics Journal POJ

POJ 7(6):445-449 (2014) ISSN:1836-3644

Genome-wide discovery and analysis of single nucleotide polymorphisms and insertions/ deletions in *Juglans regia* L. by high-throughput pyrosequencing

Zhuoyi Liao, Kai Feng, Yingnan Chen, Xiaogang Dai, Shuxian Li*, Tongming Yin

The Southern Modern Forestry Collaborative Innovation Center, Nanjing Forestry University, Nanjing 210037, China

***Corresponding author: shuxianli@njfu.com.cn**

Abstract

Walnut (*Juglans regia* L.) is one of the most widespread economic trees in the world. In this study, genome sequencing of *J. regia* was performed with Roche 454 GS-FLX sequencer. A total of 541,176 high-quality reads were produced with an average read length of 359 bp. We obtained 31,362 contigs (≥100 bp) after assembling with the Roche Newbler *de novo* Assembler (Version 2.8); the total length of the assembly was 15.1 Mb. We then mapped all the sequence reads against the *J. regia* genome assembly. Totally, 49,202 nucleotide variations were detected including 48,165 single nucleotide polymorphisms (SNPs) and 1,037 insertions/deletions (InDels). Among the detected SNPs, the number of transitions (35,480) was much more than that of transversions (12,685), with a ratio of 2.79:1. Within transitions, C/T transitions were slightly more abundant than that of G/A. Among transversions, A/T transversions were much more common than the other three types of transversions. As for the detected InDels, single nucleotide changes accounted for 52.65% of all InDels. Additionally, the InDel numbers were negatively correlated with the InDel lengths.

Keywords: SNP; *Juglans regia*; genome; transition; transversion; InDel.
Abbreviations: SNP- single nucleotide polymorphism; InDel- insertion/deletion; NGS- Next-generation sequencing.

Introduction

Single nucleotide polymorphisms (SNPs), which refer to the single-base differences among individuals, represent the most abundant source of genetic variation both in human and plant genomes (Wang et al., 1998; Brookes, 1999). The International SNP Map Working Group (2001) identified more than 1.4 million SNPs in the human genome, with approximately one SNP per kilobase (Wang et al., 1998). By contrast, the typical frequencies of SNPs in plants are in the range of one SNP every 100-300 bp (Edwards et al., 2007). Due to the enormous number, widely distribution and low mutation rate, SNP has been a useful tool in generating ultra-high-density genetic maps, constructing haplotype systems for genes or regions of interest, and map-based positional cloning (Duran et al., 2009). The assays that detect SNPs generally can also detect small insertions or deletions (InDels), which are one of the major sources of evolutionary change at the molecular level. InDels are also considered to be a valuable type of marker for quantitative trait locus mapping (Vasemagi et al., 2010), marker-assisted selection (Hayashi et al., 2006), and varietal testing (Steele et al., 2008). Although SNPs and InDels are increasingly becoming the marker of choice in genetic studies, one of the limitations is the initial cost associated with their development (Batley et al., 2003). In recent years, the development of next-generation sequencing (NGS) technology offers great opportunities for molecular marker discovery due to its ability of generating large amounts of sequence data efficiently. This high-throughput method for detecting SNPs and InDels provides the greatest potential for cost-effective SNP discovery. NGS technology has been used successfully to identify nucleotide variations for a growing number of animals and plants, such as cow (Gibbs et al., 2002), swine (Wiedmann et al., 2008), *Arabidopsis thaliana* (Ossowski et al., 2008) and rice (Subbaiyan et al., 2012).

Walnut (*Juglans regia* L.), an economically important woody plant in *Juglandaceae* family, has been widely cultivated for its valuable seed, which has abundant polyunsaturated fatty acids and proteins (Savage et al., 1999). Furthermore, other by-products derived from the walnut tree, such as shells, bark, and leaves, have been used in both pharmaceutical and cosmetic industries (Stampar et al., 2006). Although walnut has been cultivated for centuries, only a few systemic molecular studies on walnut have been reported. In the past, studies were carried out to assess genetic diversity of walnut by employing molecular markers, such as RFLP (Fjellstrom et al., 1994), RAPD (Nicese et al., 1998), and AFLP (Bayazit et al., 2007). More recently, the expressed sequence tags have been utilized for identifying SSR markers (Zhang et al., 2010). However, information on the genetic structure and diversity of walnut is still limited. In this study, we aim to develop and characterize a collection of SNPs and InDels in *J. regia* by partial genome sequencing, which would be useful for gene discovery, marker-assisted selection, and other breeding applications.

Results

Sequencing and assembly

Using the 454 GS-FLX sequencer (Roche, Inc.), a total of 541,176 reads were generated, in total bases of 194,024,314 bp. The average read length was 359 bp. The Newbler *de novo* Assembler (Roche, Inc. version 2.8) successfully assembled 39.22% (212,229) of all reads (Table 1) into 31,362

Table 1. Statistics of sequence reads generated by the 454 GS-FLX sequencer.

No. of reads	Average read length (bp)	Total bases (bp)	No. of reads in contigs	No. of reads in singletons
541,176	359	194,024,314	212,229	320,318

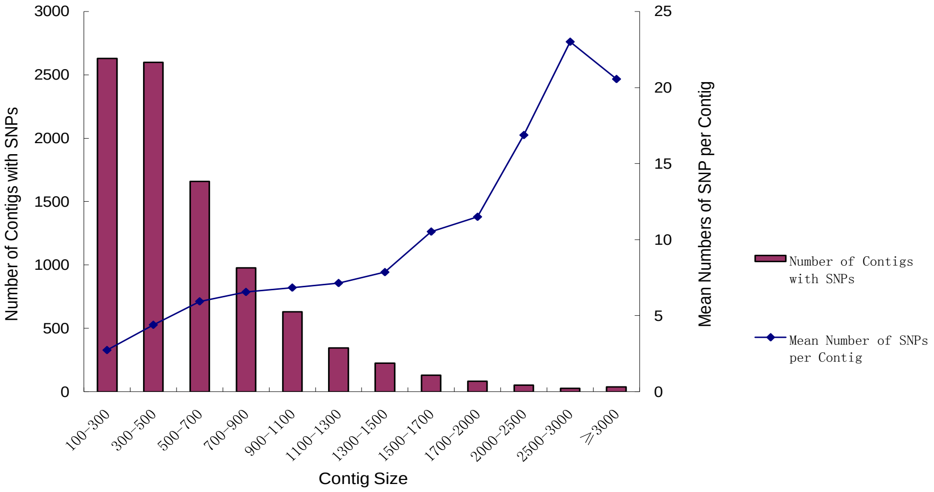

Fig 1. Distribution of numbers of contigs containing SNPs and mean numbers of SNP per contig in different size intervals.

contigs (≥100 bp) comprising 15,071,048 bp nucleotides (Table 2), and left 320,318 reads remain as singletons (Table 1). The Q40 (one miscall every 10,000 bases) plus bases showed 70.93% (6,726,255 bp) reliability among all contigs and singletons, suggesting a high accuracy of sequencing. There were a total of 9,707 contigs longer than 500 bp, which were classified as large contigs, accounting for 30.95% of all the contigs. Among the large contigs, the longest was 47,928 bp, with an average length of 957 bp, and the N50 size was 964 bp (Table 2).

Sequence alignment and SNP detection

All reads were mapped to the assembled contigs using Newbler GS Reference Mapper (Roche, Inc.). A total of 48,165 candidate SNPs were identified in 9,404 contigs, accounting for 29.99% of all contigs. The average SNP frequency was one SNP per 313 bp. Figure 1 showed the abundance of contigs containing SNPs. The majority of SNPs (90.18%) were detected in contigs ranging from 100 bp to 1700 bp. However, the number of SNPs per contig increased with contig size, indicating that larger datasets with greater contig sizes could be used to identify more SNPs.

Analysis of base substitutions

The distribution of substitution types was shown in Table 3. A greater number of transitions (35,480) than transversions (12,685) were identified, and the ratio between transitions and transversions was 2.79:1. Among the transitions, the number of C/T transitions was a little greater than that of G/A, while for transversions, C/G transversions were more infrequent than other three types.

Analysis of InDels

A total of 1,037 InDels were detected in the *J. regia* genome, including 492 insertions with length variation from 1 bp to 51 bp, and 545 deletions with length ranging from 1 bp to 46 bp. The most abundant InDels were mononucleotides, which accounted for 52.65% of total InDels. There were more deletions than insertions, and the frequency of detected InDels decreased rapidly with an increase in their lengths (Table 4). The statistics of the InDels indicated that there was a bias toward A and T nucleotides for both mononucleotide and longer InDels. There was also a low frequency of CG InDels compared with that of the other dinucleotide InDels (Table 5).

Discussion

SNP abundance

With the availability of whole-genome sequence for many woody species, SNP markers have begun to be applied to the genetic analysis of forest plants, which has illustrated their great potential (Ingvarsson, 2005). In our study, we detected 48,165 SNPs among 15,071,048 bp of walnut genome, with an average SNP density of one SNP every 313 bp. The abundance found in rice was one SNP per 678bp (Subbaiyan et al., 2012), and there was 1 SNP/1.4kb in the genome of *Arabidopsis* (Ossowski et al., 2008), and it was much lower than the density we found in *J. regia*. There were studies that compared the nucleotide diversity between woody and non-woody plant genomes based on the abundance of SNPs. For example, Brown et al. (2004) found that the nucleotide diversity of loblolly pine (*Pinus taeda*) was lower than that of maize; Savolainen and Pyhäjärvi (2007) found a lower nucleotide diversity in woody plants compared with *Arabidopsis* and rice, and suggested that it might be due to the disequilibrium in tree populations and historical changes in population size. However, a result obtained in this study was contrary to the above conclusion, but the explanation remained unresolved.

Analysis of base changes

It is known that nucleotide substitutions in the genome can be affected by various factors such as the precision of DNA duplication, the fidelity of reparation, and other physical and

Table 2. The assembled results of trimmed reads.

Large Contig (Length≥500bp)							All contigs (Length≥100bp)		
No. of Contigs	Bases(bp)	ACS[1](bp)	N50 size[2](bp)	LCS[3](bp)	Q40 Plus Bases[4](bp)	Q39 Minus Bases[5](bp)	No. of Contigs	Bases(bp)	ACS[1](bp)
9,707	9,483,541	957	964	47,928	6,726,255	2,757,286	31,362	15,071,048	481

Note:[1]Average contig size; [2]the value such that at least half of the genome is contained in contigs of size or larger; [3]Large contig size; [4]Bases that the single base quality scores≥40; [5]Bases that the single base quality scores < 40.

Table 3. Distribution of substitution types in SNPs detected in the *Juglans* genome.

Type of substitution	Number	Percentages (%)
Transitions		
A/G	17,609	36.56
C/T	17,871	37.10
Sub-total	35,480	73.66
Transversions		
A/C	3,414	7.09
A/T	4,511	9.36
C/G	1,414	2.94
G/T	3,346	6.95
Sub-total	12,685	26.34
Total	48,165	100

Table 4. Distribution of InDels in the *Juglans* genome.

InDel Size	Insertions	Percentage (%)	Deletions	Percentage (%)
1	220	21.22	326	31.44
2	143	13.79	116	11.19
3	51	4.92	36	3.47
4	22	2.12	16	1.55
5	11	1.06	10	0.95
6	10	0.96	8	0.76
7	3	0.29	2	0.19
8	2	0.19	4	0.39
9	2	0.19	4	0.39
≥10	28	2.7	23	2.23
Total	492	47.44	545	52.56

Table 5. Frequency of mononucleotide and dinucleotide InDels detected in the *Juglans* genome.

InDel	Number	InDel	Number	InDel	Number	InDel	Number
A	194	C	86	G	97	T	169
AA	22	CA	21	GA	13	TA	33
AC	19	CC	8	GC	6	TC	10
AG	22	CG	4	GG	7	TG	26
AT	35	CT	13	GT	10	TT	27

chemical factors (Timsit, 1999). In this study, there were 35,480 (73.66%) transitions, which were far more abundant than transversions (12,685). The phenomenon that transition bias has been previously reported in both plant and animal genomes (Zhang and Zhao, 2004). Wakeley (1996) has attributed this bias to a conformational advantage that results from a better tolerance of transitional mutations over transversions during natural selection, because transitions are more likely to preserve the structure of a protein than transversions. Among the transitions, there were more C/T than A/G changes, which was similar to the findings from *Arabidopsis* (Ossowski et al., 2008) and grape (Lijavetzky et al., 2007). The higher frequency of C/T transitions may be caused by nucleotide methylation (Coulondre et al., 1978). Among the transversions, A/T was the most abundant, while C/G was rarest. Similar results were found in rice (Subbaiyan et al., 2012) and Citrus (Terol et al., 2008). However, C/G transversions were the most abundant in wheat (Lai et al., 2012).

Distribution and bias of InDels

Johnson (2004) suggested that InDels were mostly located in non-coding regions, and their sizes were variable. Among the 1,037 InDels deteced in this study, mononucleotides were the most frequent. This phenomenon was also observed in *Escherichia coli* (Mo et al., 1991) and the chloroplast non-coding sequence of nine monocotyledonous plants (Golenberg et al., 1993). Among all InDels, there was a higher frequency of deletions than insertions; this was in accordance with the study on genomic InDels in 19 mammalian species (Fan et al., 2007). This deletion bias is expected on the basis of the thermodynamics of replication slippage, in which an insertion requires the melting and replication of a segment of previously duplicated DNA, whereas deletions involve only a skipping of unreplicated bases (Petrov, 2002a). The number of insertions and deletions both decreased sharply with the increasing of InDel length, which could be represented with a power law equation (Fan et al., 2007). The similar results had been reported in humans, primates, and chloroplast genomes of nine monocotyledonous species (Fan et al., 2007). InDels can be produced by errors in DNA synthesis, repair, recombination, or may be caused by the insertion and excision of transposable elements that often leave a characteristic DNA legacy of several bases, the longer bases would make the

InDels to form more difficultly (Bhattramakki et al., 2002). In our study, InDels showed a clear A/T preference. This might be occurred during the formation of InDels, bases C and G have a higher stability, and thus it is more difficult to be changed than bases A and T.

Materials and Methods

Plant material

In April 2010, fresh leaves of *J. regia* were collected on the campus of Nanjing Forestry University.

DNA extraction

Genomic DNA was extracted using a DNeasy Plant Mini kit (Qiagen, Shanghai, China). The quality and quantity of DNA were assessed by electrophoresis on 1.0% agarose gels and on a Nanodrop 2000 spectrophotometer (Thermo, USA).

Library construction

Sequencing libraries were constructed according to the manufacturer's protocol with a Roche Rapid Library kit (Roche, Inc.). Genomic DNA (1.0 µg) was first fragmented by nebulization using compressed nitrogen gas, both ends of the DNA fragments were blunted-ended and ligated to DNA PA and PB adaptors. The adaptors provided priming sequences for both amplification and sequencing of the sample library fragments as well as the sequencing key, which was a short sequence of four nucleotides used by the system software for base calling. Following the reparation of any nicks in the double-stranded library, these priming sequences also released the unbound strand of each fragment (with 5'-Adaptor A). Small fragments less than 350 bp were removed using AMPure beads (Beckman Coulter, Brea, CA, U.S.). The quality and quantity of the libraries were assessed with an Agilent 2100 Bioanalyzer (Agilent, Waldbronn, Germany) and a TBS380 (Turner Biosystems, U.S.). The library was then diluted to 1×10^7 molecules μL^{-1} and stored at 4°C before use.

DNA sequencing and assembly

454 pyrosequencing was performed with a XLR 70 Titanium sequencing kit according to the manufacturer's protocol (Roche, Inc.) on a 454 GS-FLX Sequencer (Roche, Inc.). Sequence assembly was performed using Newbler v2.8 software (Roche, Inc.) with the 'het' (heterozygous genome) and 'large' (large genome) options and default assembly parameters.

SNP and InDel discovery and analysis

For SNP and InDel identification, all assembled contigs were used as reference sequences. To detect nucleotide variations, the sequence reads were aligned to the reference sequences using GS Reference Mapper version 2.8 (Roche, Inc.). The detection parameters were as follows: minimum read length 20 bp, seed step 12 bp, seed length 16 bp, seed count 1 bp, hit-per-seed limit 70 bp, minimum overlap length 50 bp, minimum overlap identity 95%, alignment identity score 2, alignment difference score −3, and repeat score threshold 12. Candidate sequence variations were further filtered according to the following criteria: (1) at least three non-duplicated reads shared the polymorphism and (2) no other changes were detected within 5 bp on either side of the candidate sequence variations (Ma et al., 2014). The density of SNP is characterized by the formula: $D = N / L$. Where D is the density of SNP; N is the number of SNPs detected in walnut genome; and L is the length of assembled contigs.

Conclusion

In this study, we have discovered and charaterized 48,165 SNPs and 1,037 InDels in *J. regia* genome by using the 454 pyrosequencing technology. These SNPs and InDels will provide a valuable marker resource for many aspects of genetic studies for *J. regia*, including genetic diversity analysis, high resolution genetic map construction, positional cloning and so on. This study also helps to improve our understanding of the nucleotide variations in *J. regia* genome.

Acknowledgments

Funding for this work was provided by the Key Forestry Public Welfare Project (201304102), the Natural Science Foundation of China and Jiangsu Province (31270711, BK20130968). This study was also supported by the Program for Innovative Research Teams in Universities of Jiangsu Province and the Educational Department of China, and the Priority Academic Program Development (PAPD) Program at Nanjing Forestry University.

References

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Minning for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol. 132: 84–91.

Bayazit S, Kazan K, Gülbitti S, Cevik V, Ayanoğlu H, Ergül A (2007) AFLP analysis of genetic diversity in low chill requiring walnut (*Juglans regia* L.) genotypes from Hatay, Turkey. Sci Hortic. 111(4): 394–398.

Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC, Tingey SV, Rafalski A (2002) Insertion/deletion polymorphisms in 3'-regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol Biol. 48: 539–547.

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

Brown GR, Gill GP, Kuntz RJ (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci USA. 42: 15255–15260.

Coulondre C, Miller JH, Farabaugh PJ, Philip J, Walter G (1978) Molecular basis of base substitution hot spots in *Escherichia* coli. Nature. 274: 775–780.

Duran C, Appleby N, Vardy M, Imelfort M, Edwards D (2009) Single nucleotide polymorphism discovery in barley using autoSNPdb. Plant Biotechnol J. 7: 326–333.

Edwards D, Forster J W, Chagné D, Batley J (2007) What are SNPs? In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (ed) Association mapping in plants, 3rd edn. Springer, New York

Fan YH, Wang WJ, Ma GJ (2007) Patterns of insertion and deletion in mammalian genomes. Curr Genomics. 8: 370–378.

Fjellstrom R G, Parfitt D E (1994) Walnut (*Juglans* spp.) genetic diversity determined by restriction fragment length polymorphisms. Genome. 37(4): 690–700.

Gibbs R, Weinstock G, Kappes S, Schook L, Skow L, Womack J (2002) Bovine genomic sequencing initiative: Cattle-izing the human genome. Bovine sequencing white paper. 1–12 pp.

Golenberg EM, Clegg MT, Durbin ML, Doebley J, Ma DP (1993) Evolution of a noncoding region of the chloroplast genome. Mol Phylogenet Evol. 2: 52–64.

Hayashi K, Yoshida H, Ashikawa I (2006) Development of PCR-based allele-specific and InDel marker sets for nine rice blast resistance genes. Theor Appl Genet. 113: 251–266.

Ingvarsson P (2005) Nucleotide Polymorphism and Linkage Disequilibrium Within and Among Natural Populations of European Aspen (*Populus tremula* L., Salicaceae). Genetics. 169: 945–953.

Johnson KP (2004) Deletion bias in avian introns over evolutionary timescales. Mol Biol Evol. 21(3): 599–602.

Lai K, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U, Zander M, Mason AS, Batley J, Edwards D (2012) Single nucleotide polymorphism discovery from wheat by next-generation sequence data. Plant Biotechnol J. 10(6): 743–749.

Lijavetzky D, Cabezas JA, Ibáñez A, Rodríguez V, Martínez-Zapater JM (2007) High-throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics. 8(1): 424.

Ma Q, Feng K, Yang W, Chen Y, Yu F, Yin T (2014) Identification and characterization of nucleotide variations in the genome of *Ziziphus jujuba* (*Rhamnaceae*) by next generation sequencing. Mol Biol Rep. 41(5): 3219–3223.

Mo JY, Maki H, Sekiguchi M (1991) Mutational specificity of the dnaE173 mutator associated with a defect in the catalytic subunit of DNA polymerase III of *Escherichia coli*. J Mol Biol. 222: 925–936.

Nicese FP, Hormaza JI, McGranahan GH (1998) Molecular characterization and genetic relatedness among walnut (*Juglans regia* L.) genotypes based on RAPD markers. Euphytica. 101(2): 199–206.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. 18: 2024–2033.

Petrov DA (2002) Mutational equilibrium model of genome size evolution. Theor Popul Biol. 61: 531–544.

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 409:928–933.

Savage GP, McNeil DL, Osterberg K (1999) Oxidative stability of walnuts during long term in shell storage. Paper presented at the 4th International Walnut Symposium, NewYork, 591–597 September 1999.

Savolainen O, Pyhäjärvi T (2007) Genomic diversity in forest trees. Curr Opin Plant Biol.10:162–167.

Stampar F, Solar A, Hudina M, Veberic R, Colaric M (2006) Traditional walnut liqueur–cocktail of phenolics. Food Chem. 95(4): 627–631.

Steele KA, Ogden R, McEwing R, Briggs H, and Gorham J (2008) InDel markers distinguish Basmatis from other fragrant rice varieties. Field Crop Res. 105: 81–87.

Subbaiyan GK, Waters DLE, Katiyar SK, Sadananda AR, Vaddadi S, Henry RJ (2012) Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. Plant Biotechnol J. 10: 623–634.

Terol J, Naranjo MA, Ollitrault P, Talon M (2008) Development of genomic resources for *Citrus clementina*: characterization of three deep-coverage BAC libraries and analysis of 46000 BAC end sequences. BMC Genomics. 9(1): 423.

Timsit Y (1999) DNA structure and polymerase fidelity. J Mol Biol. 293:835–853.

Vasemagi A, Gross R, Palm D, Paaver T, Primmer CR (2010) Discovery and application of insertion–deletion (INDEL) polymorphisms for QTL mapping of early life-history traits in Atlantic salmon. BMC Genomics. 11(1): 156.

Wakeley J (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Tree. 11:158–162.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science. 280(5366): 1077–1082.

Wiedmann RT, Smith TP, Nonneman DJ (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet. 9(1): 81.

Zhang F, Zhao Z (2004) The influence of neighboring-nucleotide composition on single nucleotide polymorphisms (SNPs) in the mouse genome and its comparison with human SNPs. Genomics. 84:785–795.

Zhang R, Zhu A D, Wang X J, Yu J, Zhang HR, Gao JS, Cheng YT, Deng XX (2010) Development of *Juglans regia* SSR markers by data mining of the EST database. Plant Mol Biol Rep. 28(4): 646–653.