TIDE – Terra Incognita Discovery Endeavor
Comprehensive EST assignment to GeneCards^{TM} genes

Maxim Shklar[1,*], Orit Shmueli[1], Liora Strichman-Almashanu[1], Michael Shmoish[1], Tsippy Iny-Stein[1], Marilyn Safran[2] and Doron Lancet[1]

Departments of [1]Molecular Genetics and [2]Biological Services (Bioinformatics Unit),
The Weizmann Institute of Science, Rehovot 76100, Israel.

Received line

ABSTRACT

The construction of a complete EST-based gene index has been an intricate task. We present TIDE, an automated system for associating each of the >5 million human ESTs with a known or de-novo defined gene. The pipeline is heavily based on existing GeneCards links to other EST-related resources. In a specific example, we were able to provide gene identities to an additional ~15,000 unassigned EST-based Affymetrix microarray probesets, a 50% increase relative to previous annotations. TIDE is expected to help complete a comprehensive EST-associated compendium of GeneCards genes.

Availability: http://genecards.weizmann.ac.il/TIDE/
Contact: maxim.shklar@weizmann.ac.il

INTRODUCTION

To date, gene indices are based mainly on full length mRNA sequences and genes predicted from genomic data. Since sequencing full length mRNA is both time consuming and costly, the mRNA sequences for many genes are not yet available. This results in the absence of numerous genes from major gene compendia including GeneCards (Rebhan et al., 1998; Safran et al., 2002), LocusLink (Pruitt and Maglott, 2001) and Ensembl (Hubbard et al., 2002).

High throughput methods have generated a substantial number (>5 million) of Expressed Sequence Tags (ESTs) (Adams et al., 1991), which now offer the most extensive window to the entire human transcriptome, and to the genes coded within it. Unfortunately, given their fragmentary nature (typically 400-600 bases) and inaccurate information (1-3% sequencing errors) (Schuler, 1997), assigning each of these ESTs to genes has been elusive. Previous and ongoing projects designed to address this problem, such as UniGene (Wheeler et al., 2004), DoTs (http://www.cbil.upenn.edu/downloads/DoTS/), AceView (Thierry-Mieg et al.) and TIGR (Quackenbush et al., 2001), which employed different strategies, including filtering EST data for contaminants (Sorek and Safer, 2003), clustering by sequence similarity, and generating consensus sequences and aligning them against the genome, have resulted in various gene lists that exhibit only partial overlap.

Our goal is to achieve association between the set of all ~5.5 million human ESTs to the set of ~35,000 human genes as defined in GeneCards. Heretofore unassociated ESTs can be proven to belong to an existing gene or, on the other hand, help define a new gene. Alternately, an EST may be demonstrated to be an artifact or contaminated, and should then be discarded.

To this end we are developing a software system, TIDE (Terra Incognita Discovery Endeavor) that offers association between ESTs and GeneCards genes. This system is based on the same concept underlying GeneCards: to sift, merge and integrate data retrieved from various external resources together with in-house generated experimental results. We applied TIDE to a sample subset of human ESTs and demonstrate its promise for generating a more comprehensive, integrative and accurate gene index than has been previously available.

RESULTS AND DISCUSSION

TIDE workflow

The workflow for EST annotation includes two aspects. For the first (Figure 1, left), EST clusters grouped by UniGene (build 162, Nov 2003) and DoTS (build 7, Nov 2003), as well as gene associations by AceView (build 33, July 2003), were retrieved. For each EST the gene associations (according to these resources) in terms of a LocusLink identifier were recorded. This identifier was later used to associate each EST with a specific GeneCards gene.

In the second aspect (Figure 1, right), genomic locations for the ESTs in question, which were obtained using *BLAT* (Kent, 2002), were downloaded from UCSC's genome

[*] To whom correspondence should be addressed

browser database (Human Apr. 2003, hg15, NCBI Build 33) (Karolchik et al., 2003) and compared with data from GeneLoc (Rosen et al., 2003), an exon-based system which forms part of the GeneCards suite of databases and integrates data from LocusLink and Ensembl to create a unified location for each gene. The genes that are located on the same genomic region as found by *BLAT* for a specific EST were recorded. In order to overcome the problem of unknown orientation of ESTs deposited into GenBank, especially those taken from the 3' ends of transcripts, UCSC's *polyInfo* program was used. This program checks for canonical splice signals at the genomic location of ESTs that undergo splicing according to their *BLAT* alignment, and determines the most probable alignment orientation of the EST accordingly. In addition, our GeneAnnot (Chalifa-Caspi et al., 2003) system, another GeneCards-related database, which links Affymetrix HGU95A-E GeneChip probe sets and GeneCards genes by aligning the probes sequences against full length mRNAs, was used to annotate ESTs with the same gene annotation as their associated probe sets .

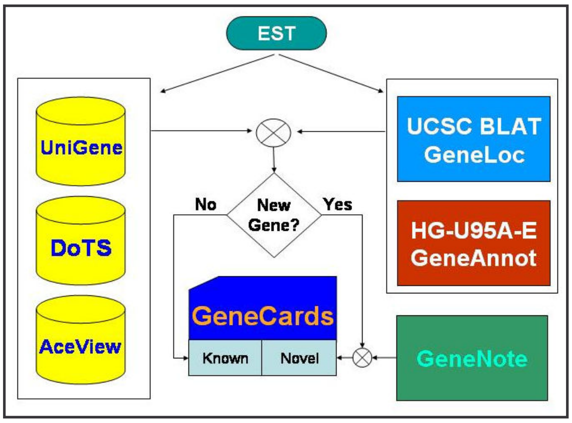

Figure 1. TIDE EST annotation workflow. External databases are queried for the possible gene origin of the EST in question. Genomic location data, and probe set to full length mRNA alignment were added for full annotation. ESTs with no clear link to any previously known genes were checked for expression patterns, and used to help define novel genes.

Integration and scoring scheme

The various gene annotations obtained for each EST by the five aforementioned methods (UniGene, DoTS, AceView, *BLAT* & GeneLoc, and GeneAnnot) are integrated into a single scoring scheme which is composed of two major parameters each ranging between 0 and 1. Consensus (*Co*) and Uniqueness (*Uq*) are defined for a pair (*E,G*), where *E* is an EST and *G* is a GeneCards gene; *Nresources(E,G)* is the number of resources supporting the annotation of *E* as originating from *G* ; W_R is a weight that can be assigned to each of the resources according to quality considerations, as follows:

$$Co(E,G) = \frac{Nresources(E,G)}{Ntotal\ resources} \quad (1)$$

$$Uq(E,G) = \frac{\sum_{R \in \substack{resources \\ annotating\ E}} \left(\frac{W_R}{Ngenes\ assigned\ to\ E\ by\ R} \right)}{\sum_{R \in resources\ annotaing\ E} W_R} \quad (2)$$

An integrative value designed to recapitulate these two previous parameters into one, ranging from 0 to 1 as well, was termed *score* and defined as:

$$Score(E,G) = \sqrt{\frac{Co^2 + Uq^2}{2}} \quad (3)$$

Last, the highest scoring gene was used for the final annotation of an EST.

Annotation of a test work set

As a test work set, we selected the ESTs used by Affymetrix to derive probesets for its HG-U95A-E GeneChips. This work set is of particular interest, as it has been used in the GeneNote (Shmueli O., 2003) project to determine the mRNA expression levels in 12 representative normal human tissues. Hence, the task of assigning a genetic meaning to each of these readings plays a key role in the analysis of the data.

In order to obtain the most comprehensive level of mapping between Affymetrix HG-U95A-E GeneChip probe sets and GeneCards genes, we applied the TIDE system to the set of transcripts (ESTs + mRNA) from which these probe sets were extracted. This initial work set contained 59,480 different transcribed sequences from which 62,667 probe sets were derived. The process led to the association of 43,468 transcripts to 23,271 GeneCards genes. These constitute nearly 75% of the probe sets in question, which is a significant increase in comparison to the annotation of 50% of the probe sets achieved by the latest versions of GeneAnnot (v0.3, Nov 2003) and by Affymetrix annotation. These results suggest that when applied to the entire set of all publicly available ESTs (work in progress) TIDE will be able to obtain more accurate and robust gene annotation for a significantly larger number of ESTs than is currently available at any of the individual resources.

Unannotated ESTs and de novo genes

Out of the remaining 16,012 unannotated transcripts, 3,016 are no longer members of any UniGene cluster in the currently used build #162, contrary to their status at the time UniGene build #95 was used to construct GeneChips HG-U95A-E. These probably correspond to genomic DNA contamination, to intronic sequences stemming from unprocessed RNA, or to other possible contaminants.

Among the probesets derived from the unannotated ESTs, 2,347 were of particular interest, as their GeneNote expression patterns varied significantly among the 12 tissues. An additional 1,035 probesets display a housekeeping gene expression pattern (Fig. 2). These two groups of probesets are currently being investigated via the integration of GeneNote expression results with information from other resources, so as to gain more insights towards defining them as potentially novel genes.

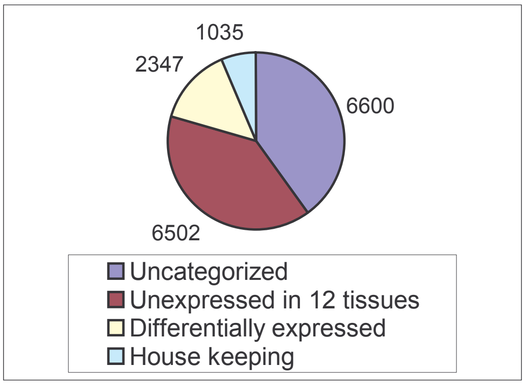

Figure 2. Distribution of unannotated probe sets from Affymetrix HG-U95A-E GeneChips into expression categories. Differentially expressed – 2,347 probe sets exhibit significantly more variation among tissues than among replicates of the same tissue. Another 1,035 show a housekeeping gene pattern. These two groups are strong candidates for being the most significant proof for the existence of genes yet unknown. The remaining probe sets were either not expressed in any of the GeneNote tissues (and thus require further analysis), or did not pass our internal quality checks (uncategorized).

In a specific example, the probeset 47305_at, derived from the EST D51371 (GenBank accession) exhibits a brain specific gene expression pattern, and shows high sequence identity (92%) to a brain related full length mRNA (GenBank accession AB060887) in a monkey (*Macaca fascicularis*). Aligned against the genome, this EST resides on the minus strand of chromosome 5 along with 10 other ESTs in a genomic region where no full length mRNA or genes have been found. A "same gene origin" hypothesis is confirmed for seven of these ESTs by UniGene, DoTS and AceView - strengthening the assumption that this is indeed a gene, which is even conserved in another species, despite not being defined as such by either LocusLink or Ensembl.

Our results suggest that TIDE, in conjunction with GeneNote tissue expression patterns, holds the potential to uncover a large number of novel genes, and could significantly accelerate the elucidation of an inclusive EST-annotated compendium of human genes.

ACKNOWLEDGEMENTS

The authors wish to thank Itai Yanai for his important suggestions. This work was funded by the Weizmann Institute Crown Human Genome Center, the Abraham and Judith Goldwasser Foundation, the Yeda Fund and XenneX Inc.

REFERENCES

Affymetrix - http://www.affymetrix.com/analysis/index.affx.

Adams, M. D., Kelley, J. M., et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." Science, **252**(5013): 1651-6.

Chalifa-Caspi V., Shmueli, O., et al. (2003). "GeneAnnot: Interfacing GeneCards with high throughput gene expression compendia." Brief. Bioinform., **4**(4): 349-360.

The Computational Biology and Informatics Laboratory. DoTS: a database of transcribed sequences for human and mouse genes. Center for Bioinformatics, University of Pennsylvania. http://www.cbil.upenn.edu/downloads/DoTS/.

Danielle and Jean Thierry-Mieg, Potdevin, M., et al. (unpublished). Identification and functional annotation of cDNA-supported genes in higher organisms using AceView. http://www.humangenes.org .

Hubbard, T., Barker, D., et al. (2002). "The Ensembl genome database project." Nucleic Acids Res., **30**(1): 38-41.

Karolchik, D., Baertsch, R., et al. (2003). "The UCSC Genome Browser Database." Nucleic Acids Res., **31**(1): 51-4.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res., **12**(4): 656-64.

Pruitt, K. D. and Maglott, D. R. (2001). "RefSeq and LocusLink: NCBI gene-centered resources." Nucleic Acids Res., **29**(1): 137-40.

Quackenbush, J., Cho, J., et al. (2001). "The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species." Nucleic Acids Res., **29**(1): 159-64.

Rebhan, M., Chalifa-Caspi, V., et al. (1998). "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support." Bioinformatics, **14**(8): 656-64.

Rosen, N., Chalifa-Caspi, V., et al. (2003). "GeneLoc: exon-based integration of human genome maps." Bioinformatics, **19 Suppl 1**: I222-I224.

Safran, M., Solomon, I., et al. (2002). "GeneCards 2002: towards a complete, object-oriented, human gene compendium." Bioinformatics, **18**(11): 1542-3.

Schuler, G. D. (1997). "Pieces of the puzzle: expressed sequence tags and the catalog of human genes."J. Mol. Med., **75**(10): 694-8

Shmueli O, Horn-Saban. S.et al. (2003). "GeneNote: whole genome expression profiles in normal human tissues." Proc. French Acad. Sci., Comptes Rendus Biologies **326**(10-11): 1067-1072.

Sorek, R. and Safer, H. M. (2003). "A novel algorithm for computational identification of contaminated EST libraries." Nucleic Acids Res., **31**(3): 1067-74.

Wheeler, D. L., Church, D. M., et al. (2004). "Database resources of the National Center for Biotechnology Information: update." Nucleic Acids Res., **32**(1): D35-40.