Journal of Computer Science & Systems Biology - Open Access
Research Article JCSB/Vol.1 2008

Bioinformatics Analysis of Distribution of Microsatellite Markers (SSRs) / Single Nucleotide Polymorphism (SNPs) in Expressed Transcripts of *Prosopis Juliflora*: Frequency and Distribution

Sablok G* and N.S.Shekhawat

Computational Unit, Biotechnology Center, Jai Narain Vyas University, Jodhpur-342033

*Corresponding author: Sablok G, Computational Unit, Biotechnology Center, Jai Narain Vyas University, Jodhpur-342033, E-mail: sablokg@gmail.com

Received September 27, 2008; Accepted November 20, 2008; Published December 26, 2008

Citation: Sablok G, Shekhawat NS (2008) Bioinformatics Analysis of Distribution of Microsatellite Markers (SSRs) / Single Nucleotide Polymorphism (SNPs) in Expressed Transcripts of *Prosopis Juliflora*: Frequency and Distribution. J Comput Sci Syst Biol 1: 087-091. doi:10.4172/jcsb.1000008

Abstract

The present paper aims to identify the distribution of Microsatellite markers (SSRs)/Single Nucleotide Polymorphism (SNPs) as resource tools for the analysis of interspecies hypervariability in *Prosopis spp*. Microsatellites are ubiquitously repeated extension of 1-5 bp motif extended throughout the genome. The dbEST division of Genbank contains 1467 ESTs of *Prosopis juliflora* which have been utilized for the development of genic microsatellite markers (SSRs). Locally installed assembling program was used for the cluster analysis of the EST. Analysis was performed on a Linux Cluster system. The analysis of the putative unigenes has been shown to have most abundant motif of A/T followed by dinucleotide AG/CT and trinucleotide repeat AAG/CTT. EST-derived SNPs are becoming the resources for the development of SNP markers. The relative rates of development of the high throughput computational methods for the detection of SNPs (Single Nucleotide Polymorphism) and small indels (insertion / deletion) has gained wide applications in the field of the molecular markers.

Keywords: *Prosopis juliflora*; *Insilico* analysis; Expressed sequence tag (EST); Microsatellite markers; SNP markers

Introduction

Microsatellites, often abbreviated as simple sequence repeats (SSRs), are 1 to 5 bp repeat mutation prone motifs. Due to mutation prone specific and high rate of polymorphism, microsatellites (SSR) have become a modern genetic resource tool for the genetic mapping studies, developing and establishing new patterns of molecular breeding and in inferring the phylogenetic and comparative genome analysis. These repetitive stretches are distributed in the coding and non-coding region with a bias towards the non-coding regions. The variation of SSR is controlled by a complex two-level system allowing changes in the number of repeat units and changes in the repeat sequences by point mutations (Natalya, 2008). Simple Sequence Repeats have attracted relatively more attention because of their abundance in plants genome, reproducibility, high level of polymorphism, co-dominant inheritance and hypervariability. The 'hypervariability' can be documented and supported by illustrating the fact that the SSR differs in the number of repeats from genotype to genotype making them as the best suitable model genetic markers. The linkage maps of nuclear genomes are constructed through the analysis of di-, tri- or tetra-nucleotide SSRs. Single nucleotide repeats have been used in the population genetic analyses of chloroplast genomes (Powell et al., 1995). SSR are also finding applications in the field of comparative genomics where SSR of one species are finding the utilization in the (genetic mapping of the other species. Cordeiro et al., 2001; Peakall et al., 1998; Rallo et al., 2003). The frequency of the SSR reported is 1 SSR /6 kb in plant genomes (Cardle et al., 2000). Recent investigation and analysis on several plant genomes have also demonstrated that the frequencies of SSRs were significantly higher in ESTs

than in genomic DNA (Morgante et al., 2002). Bio informatics is currently acting as a resource tool for the development of SSR with the availability of the genomic sequences and Expressed Sequence Tags (ESTs) as molecular databases. Expressed Sequence Tags (ESTs) are currently the most widely sequenced nucleotide commodity in the terms of number of sequences and total nucleotide count. These provide a robust sequence resource that can be used for genome annotation, gene discovery and comparative genomics (Rudd, 2003). The concept of using cDNA as a route to expedited gene discovery was first discovered in early 1980s (Putney et al., 1983). Brenner, (1990) proposed that an obvious method for characterizing the important part of human genome will involve looking at messengers from the expressed genes thus advocating the application of high-throughput methods for transcriptome sampling. Mark Adams, (1991) first used the term EST in relation to gene discovery and the Human Genome Project. The first step in deriving a gene index from an EST set is to remove redundancy by clustering ESTs representing the same native transcripts (Kalyanaraman et al., 2003). This strategy is implicit in DNA assembly tools such as PHRAP (http://www.phrap.org/), TIGR Assembler (Sutton et al., 1996) or CAP3 (Huang and Madan, 1999) that are widely used for EST clustering. The Nucleotide Redundancy Index is calculated at the whole organism level, and represents the ratio of the total length of all clusters for that organism to the total length of all ESTs for that organism; at the extreme case, a value of 1 would indicate that every cluster was a singleton containing one EST. This measure should provide some indication of when newly sequenced ESTs are not finding new gene products. In this study we screened SSRs based upon ESTs that were publicly available for *Prosopis* species. We evaluated different motifs of mono, di-, tri-, tetra-, and pentanucleotide SSRs for variation in length and location, relative abundance, distribution for further usage as molecular markers. We have also analyzed the distribution of the SNP and the frequency of the SNP distribution.

Materials and Methods

All the *Prosopis* ESTs (1467) used are downloaded from the public database NCBI dbEST division (http://www.ncbi.nlm.nih.gov/dbEST/index.html) in May of 2008. ESTs downloaded from dbEST division (http://www.ncbi.nlm.nih.gov/dbEST/index.html) of NCBI were analyzed for the nucleotide redundancy index and for putative gene mining. The ESTs of *Prosopis* used were directly downloaded from the NCBI dbEST database. Cluster analysis was employed to define groups of overlapping EST sequences and by so doing eliminate EST redundancy while simultaneously improving the length of sequences in the local database. This type of clustering also assists in the estimation of genome coverage and SSR frequency per genome. ESTs were respectively clustered using PHRAP obtained under an academic License. The putative unigenes were saved as dual resource databases in FASTA-formatted files for SSR detection. The number of sequences per cluster varied widely. All unigene databases were used to identify and characterize SSRs using a Perl Script for SSR identification. The minimum repeat unit was defined as ten for mononucleotides, six for dinucleotides, and five for all the higher order motifs including tri-, tetra-, penta-, and hexanucleotides The FASTA-formatted sequence file was allowed to search each sequence for all possible combination of mononucleotide, dinucleotide, trinucleotide, tetranucleotide, and pentanucleotide repeats. The SNP were identified using the alignment with the consensus sequence and in house perl script was used to validate it.

Occurrence and Frequency Analysis of Different SSR Motifs and SNP Distribution

The increasing numbers of genomic and expressed sequences are acting as valuable resources for developing a new class of molecular markers. These data sets have been used to study comparative genomics, assigning gene function and in protein evolution. Comparative genome analysis requires the same sets of genes (i.e. cross-reference genes) to be mapped to chromosomes in the species compared. Thus, comparative maps with sets of expressed sequence tag (EST) derived markers (i.e. cross-species markers) are essential for comparative genome analysis. Several studies have utilized publicly available ESTs to mine simple sequence repeats (SSRs) or microsatellites markers for plants and human. The EST derived SSR markers (EST-SSRs) have proved very useful for construction of genetic and comparative maps. Cluster analyses were performed on *Prosopis* ESTs using the PHRAP program. A total of 1467 *Prosopis* ESTs analyzed fell into 953 putative Unigenes showing a redundancy of 55.62 %. The 953 putative Unigenes have a total base pair of 402624 bp were then used as a source database in FASTA-formatted text files for the for SSR identification. The fasta formatted files were allowed to search in the clustered databases of *Prosopis juliflora* leaf cDNA constructs for mono, di-, tri-, tetra-, and penta-nucleotide SSRs respectively. The results depicted 199 EST- SSRs in the 179 putative unigenes out of 953 showing an occurrence of 18.78 % with the frequency of 1 SSR/2.021 kb. The eSSR's identified in the present study is clustered into 10 different types of motifs in *Prosopis juliflora* leaf cDNA constructs. The frequency distribution of 10 different types of motifs has been shown (Figure 1).

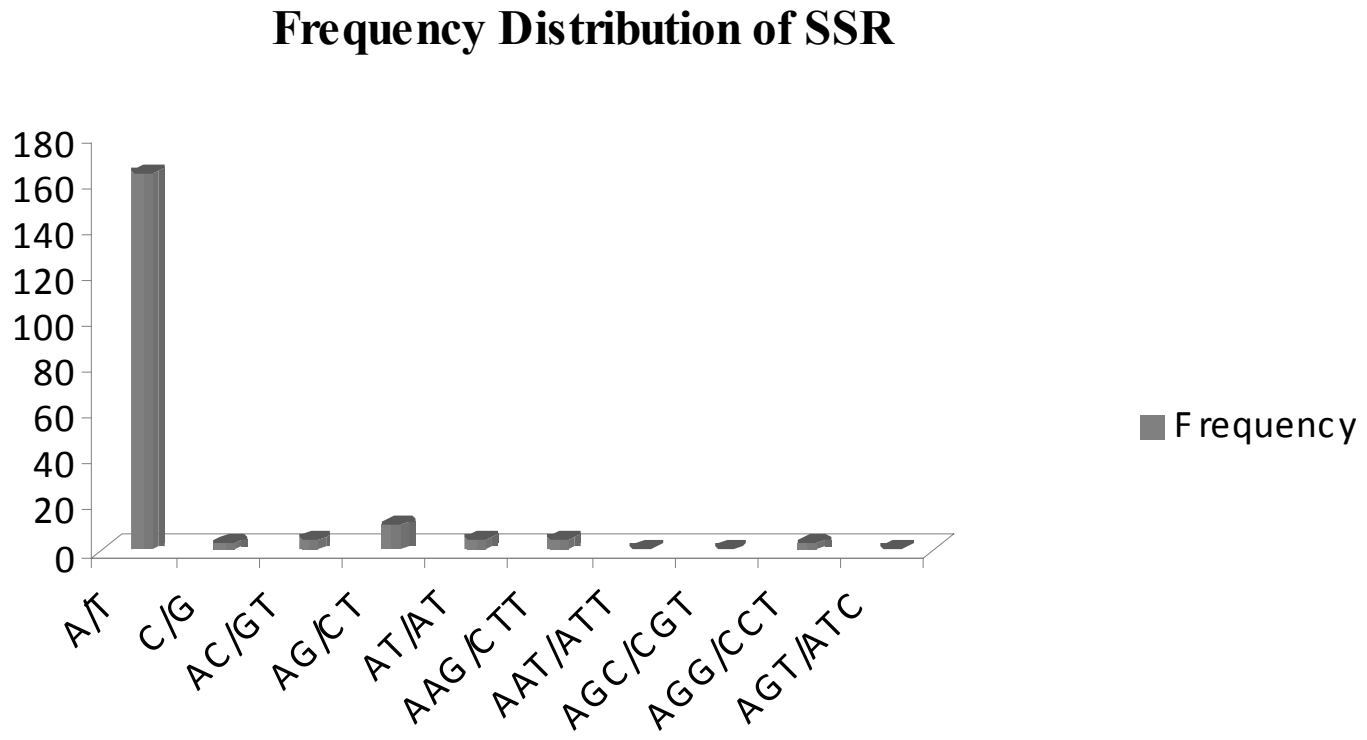

Figure 1

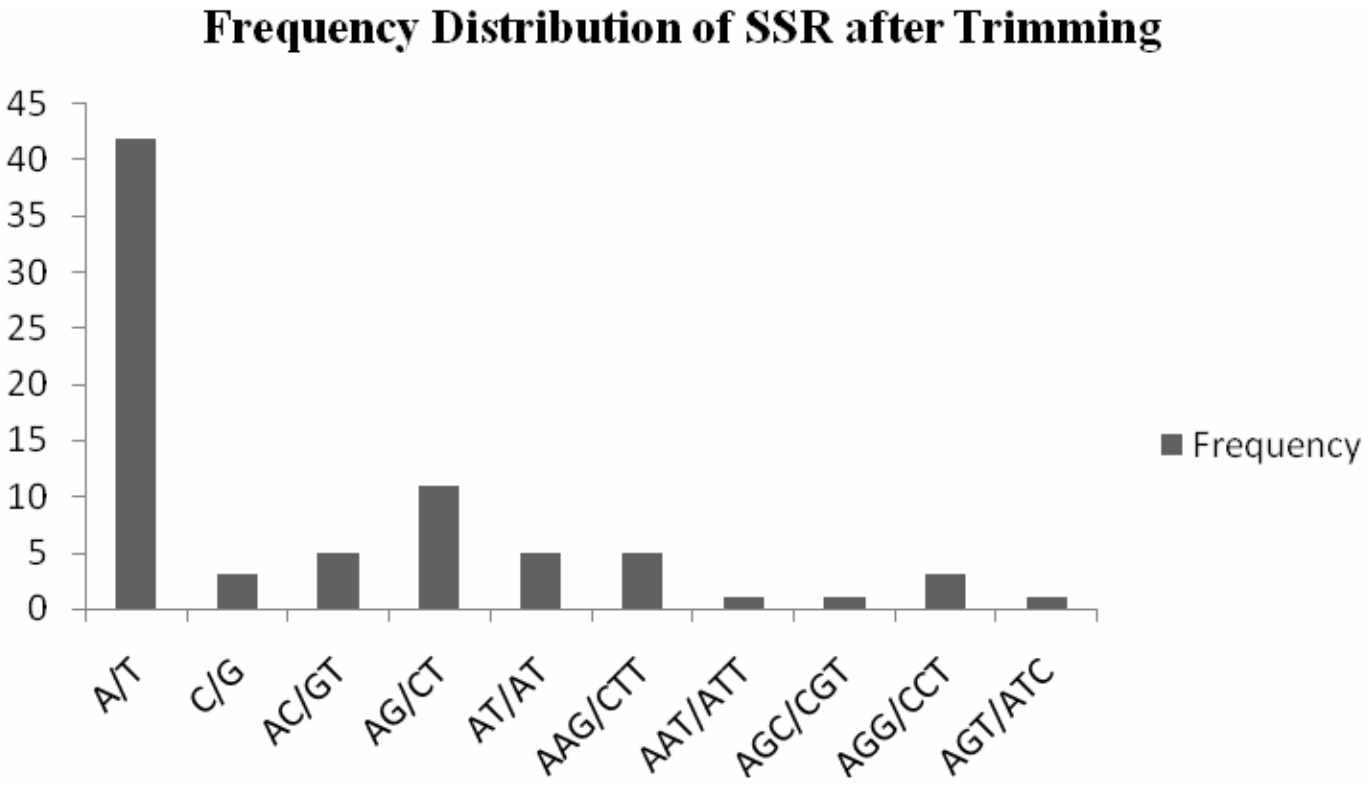

Figure 2

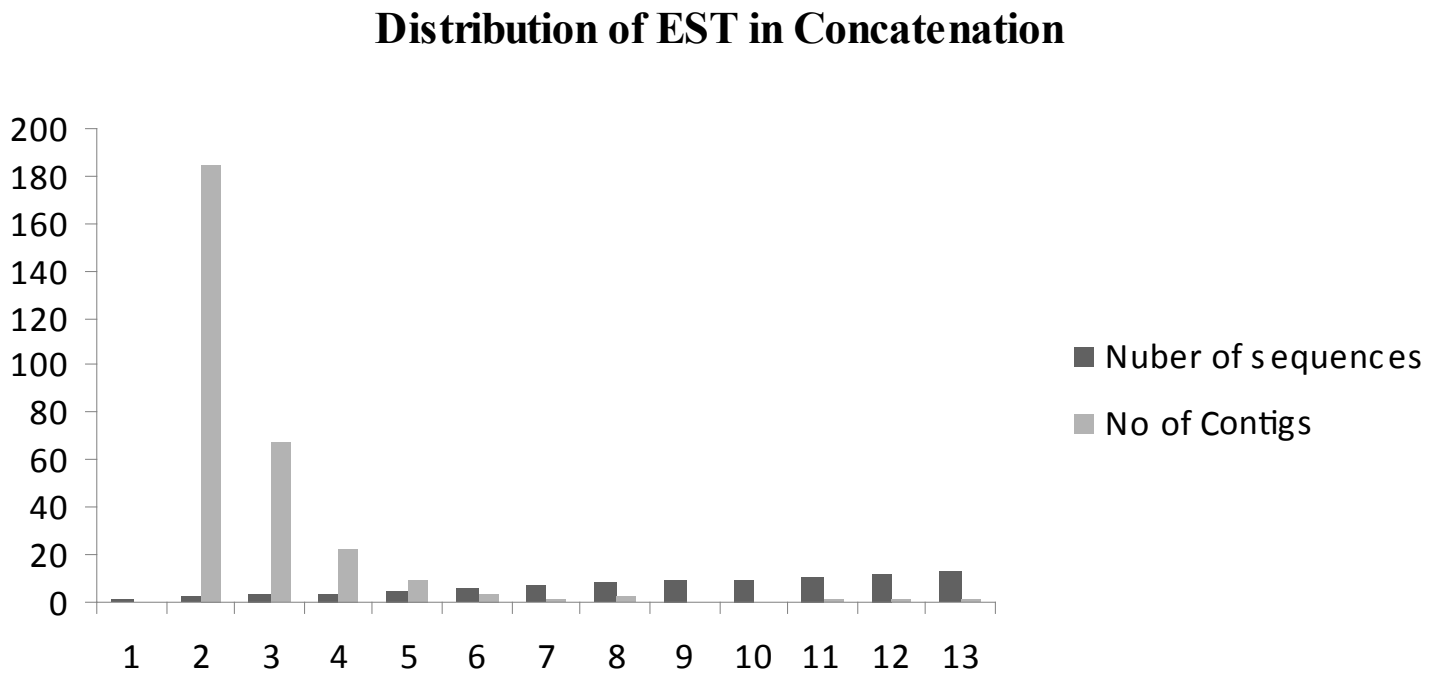

Figure 3

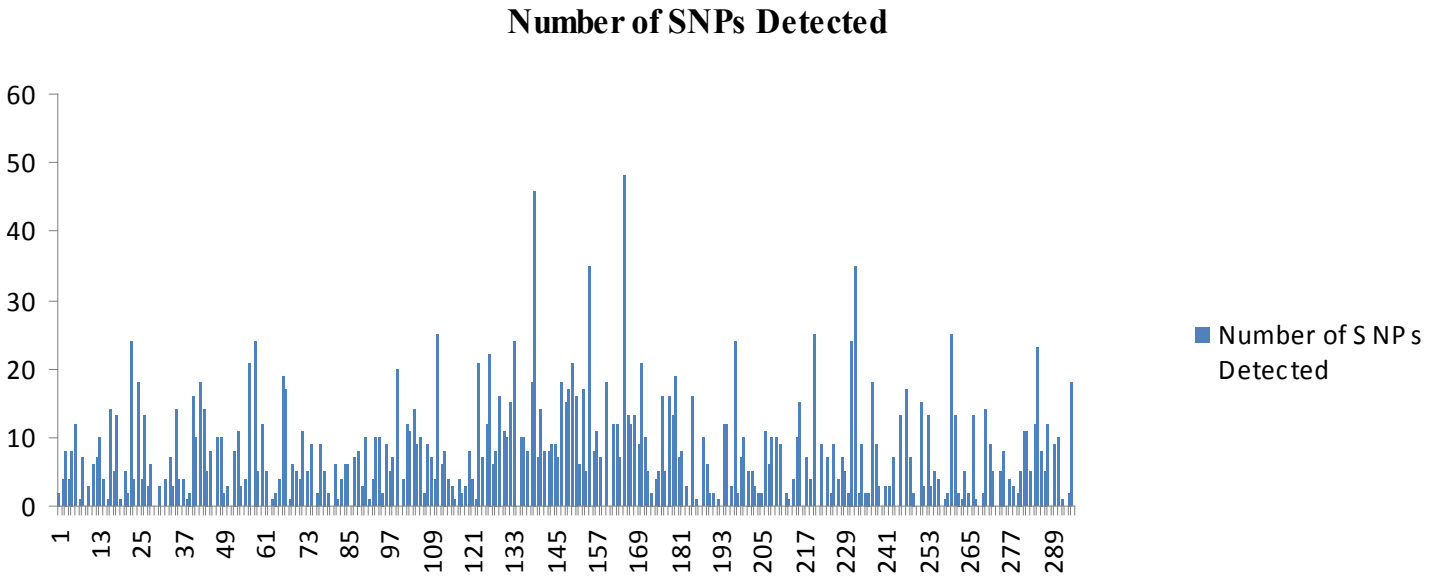

Figure 4

The analysis was again performed on the Unigene dataset for the identification of the potent SSR by removing the Poly A tail which are added during post transcriptional modifications. The Unigene dataset was trimmed using trimEST program of EMBOSS package and the results showed a remarkable decrease in the SSR frequency. Earlier the SSR frequency was 1 SSR /2.021kb but now the frequency changed to 1 SSR/5.20kb indicating the presence of the detection of the Poly A tail as Mononucleotide repeat motif. The frequency Distribution has been shown after trimming (Figure 2). The number of SSR falls from 199 to 77 showing an occurrence of 7.24% in a total 401042 bp analyzed. For the identification of the putative SNP we have used CAP3 and then on the basis of alignment with the consensus sequence we have identified 2284 SNP sites and 1008 indel polymorphisms with frequency 1.60 SNPs / 100 bp. The reported frequency of SNP has also shown the smallest contig containing the concatenation of the two EST .The relative rate of the transition and transversion were 589 and 687 respectively. To identify the SNP frequency the concatenation of the 795 sequences were done to generate a consensus sequence of 142711 bp. The distribution of the EST concatenation is shown (Figure 3). The SNP distribution is shown (Figure 4).

Conclusion

This work is done to implement the clustering algorithm on the identification of the SSR in the *Prosopis juliflora spp.* The work described the identification of the SSR and the removal of the Poly A tail in the identification of true mononucleotide tracts. The work also helps in the identification of SNP as a resource tool for further usage as a molecular marker.

Acknowledgements

We thank Biotechnology Center, Jai Narain Vyas University for providing the computational facilities. This work was supported by a grant from Department of Science and Technology, Rajasthan, India.

References

1. Brenner S (1990) The human genome: the nature of the enterprise. CIBA Found Symp 149: 6-17. » Pubmed » Google Scholar
2. Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, et al. (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156: 847-854. » CrossRef » Pubmed » Google Scholar
3. Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to *erianthus* and *sorghum*. Plant Sci 160: 1115-1123. » CrossRef » Pubmed » Google Scholar
4. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868-877F. » CrossRef » Pubmed » Google Scholar
5. Kalyanaraman A, Aluru S, Kothari S, Brendel V (2003) Efficient clustering of large EST data sets on parallel computers. Nucleic Acids Res 31: 2963-2974. » CrossRef » Pubmed » Google Scholar
6. Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genet 30: 194-200. » CrossRef » Pubmed » Google Scholar
7. Natalya S (2008) Plant simple sequence repeats: distribution, variation, and effect on gene expression. Genome 51: 79-90.
8. Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A (1998) Cross-species amplification of Soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol Biol Evol 15: 1275-1287. » CrossRef » Pubmed » Google Scholar
9. Powell W, Morgante M, McDevitt R, Vendramin G, Rafalski J (1995) Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines. Proc Natl Acad Sci USA 92: 7759-7763. » CrossRef » Pubmed » Google Scholar
10. Putney SD, Herlihy WC, Schimmel P (1983) A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. Nature 302: 718-721. » CrossRef » Pubmed » Google Scholar
11. Rallo P, Tenzer I, Gessler C, Baldoni L, Dorado G, et al. (2003) Transferability of olive microsatellite loci across the genus Olea. Theor Appl Genet 107: 940-946. » CrossRef » Pubmed » Google Scholar
12. Rudd S, Mewes HW, Mayer KF (2003) Sputnik: a database platform for comparative plant genomics. Nucleic Acids Res 31: 1280-132. » CrossRef » Pubmed » Google Scholar
13. Sutton G, White O, Adams M, Kerlavage A (1996) TIGR Assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1: 9-19. » Google Scholar