Nucleic Acids Research, 2005, Vol. 33, Database issue ***D71–D74***
doi:10.1093/nar/gki064

The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes

Y. Lee*, J. Tsai, S. Sunkara, S. Karamycheva, G. Pertea, R. Sultana, V. Antonescu, A. Chan, F. Cheung and J. Quackenbush

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

Received September 14, 2004; Revised and Accepted October 5, 2004

ABSTRACT

Although the list of completed genome sequencing projects has expanded rapidly, sequencing and analysis of expressed sequence tags (ESTs) remain a primary tool for discovery of novel genes in many eukaryotes and a key element in genome annotation. The TIGR Gene Indices (http://www.tigr.org/tdb/tgi) are a collection of 77 species-specific databases that use a highly refined protocol to analyze gene and EST sequences in an attempt to identify and characterize expressed transcripts and to present them on the Web in a user-friendly, consistent fashion. A Gene Index database is constructed for each selected organism by first clustering, then assembling EST and annotated cDNA and gene sequences from GenBank. This process produces a set of unique, high-fidelity virtual transcripts, or tentative consensus (TC) sequences. The TC sequences can be used to provide putative genes with functional annotation, to link the transcripts to genetic and physical maps, to provide links to orthologous and paralogous genes, and as a resource for comparative and functional genomic analysis.

INTRODUCTION

The TIGR Gene Index databases (TGI) (http://www.tigr.org/tdb/tgi) are constructed using all publicly available expressed sequence tags (EST) and known gene sequence data stored in GenBank for each target species. Sequences are first cleaned to identify and remove contaminating sequences, including vector, adaptor, mitochondrial, ribosomal and chimeric sequences. These sequences are then searched pairwise against each other and grouped into clusters based on shared sequence similarity. The clusters are assembled at high stringency to produce tentative consensus (TC) sequences. The virtual transcripts represented in the TCs are annotated using a variety of tools for open reading frame (ORF) prediction, single nucleotide polymorphism (SNP) prediction, long oligo prediction for microarrays, putative annotation using a controlled vocabulary, Gene Ontology (GO) and Enzyme Commission (EC) number assignments and maps onto complete or drafted genomes or available genetic maps. The TCs are used to construct a variety of other databases, including the Eukaryotic Gene Orthologs (EGO) database and RESOURCERER, a database that annotates and cross-references microarray resources for plants and animals.

At present, 77 species are represented in the Gene Index databases, including 29 animals, 25 plants, 8 fungi and 15 protists; this includes most species for which public EST projects have released more than 50 000 ESTs. Current release information for each species-specific database is summarized in Table 1. Individual databases are updated and released three times yearly, on February 1, June 1 and October 1, if the number of available ESTs for that species has increased by either 25 000 or >10%, whichever is less.

RECENT DEVELOPMENTS

Construction of the Gene Indices

The process used to assemble each Gene Index is similar to that described previously (1–3), although some modifications have been made to improve the efficiency and accuracy of the process. mgBLAST, a modified version of the Megablast (4) program, is now used for the pairwise sequence comparisons that are the basis for defining the sequence clusters which form the basis for assembly. For large clusters containing hundreds or thousands of sequences (e.g. highly expressed genes such as actin), sequence representation is reduced prior to assembly using a variety of multilayer approaches, including transitive clustering, containment clustering and seeded clustering with known genes. Following clustering, the Paracel Transcript Assembler (PTA), a

*To whom correspondence should be addressed. Tel: +1 301 795 7831; Fax: +1 301 838 0208; Email: dlee@tigr.org

Table 1. Summary of the current release of TIGR Gene Indices (TGI)

Species	Species_name	TGI	TC	sET	sEST
Animals (29)					
Human	*Homo sapiens*	HGI 15.0	221 418	19 740	594 468
Mouse	*Mus musculus*	MGI 14.0	167 694	7499	602 312
Rat	*Rattus norvegicus*	RGI 13.0	56 933	2131	87 992
Cattle	*Bos Taurus*	BtGI 10.0	38 760	413	56 644
Pig	*Sus scrofa*	SsGI 9.0	33 963	519	50 376
Dog	*Canis familiaris*	DogGI 4.0	6613	684	11 506
Chicken	*Gallus gallus*	GgGI 8.0	42 988	848	72 941
Frog	*Xenopus laevis*	XGI 9.0	39 724	626	37 249
Zebrafish	*Danio rerio*	ZGI 15.0	32 889	395	53 940
Catfish	*Ictalurus punctatus*	Cfgi 5.0	3254	156	16 694
R.trout	*Oncorhynchus mykiss*	RtGI 4.0	23 135	190	27 448
A.salmon	*Salmo salar*	AsGI 2.1	12 277	93	18 971
C.intestinalis	*Ciona intestinalis*	CinGI 3.0	20 616	39	30 690
Medaka	*Oryzias latipes*	OlGI 5.0	12 849	171	13 669
Fugu	*Takifugu rubripes*	FGI 1.0	3120	448	7667
A.burtoni	*Astatotilapia burtoni*	AbGI 1.0	402	15	2300
H.chilotes	*Haplochromis chilotes*	HchGI 1.0	2147	0	4030
H.red_tail_sheller	*Haplochromis* sp. 'rts'	HsGI 1.0	1883	0	4422
Killifish	*Fundulus heteroclitus*	FhGI 1.0	3540	57	11 941
Honeybee	*Apis mellifera*	AMGI 4.0	3700	53	7571
A.aegypti	*Aedes aegypti*	AeGI 4.0	15 888	32	5075
Drosophila	*Drosophila melanogaster*	DGI 9.0	20 693	1104	6662
Mosquito	*Anopheles gambiae*	AgGI 7.0	17 120	6847	14 940
A.variegatum	*Amblyomma variegatum*	AvGI 2.0	478	0	1631
R.appendic	*Rhipicephalus appendiculatus*	RaGI 1.0	2543	19	4797
C.elegans	*Caenorhabditis elegans*	CeGI 8.0	17 728	5034	5678
B.malayi	*Brugia malayi*	BmGI 4.0	2060	44	6841
O.volvulus	*Onchocerca volvulus*	OvGI 3.0	1065	23	2942
S.mansoni	*Schistosoma mansoni*	SmGI 5.0	12 912	39	20 753
Plants (25)					
Pine	Pinus	PGI 4.0	13 622	205	17 944
Cocoa	*Theobroma cacao*	TcaGI 1.0	754	26	1759
Cotton	Gossypium	CGI 5.0	6812	142	17 396
Arabidopsis	*Arabidopsis thaliana*	AtGI 11.0	28 010	5188	12 485
L.japonicus	*Lotus japonicus*	LjGI 3.0	12 485	56	15 919
Lettuce	*Lactuca sativa*	LsGI 2.0	7961	56	14 168
Sunflower	*Helianthus annuus*	HaGI 3.0	6038	110	14 372
Tomato	*Lycopersicon esculentum*	LeGI 9.0	20 530	164	14 923
Pepper	*Capsicum annuum*	CaGI 1.0	3203	47	7462
Potato	*Solanum tuberosum*	StGI 9.0	19 225	102	13 226
Tobacco	*Nicotiana tabacum*	NtGI 1.0	897	806	8529
N.benthamiana	*Nicotiana benthamiana*	NbGI 1.1	3819	44	3735
Soybean	*Glycine max*	GmGI 12.0	30 084	141	37 601
Medicago	*Medicago truncatula*	MtGI 7.0	17 610	25	19 341
Ice_plant	*Mesembryanthemum crystalline*	McGI 4.0	2851	47	5557
Grape	*Vitis vinifera*	VvGI 3.1	13 218	54	9837
Rice	*Oryza sativa*	OsGI 15.0	33 089	17 776	37 900
Maize	*Zea mays*	ZmGI 14.0	29 414	524	26 426
Wheat	*Triticum aestivum*	TaGI 8.0	44 630	169	79 008
Sorghum	*Sorghum bicolor*	SbGI 8.0	20 029	143	18 976
Barley	*Hordeum vulgare*	HvGI 9.0	21 981	168	27 041
S.cereale	*Secale cereale*	RyeGI 3.0	1391	66	3890
S.officinarum	*Saccharum officinarum*	SoGI 1.0	23 596	7	72 281
A.cepa	*Allium cepa*	OnGI 1.0	3838	18	7870
C.reinhardtii	*Chlamydomonas reinhardtii*	ChrGI 4.0	10 777	96	19 466
Fungi (8)					
A.flavus	*Aspergillus flavus*	AfGI 4.0	3749	10	3459
C.posadasii	*Coccidioides posadasii*	CpoGI 2.0	6275	0	3037
S.cerevisiae	*Saccharomyces cerevisiae*	ScGI 3.0	4107	2005	198
S.pombe	*Schizosaccharomyces pombe*	SpGI 3.0	2449	2974	510
Cryptococcus	*Filobasidiella neoformans*	CrGI 7.0	2384	59	3231
N.crassa	*Neurospora crassa*	NcrGI 3.0	4389	6547	1586
A.nidulans	*Aspergillus nidulans*	AnGI 4.0	3532	6664	2904
M.grisea	*Magnaporthe grisea*	MgGI 5.0	6375	6195	8320
Protists (15)					
P.berghei	*Plasmodium berghei*	PbGI 5.0	1168	41	3980
P.falciparum	*Plasmodium falciparum*	PfGI 7.0	3978	2487	3142
P.vivax	*Plasmodium vivax*	PvGI 0.5	158	175	567

Table 1. *Continued*

Species	Species_name	TGI	TC	sET	sEST
P.yoelii	*Plasmodium yoelii*	PyGI 5.0	3611	3784	2418
E.tenella	*Eimeria tenella*	EtGI 4.0	2077	29	3066
T.gondii	*Toxoplasma gondii*	TgGI 6.0	6977	31	11 401
N.caninum	*Neospora caninum*	NcGI 5.0	1980	3	3715
S.neurona	*Sarcocystis neurona*	SnGI 4.0	665	0	1644
C.parvum	*Cryptosporidium parvum*	CpGI 4.0	171	485	254
T.vaginalis	*Trichomonas vaginalis*	TvGI 1.0	87	109	704
Leishmania	*Leishmania*	LshGI 4.0	600	1454	1120
T.cruzi	*Trypanosoma cruzi*	TcGI 4.0	2189	164	4749
T.brucei	*Trypanosoma brucei*	TbGI 5.0	734	1287	2018
D.discoideum	*Dictyostelium discoideum*	DdGI 4.0	6826	172	6392
T.thermophila	*Tetrahymena thermophila*	TtGI 3.0	1436	165	2626

TIGR Gene Indices are a collection of species-based databases which assemble the ESTs and the Expressed Transcripts (ETs) into TC sequences. Singletons (sET and sEST) are the ET/EST sequences that are not incorporated into a TC during assembly. TCs, sET and sEST are the unique sequences in TGI. There are 77 gene indices in total (data until September 1, 2004). Each line includes species, species name, gene index name and version, total number of TCs within current release, number of singleton ETs and number of singleton ESTs. For Leishmania, pine and cotton, the ESTs were pooled from dbEST for the genus, not a single species. The table was arranged by grouping the total 77 gene indices into animals (29), plants (25), fungi (8) and protists (15).

modified version of CAP3 assembly program (5), is used to assemble each TC. An open source set of software tools that embody this process, TGICL, is available (http://www.tigr.org/tdb/tgi/software) with other open-source utilities for users interested in performing a similar analysis on their own datasets (6).

New features of the TC report

The central element of the TGI databases are the TC sequences and the TC reports that are presented through the project website. Each TC report presents a summary of the assembly and annotation process, including the consensus TC sequence in the FASTA format with a history from previous builds in the header, a map showing component EST and gene sequences, and a table providing links to the primary sequences, putative annotation, an expression summary based on the number of ESTs from various libraries, genomic locations and links to tentative orthologs in EGO. Since the last presentation of the TGI databases in *Nucleic Acids Research*, several new features have been added to the TC report. Putative polyadenylation signals are identified and shaded in the consensus sequence and putative poly(A/T) trimming sites are shown in sequence map for each of the component ESTs. Potential ORFs are predicted for each TC using a variety of software tools including the NCBI ORF Finder, ESTScan (7) and FrameFinder; predicted ORFs can be searched against a variety of databases using WU-BLAST. Assembly of the TCs can result in incorrect orientations for the consensus and an attempt is now made to determine the proper orientation using the annotated direction of component gene and EST sequences as well as BLAST search results. Putative SNP sites are found by analyzing the multiple sequence alignments that are produced in the assembly stage; putative SNPs are reported only if a variant is found in multiple sequences from independent libraries. Unique 70mers are predicted for each TC using OligoPicker (8). GO terms and metabolic pathway in KEGG are provided for each TC based on protein database searches. Where possible, TCs are aligned with draft genomes and displayed using TGIviewer, gbrowse, EnsEMBL and the UCSC genome viewers.

New databases and tools

The EGO (http://www.tigr.org/tdb/tgi/ego) (9) database, previously known as TIGR Orthologous Gene Alignments (TOGA), uses pairwise sequence similarity searches and a transitive, reciprocal closure process to identify Tentative Ortholog Groups (TOGs) in eukaryotes (9). EGO has expanded its representation to include all 77 species represented in the TGI and TOGs have been cross-referenced to the Online Mendelian in Man (OMIM) (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) database of human disease genes.

RESOURCERER (10) provides annotation based on the TIGR Gene Indices for widely available microarray resources in human, mouse, rat, zebrafish and Xenopus, including widely used clone sets and Affymetrix GeneChips^{TM} as well as a variety of other sequence-based resources such as RefSeq. RESOURCERER provides a wide range of annotation and integration with genomic and other resources, including gene name assignments, GO term and EC number assignments, chromosomal localization, integration with genetic and quantitative trait locus (QTL) maps, ortholog identification, lists of relevant abstracts in PubMed and promoter region identification. Owing to its integration with the TGI and EGO, RESOURCERER also provides links between microarray platforms both within and between species. Users can also submit a list of GenBank accessions corresponding to their microarray databases for annotation and functional analysis. A plant-specific version, Plant RESOURCERER, was released in September 2004 with microarray resources from *Arabidopsis*, potato, tomato, maize and rice.

Genomic maps align TCs to available complete or draft genomes, including human, mouse, rat, zebrafish, fly, worm, Fugu, mosquito, *Arabidopsis*, yeast, fission yeast and rice. Also these alignments can be viewed using either TGIviewer or gbrowse or through a number of distributed annotation system (DAS) viewers (11), including one developed at TIGR. Each Gene Index also includes graphical metabolic pathway maps linked to TCs associated with specific pathways through GO term and EC number annotation. Comparisons between TCs are also used to identify putative alternative splice forms based on shared blocks of sequence similarity.

Using the TIGR Gene Indices

There are many ways in which users can access the TIGR Gene Index databases. Nucleotide or protein sequences can be searched using WU-BLAST against individual TGI databases, EGO or pre-selected classes of species, such as animals or plants. The TGI can be searched using unique identifiers (GB and TC Accessions, EST identifiers and ET numbers from the TIGR PREEGAD database), gene product names, functional classifications based on GO terms, metabolic pathways, library-related expression analysis, map position within various sequenced genomes, TOGs in the EGO database and alternative splice forms. Complete annotations for all of the ESTs and TCs in each TGI database are now also provided through the EST Annotator and TC Annotator features which provide comprehensive lists of sequences within each species-specific database.

All of the TIGR Gene Indices are available for download through the main page for each species. Downloads consist of six files, including a FASTA file for all unique sequences, the TC list, the component ESTs in each TC, GO analysis, predicted oligos and a README file.

Software

Many of the software tools used to create the TGI are available with source code to the research community through the TGI software tools website (http://www.tigr.org/tdb/tgi/software). The TGI Clustering tool (TGICL) (6) is a software system for fast clustering and assembly of large EST datasets. TGICL starts with a large multi-FASTA file (and an optional quality value file) and outputs the assemblies produced by CAP3 (5). Both clustering and assembly phases can be parallelized by distributing the searches and the assembly jobs across multiple CPUs, as TGICL can take advantage of either SMP or PVM (Parallel Virtual Machine) clusters. Other available software includes clview for viewing sequence assemblies in .ace format, SeqClean which is used to remove contaminating sequences from EST and gene sequences and cdbfasta/cdbyank which index FASTA-formatted files and can be used to rapidly extract sequences from them.

ACKNOWLEDGEMENTS

The authors wish to thank TIGR IT group for their database and computer system support. This work was supported by the US Department of Energy, grant DE-FG02-99ER62852 and the US National Science Foundation, grant DBI-9983070.

REFERENCES

1. Quackenbush,J., Liang,F., Holt,I., Pertea,G. and Upton,J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.*, **28**, 141–145.
2. 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.
3. Liang,F., Holt,I., Pertea,G., Karamycheva,S., Salzberg,S.L. and Quackenbush,J. (2000) An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.*, **28**, 3657–3665.
4. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
5. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
6. Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F., Parvizi,B., Tsai,J. and Quackenbush,J. (2002) TIGR Gene Indices Clustering Tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.
7. Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 138–148.
8. Wang,X. and Seed,B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, **19**, 796–802.
9. Lee,Y., Sultana,R., Pertea,G., Cho,J., Karamycheva,S., Tsai,J., Parvizi,B., Cheung,F., Antonescu,V., White,J., Holt,I., Liang,F. and Quackenbush,J. (2002) Cross-referencing eukaryotic genomes: TIGR Othologous Gene Alignments (TOGA). *Genome Res.*, **12**, 493–502.
10. Tsai,J., Sultana,R., Lee,Y., Pertea,G., Karamycheva,S., Anonescu,V., Cho,J., Parvizi,B., Cheung,F. and Quackenbush,J. (2001) RESOURCERER: a database for annotating and linking microarray resources with and cross species. *Genome Biol.*, **2**, software0002.1–software0002.4.
11. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.