

The TIGR Plant Transcript Assemblies database

Kevin L. Childs, John P. Hamilton, Wei Zhu, Eugene Ly, Foo Cheung, Hank Wu,
Pablo D. Rabinowicz, Chris D. Town, C. Robin Buell and Agnes P. Chan*

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received August 15, 2006; Revised September 21, 2006; Accepted September 29, 2006

ABSTRACT

The TIGR Plant Transcript Assemblies (TA) database (<http://plantta.tigr.org>) uses expressed sequences collected from the NCBI GenBank Nucleotide database for the construction of transcript assemblies. The sequences collected include expressed sequence tags (ESTs) and full-length and partial cDNAs, but exclude computationally predicted gene sequences. The TA database includes all plant species for which more than 1000 EST or cDNA sequences are publicly available. The EST and cDNA sequences are first clustered based on an all-versus-all pairwise sequence comparison, followed by the generation of consensus sequences (TAs) from individual clusters. The clustering and assembly procedures use the TGICL tool, Megablast and the CAP3 assembler. The UniProt Reference Clusters (UniRef100) protein database is used as the reference database for the functional annotation of the assemblies. The transcription orientation of each TA is determined based on the orientation of the alignment with the best protein hit. The TA sequences and annotation are available via web interfaces and FTP downloads. Assemblies can be retrieved by a text-based keyword search or a sequence-based BLAST search. The current version of the TA database is Release 2 (July 17, 2006) and includes a total of 215 plant species.

INTRODUCTION

Despite the best efforts of gene prediction software, expressed sequence tag (EST) and other cDNA sequences are among the most reliable evidence for gene expression and gene identification. Many EST sequencing projects have generated millions of sequences, typically representing partial transcripts derived from single-pass sequences of the 5' or 3' end of random cDNA clones. High-throughput full-length cDNA sequencing projects have generated full-length transcript sequences for selected model plants

including *Arabidopsis*, rice and maize. Experimentally derived cDNAs from individual gene cloning studies generally represent both partial and full-length gene transcripts.

Over 10 million plant EST sequences are present in the NCBI GenBank Nucleotide database. However, the number of annotated cDNA sequences corresponds to <2% of the EST sequences. The large volume, short lengths and lack of functional annotation are common obstacles to using EST sequences for gene modeling and gene structure identification. One approach to reduce the redundancy and to consolidate the EST sequence data, as well as to potentially maximize the sequence information of the gene transcripts is to assemble the EST and cDNA sequences using highly stringent assembly criteria.

Several groups have undertaken efforts to assemble and annotate ESTs and mRNAs from plants, and each uses a different procedure for selecting the plant species and types of sequences that are to be included in the databases. The first such effort was the NCBI UniGene project (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>), which currently contains transcription loci from 29 plant species (1,2). Each transcription locus is represented by a set of related but unassembled transcript sequences. The Gene Index project (<http://biocomp.dfci.harvard.edu/tgi>) generated tentative consensus sequences for approximately 30 plant species (3,4). Species with more than 50 000 ESTs are included in the Gene Index database. However, other plant gene indices have been built on request, and several important plants containing only a few thousand ESTs are also present. The assembled tentative consensus sequences in the Gene Index database are derived from ESTs, cDNAs and predicted gene transcripts using the CAP3 assembler. The PlantGDB EST assembly project (<http://www.plantgdb.org/prj/ESTCluster>) contains a collection of EST assemblies from 107 plant species, using the CAP3 assembler (5,6). A species is included in the PlantGDB EST assemblies if there are more than 10 000 ESTs available for that species. The HarvEST project (<http://harvest.ucr.edu>) provides CAP3-assembled EST assemblies for eight plant species including selected cereal and crop species.


With the goal of generating a comprehensive resource of assembled and annotated gene transcripts derived from experimentally-derived EST and cDNA sequences, we have developed the TIGR Plant Transcript Assemblies (TA)

*To whom correspondence should be addressed: Tel: +1 301 795 7862; Fax: +1 301 838 0208; Email: achan@tigr.org

database (<http://plantta.tigr.org>). To enable rich, informative cross-species comparative studies within the Plant Kingdom such as the discovery of putative orthologues from among the represented species and also to simplify and broaden the range of species used for genome annotation, the TA database includes all plant species with more than 1000 publicly available EST and cDNA sequences from the GenBank Nucleotide database. The input EST and cDNAs sequences are clustered and assembled based on sequence similarity to generate a consensus sequence using the TGICL tool (7). The TGICL tool is a wrapper script which first clusters the input sequences from individual plant species based on an all-versus-all pairwise comparison using Megablast (8), and subsequently creates the final assemblies using CAP3 (9). Functional annotation and orientation of the TAs are determined using the UniProt Reference Clusters (UniRef100) protein database as the reference resource (10). The TAs can be retrieved using keyword or sequence searches. A tree-view is available to allow direct access to the 215 TAs according to species, genus or higher order taxonomic classification. The TAs and their annotation are available through web interfaces and FTP downloads. The TA database will be regularly updated to include new EST and cDNA submissions to the GenBank Nucleotide database. The current version of the TA database is Release 2 (July 17, 2006).

TA construction

The TA database was set up using a MySQL database management system. The database schema was designed to handle the particular needs of the TA database including the ability to track assembly history and annotation. To set up the first release of the TA database, we generated a list of plant species that satisfied the criteria of having more than 1000 available EST or cDNA sequences. For each species, the corresponding EST as well as partial and full-length cDNA sequences were collected from GenBank. Computationally derived or predicted transcripts from genome annotation projects were excluded so that any given TA sequence represents experimental proof that such sequence is actually transcribed. EST sequences were directly collected from the GenBank Nucleotide EST division. Full-length and experimentally derived cDNA sequences but not predicted transcripts were selectively retrieved by collecting GenBank Nucleotide records labeled as 'mRNA biomolecules' and by limiting the source databases to DDBJ, EMBL and GenBank. The collected EST and cDNA sequences were first cleaned of potential vector contamination using the TIGR SeqClean tool (<http://www.tigr.org/tdb/tgi/software>) and the NCBI UniVec database (<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>). The SeqClean tool also screens for low-complexity regions, trims poly A and poly T tracts and sequence ends rich in undetermined



TIGR
Plant Transcript Assemblies

Home Current Release Plant TA Search Blast Search Download TAs Contact TIGR Home

Current Release Summary

TIGR Plant Transcript Assemblies - latest update July 17th, 2006

Showing assembly statistics for 6 species out of a total of 215. Click on a column header to sort by that column and again to sort in reverse order.

Taxon ID	Scientific Name	Common Name	EST Retrieval Date	Release	Transcript Assemblies			Transcript Assembly Components			Download FASTA
					Assemblies	Singletons	Total	EST	fl-cDNA	mRNA	
3702	<i>Arabidopsis thaliana</i>	Mouse-ear Cress	2006-06-05	2	27983	120385	148368	616579	65976	2204	Download
3880	<i>Medicago truncatula</i>	Barrel Medic	2006-06-05	2	20414	34768	55182	217151	251	439	Download
4513	<i>Hordeum vulgare</i>	Barley	2006-06-05	2	30171	93180	123351	456411	860	448	Download
4530	<i>Oryza sativa</i>	Rice	2006-06-05	2	49870	197646	247516	1170755	34559	888	Download
4565	<i>Triticum aestivum</i>	Bread (common) wheat	2006-06-05	2	62121	257828	319949	840900	1833	554	Download
4577	<i>Zea mays</i>	Maize	2006-06-05	2	50465	118622	169087	744330	4676	9725	Download

021811

Figure 1. An example of the TA current release page based on six selected TA species, including rice, maize, wheat, barley, *Arabidopsis* and *Medicago*, out of a total of 215 available TA species. The table shows the GenBank taxon id, the scientific and common names of the species, the date of sequence retrieval from GenBank, the current release number, the numbers of input sequences obtained as ESTs and partial and full-length cDNAs, and the number of TA and singleton sequences.



TIGR Plant Transcript Assemblies



[Home](#) [Current Release](#) [Plant TA Search](#) [Blast Search](#) [Download TAs](#) [Contact](#) [TIGR Home](#)

Plant Transcript Assembly Keyword Search Results

9 results were found matching the query *gibberellin and oxidase*.

Plant TA Accession	Species	Plant TA Annotation	% Identity	% Coverage	Release Version
C0486173	<i>Picea glauca</i>	Gibberellin 20-oxidase [Lolium perenne (Perennial ryegrass)]	63.89	27.84	2
DR553915	<i>Picea glauca</i>	Gibberellin 20-oxidase [Citrus sinensis x Poncirus trifoliata]	60.42	44.51	2
DR575289	<i>Picea glauca</i>	Gibberellin 20-oxidase2 [Daucus carota (Carrot)]	70.83	47.79	2
TA7540_3332	<i>Picea sitchensis</i>	Gibberellin 20-oxidase2 related cluster	87.5	42.02	1
DR543176	<i>Picea sitchensis</i>	Gibberellin 20-oxidase2 related cluster	73.49	49.5	1
TA11584_3352	<i>Pinus taeda</i>	Gibberellin 2-oxidase 1 [Nicotiana tabacum (Common tobacco)]	66.48	46.19	2
TA13335_3352	<i>Pinus taeda</i>	Gibberellin 20-oxidase [Gossypium hirsutum (Upland cotton)]	48.96	34.37	2
CF401584	<i>Pinus taeda</i>	Gibberellin 2-oxidase 2 [Nicotiana tabacum (Common tobacco)]	66.36	45.02	2
BX681458	<i>Pinus pinaster</i>	Gibberellin 2-oxidase 1 [Nicotiana tabacum (Common tobacco)]	73.83	56.32	2



Figure 2. An example of the TA keyword search result page using the keywords 'gibberellin oxidase' to search from the nine Coniferopsida species available. The result page lists all matching TA and singletons from the user-defined species. The species name, annotation description and release number are provided with links to the TA report page and the UniRef protein entry.

bases, as well as removes low quality sequences. The TGICL tool (7), Megablast (8) and the CAP3 assembler (version:4/15/05) (9) were used for the clustering and assembly steps (command: `tgicl <fasta_file> -p 95 -l 50 -v 20 -s 10 000 -O 'p 95 -y 20 -o 50'`). TGICL is a wrapper script which first clusters the vector-trimmed sequences based on an all-versus-all pairwise sequence comparison using Megablast, and subsequently assembles each cluster by CAP3 using the following highly stringent criteria: (i) the overlap between two sequences must be longer than 50 bp; (ii) the sequence identity between the overlapping region of the two sequences must be >95%; (iii) the number of overhanging, unaligned bases at the ends of the sequences must be <20 bases; (iv) large clusters with >10 000 component sequences are avoided. The resulting consensus sequences for each species are assigned a TA identifier number. The format of a TA identifier is *TA**number**_taxonID* where, *number* is a unique numerical identifier for the TA within a species and *taxonID* is the NCBI taxonomy identifier of the species from which the sequence was obtained. EST and cDNA sequences that are not included in a TA assembly are stored as singletons in the TA database and retain their original GenBank accession number as their unique identifier.

TA annotation

Functional annotation for each TA was assigned using the BLAT search program due to its speed (11). The UniRef protein database was used as the reference due to its high-quality annotation and compact size (10). The current TA release (Release 2) uses the UniRef100 database. Low complexity regions of the protein database were masked using the SEG filter program (<http://blast.wustl.edu>) prior to carrying out the BLAT searches to minimize biologically non-significant hits. For each TA and singleton sequence, functional annotation was assigned based on the best UniRef protein hit ranked by alignment score which satisfied a minimum threshold of 20% identity and 20% coverage between the TA and the protein sequence. The transcription orientation for each TA was determined by the relative orientation to the protein hit with the best alignment score. Additional functional annotation such as protein families and domains [e.g. InterPro (12)] and gene ontology (13) will be included in future releases. Owing to the extensive coverage of the plant species by the TAs, new features such as cross-referencing of putative gene orthologues among the plant TAs will allow a thorough analysis of the functional similarity or conservation of the orthologues.

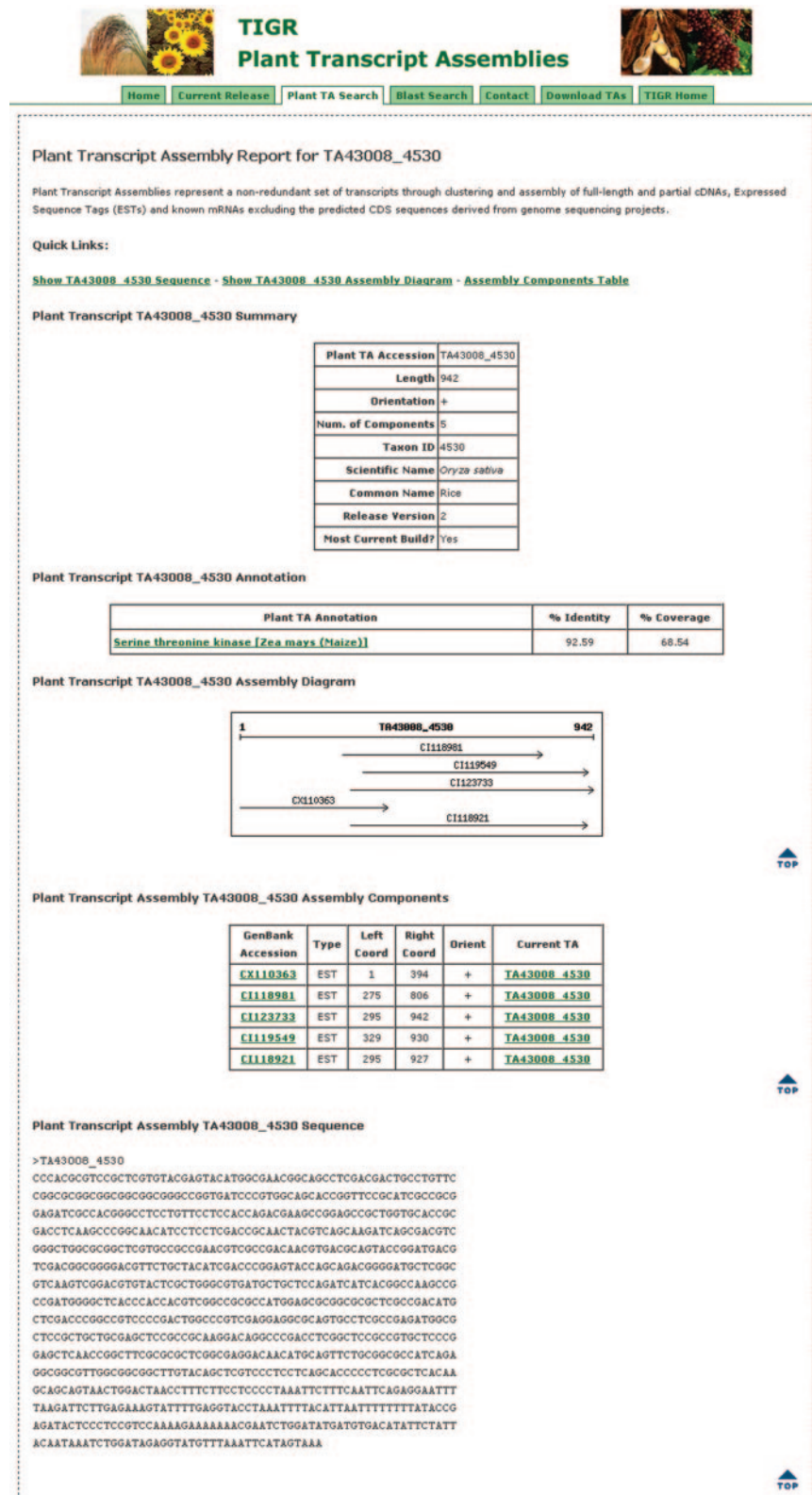


Figure 3. The TA report page for a rice TA from release 2 (TA43008_4530). The TA report pages consist of five sections. The first section provides general information about the TA: accession number, length, transcription orientation, number of component sequences, GenBank taxon id, species name, TA version number and whether the version is the most current. A second section gives the description of the best UniRef protein hit based on BLAT alignment. The percent identity and percent coverage of that alignment are also shown. The third section of the report is a diagram of the relative position and orientation of each member sequence in the TA. The fourth part of the report lists the GenBank accession number of each component of the TA, the start and end coordinates, and the orientation for each sequence within the TA. A link is also provided to the most current TA versions to which individual component reads belong. The last section provides the consensus sequence of the TA.

Data access

The TA database is accessible via the project website (<http://plantta.tigr.org>) and FTP downloads (<ftp://ftp.tigr.org/pub/data/plantta>). A current release page (http://plantta.tigr.org/cgi-bin/plantta_release.pl) summarizes the number of species included in the current data release and associated information such as the number of input EST and cDNA sequences and the number of TAs generated. A summary of the TA release information for six selected model plants including rice, maize, wheat, barley, *Arabidopsis* and *Medicago* is shown in Figure 1. The release table can be sorted by species name, taxon *id*, release number and the number of TAs in the collection. For each species, a link out from the release table to the FTP site for TA and annotation download is also available.

Users can access the TA database using keyword searches or through the TA BLAST server for sequence-based searches (http://tigrblast.tigr.org/euk-blast/plantta_blast.cgi). A dynamically expandable taxonomic tree view allows users to specifically select species, genera or higher order taxonomic groups to narrow the keyword and BLAST searches. An example search using the keyword option to search for gibberellin oxidase-related TAs from the nine available Coniferopsida species is shown in Figure 2. The search results display the TA identifier number, species name, annotation description, percent identity and percent coverage between the TA and the best matching UniRef protein entry. The TA identifier number provides a link to the TA report page. For TAs from Release 2 and later, the annotation description links to the UniRef protein database. The TA BLAST server allows a user-provided sequence to be searched against the TAs using the BLAST programs (14): *blastn*, *tblastn* and *tblastx*. All BLAST searches are performed with WU-BLAST 2.0 (<http://blast.wustl.edu>). The search result provided is a standard BLAST output with sequence alignments and links to TA report pages.

A TA summary report can be retrieved using different identifiers including the TA identifier number, the GenBank accession of the component EST or cDNA sequences of an assembly, or the GenBank accession of a singleton sequence. The database schema was specifically designed to allow for the retrieval of a TA using the sequence identifiers of the component sequences. Figure 3 shows an example of a TA report page for TA43008.4530, a serine threonine kinase-related rice TA from Release 2. The assembly report displays the TA identifier number, TA functional annotation, an assembly diagram, a list of component sequences and their corresponding coordinates within the assembly, and the TA sequence. The list of component EST and cDNA sequences also provides links to the most current assembly to which individual component reads belong. This allows forward tracking of deprecated TAs from previous releases.

TA updates

As new EST and cDNA sequences become available for species with existing TAs, it will be necessary to incorporate the new sequences by rebuilding the transcript assemblies. When the total number of EST and cDNA sequences for a particular species increases by 10% and that the increase is greater than 1000 sequences, the TA will be reconstructed. The TA identification numbers are not reused but increase

sequentially. The TA database will be updated approximately every 3 months.

CONCLUSIONS

We have created the TA database as a combined effort to support in-house genome annotation work and also to provide the plant genomics and molecular genetics community with a comprehensive resource of assembled and annotated gene transcripts from available plant species. In particular, the inclusion of all plant species for which more than 1000 EST and cDNA sequences exist will enable future comparative analysis across multiple taxonomic clades such as the discovery of gene orthologues and novel species-specific genes. In Release 2, we fully implemented the use of stringent assembly parameters and UniRef100 for annotation instead of UniRef90 which was used in Release 1.

Although the TA database uses the TGICL tool, an open source developed by the Gene Index project, the major differences of the TA database and the Gene Index project are as follows. First, the TA database only includes expressed transcripts such as EST, partial and full-length cDNA from GenBank as input sequences, but not computationally derived gene transcripts originated from whole genome annotation projects. Second, all TAs are clustered solely based on individual plant species but not at higher levels (e.g. genus) of taxonomic classification. Third, the TA database uses a newly developed database schema different from the Gene Index project. Furthermore, the wide collection of 215 plant species in the TA database is also a unique feature not represented by other plant transcript assembly projects. In future releases, we intend to develop new features and functionality for the TA database including additional filtering of input sequences to avoid microbial sequences and repetitive sequences such as the rRNAs, checking for potential chimeric transcript assemblies, functional annotation of protein domains and gene ontology of the TAs, cross-referencing among putative orthologues, addition of cDNA library source information associated with the EST and cDNA sequences to provide spatial and temporal expression evidence for the TAs, as well as backward version tracking of the TAs to enable cross-referencing between releases.

ACKNOWLEDGEMENTS

The efforts of the TIGR Bioinformatics Department and IT Department in creating infrastructure to support this project are appreciated. This work was supported in part by grants from the National Science Foundation Plant Genome Research Program to C.R.B. (DBI-0218166, DBI-0321538, DBI-0321663, DBI-313887) and C.D.T. (DBI-0321460) and from the U.S. Department of Agriculture to P.D.R. (USDA 59-1275-5-364). Funding to pay the Open Access publication charges for this article was provided by National Science Foundation Plant Genome Research Program.

Conflict of interest statement. None declared.

REFERENCES

1. Pontius, J.U., Wagner, L. and Schuler, G.D. (2003) *UniGene: A Unified View of the Transcriptome*. National Center for Biotechnology Information, Bethesda, MD.

2. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
3. Lee,Y., Tsai,J., Sunkara,S., Karamycheva,S., Pertea,G., Sultana,R., Antonescu,V., Chan,A., Cheung,F. and Quackenbush,J. (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.*, **33**, D71–D74.
4. Quackenbush,J., Cho,J., Lee,D., Liang,F., Holt,I., Karamycheva,S., Parvizi,B., Pertea,G., Sultana,R. and White,J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.
5. Dong,Q., Lawrence,C.J., Schlueter,S.D., Wilkerson,M.D., Kurtz,S., Lushbough,C. and Brendel,V. (2005) Comparative plant genomics resources at PlantGDB. *Plant Physiol.*, **139**, 610–618.
6. Dong,Q., Schlueter,S.D. and Brendel,V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.*, **32**, D354–D359.
7. Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F., Parvizi,B. *et al.* (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
8. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
9. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
10. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
11. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
12. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P. *et al.* (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform.*, **3**, 225–235.
13. Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.
14. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.