

GeneTide – Terra Incognita Discovery Endeavor Mining ESTs and Expression Data to Elucidate Known and De-Novo GeneCards® Genes

Maxim Shklar¹, Orit Shmueli¹, Liora Strichman-Almashanu¹, Michael Shmoish¹,
Doron Lancet¹ and Marilyn Safran²

Departments of¹Molecular Genetics and²Biological Services (Bioinformatics Unit),
The Weizmann Institute of Science, Rehovot 76100, Israel

maxim.shklar@weizmann.ac.il, marilyn.safran@weizmann.ac.il

Abstract

The construction of a complete EST-based gene index is an intricate task yet to be accomplished. GeneTide [1], the Gene Terra Incognita Discovery Endeavor (<http://genecards.weizmann.ac.il/genetide/>), which is the newest addition to the GeneCards [2] [3] suite of databases, comprehensively maps >4.5 of the ~5.5 million human ESTs currently available at dbEST with either known or newly defined putative human genes. The association is accomplished via data mining genomic resources, and integrating using a unified scoring scheme. Groups of unassociated transcripts serve as a basis for defining EST-based Gene Candidates (EGCs). These EGCs are annotated with various parameters, including expression data, to determine their validity as possible de-novo genes. An immediate application of GeneTide to microarray annotation has increased, in a specific example, the number of annotated Affymetrix HGU95A-E probe sets by 50% in comparison to previous attempts.

1. Introduction

High throughput methods have generated a substantial number (>5 million) of Expressed Sequence Tags (ESTs) [4], 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) [5], assigning each of these ESTs to genes has been elusive. Previous and ongoing projects designed to address this problem, such as UniGene [6], DoTs [7],

and AceView [8], employed different strategies, and have resulted in various gene lists that exhibit only partial overlap.

We have developed GeneTide, an automated system that offers association between ESTs and GeneCards genes, as well as elucidation of candidate genes based on EST clusters supported by expression data. This system is founded on the same concept underlying GeneCards: to sift, merge and integrate data retrieved from various external resources together with in-house generated experimental results. GeneTide significantly decreases the number of previously unassociated ESTs, offers a large number of new putative groups of transcripts as candidate genes for further research, and improves the quality of microarray annotation.

2. Results and discussion

GeneTide's workflow consists of two stages – association with existing GeneCards genes, and defining new ones (Figure 1). For the first stage, EST clusters grouped by UniGene and DoTs, as well as gene associations by AceView, were retrieved. For each EST the gene associations via LocusLink identifier were recorded. This identifier was later used to associate each EST with a specific GeneCards gene.

Next, genomic locations for the ESTs in question, which were obtained using *BLAT* [9], were downloaded from UCSC's genome browser database [10] and compared with data from GeneLoc [11], our exon-based system which integrates data from LocusLink and Ensembl to create a unified location for each gene. Genes located on the same genomic region

as found by BLAT for a specific EST were recorded. In addition, the GeneAnnot [12] system, the GeneCards family database which links Affymetrix 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.

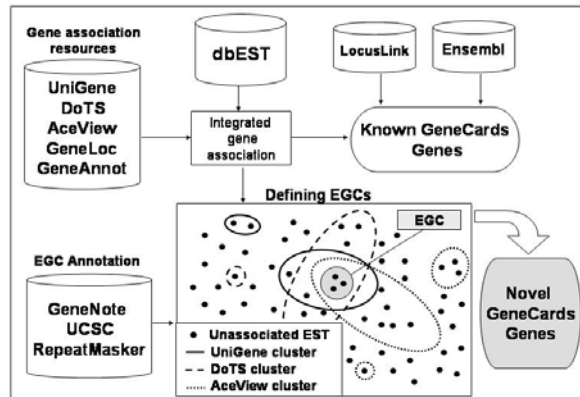


Figure 1. GeneTide workflow. Transcripts are either associated with known genes, used to define EST based GeneCards Candidates (EGCs), or discarded as artifacts.

The various gene annotations obtained for each EST by the five aforementioned methods (UniGene, DoTS, AceView, BLAT & GeneLoc, and GeneAnnot) were integrated into a consensus and uniqueness based scoring scheme [1], which ranks the possible gene annotations by quality.

In defining de-novo genes, non-singleton sets of previously unassociated ESTs, which were supported by UniGene, DoTS and AceView as originating from the same gene, were defined as EST based GeneCards Candidates (EGCs), yielding nearly 20,000 such sets in GeneTide's version 0.2. These EGCs were further annotated by extracting splicing site evidence, and validated by querying GeneNote [13] for the normal tissue expression of probe sets derived from the EGC's underlying transcripts. The information collected for each EGC was used for testing the validity of its candidacy as a true gene. We are working on developing a ranking mechanism that will lead to a list of putative genes sorted by their validity, so that future experimental efforts can be directed at the most likely genes first.

An additional application of GeneTide is in defining the gene origin of microarray probe sets. By assigning probe sets with the gene origin annotation given (via

the process described above) to the transcript from which they were derived, we were able to increase by 50% the number of Affymetrix HGU95A-E probe sets annotated in comparison to previous attempts by GeneAnnot [12] (version 0.3).

These results suggest that GeneTide 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.

2. Acknowledgements

This work was funded by the Weizmann Institute Crown Human Genome Center and the Abraham and Judith Goldwasser Foundation, and by Xenex Inc.

3. References

- [1] M. Shklar, et al., "TIDE – Terra Incognita Discovery Endeavor. Comprehensive EST assignment to GeneCards genes.", presented at ISMB/ECCB, 2004.
- [2] M. Safran, et al., "GeneCards 2002: towards a complete, object-oriented, human gene compendium," *Bioinformatics*, vol. 18, pp. 1542-3, 2002.
- [3] M. Rebhan, et al., "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support," *Bioinformatics*, vol. 14, pp. 656-64, 1998.
- [4] M. D. Adams, et al., "Complementary DNA sequencing: expressed sequence tags and human genome project," *Science*, vol. 252, pp. 1651-6, 1991.
- [5] G. D. Schuler, "Pieces of the puzzle: expressed sequence tags and the catalog of human genes," *J Mol Med*, vol. 75, pp. 694-8, 1997.
- [6] D. L. Wheeler, et al., "Database resources of the National Center for Biotechnology Information: update," *Nucleic Acids Res*, vol. 32, pp. D35-40, 2004.
- [7] "DoTS: a database of transcribed sequences for human and mouse genes. Center for Bioinformatics, University of Pennsylvania."
- [8] M. Potdevin, et al., "Identification and functional annotation of cDNA-supported genes in higher organisms using AceView (unpublished)."
- [9] W. J. Kent, "BLAT--the BLAST-like alignment tool," *Genome Res*, vol. 12, pp. 656-64, 2002.
- [10] D. Karolchik, et al., "The UCSC Genome Browser Database," *Nucleic Acids Res*, vol. 31, pp. 51-4, 2003.
- [11] N. Rosen, et al., "GeneLoc: exon-based integration of human genome maps," *Bioinformatics*, vol. 19 Suppl 1, pp. I222-I224, 2003.
- [12] V. Chalifa-Caspi, et al., "GeneAnnot: interfacing GeneCards with high-throughput gene expression compendia," *Brief Bioinform*, vol. 4, pp. 349-60, 2003.
- [13] O. Shmueli, et al., "GeneNote: whole genome expression profiles in normal human tissues," *C R Biol*, vol. 326, pp. 1067-72, 2003.