ONLINE RESOURCES

Computational identification and analysis of single-nucleotide polymorphisms and insertions/deletions in expressed sequence tag data of *Eucalyptus*

TIRATHA RAJ SINGH[1], ARUN GUPTA[2,3], AYKKAL RIJU[6], M. MAHALAXMI[4], ABHIK SEAL[5] and V. ARUNACHALAM[6,7]*

[1] *Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Waknaghat, Teh Kandaghat, Solan 173 234, India*
[2] *School of Computer Science and Information Technology, Devi Ahilya Vishwavidyalaya (DAVV), Indore 452 013, India*
[3] *Computational Biology Group, Abhyudaya Technologies, Indore 452 003, India*
[4] *Department of Biotechnology and Bioinformatics, Kuvempu University, Jnaneshwari, Shankaraghatta 577 451, India*
[5] *DOEACC Society, Jadavpur University Campus, Kolkata 700 032, India*
[6] *Molecular Biology and Bioinformatics Laboratory, Central Plantation Crops Research Institute, Indian Council of Agricultural Research, Kasaragod 671 124, India*
[7] *ICAR Research Complex for Goa, Ela, Old Goa 403 402, India*

[Singh T. R., Gupta A., Riju A., Mahalaxmi M., Seal A. and Arunachalam V. 2011 Computational identification and analysis of single-nucleotide polymorphisms and insertions/deletions in expressed sequence tag data of *Eucalyptus*. *J. Genet*. **90**, e34–e38. Online only: http://www.ias.ac.in/jgenet/OnlineResources/90/e34.pdf]

Introduction

Molecular markers are widely employed in plant research and breeding as genes of scientific and agronomic importance can be isolated solely on the basis of their position on the genetic map. In plant genetic research, molecular markers are also being used for population and evolutionary studies (Bomblies and Weigel 2007). Single nucleotide polymorphism (SNP) markers have gained much interest, becoming the marker of choice in genetic analysis in the scientific and breeding community (Gupta *et al.* 2001). For myriad crop plant species, large numbers of expressed sequence tags (ESTs) have been generated which could serve as valuable source for SNP identification and analysis. The emergence of many novel molecular markers is changing and accelerating the process of analysing mutations in plant molecular biology research. SNPs represent the most frequent type of genetic polymorphism and thus provide a high density of markers near the locus of interest. SNPs are highly stable, reliable and have a fine resolution (Syvanen 2001). The development of high throughput methods for the detection of SNPs and small indels (insertion/deletion) has led to a revolution in their use as molecular markers (Novaes *et al.* 2008).

Here we report the results of computational identification and analysis of SNPs in expressed sequence data of *Eucalyptus*. *Eucalyptus* is an important short rotation pulpy woody plant, grown widely in tropics. Many genomic programmes are underway, worldwide, for linkage and physical mapping of *Eucalyptus* which will be useful for the scientific community to address several genetic improvement in *Eucalyptus*. *Eucalyptus* is a native Australian genus comprises of more than 700 species. It is one of the world's main sources of 32 woody biomass and is the main hardwood used for plywood and timber with more than 19 million hectares spread over 37 countries, acclaiming for 16% of forest plantation areas (FAO 2000).

The other properties that make *Eucalyptus* a perfect model for adaptability and ecological conditions are their good quality wood fiber, essential oil synthesis, resistance to insects and tolerance to abiotic stresses such as salt or cold (Teulières *et al.* 2007). Cold acclimation, one of the adaptive responses, occurs in many plant species such as *Eucalyptus* and is expected to be more complex in woody species than in herbaceous plants (Savitch *et al.* 2002). *Eucalyptus* was shown to develop specific gene regulation suggesting its ability as a model for studying the adaptive biology of trees (El Kayal *et al.* 2006b). Genomic resources are being developed for *Eucalyptus* wood related unigenes viz., simple sequence repeats and single nucleotide polymorphisms

*For correspondence. E-mail: v.arundevi@gmail.com.

Keywords. ESTs; indels; SNPs; nucleotide diversity; Shannon index; molecular markers; *Eucalyptus*.

(Rengel *et al.* 2009) and a database is available in public domain as EUCAWOOD (http://www.polebio.scsv.ups-tlse.fr/eucalyptus/eucawood/).

SNPs analysed in *Eucalyptus* are associated with wood properties of an important trait such as *microfibril angle* (MFA), cellulose content (Thumma *et al.* 2009) and related biochemical pathways of four species of *Eucalyptus* (Külheim *et al.* 2009). In the assembled EST sequences of 148 Mbp of *Eucalyptus grandis*, 23,742 SNPs were predicted (Novaes *et al.* 2008). Our objective in this work was to mine the *Eucalyptus* transcriptome for SNP and indel variation and their types, nucleotide diversity and to provide information as public domain database. A database of all identified SNPs based on tissue types has been developed which is available for academic use at http://www.bioinfoindia.org/eusnpdb.

Materials and methods

Comparative bioinformatics analyses were conducted to assign functional annotation to the available sequences with the aim of identifying all possible SNPs in *Eucalyptus* species. All patent sequences were removed from the analysis as most of the *Eucalyptus* projects generally involve private companies due to its commercial importance.

EST sequences in FASTA format were retrieved from dbEST (Boguski *et al.* 1993), which contained 35,320 sequences from six tissues namely mesophyll leaves, differentiating xylem, flower, shoot apex, woody tissue and root. Candidate SNPs were detected using the method given in Batley *et al.* (2003). In brief, contigs were scanned for putative polymorphism candidates. Spacing characters (-) added during sequence alignment were considered as a fifth element in addition to A, C, G and T.

When several SNPs are present in an alignment, a redundant co-segregation score is calculated for each SNP. This is measured as the frequency of that SNP pattern occurring among each of the SNPs identified in the alignment. This figure is then normalized to the number of sequences and number of SNPs detected in the alignment to produce a standard co-segregation score. Minimum similarity threshold of 95% and minimum overlap of 100 bases were specified in the Cap3 program (Huang and Madan 1999).

We used Shannon index (Shannon 1948) towards the analysis of the distribution of SNPs/indels among 10 possible categories. Frequency of each of the 10 types of SNPs/indels sites was scored, from which proportion of occurrence (Pi) of each type (transition/transversion/indel) to the total SNPs/indels in each tissue library was calculated. Shannon index estimate analysis was performed according to Riju *et al.* (2007).

Measuring nucleotide diversity can reveal differences in selection pressure acting on different genomic sequences. Relative measurement of nucleotide diversity (β) and its analysis were performed according to Novaes *et al.* (2008). In brief, β is estimated by the equation:

$$\beta = \frac{\left[\frac{S+1}{L}\right]}{\sum_{i=1}^{D-1} (1/i)}$$

Where S is the number of SNPs detected in the contigs, L is the contig sequence length and D is the sequence depth estimated by the average number of read and aligned to each nucleotide position during contigs assembly.

Results and discussion

There is evidence for occurrence of SNPs with variations among different species of *Eucalyptus* and many other plant species. To name a few, SNPs with the frequencies (number per 100 bp) in the range 3.83–7.3 were found among four species of *Eucalyptus* (Külheim *et al.* 2009). Similar results are reported now and also earlier in other studies too, such as in beetroot (Schneider *et al.* 2001), maize (Batley *et al.* 2003), oil palm (Riju *et al.* 2007), and citrus (Dong *et al.* 2010) with frequencies 0.77, 1, 1.36 and 0.61, respectively. Such studies discovered variation in diversity among these species and suggest for comprehensive analysis.

We found a total 33,466 SNP sites and 5874 indel polymorphisms in 26,026 ESTs analysed. Results of the tissue wise SNP and indel discovery are given in table 1. Among the six tissues from which the EST libraries have been generated, shoot apex has the highest frequency of 2.89 SNPs per 100 bp, whereas woody tissue has the lowest frequency of 0.65 SNPs per 100 bp (table 1). Candidate SNPs were categorized according to nucleotide substitution as either transition (C↔T or G↔A) or transversion (C↔G, A↔T, C↔A or T↔G). We found transversions (14,026) as slightly higher than transitions (13,666) in available *Eucalyptus* transcriptome. However, considering the individual substitutions bias, the transition type substitutions G↔A (7194) and C↔T (6472) were found to be slightly lesser than transversion type substitutions (table 1).

In general, transitions occur at a higher frequencies than transversions such as beetroot (Schneider *et al.* 2001), maize (Batley *et al.* 2003) and oil palm (Riju *et al.* 2007). A recent study on grasshopper genome revealed that the majority of transitions of cytosine residues are at methylated sites (CpG dinucletoide). After accounting for those methylation effects, there was no significant difference observed among the transition and transversion rates reported in this study (Keller *et al.* 2007). Recent report on SNP analysis revealed higher transversion sites than transitions sites, Ts/Tv ratio < 1 (more transversions than transitions) was seen in regulatory genes such as endonuclease reverse transcriptase and Tc1-like transposase (Hale *et al.* 2009). Transversions were seen at higher frequencies than transitions in ginger (*Zingiber officinale* Rosc) EST-SNPs (Chandrasekar *et al.* 2009). The

Table 1. *Eucalyptus* SNP data.

Results	Mesophyll leaves	Differentiating xylem	Flower	Shoot apex	Woody tissue	Root	Total
Total no. of ESTs	23620	10276	542	778	44	60	35320
Total sequences analysed	17590	8382	32	14	6	2	26026
No. of contigs	3003	1945	15	7	3	1	4974
Total SNPs detected	24948	8242	94	169	7	6	33466
Total consensus size (bp)	2534368	1252795	7383	5841	1072	464	3801923
Frequency of SNP per 100 bp	0.98	0.66	2.63	2.89	0.65	1.29	9.1
Transitions	9945	3533	88	99	1	0	13666
C/T	4826	1542	49	55	0	0	6472
G/A	5119	1991	39	44	1	0	7194
Transversions	10175	3700	83	61	5	2	14026
A/T	2774	976	17	27	0	1	3795
C/G	2339	816	31	11	1	0	3198
G/T	2356	954	11	14	3	0	3338
A/C	2706	954	24	9	1	1	3695
Ts/Tv ratio	0.978	0.955	1.06	1.62	0.2	0.0	0.974
Indels	4828	1009	23	9	1	4	5874
A	1343	298	8	3	0	3	1655
C	1115	214	5	1	0	0	1335
T	1249	282	2	4	1	0	1538
G	1121	215	8	1	0	1	1346
Shannon index	3.114	2.976	2.373	2.510	1.203	1.792	3.083

much lower Ts/Tv ratio was observed in MA-line genomes of *Caenorhabditits elegans,* which suggests that, genome wide transversions might be more susceptible to selective purging than transitions in *C. elegans* natural populations (Denvera *et al.* 2009).

It was observed that indels occurred at a very low frequency (1.5 indel/1000 bp) in *Eucalyptus*. Indels may be produced by errors in DNA synthesis, repair, recombination or due to the insertion and excision of transposable elements that often leave a characteristic DNA footprint of several nucleotide bases. Adenine involved indels (1655) were found to be more abundant followed by thymine (1538) and the other possible indels occurring in same fashion.

Shannon information index was used to analyse the proportion of 10 possible types of SNPs/indels. ESTs from tissues of mesophyll leaves showed highest value of Shannon index (3.114) whereas woody tissue had the least value (1.203). On all available tissue types our study with respect to Shannon index analysis reveals that more genomic variation is found in genes expressed specifically in mesophyll leaves than other tissue types.

Ratio of transition to transversion (Ts/Tv) was very useful to compare the genotypes of hepatitis virus C and also to identify differences among the mitochondrial genomes of animals (Belle *et al.* 2005). Our study compare 10 possible types of SNPs/indels in a single Shannon index and such analysis can be applied in future research on EST-SNPs data. Transition to transversion ratio was calculated for all tissues involved in this study and this ratio is found to be in the range of 0.2–1.62 (table 1). Ts/Tv for root is observed to be 0 as

Table 2. Summary of diversity in EST data of *Eucalyptus.*

	Parameter β		
Tissue type	Mean	Median	Ideal range
Shoot apex	3.4×10^{-2}	2.24×10^{-2}	6.28×10^{-3}–8.56×10^{-2}
Woody tissue	9.42×10^{-3}	9.56×10^{-3}	3.50×10^{-3}–1.52×10^{-2}
Root	1.51×10^{-2}	$1. 51 \times 10^{-2}$	–
Mesophyll leaves	8.69×10^{-3}	5.16×10^{-3}	9.54×10^{-5}–9.62×10^{-2}
Differentiating xylem	5.69×10^{-3}	3.17×10^{-3}	1.81×10^{-4}–5.76×10^{-2}
Flower	2.52×10^{-2}	1.91×10^{-2}	3.59×10^{-3}–7.19×10^{-2}

Distribution summary of diversity parameter estimated for 4974 contigs. Only one contig has been generated for root.

there are no transitions found in its available data whereas maximum Ts/Tv is observed as 1.62 for shoot apex in our study. Ratio is found least in woody tissues.

As number of independent haplotypes sampled is unknown in our study, we used modified nucleotide diversity estimator β. We estimated β for each tissue for their respective contigs and in total for collective 4974 contigs used in this study (table 2). The most commonly used measure of genetic diversity theta (θ) remains overestimated than β, as the number of readings at a given SNP position is always equal or smaller than the effective number of genotypes. Diversity values (expected heterozygosity) for SNPs are generally low due to their bi-allelic nature. The average β reported here is in the range of 0.00982–0.034 (table 2), and is much lower as expected, than the average θ reported in previous studies of *Populus* (Ingvarsson 2005), loblolly pine (Gonzalez-Martinez *et al.* 2006), (Norway spruce) (Heuertz *et al.* 2006), Douglas fir (Krutovsky and Neale 2005).

All SNPs identified and analysed in this study have been made available through interconnected web pages and hyperlinked data with their respective parental sequences and contigs. A web interface has also been developed to allow users to visualize this data tissue wise and is made available at http://bioinfoindia.org/eusnpdb.

Conclusion

We have predicted and analysed several putative SNP/indel markers in EST data of *Eucalyptus*. Our study compares 10 possible types of SNP/indels in a single Shannon index and such analysis can be applied in future research on such EST-SNP data. Predicted SNPs are maintained and are presented in a custom, web-accessible database named EUSNPDB. This database will serve as a reliable resource of annotated markers in genetic diversity analysis, population genetics, phylogenetic analysis, high resolution genetic map construction and in comparative genomics.

Acknowledgements

We thank David Edward for providing AutoSNP script for automated discovery of SNPs. We thank Bioinfoindia (www.bioinfoindia.org) for providing database space and hosting this project.

References

Batley J., Barker G., Helen O'Sullivan, Edwards K. J. and Edwards D. 2003 Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* **132**, 84–91.

Belle E. M., Piganeau G., Gardner M. and Eyre-Walker A. 2005 An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene* **355**, 58–66.

Boguski M. S., Lowe T. M. and Tolstoshev C. M. 1993 dbEST–database for expressed sequence tags. *Nat. Genet.* **44**, 332–333.

Bomblies K. and Weigel D. 2007 Arabidopsis: a model genus for speciation. *Curr. Opin. Genet. Dev.* **17**, 500–504.

Chandrasekar A., Riju A., Sithara K., Anoop S. and Eapen S. J. 2009 Identification of single nucleotide polymorphism in ginger using expressed sequence tags. *Bioinformation* **4**, 119–122.

Denvera D. R., Dolan P. C., Wilhelm L. J., Sung W., Lucas-Lledo J. I., Howe D. K. *et al.* 2009 A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc. Natl. Acad. Sci. USA* **106**, 16310–16314.

Dong J., Qing-liang Y. E., Fu-Sheng W. and Li C. 2010 The Mining of citrus EST-SNP and its application in cultivar discrimination. *Agric. Sci. China.* **9**, 179–190.

El Kayal W., Navarro M., Marque G., Keller G., Marque C. and Teulières C. 2006b Expression profile of CBF-like transcriptional factor genes from Eucalyptus in response to cold. *J. Exp. Bot.* **57**, 2455–2469.

FAO 2000 *Global forest resource assessment, main report*. United Nations Food and Agriculture Organisation, Rome.

Gonzalez-Martinez S. C., Ersoz E., Brown G. R., Wheeler N. C. and Neale D. B. 2006 DNA sequence variation and selection of tag SNPs at candidate genes for drought-stress response in Pinus taeda L. *Genetics* **172**, 1915–1926.

Gupta P. K., Roy J. K. and Prasad M. 2001 Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr. Sci.* **80**, 524–535.

Hale M. C., McCormick C. R., Jackson J. R. and DeWoody J. A. 2009 Next-generation pyrosequencing of gonad ranscriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics* **10**, 203–213.

Heuertz M., DePaoli E., Kallman T., Larsson H., Jurman I., Morgante M. *et al.* 2006 Multilocus patterns of nucleotide diversity, linkage disequilibrium and demographic history of Norway spruce [*Picea abies* (L.) Karst]. *Genetics* **174**, 2095–2105.

Huang X. and Madan A. 1999 CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877.

Ingvarsson P. K. 2005 Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European Aspen (*Populus tremula* L., Salicaceae). *Genetics* **169**, 945–953.

Keller I., Bensasson D. and Nichols R. A. 2007 Transition-transversion bias is not universal: A counter example from grasshopper pseudogenes. *PLoS Genet.* **3**, e22.

Krutovsky K. V. and Neale D. B. 2005 Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. *Genetics* **171**, 2029–2041.

Külheim C., Hui Yeoh S., Maintz J., William J. F. and Moran G. F. 2009 Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways. *BMC Genomics* **10**, 452.

Novaes E., Drost D. R., Farmerie W. G., Pappas G. J., Grattapaglia D., Sederoff R. R. *et al.* 2008 High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**, 312–325.

Rengel D., San Clemente H., Servant F., Ladouce N., Paux E., Wincker P. *et al.* 2009 A new Eucalyptus genomic resource dedicated to wood formation: a comprehensive survey. *BMC Plant Biol.* **9**, 36–49.

Riju A., Chandraseker A. and Arunachalam V. 2007 Mining for single nucleotide polymorphisms and insertions/deletions in expressed sequence tag libraries of oil palm. *Bioinformation* **2**, 128–131.

Savitch L. V., Leonardos E. D., Krol M., Jansson S., Grodzinski B., Huner N. P. A. *et al.* 2002 Two different strategies for light utilization in photosynthesis in relation to growth and cold acclimation. *Plant Cell Environ.* **25**, 761–771.

Schneider K., Weisshaar B., Borchardt D. C. and Salamini F. 2001 SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. *Mol. Breed.* **8**, 63–74.

Shannon C. E. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423.

Syvanen A. C. 2001 Accessing genetic variation. Genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* **2**, 930–942.

Teulières C., Bossinger G., Moran G. and Marque C. 2007 Stress studies in Eucalyptus. *Plant Stress* **1**, 197–215.

Thumma B. R., Matheson B. A., Zhang D., Meeske C., Meder R., Downes G. M. *et al.* 2009 Identification of a cis-acting regulatory polymorphism in a eucalypt COBRA-like gene affecting cellulose content. *Genetics* **183**, 1153–1164.

Received 30 October 2010; accepted 10 December 2010
Published on the Web: 18 August 2011