D556–D561 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki122

GeneTide—Terra Incognita Discovery Endeavor: a new transcriptome focused member of the GeneCards/GeneNote suite of databases

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

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

Received August 2, 2004; Revised and Accepted October 21, 2004

ABSTRACT

GeneCards® is an automatically mined database of human genes that strives to create, along with its auxiliary databases—GeneLoc, GeneNote and GeneAnnot—the most inclusive resource of gene-centered information of the human genome. GeneTide, the Gene Terra Incognita Discovery Endeavor (http://genecards.weizmann.ac.il/genetide/), the newest addition to this family, is a transcriptome-focused database which aims to enhance GeneCards with additional expressed sequence tag (EST)-based genes. This is achieved by comprehensively mapping >85% of the ~5.6 million human ESTs currently available at dbEST to known genes by means of data mining and integration of genomic resources including UniGene, DoTS, AceView and in-house resources. GeneTide thus creates comprehensive links between ESTs and GeneCards genes. Furthermore, groups of unassociated transcripts serve as a basis for defining novel EST-based GeneCards Candidates (EGCs). These EGCs, nearly 25 000 of which were defined in version 0.3 of GeneTide, are further annotated with various parameters, including splicing evidence and expression data extracted from the GeneNote database, to determine their validity as possible *de novo* genes.

INTRODUCTION

GeneCards (http://bioinfo.weizmann.ac.il/genecards/) (1) is an automated, integrated database of human genes, proteins and associated phenotypes, which strives to provide a comprehensive and concise human gene compendium. The GeneCards suite of databases includes (i) GeneLoc (http://genecards.weizmann.ac.il/geneloc/) (2), an exon-based system that integrates data from LocusLink (3) and Ensembl (4) (which together with HUGO constitute the core of the GeneCards gene set) to create a unified location for each gene; (ii) GeneAnnot (http://genecards.weizmann.ac.il/geneannot/) (5), which links Affymetrix GeneChip probe sets with GeneCards genes, by aligning the probe sequences to transcripts; and (iii) GeneNote (http://genecards.weizmann.ac.il/genenote/) (6), the Gene Normal Tissue Expression project, which portrays gene expression profiles in healthy human tissues using Affymetrix GeneChip experiments as well as serial analysis of gene expression (SAGE) and electronic northern results. GeneNote additionally offers gene-centered expression attributes such as figurative expression profiles, novel tissue specificity indices and classification of binary patterns.

To date, gene resources are based mainly on full-length mRNA sequences, and on 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. Although new genes are continuously being uncovered via ongoing full-length sequencing projects (7), it appears that numerous genes are still absent from major gene compendia including GeneCards, HUGO, LocusLink and Ensembl.

Since the early 1990's, high-throughput methods were able to generate a substantial number (>5.6 million) of human expressed sequence tags (ESTs) (8), which now offer the most comprehensive 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) (9), assigning each of these ESTs to a gene has been incomplete. Previous and ongoing projects designed to address this problem, such as UniGene (http://www.ncbi.nlm.nih.gov/UniGene) (10), AceView (http://www.aceview.org/) and

*To whom correspondence should be addressed. Tel: +972 8 934 3455; Fax: +972 8 934 4113; Email: maxim.shklar@weizmann.ac.il
Correspondence may also be addressed to Marilyn Safran. Tel: +972 8 934 3455; Fax: +972 8 934 4113; Email: marilyn.safran@weizmann.ac.il

DoTS (http://www.allgenes.org/), have employed different strategies, which resulted in various gene collections exhibiting only partial mutual overlap. The relevant counts are 106 937 UniGene gene-oriented clusters (build 171), 652 095 DoTS genes (release 8) and 269 134 AceView genes (build 34), many of which are singletons or contain very few member transcripts; all vastly exceed recent estimates of the true count of human genes (25 000–40 000) (11). Only a small number of these clusters (for instance, 21 638 out of 106 937 in UniGene) are associated with known genes; this leaves the majority of such clusters with no related indexed gene. Further, no attempt has yet been made to compare and integrate the different clusters created by the various databases, or to settle the discrepancies among them. Hence, an efficient mechanism for integrating data from such transcriptome-centered resources is clearly called for.

To this end we have developed GeneTide, a transcriptome-focused automated system for determining association between ESTs and GeneCards genes, as well as for the elucidation of new putative 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, improves the quality of microarray annotation and offers a large number of new groups of transcripts (∼ 25 000 in version 0.3 of GeneTide) as candidate genes for further research.

RESULTS AND DISCUSSION

GeneTide is an automated data integration system, which consists of two major phases (Figure 1). In the first phase, transcripts (ESTs and mRNAs) are associated with existing GeneCards genes using a unified scoring scheme to help overcome differences between resources. In the second phase, new candidate genes are defined based on the previously unassociated transcripts. These candidate genes are further annotated and ranked in order to help focus one's attention on those that hold the greatest promise for discovering valid novel genes. GeneTide, like other members of the GeneCards suite of databases can be expected to be updated roughly on a quarterly basis, synchronized with the NCBI and Ensembl databases whenever possible.

Transcript-gene association

Data resources. For GeneTide's first phase, we retrieved transcript clusters created by UniGene, DoTS and AceView, and recorded gene associations (via LocusLink identifier). These identifiers were later used to link all of the transcripts contained within each cluster to the corresponding GeneCards gene.

Next, genomic locations for the transcripts, which were obtained using BLAT (12), were downloaded from the UCSC's genome browser database (13) and compared with data from GeneLoc. Genes with exons located at a genomic region found by BLAT for a specific transcript were recorded. In order to overcome the problem of unknown orientation for some of the ESTs deposited into GenBank, UCSC's polyInfo program results were retrieved. 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. Finally, transcripts that served as a basis for deriving Affymetrix GeneChip HGU95A-E probe sets were associated with GeneCards genes according to the GeneAnnot annotation of such probe sets.

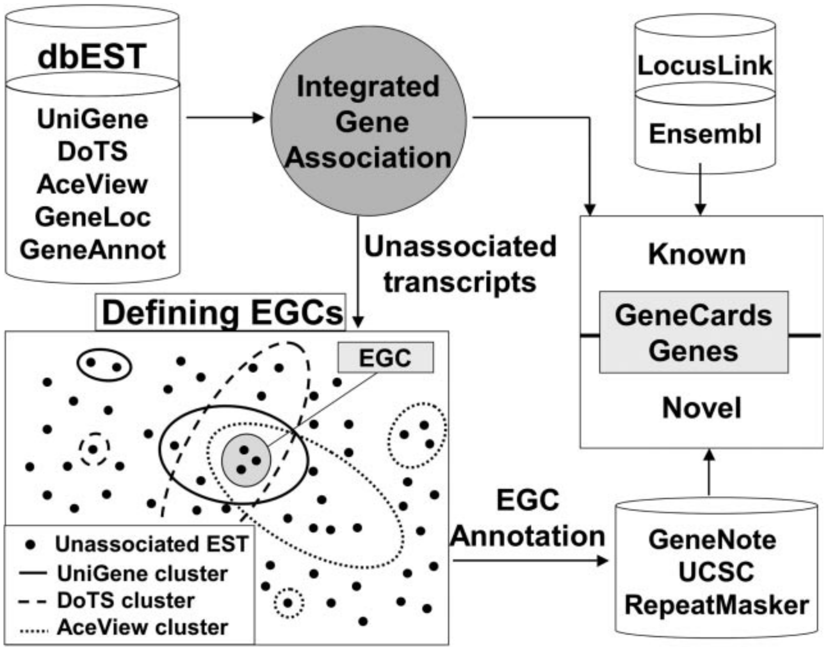


Figure 1. GeneTide workflow. Transcripts are either associated with known genes, used to define EST-based GeneCards Candidates (EGCs), or discarded as artifacts.

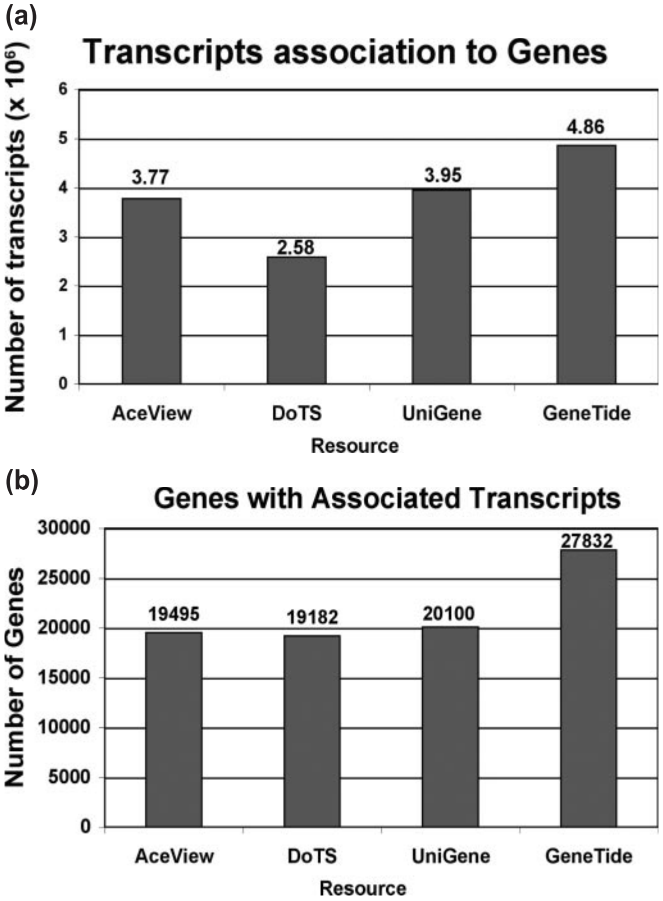


Figure 2. Association between transcripts and genes. GeneTide maps ~4.8 million transcripts (**a**) to 27 832 genes (**b**). This is a significant increase in the number of associated transcripts and genes in comparison to UniGene, DoTS and AceView according to information extracted from their respective websites.

Transcript validation. In order to assess the probability that transcripts are genuine gene products rather than some genomic artifact or contamination, a 'validation level' ranging from 0 (lowest) to 4 (highest) is assigned to each transcript. A point is deducted from the transcript's initial maximum for each of the resources (UniGene, DoTS and AceView) that filtered out the transcript according to its own validation criteria. An additional point is deducted if a repetitive element was identified by RepeatMasker along the transcript. The distribution of the validation level among GeneTide's transcripts is offered in the Supplementary Material.

Integration and scoring scheme. The various gene annotations obtained for each transcript from the five afore-mentioned sources (UniGene, DoTS, AceView, BLAT and GeneLoc, and GeneAnnot) were integrated into a single scoring scheme, which ranks possible gene annotations by quality. Consensus (Co) (1) and Uniqueness (Uq) (2), both ranging from 0 to 1, are defined for a pair (E,G), where E is an EST and G is a GeneCards gene as follows:

$$\mathrm{Co}(E,G) = \frac{N_{\mathrm{resources}(E,G)}}{N_{\mathrm{total\ resources}}}, \qquad 1$$

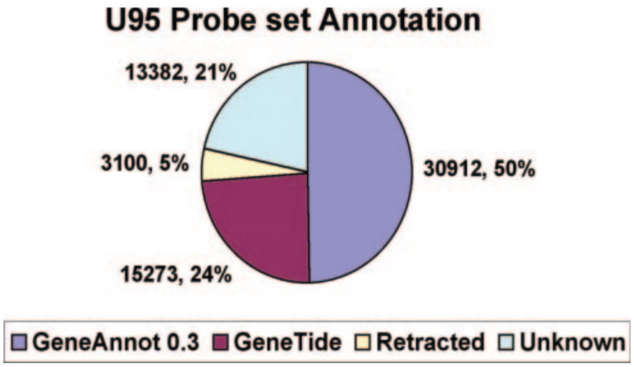


Figure 3. Annotation of Affymetrix 62 667 probe sets on GeneChip arrays U95A-E. GeneTide has been able to increase the number of probe sets with an assigned putative gene origin by ~50% in comparison to other resources. Another 5% of the probe sets have been retracted from UniGene (where they were originally taken from) since the production of the U95 set.

$$\mathrm{Uq}(E,G) = \frac{\sum_{\mathrm{R}\in \text{ resources annotating } E}\left(\frac{W_{\mathrm{R}}}{N_{\text{genes assigned to } E \text{ by R}}}\right)}{\sum_{\mathrm{R}\in\text{resources annotating } E} W_{\mathrm{R}}}, \qquad 2$$

where $N_{\mathrm{resources}(E,G)}$ is the number of resources supporting the annotation of E as derived from G, and W_{R} is a weight assigned to each of the resources according to quality considerations. An integrative value (Score) designed to collapse these two previous parameters into one, ranging from 0 to 1 was defined as follows:

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

Finally, the gene that achieved the highest score among all related genes was defined as the gene from which the EST originated. More than 4.8 million transcripts out of an initial number of nearly 5.6 million were associated in this manner to 27 832 GeneCards genes in GeneTide's version 0.3. This is a significant increase in both quantity (Figure 2) and quality of association, between transcripts and genes, in comparison to any of the other individual projects, which we attribute to the unique integrative nature of GeneTide.

Microarray annotation. An immediate application of GeneTide is defining the gene of origin for microarray probe sets. Affymetrix (www.affymetrix.com) probe sets are short subsequences derived from transcripts for the purpose of defining tags for those transcripts in hybridization experiments. Unfortunately, in many cases, the identity of the gene remains unknown, rendering the valuable experimental expression data devoid of its genetic context. In order to overcome this problem, we annotated probe sets with the same gene origin as given by GeneTide (via the process described above) to the transcripts from which they were derived. In a specific example (Figure 3), this has allowed us to increase by nearly 50% (from 30 912 to 46 185 out of the entire set of 62 667 probe sets) the number of Affymetrix HGU95A-E probe sets annotated in comparison with previous attempts by GeneAnnot version 0.3.

Defining novel putative genes

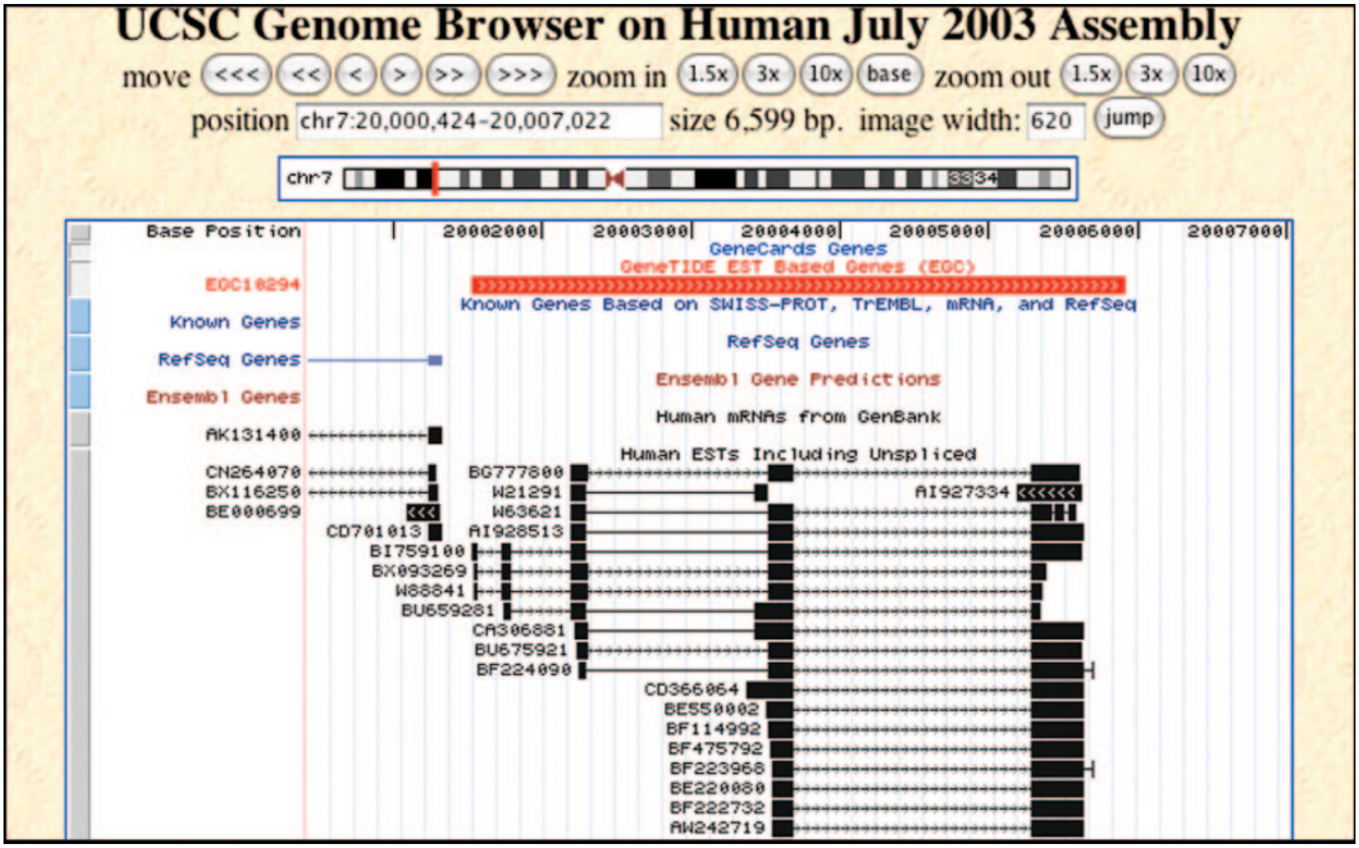
For the purpose of defining *de novo* genes, non-singleton sets of transcripts found to belong to the same cluster by all sources and not associated previously with a known GeneCards gene, were defined as EST-based GeneCards Candidates (EGCs) (Figure 1, bottom). Altogether, 25 000 EGCs were defined, and associated with 110 000 transcripts in GeneTide.

To distinguish the bona fide novel genes from other possible sources of EGCs, such as artifacts or exons of known genes, we have taken the EGCs through an additional phase of annotation. To assess the probability of a certain EGC representing a valid gene rather than being an artifact, several parameters were considered. These include the number of transcripts the EGC is composed of (n), and the number of them which undergo splicing (s) or contain repetitive elements (r)—information extracted from the UCSC genome browser database. Expression patterns of probe sets, derived from transcript members of EGCs, retrieved from the GeneNote database, helped to confirm the validity of the transcribed sequences, and focus attention on the tissues where these putative genes are most likely expressed; they also provide initial insight into its possible function. The number of derived probe sets found by the GeneNote project to be expressed is denoted (x). All these parameters were integrated into a single *Validity* function V (4).

$$V(n,s,r,x) = \sqrt{n} \times \left(0.5 + \frac{s}{n}\right) * \left(1.5 - \frac{r}{n}\right) \times 2^{\sqrt{x}}. \qquad 4$$

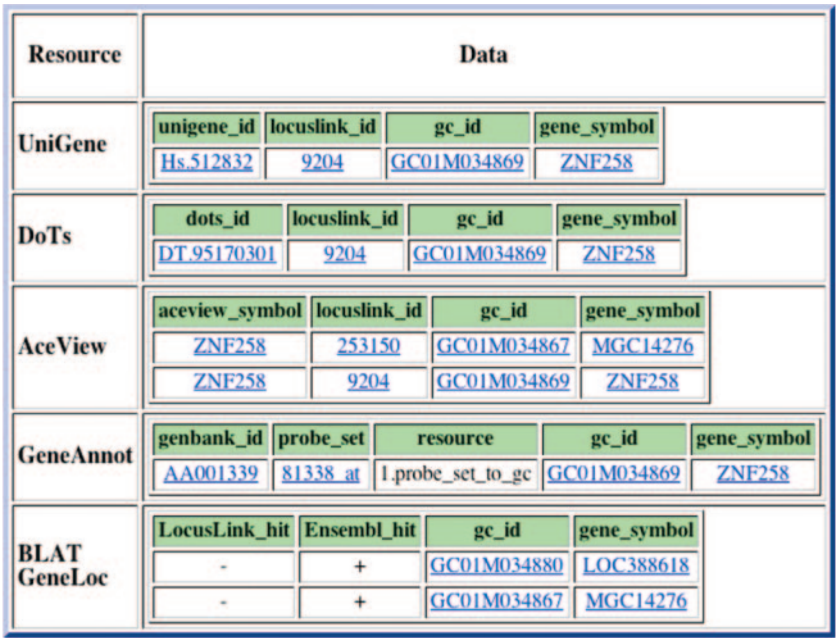
Taking into account the possibility that many of the EGCs may constitute previously unidentified exons of known genes, the average distance (d) of the transcripts constituting the EGC from their nearest known gene was also noted. In order to determine whether the resulting distance is typical of an intron, the distribution of intron length (of 185 175 introns from 26 602 LocusLink genes included in GeneCards version v2.30, which corresponds to NCBI build #34) was analyzed (see Supplementary Material). The analysis shows that human introns in general tend to be short, so the probability that an EGC is an exon of its neighboring gene decreases as the distance from that gene increases. To quantify this, we defined the *Novelty* function $N(d)$ to be the percentage of introns in the above mentioned dataset that are shorter than the distance to the nearest known gene (d).

The product of the *Validity* of an EGC and its *Novelty* was defined as the *Quality* of the EGC. These quality values provide an efficient manner for highlighting those EGCs (see Supplementary Material for the distribution of the quality score across EGCs) which hold the greatest potential to be


UCSC Genome Browser on Human July 2003 Assembly

position chr7:20,000,424-20,007,022 size 6,599 bp. image width: 620

Figure 4. EGC10294 displayed on UCSC's genome browser. The EGC is located on the plus strand of chromosome 7 at the genomic region of 20.001–20.007 Mb, in a region where no LocusLink, Ensembl or GeneCards gene is known. The graphic display shows the genomic location of EGCs along with predicted transcription orientations, and is available for each EGC via GeneTide's website along with a track of all currently known GeneCards genes. This display was made possible thanks to UCSC's custom track mechanism (http://genome.ucsc.edu/goldenPath/help/customTrack.html).



Resource	Data				
UniGene	unigene_id	locuslink_id	gc_id	gene_symbol	
	Hs.512832	9204	GC01M034869	ZNF258	
DoTs	dots_id	locuslink_id	gc_id	gene_symbol	
	DT.95170301	9204	GC01M034869	ZNF258	
AceView	aceview_symbol	locuslink_id	gc_id	gene_symbol	
	ZNF258	253150	GC01M034867	MGC14276	
	ZNF258	9204	GC01M034869	ZNF258	
GeneAnnot	genbank_id	probe_set	resource	gc_id	gene_symbol
	AA001339	81338_at	1.probe_set_to_gc	GC01M034869	ZNF258
BLAT GeneLoc	LocusLink_hit	Ensembl_hit	gc_id	gene_symbol	
	-	+	GC01M034880	LOC388618	
	-	+	GC01M034867	MGC14276	

Figure 5. Transcript query results in GeneTide. GeneTide provides detailed information regarding known genes that the queried transcript is associated with according to UniGene, DoTS, AceView, GeneAnnot and GeneLoc. This includes the identifier of the relevant cluster in each resource, the GeneCards identifier (GC ID) and symbol along with hyperlinks to the resources. Additional information available via GeneTide's website for each transcript includes the validation level of the transcript, its genomic location, known splicing sites, existence of repetitive elements and neighboring genes.

new genes—a higher score indicating a higher likelihood of being a genuine gene. Lower scoring EGCs might turn out to be either alternative exons of known genes or artifacts/ contaminations.

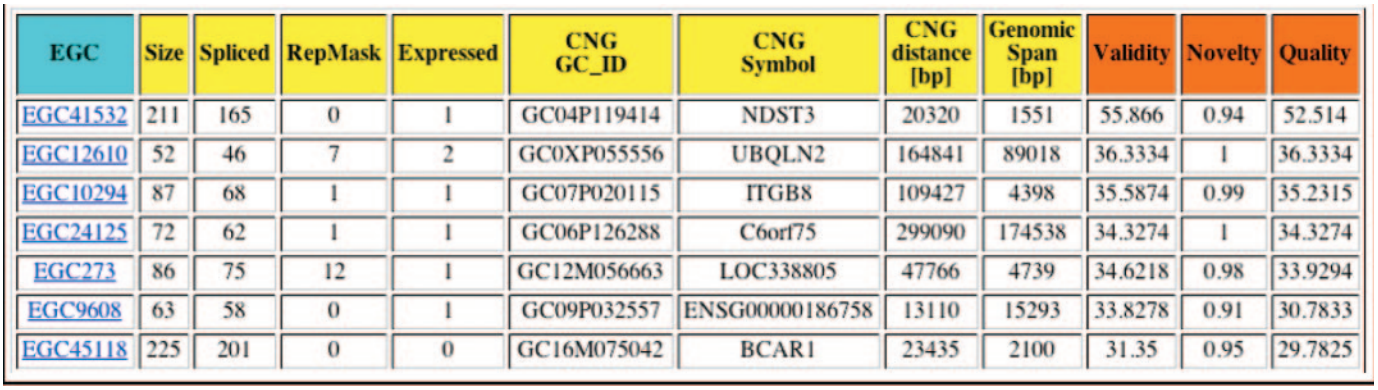
An example for a high scoring EGC is EGC10294, one of the top five scoring EGCs in GeneTide (e.g. see EGCs at http://genecards.weizmann.ac.il/genetide/example_EGC.html), which consists of 87 ESTs, 68 of which have known splice sites. This EGC was constructed from UniGene, DoTS and AceView gene clusters Hs.59203, DG.55117 and 'roky', respectively. The EGC is located on the plus strand of chromosome 7 at the genomic region of 20.001–20.007 Mb, and the structure delineated by the underlying multiple ESTs exhibits a clear pattern of introns and exons (Figure 4). The average distance of the ESTs of this EGC to the nearest known gene on the same strand (ITGB8) is 109 427 bp, which corresponds to a very low probability of 0.01 (calculated by using the *Novelty* function) of this EGC being an exon of the known gene. In addition, the U95 probe set 53201_at derived from the EST AI928513 exhibits an expression pattern in which it is highly expressed in the kidney and pancreas, moderately expressed in the lung and spleen and expressed in a lower level in the remaining tissues. This EGC is also supported by independent gene predictions by TIGR (human_THC1807481) and partially by Ensembl EST genes (ENSESTG00000015877).

In another example, EGC24125, a cluster of 72 transcripts, 62 of which have splice sites, aligned on the plus strand of chromosome 6 at the genomic location of 126.47–126.64 Mb. These transcripts are contained in each of the following gene oriented clusters—UniGene cluster Hs.35962, DoTS gene DG.73559 and AceView gene 'garjer', but are nevertheless not associated with any of the genes in GeneCards, LocusLink or Ensembl. Gene predictions corresponding to this EGC were independently confirmed by other EST-based gene prediction projects such as Ensembl's Vega (EM: Em:AC020559.3) and EST gene (ENSESTG00000003851). Several other EGCs defined in this genomic region appear to be the result of alternative transcripts of the predicted gene. The nearest known gene to this EGC is nearly 300 000 bp away, making the probability very low that this is one of its exons. In addition, one of the ESTs—accession W55924, which had a U95 probe set (51263_at) derived from it, exhibits, according to GeneNote data, an immune system specific expression pattern, since it is highly expressed in the thymus and bone marrow, but not expressed in any of the other tissues. This expression data offers initial insight as to the possible function of the suggested EGC, information that will facilitate future examinations of this putative gene.

GeneTide website

The GeneTide website offers extensive detailed information regarding queried transcripts. This information includes the gene associations of a transcript according to a variety of resources as described above (Figure 5). In addition, the genomic location of the transcript as well as the genes that reside in the same or nearest area are presented. Queries submitted to the GeneTide database include Affymetrix probe set identifiers (currently GeneChips U95 and U133 are supported), which result in all the information related to the transcript from which the queried probe set was derived.

A batch query mechanism incorporated into GeneTide allows for querying a large number of ESTs simultaneously, by uploading a list of ESTs to the site. The result is a table of highest scoring GeneCards genes associated with each transcript, or the identity of an EGC in which the transcript is



EGC	Size	Spliced	RepMask	Expressed	CNG GC_ID	CNG Symbol	CNG distance [bp]	Genomic Span [bp]	Validity	Novelty	Quality
EGC41532	211	165	0	1	GC04P119414	NDST3	20320	1551	55.866	0.94	52.514
EGC12610	52	46	7	2	GC0XP055556	UBQLN2	164841	89018	36.3334	1	36.3334
EGC10294	87	68	1	1	GC07P020115	ITGB8	109427	4398	35.5874	0.99	35.2315
EGC24125	72	62	1	1	GC06P126288	C6orf75	299090	174538	34.3274	1	34.3274
EGC273	86	75	12	1	GC12M056663	LOC338805	47766	4739	34.6218	0.98	33.9294
EGC9608	63	58	0	1	GC09P032557	ENSG00000186758	13110	15293	33.8278	0.91	30.7833
EGC45118	225	201	0	0	GC16M075042	BCAR1	23435	2100	31.35	0.95	29.7825

Figure 6. Members of the EGC list with highest *quality* values. A full list of EGCs is provided in GeneTide's website.

included. This greatly facilitates microarray annotation, as it allows for uploading the list of all transcripts from which the probe sets of a certain microarray were derived. These transcripts are then associated by GeneTide to genes, an association which is transitively applied back to the probe sets.

From a gene-centered perspective, querying for a specific GeneCards gene provides a list of all transcripts associated with that gene ordered by the quality of the association, so that the best matching sequences appear first. These sequences are also provided through the GeneCards website.

GeneTide's website offers a list of all EGCs (Figure 6) sorted by their *quality*, as described above. For each EGC, details about the identity, genomic location and other attributes of the contained transcripts, including expression patterns of derived probe sets, are shown.

SUMMARY

GeneTide offers a significant improvement, both in quantity and in quality, of association between members of the human transcriptome and the set of known genes. GeneTide's website provides an integrated resource of data regarding individual transcripts as well as an efficient means for microarray annotation. In addition, GeneTide's EGC concept offers an important step toward elucidating a large number of novel genes.

Future plans for GeneTide include the addition of sources such as Ensembl's Vega and EST genes, TIGR and STACKDB for both phases of associating transcripts with known genes and defining novel genes. Annotation of putative genes will be further enhanced with the addition of expression data from other databases besides GeneNote such as GNF. Introduction of homology comparisons as well as validation with other gene prediction programs are expected to boost GeneTide's performance even further. This will help to diminish the terra incognita, toward making GeneCards a truly comprehensive compendium of human genes.

ACKNOWLEDGEMENTS

The authors wish to thank Ron Ophir, Itai Yanai, Tsippi Iny Stein and the reviewer for their important suggestions and help. This work was funded by the Weizmann Institute Crown Human Genome Center and the Abraham and Judith Goldwasser Foundation, and by XenneX Inc.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

REFERENCES

1. Safran,M., Chalifa-Caspi,V., Shmueli,O., Olender,T., Lapidot,M., Rosen,N., Shmoish,M., Peter,Y., Glusman,G., Feldmesser,E. *et al*. (2003) Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*., **31**, 142–146.
2. Rosen,N., Chalifa-Caspi,V., Shmueli,O., Adato,A., Lapidot,M., Stampnitzky,J., Safran,M. and Lancet,D. (2003) GeneLoc: exon-based integration of human genome maps. *Bioinformatics*, **19** (Suppl. 1), I222–I224.
3. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*, **29**, 137–140.
4. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al*. (2004) Ensembl 2004. *Nucleic Acids Res*., **32**, D468–D470.
5. Chalifa-Caspi,V., Yanai,I., Ophir,R., Rosen,N., Shmoish,M., Benjamin-Rodrig,H., Shklar,M., Stein,T.I., Shmueli,O., Safran,M. *et al*. (2004) GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes. *Bioinformatics*, **20**, 1457–1458.
6. Shmueli,O., Horn-Saban,S., Chalifa-Caspi,V., Shmoish,M., Ophir,R., Benjamin-Rodrig,R., Safran,M., Domany,E. and Lancet,D. (2003) GeneNote: whole genome expression profiles in normal human tissues. *Proc. French Acad. Sci. Comptes. Rendus. Biologies*, **326**, 1067–1072.
7. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al*. (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nature Genet*., **36**, 40–45.
8. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.
9. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med*., **75**, 694–698.
10. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **32**, D35–D40.
11. Pennisi,E. (2003) Bioinformatics. Gene counters struggle to get the right answer. *Science*, **301**, 1040–1041.
12. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
13. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.