

# The GeneCards™ Family of Databases: GeneCards, GeneLoc, GeneNote and GeneAnnot

Marilyn Safran<sup>2</sup>, Vered Chalifa-Caspi<sup>2</sup>, Orit Shmueli<sup>1</sup>, Naomi Rosen<sup>1</sup>, Hila Benjamin-Rodrig<sup>1</sup>, Ron Ophir<sup>2</sup>, Itai Yanai<sup>1</sup>, Michael Shmoish<sup>1</sup>, and Doron Lancet<sup>1</sup>

Depts of <sup>1</sup>Molecular Genetics and <sup>2</sup>Biological Services (Bioinformatics and Biological Computing Unit), the Weizmann Institute of Science, 76100 Rehovot, Israel

Contact: [marilyn.safran@weizmann.ac.il](mailto:marilyn.safran@weizmann.ac.il)

## Abstract

The popular **GeneCards™** integrated database of human genes, genomic maps, proteins and diseases [1,2] has recently spawned three related functional genomics efforts. As sequence data rapidly accumulates, the bottleneck in biology shifts from data production to analysis; researchers seek a profound understanding of the role of each gene, and of the way genes function together. **GeneLoc** [3] integrates human gene collections by comparing genomic coordinates at the exon level, eliminating redundancies, and assigning unique and meaningful location-based identifiers. **GeneCards** expression tissue vectors are provided by **GeneNote** [4], the first effort to present sophisticated expression analyses for a variety of normal human tissues using the full complement of gene representations (Affymetrix arrays HG-U95A-E). The **GeneAnnot** system [5] aligns probe-sets with the major public repositories of human mRNA sequences, and provides detailed annotation for each probe-set, with links to GeneCards.

## 1. GeneCards and Companion Databases

GeneCards (<http://bioinfo.weizmann.ac.il/genecards/>) is an automated, integrated database of human genes, genomic maps, proteins and diseases, with retrieval, consolidation, search and display software, striving to provide *just the right mix* of explicit textual and graphical information on a page on the one hand, with links to more detail on the other. Its infrastructure is being upgraded to use object-oriented Perl and XML with schema-driven display code and context-specific searches. Currently, GeneCards sifts and/or links to data from 42 resources. Figure 1 shows the functional groupings presented in GeneCards, with arrows linking to the synergistic contributions of the companion Weizmann Institute of Science Genome Center databases described in more detail below.

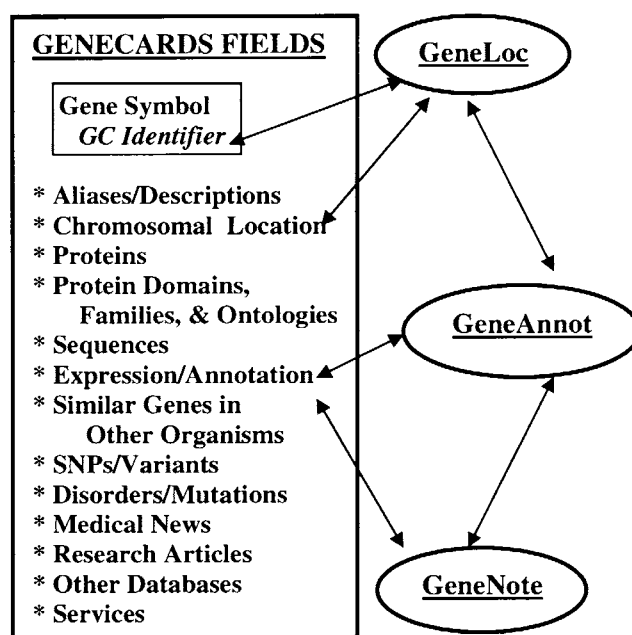


Figure 1: GeneCards/companion databases interactions

## 2. GeneLoc

GeneLoc (<http://genecards.weizmann.ac.il/geneloc/>) unifies gene collections from LocusLink [6] and Ensembl [7], eliminates redundancies, and assigns each gene a meaningful location-based GeneCards identifier (GC ID). Since both collections use the same genomic assembly and coordinate scheme, GeneLoc effects this gene integration by comparing genomic locations. The resulting 'gene territory' reflects the range of the unified genes, taking into account every exon.

When upgrading to new versions, we have grappled with keeping these GC IDs both persistent and consistent, while factoring in changes in gene location due to new elucidations of the human genome. New identifiers are assigned only if a gene's position has changed significantly by moving to a different chromosome or strand, or beyond a threshold (currently 100kb). Retired identifiers remain (as aliases) with their original genes

### 3. GeneNote

GeneNote (<http://genecards.weizmann.ac.il/genenote>), is a database of normal human gene tissue expression based on in-house DNA array experiments using the full complement of gene representations (Affymetrix GeneChip HG-U95A-E). The goals of the project are: to create a tissue expression profile for each human gene; to learn about the function of unknown genes from the similarities in their expression profiles to known genes using clustering and simplified k-nearest neighbors [8] of genes; and to create tissue specificity scores for each gene. GeneNote currently provides expression *tissue vectors* for each GeneCards gene for the following tissues: bone marrow, brain, heart, kidney, liver, lung, pancreas, prostate, skeletal muscle, spinal cord, spleen, and thymus, all color-coded into the immune, nervous, muscle and secretory system groups. As a quality measure, variation plots are also presented, which visualize Pearson's correlations between individual probe-set vectors and the average tissue vector for each gene, as well as relative scalar lengths of individual vectors. In addition to experimental results, *in silico* expression was calculated by mining the Unigene database for information about the number of unique clones per gene per tissue, and the data presented on a root scale designed to be visually parallel with the experimental tissue vectors. Expanding beyond the GeneCards gene-centric view, the GeneNote web site provides additional detailed information on a probe-set basis including links to GeneAnnot, and offers a variety of specialized searching and clustering options including tissue specificity and depiction of different splice variants.

### 4. GeneAnnot

GeneAnnot (<http://genecards.weizmann.ac.il/geneannot>) provides annotation for each GeneCards gene (in the *Expression in Human Tissues* section, above the graphics provided by GeneNote) with relevant probe-set associations, including their identifiers, arrays, sensitivity and specificity scores, and genes-to-probe-sets counts. Using *blat* [9], all probe sequences were compared to publicly available human mRNA sequences: Whenever

possible, the mRNAs are matched with GeneCards genes using data and algorithms from GeneLoc. Each probe set to gene pairing is scored to indicate the sensitivity and specificity of the relation. When the gene association is not known, probe sets are annotated in the GeneAnnot web site with GenBank accession numbers and corresponding Unigene cluster identifiers. Future work will support the HG-U133 array set, other organisms starting with the mouse, and more direct interfaces for interacting with experimental results. And, together with GeneLoc, GeneAnnot will improve the gene matching algorithms, and help elucidate *de novo* GeneCards genes.

### 4. Acknowledgements

This work was supported by the Crown Human Genome Center, the Judith and Abraham Goldwasser foundation, and the Yeda Fund.

### 5. References

- [1] M. Safran, I. Solomon., O. Shmueli, M. Lapidot, S. Shen-Orr, A. Adato, U. Ben-Dor, N. Esterman, N. Rosen, I. Peter, T. Olender, V. Chalifa-Caspi, and D. Lancet, "GeneCards™ 2002: towards a complete, object-oriented, human gene compendium", *Bioinformatics*, Oxford University Press, 2002, pp. 1542-1543.
- [2] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. "GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support", *Bioinformatics*, Oxford University Press, 1997, pp. 656-664.
- [3] N. Rosen, V. Chalifa-Caspi, O. Shmueli, A. Adato, M. Lapidot, J. Stampnitzky, M. Safran, and D. Lancet (2003) "GeneLoc: Exon-based integration of human genome maps." *Bioinformatics*, **19**,S1:pp. 222-224 (*in press*).
- [4] O. Shmueli, S. Horn-Saban., V. Chalifa-Caspi, M. Shmoish, R. Ophir, H. Benjamin-Rodrig, M. Safran, E. Domany, and D. Lancet (2003) "GeneNote: whole genome expression profiles in normal human tissues." *Proc. French Acad. Sci.* (*in press*).
- [5] V. Chalifa-Caspi, I. Yanai, R. Ophir, N. Rosen, O. Shmueli, M. Shmoish, H. Benjamin-Rodrig, T. Iny Stein, M. Safran and D. Lancet "GeneAnnot: Elucidating the many-to-many relationship between genes and oligonucleotide array probes" (*in preparation*).
- [6] D. L. Wheeler, D. M. Church et al (2002) "Database Resources of the NCBI: 2002 update.", *Nucleic Acids Research*, **30**(1) :pp. 6-13.
- [7] T. Hubbard, D. Barker, et al. (2002). "The Ensembl genome database project." *Nucleic Acids Research* **30**(1) :pp. 38-41.
- [8] S.A. Dudani, (1976). "The distance-weighted k-nearest-neighbor rule." *IEEE Trans. Syst. Man Cyber.*, **6**:325--327.
- [9] W. J. Kent (2002) "BLAT - The BLAST-Like Alignment Tool", *Genome Research* pp. :656-66