NRESTdb: Access to the Transcriptome of Natural Rubber Latex


Current technological advancement of DNA sequencing has afforded opportunities to analyse genes on a larger scale. In natural rubber research, DNA sequencing technology has been used in the generation of approximately 10,000 DNA sequences representing genes expressed in the latex of Hevea brasiliensis (Muell. Arg.) clone RRIM600. These sequence data, more popularly known as expressed sequence tags (ESTs), can be analysed using bioinformatics tools to facilitate gene discovery and provide information regarding the putative function and expression profile of latex genes. We have developed NRESTdb, a Natural Rubber EST database, to provide easy access and rapid analysis of these data. To our knowledge, this is the first publicly available EST database for H. brasiliensis. In addition to serving as a repository for individual latex EST and clustered sequences, this database allows users to view, mine and retrieve useful information regarding: the transcriptome of natural rubber latex including matches with existing domains and gene sequences available in public databases; description based on the Gene Ontology nomenclature; homologues identified via comparative sequence analysis; and enhanced data sets of special interest, namely rubber biosynthesis and latex allergy.

Keywords: EST; bioinformatics; database; transcriptome; genomics; unique transcript; Hevea brasiliensis

Hevea brasiliensis (Muell. Arg.) is the commercially cultivated tree species from which latex is extracted for the production of natural rubber. In the last 20 years, molecular biology and recombinant DNA techniques have facilitated the characterisation of specific gene sequences in natural rubber and the molecular genetics of these genes. These include genes regulating rubber biosynthesis1–4, various types of molecular markers5–7 and allergenic latex proteins8,9. Until now, the isolation of rubber genes of interest has largely been on a
gene-by-gene basis using conventional cloning strategies. With the development of genomics technologies, analysis of rubber genes and their functions has advanced to a high-throughput scale. One of the methods employs the use of large-scale DNA sequencing to generate partial DNA sequence, usually at the 5’ end, of randomly selected clones from a cDNA library. This produces a vast number of sequence reads known as expressed sequence tags (ESTs), to which putative functions can be assigned based on sequence similarity with known genes. In addition, assembly of the ESTs into clusters allows profiling of gene expression in a particular tissue or organism. Taking advantage of this technology, we have previously sequenced approximately 10,000 ESTs from natural rubber latex that formed the basis of NRESTdb, a Natural Rubber EST database. In this paper, we describe the development of NRESTdb, which contains detailed information on these ESTs, and its features and functionalities. In addition, we highlight the usefulness of the database by providing specific examples relating to rubber research.

METHODOLOGY

Sequence Generation and Clustering

Latex ESTs in NRESTdb were obtained from sequencing randomly selected clones from a latex cDNA library of *Hevea brasiliensis* (Muell. Arg.) tree clone RRIM600. Prior to clustering, all raw sequence data generated were subjected to base calling and quality trimming based on a cutoff Phred value of 20. Vector sequences were identified and masked using Cross_match (Green 1996, unpublished; http://www.phrap.org) and trimmed using a Perl script (VectorTrim.pl). Cleaned sequences were clustered based on the d2_cluster (generates clusters), Phrap (generates contigs) (Green 1996, unpublished; http://www.phrap.org) and Craw (generates consensus sequences) algorithms within the StackPACK version 2.2 package. Individual ESTs that did not show sufficient identity to be included in the assemblies were designated as singletons. The non-redundant latex transcripts, termed as unique transcripts (UTs), consisted of consensus sequences and singletons.

Functional Annotation

Assignment of putative function for the generated UTs was carried out by sequence comparison using various bioinformatics tools. Searches were performed against the non-redundant GenBank and SwissProt databases using BlastX with a cutoff E-value of 1e-05. The UTs were also compared against the *Arabidopsis thaliana* data set of the KOG and InParanoid databases to identify potential homologues. The InterProScan was utilised for searches against Gene Ontology (GO), InterPro and Pfam databases. Protein function annotation, including those based on domain and motif analyses, was derived from the results of these InterProScan searches. Possible functionality was assigned for each query sequence if significant database matches were found. The results of all similarity searches were parsed and stored in MySQL tables. Figure 1 summarises the analysis pipeline implemented in NRESTdb.

Database Architecture

NRESTdb was built based on the Red Hat Enterprise Linux 4 (Kernel version 2.6.18) platform and was developed with MySQL 5.0 Relational Database Management System. The web interface was developed using PHP4. Perl scripts were used to parse and fill the data into the MySQL tables. All database search outputs were directly linked to their respective original source databases to facilitate retrieval.
of further information. NRESTdb is available online at HYPERLINK http://genome.ukm.my/nrestdb/.

RESULTS AND DISCUSSION

Database Features and Functions

The main content of NRESTdb constitutes individual latex EST and clustered sequences, and their curated biological functions. The database currently contains 9,896 latex EST sequences, which have been clustered to produce 3,198 UTs. Annotations for these sequences include output of searches carried out with BlastX and InterProScan against a number of databases. In addition to being a repository, the database includes features for browsing, mining, accessing and analysis of these data.

The content of the database may be viewed by clicking on appropriate links of the main menu (Figure 2). In the Introduction menu, general description of the contents of the database and the details of the latex cDNA library used for EST generation are provided (Figure 3). On various NRESTdb pages, sequence ID (i.e. EST ID and UT ID) assigned
The Natural Rubber EST database (NRESTdb) is a publicly available website documenting rubber EST projects carried out in collaboration between MGI and MRB.

Figure 2. NRESTdb homepage. The database menu is divided into specific content sections.

Figure 3. General and background information in the Introduction menu of NRESTdb.
to each EST and UT, respectively, links users to information such as the nucleotide sequence, sequence length and where applicable, number of members in a consensus (Figure 4). Sequences can be downloaded or copied directly for further sequence manipulation. The BlastX output provides a list of putative gene identity based on comparisons with the GenBank and SwissProt databases (Figure 5). Similarly, a list of Arabidopsis thaliana homologues corresponding to the latex UTs is listed in the InParanoid output. A simple search of these lists is made possible by using the UT ID or keyword search box. Additionally, search of gene annotations based on KOG and GO functional classes are available in the Database menu.

For easy cross-referencing and database navigation, sequences and annotations may be accessed through more than one NRESTdb menu option. For example, using the UT...
ID, users are able to directly access KOG, GO, Pfam, InterPro and BlastX annotations (Figure 4). Similarly, simple search boxes are available on most pages (examples in Figures 4 and 5). However, for more refined searches, NRESTdb is equipped with sequence and keyword based search tools that can be found in the Search menu. The first is equivalent to performing a local BLAST on the EST or UT data sets using a BLAST algorithm of choice while the second uses keywords to query functional descriptions (Figure 6). In addition to sequence and function information, NRESTdb contains enhanced data sets in two research areas where EST analysis has been applied, i.e. rubber biosynthesis and latex allergy (see Special Interest menu).

**Gene Discovery**

ESTs, as the name indicates, act as “tags” or leads to the isolation of genes of interest including novel unknown genes. Clustering cleaned sequences is a crucial step as it enables groups of ESTs sharing sufficient DNA identity to be assembled into contiguous transcripts. This reduces the complexity of the initial EST data set, particularly for highly abundant transcripts, by generating...
one consensus sequence from each assembly. The availability of longer transcripts generated from partial sequences, i.e. ESTs, provides several approaches for gene discovery. Firstly, the list of genes serves as a catalogue from which query for sequences of interest may be made using search tools available in NRESTdb. In the study on rubber biosynthesis, this approach was found to be useful for confirming the presence of the plastidic 1-deoxy-D-xylulose 5-phosphate/2-C-methyl-D-erythritol 4-phosphate (MEP) pathway in latex through the identification of an EST for 1-deoxy-D-xylulose 5-phosphate synthase (DXPS), the enzyme involved in the first step of this pathway. Secondly, the need for conventional full-length cDNA cloning methods, such as rapid amplification of cDNA ends (RACE) and repeated library screening, is also greatly reduced as the consensus sequences would very likely consist of full-length transcripts. Thirdly, the algorithm of clustering in StackPACK is capable of separating variant members (isoforms) belonging to the same gene family. This is done by generating separate consensus sequences from assemblies that were deemed to be sufficiently distinct although having the same gene identity.
Additionally, EST analysis is useful for deriving markers such as simple sequence repeats (EST-SSRs) and single nucleotide polymorphisms (SNPs). Recently, the use of the NRESTdb EST collection for generating EST-SSRs for rubber marker-assisted breeding and pedigree analysis was reported.

Expression Profiling

Each latex EST represents a specific mRNA that is expressed in the rubber tree laticifer. The benefit of redundant copies of ESTs is that it may be used as a measure of transcript expression level. Transcript abundance or EST frequency is expressed as the number of EST members belonging to a particular UT. The frequency of a singleton is one. However, nondetection of an EST for a gene of interest may not mean that it is not expressed but that it was not detectable due to insufficient depth of EST sequencing. In previously published research, we have used EST frequency analysis to determine the profile of abundance of latex allergens and rubber particle membrane proteins, and found that it is a feasible approach for gene expression profiling.

CONCLUSION

The discovery of novel genes and elucidation of gene expression levels have demonstrated the usefulness of NRESTdb as a genomics resource for rubber research. The availability of NRESTdb narrows the gap between gene isolation and functional characterisation. As part of the effort to make NRESTdb publicly available, NRESTdb has also been listed in the Molecular Biology Database Collection of the Nucleic Acids Research Journal. Future work will include supplementing NRESTdb with more ESTs and analytical tools to facilitate studies on rubber genes and proteins.

ACKNOWLEDGEMENT

We thank the National Biotechnology and Bioinformatics Network (NBBNet: http://www.nbbnet.gov.my) for hosting NRESTdb.

Date of receipt: April 2009
Date of acceptance: September 2009

REFERENCES


5. LOW, F.C., SAFIAH ATAN, HAFSAH JAAFAR AND TAN, H. (1996) Recent Development of Molecular Markers for


20. YEANG, H.-Y., HAMILTON, R.G., BERNSTEIN, D.I., ARIF, S.A.M, CHOW,