

**Vered Chalifa-Caspi, PhD** was a research associate at the Bioinformatics and Biological Computing Unit of the Department of Biological Services at the Weizmann Institute of Science, Rehovot, Israel. Being one of the founders of the GeneCards team, Vered started GeneLoc (formerly UDB), developed GeneAnnot and took part in GeneNote. She currently heads the Bioinformatics Support Unit at Ben-Gurion University, Beer-Sheva, Israel.

**Marilyn Safran, MSc** is head of the GeneCards development team. She is a research engineer at the Bioinformatics and Biological Computing Unit of the Department of Biological Services at the Weizmann Institute of Science, Rehovot, Israel.

**Doron Lancet, PhD** is Ralph & Lois Silver Professor of Human Genomics and Head of the Crown Human Genome Center, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel. He is the scientific supervisor for the team that develops the GeneCards suite of databases.

**Orit Shmueli, PhD, Naomi Rosen, MSc, Michael Shmoish, PhD and Itai Yanai, PhD** are from the Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.

**Keywords:** *DNA arrays, gene annotation, human genome, databases*

Vered Chalifa-Caspi, PhD,  
Bioinformatics Support Unit,  
Life Sciences Dept.,  
Ben-Gurion University of the  
Negev,  
Beer-Sheva 84105, Israel

Tel: +972 08 6477 917  
Fax: +972 08 6472 890  
E-mail: veredcc@bgumail.bgu.ac.il

# GeneAnnot: Interfacing GeneCards with high-throughput gene expression compendia

Vered Chalifa-Caspi, Orit Shmueli, Hila Benjamin-Rodrig, Naomi Rosen, Michael Shmoish, Itai Yanai, Ron Ophir, Pavel Kats, Marilyn Safran and Doron Lancet

Received (in revised form): 22nd September 2003

## Abstract

The interpretation of microarray expression results often includes extensive efforts to identify and annotate the gene representatives immobilised on the arrays. In this paper we describe the usage of our automatic GeneAnnot system, which links between Affymetrix arrays and the rich human gene annotations available in GeneCards. We explain GeneCards search options and results display; elaborate on the presentation of expression information in GeneCards, including both our whole-genome GeneNote project and external expression resources; describe the various parameters and displays used by GeneAnnot to assess the annotation quality and probeset specificity; and show how to search GeneAnnot and GeneNote websites directly.

## INTRODUCTION

Microarray experiments produce a large volume of experimental results, encompassing tens of thousands of genes. Beyond the demanding task of statistical analysis, including noise reduction, data normalisation and pattern extraction, lies the challenge of discovering new biological knowledge by linking expression data with accurate gene identities and annotations. For known genes, this task may be relatively straightforward, using currently available mutual links between sequence identifiers (eg GenBank accession numbers) and annotation resources such as UniGene and LocusLink. However, some of the current array sets also strive to encompass the 'terra incognita' of newly discovered or tentatively identified genes, such as those defined by expressed sequence tags (ESTs), whole-genome full-length mRNAs and putative genes predicted from genomic sequences. The definition of these genes in the public databases is much sparser, and is less consistent among different databases and between successive

updates of each database. Consequently, the annotation of these genes is much more demanding. An additional annotation problem is presented by oligonucleotide arrays (eg those produced by Affymetrix), where each gene is represented by a set of short oligonucleotides, derived from transcript sequences available at the time the array is manufactured. The uniqueness of the oligonucleotides is not always guaranteed and they do not always represent the same gene.

The GeneCards database and the GeneAnnot system have recently taken up the challenge of automating the biological annotation of expression arrays, using two parallel strategies:

- Extending GeneCards to include nearly all human genes, while keeping track of persistent identifiers for both the known and newly discovered genes and maintaining links to UniGene, LocusLink, Ensembl and other gene-centred databases.

**Hila Benjamin-Rodrig, BSc** is from the Department of Physics of Complex Systems and the Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel.

**Ron Ophir, PhD** is from the Bioinformatics and Biological Computing Unit of the Department of Biological Services, the Weizmann Institute of Science, Rehovot, Israel.

**Pavel Kats, BSc** worked as a volunteer software engineer with the GeneNote and GeneAnnot projects.

### The GeneCards database integrates information for nearly all human genes

- Constantly updating the links between oligonucleotide arrays and GeneCards genes, by direct sequence comparison of array oligonucleotides and gene transcripts.

In this paper we explain the principles behind these systems, and their use in analysing experimental results.

## GeneCards

GeneCards<sup>1-4</sup> is an integrated digital compendium of human genes, generated about quarterly by dedicated software, which automatically retrieves gene-related information from over 40 external databases. GeneCards organises the information as an online 'card' for each gene, which includes the categories indicated in Table 1. The cards are designed to enable fast screening of a large number of genes, a feature that is particularly useful in large-scale microarray analyses.

In an attempt to include all human genes in GeneCards, we have recently developed the GeneLoc algorithm and

database.<sup>5,6</sup> GeneLoc merges the major non-redundant human gene indices, created independently by NCBI (the RefSeq/LocusLink project<sup>7</sup>) and Ensembl,<sup>8</sup> by means of shared gene symbols and IDs, as well as overlapping exons at the genomic level. Each GeneCards gene receives a meaningful position-based identifier, which is kept consistent when upgrading to new versions.

## High-density oligonucleotide arrays and their annotations

Affymetrix array set U95A-E has been designed to contain a maximal representation of a large majority of human genes. Each probeset, composed of 16 25-long oligonucleotide probes, is derived from a several hundred bases 'target sequence', usually in the 3' untranslated region of the mRNA. The probesets were designed according to cDNAs, predicted genes and expressed sequence tag (EST) sequences that were available at the time the chip set was produced. Whereas chip U95A includes

**Table 1:** GeneCards categories and resources. Short descriptions and URLs of the resources are available at the GeneCards site:

<http://bioinfo.weizmann.ac.il/cards/background.html#where>

Category	Resources
Gene symbol	HUGO, LocusLink, Ensembl
Gene identifier	GeneLoc
Aliases/descriptions	GDB, HUGO, LocusLink, SWISS-PROT, GeneLoc
Chromosomal location	GeneLoc, HUGO, LocusLink, UCSC, Ensembl
Proteins	SWISS-PROT, MIPS
Protein domains/families/ontologies	InterPro, Gene Ontology Consortium, BLOCKS
Sequences	UniGene, GenBank, LocusLink, MIPS, DOTS
Expression in human tissues	GeneNote, Affymetrix, GeneAnnot, UniGene, SOURCE, SWISS-PROT
Similar genes in other organisms	MGD, HomoloGene, Stony Brook <i>C. elegans</i> - <i>H. sapiens</i> alignment database, euGenes, FlyBase, WormBase
SNPs/variances	dbSNP, SWISS-PROT
Disorders and mutations	OMIM, SWISS-PROT, GENATLAS, GeneTests, HGMD, BCGD, TGDB
Medical news	Doctor's Guide
Research articles	PubMed
The gene in other genome-wide resources	GDB, LocusLink, euGenes, Ensembl, GeneLynx
The gene in general databases, limited scope	HUGE
The gene in specialised databases	ATLAS, GENATLAS, HORDE, IMGT, MTDB, LEIDEN, SWISS-PROT
Services	RZPD

**GeneAnnot explores the many-to-many relationship between Affymetrix probesets and GeneCards genes**

mostly known genes, chips U95B-E consist of probesets derived mostly from EST sequences, and are thus of lower quality and with a considerably lower degree of annotation.

To keep probeset annotation up to date, Affymetrix constantly redefines the links between probesets and their corresponding genes, by determining the UniGene cluster that contains the probeset's representative sequence, and retrieving the gene symbol and LocusLink ID from the UniGene record when available.<sup>9</sup> This typically results in the assignment of no more than one gene per probeset, without any indication for the annotation reliability and probeset specificity. Similar annotation procedure is performed by most of the other publicly available array annotation resources<sup>10-17</sup> (Table 2). In Ensembl, on the other hand, the individual probe sequences are directly compared with all Ensembl transcripts, consequently depicting those cases in which more than one gene matches the same probeset. The probeset-to-Ensembl gene matches are accessible both from the genome browser and from the EnsMart data-mining tool. However, no information is provided for the match quality and specificity (eg the fraction of probes in a probeset that match the respective gene, and to what extent these

probes bind other genes) and there is no access to the raw sequence analysis data (eg probe positions on the transcripts).

### GeneAnnot at a glance

The GeneAnnot system was developed to further explore and document the many-to-many relationship between probesets and genes. This is done by directly comparing the individual probe sequences with publicly available cDNAs and predicted genes from GenBank, RefSeq and Ensembl. A central aim is to provide a comprehensive link between a gene compendium such as GeneCards, and genome-wide expression studies, as exemplified by GeneNote (next section). The transcript sequences are identified as GeneCards genes using various tools, including the GeneLoc genome localisation system.<sup>5</sup> Furthermore, each probeset/gene pair is assigned sensitivity and specificity scores, respectively reflecting the number of probes that match the gene, and the matching of other genes to the same probeset.<sup>18</sup> GeneAnnot website displays all levels of the annotation process, from summary tables down to probe-level views.

### The GeneNote project

The GeneNote project, conducted in our laboratory, measures human gene

**Table 2:** Summary of the main features of other array annotation resources (not including resources reviewed previously<sup>10</sup>)

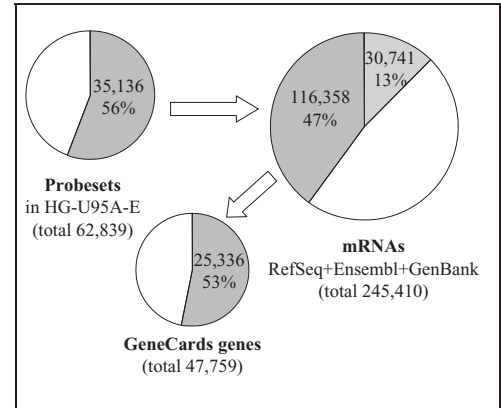
Resource	Annotation strategy		Links to external databases	Graphical display of representation of annotation categories within specified lists of genes/probesets			Promoter analysis
	Cross-referencing of database IDs	Direct probe/mRNA sequence comparison		GO	Pathways	Protein domains	
Ensembl <sup>8</sup>		+	+				
NetAffx <sup>9</sup>	+		+	+			
Chiplnfo <sup>11</sup>	+		+	+			
Onto-Express <sup>12,13</sup>	+		+	+			
GenePublisher <sup>14</sup>	+		+	+	+		+
DAVID <sup>15</sup>	+		+	+	+	+	
AnnBuilder <sup>16</sup>	+		+				
ARROGANT <sup>17</sup>	+		+				

**GeneAnnot was used to link between normal human tissue expression profiles obtained in the GeneNote project and GeneCards genes**

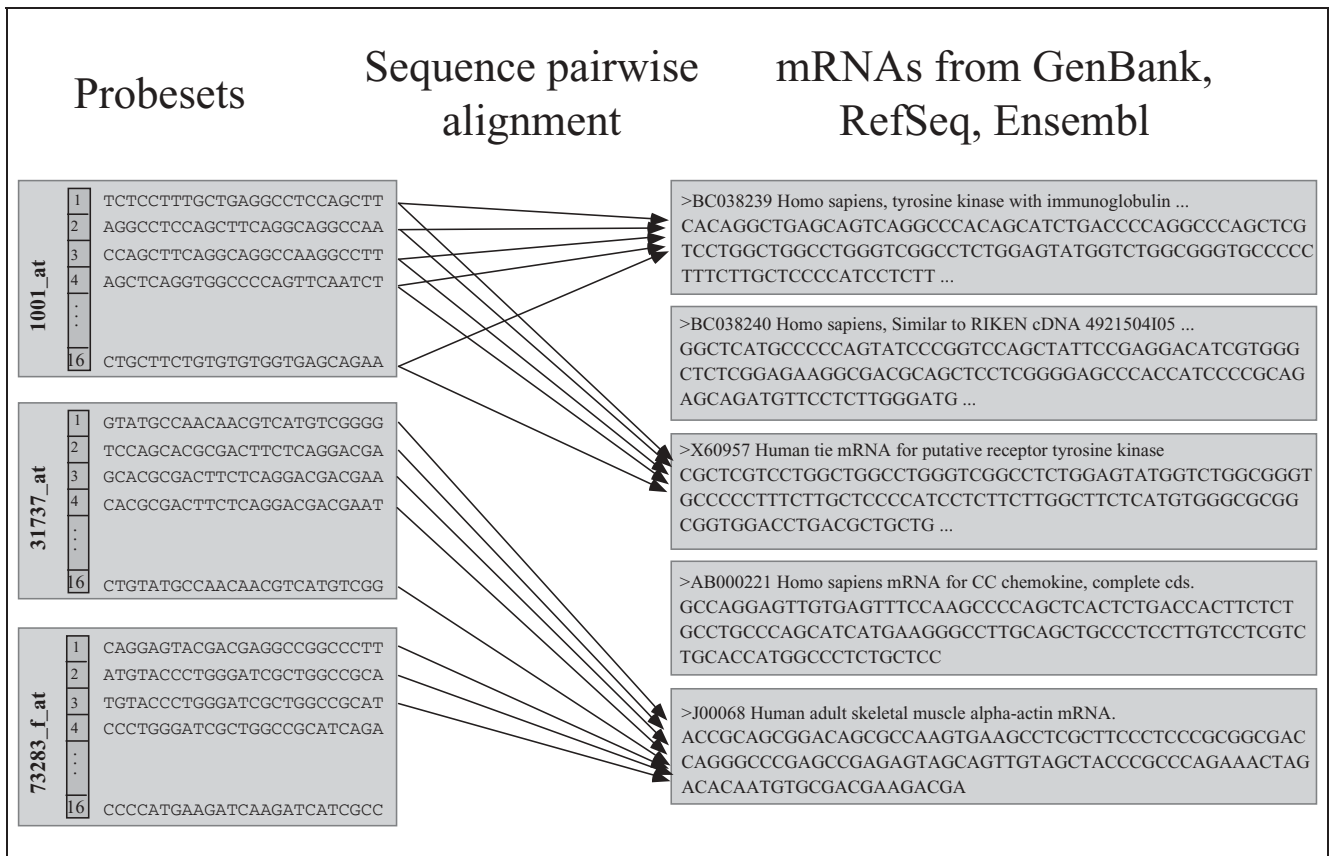
expression in 12 normal human tissues, using the entire Affymetrix U95A-E array set.<sup>19</sup> This is aimed at providing normal tissue expression patterns for a maximal coverage of the gene gamut in the human genome. GeneAnnot plays a crucial role in this project, as comprehensive probeset annotation and accurate links to GeneCards are indispensable for a study of this magnitude. This allows the assessment of tissue specificity indices for genes in the ‘terra incognita’ of the human genome, thus acquiring novel information about relatively uncharted genes.

**GENEANNOT CONCEPT AND STATISTICS**

The GeneAnnot procedure includes two steps. In the first, the association between probesets and genes is established at the sequence level, by aligning the probe sequences with sequences of cDNAs and



**Figure 2:** Overall GeneAnnot statistics. 35,136 probesets from Affymetrix array set HG-U95A-E matched, in a many-to-many relationship, 147,099 mRNAs from the major public sequence databases, out of which 116,358 mRNAs were mapped to 25,336 GeneCards genes. Data are from GenBank flat file release 134.0 (15th February, 2003) and GeneCards version 2.27 (5th May, 2003)



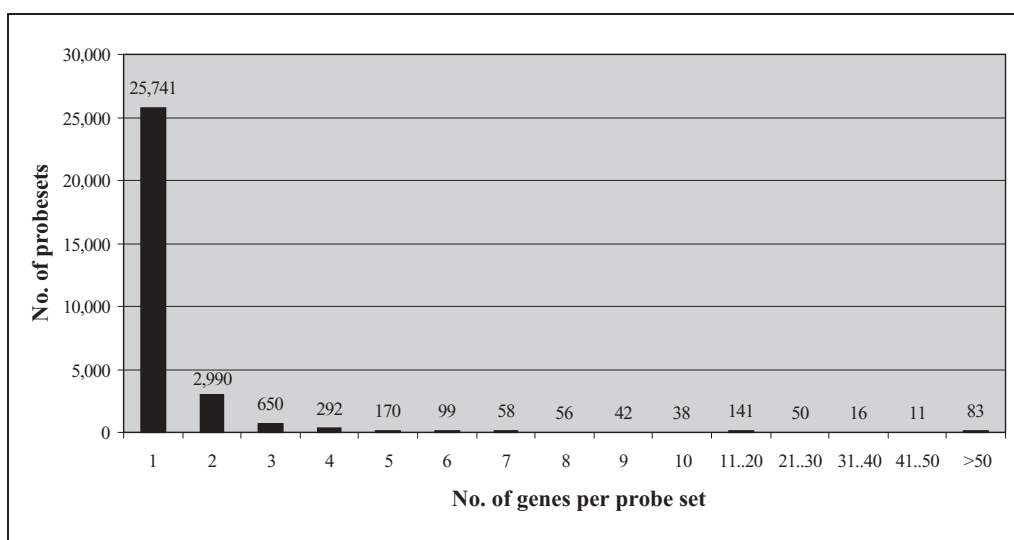
**Figure 1:** Direct sequence comparison of Affymetrix probes and public transcript sequences reveals many-to-many relationship. Sequence alignment was performed using BLAT,<sup>20</sup> allowing up to one mismatch

**Affymetrix probe sequences were directly aligned with GenBank, RefSeq and Ensembl mRNAs, which were further mapped to GeneCards genes**

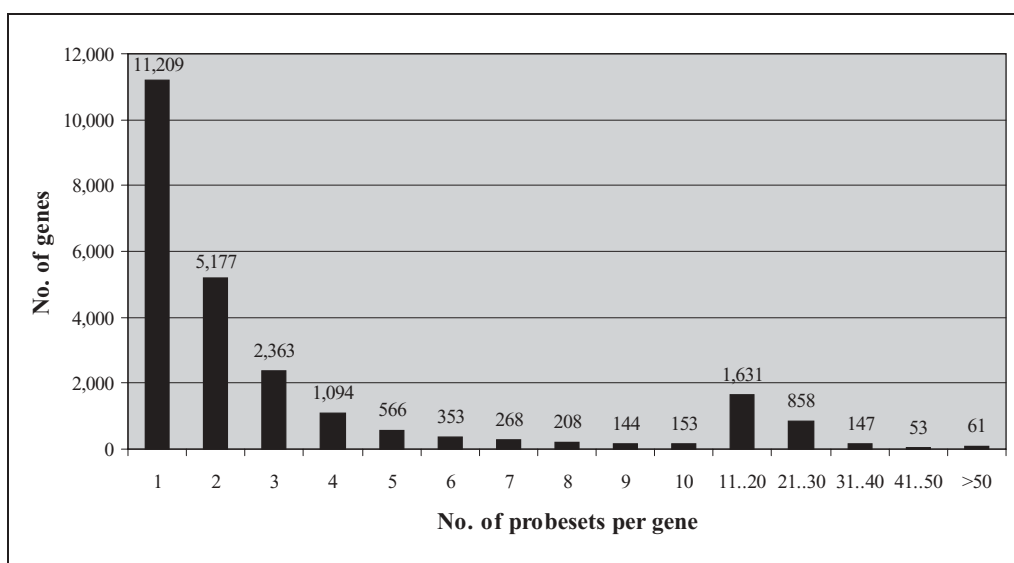
predicted genes from GenBank, RefSeq and Ensembl. As demonstrated in Figure 1, a given probeset could match one or more mRNA sequences, each with a certain subset of the probes. On the other hand, there are cases where the same mRNA sequence matches more than one probeset. In the second step, the mRNAs are mapped to genes, as comprehensively represented within GeneCards, through the association of IDs from GenBank/Refseq to LocusLink and from Ensembl. If mapping is successful, the annotation is marked with 'quality' 1. GenBank

mRNAs that cannot be mapped to LocusLink are mapped directly to GeneCards genes by comparing the genomic positions of those mRNAs and the positions of GeneCards gene exons. This procedure may be somewhat less reliable than the ID-based mapping, as errors may be introduced in the genomic alignment of both the mRNA sequences (retrieved from UCSC<sup>21</sup>) and the GeneCards genes (retrieved from Ensembl<sup>8</sup> and NCBI's MapView<sup>7</sup>). Probesets annotated this way were thus marked with annotation quality 2. Figure

**Figure 3:** Distribution of the number of GeneCards genes per probeset. A probeset–gene match is considered when at least one probe from the probeset matches the gene



**Figure 4:** Distribution of the number of probesets per GeneCards gene. A probeset–gene match is considered when at least one probe from the probeset matches the gene



2 shows the overall statistics of this process, whereby a ‘match’ between a probeset and a gene is considered if at least one probe from the probeset matched the gene.

**Each probeset/gene pair received sensitivity and specificity scores**

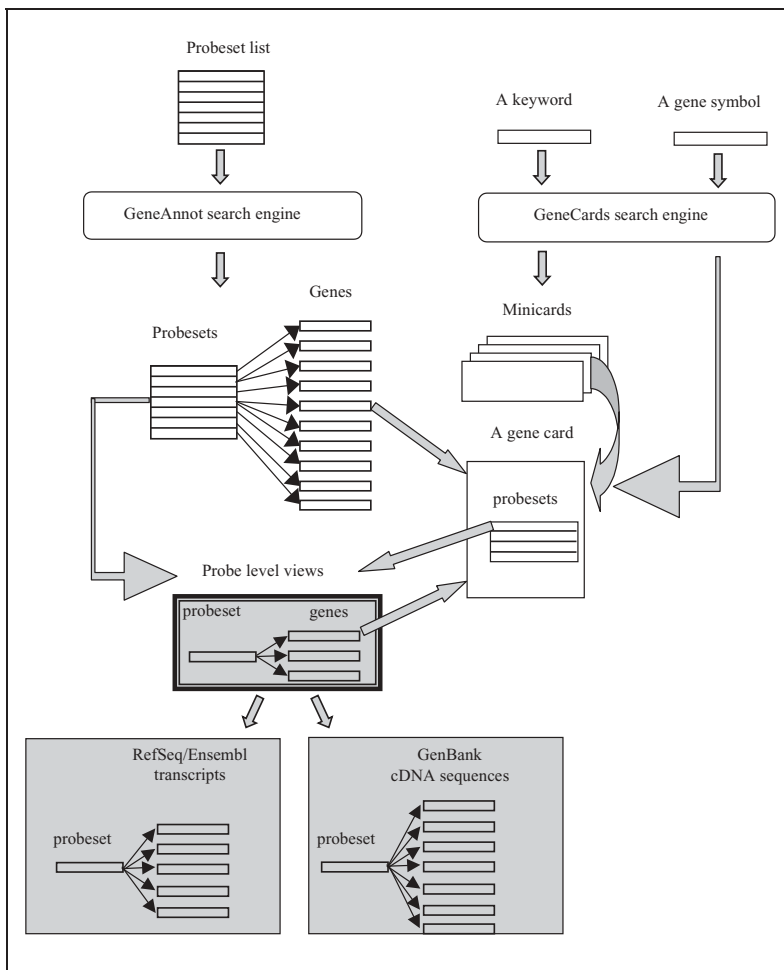
Each probeset/gene pair receives a ‘match score’, consisting of three numbers:  $S_n$  – the sensitivity score, indicating the number of matching

probes, divided by the total number of probes in the probeset (usually 16);  $S_p$  – the specificity score, which sums up the number of matching probes while giving lower weight to probes that match additional genes, and eventually dividing by the total number of probes that matched any gene;  $N_g$  – the total number of genes that match a given probeset. This last simple measure helps distinguish between, for example, a probeset that matches 2 genes with most of its probes, and a probeset that matches one gene with most of its probes and many additional genes with one or two probes.

As seen in Figure 3, most of the probesets link to only one gene (linking considered when there is at least one matching probe). Still the  $S_p$  and  $N_g$  scores help identify the other, non-specific, probesets. Practically, when we study a gene that links to a non-specific probeset, we should bear in mind that the observed expression value may not necessarily reflect the expression of this gene alone. However, knowing the identity of the other genes that match this probeset may help recognise the relevant gene family, and thus relate the observed expression to this family.

Conversely, a gene may match more than one probeset. As shown in Figure 4, this is a fairly common situation. As explained below, GeneNote and GeneAnnot can be used to compare the expression patterns of individual probesets, in order to try to relate them to gene splice variants, and to filter out less specific probesets.

For a summarised display of probesets associated with a given gene (in GeneCards), and the genes associated with a given probeset (in GeneAnnot), a graded cutoff was applied based on the number of matching probes, as follows: a probeset is considered associated with a GeneCards gene if it matches the gene with more than 12 probes. This may link more than one gene to a probeset. If a probeset remains without any matching gene after this cutoff is applied, the



**Figure 5:** Accessing GeneAnnot results through GeneCards and GeneAnnot web interfaces. Left: the GeneAnnot search engine accepts a probeset list and retrieves a table in which each probeset points to one or more genes. Each probeset is linked to three probe level views (grey), showing the relationships between this probeset and all linked genes (which in turn point back to GeneCards), RefSeq and Ensembl transcripts, and GenBank cDNAs. Right: the GeneCards search engine accepts either a keyword, which retrieves all relevant minicards, each pointing to a GeneCard, or a gene symbol which retrieves the GeneCard directly. The GeneCard contains, among other things, a list of all probesets linked to this gene, each pointing to the corresponding probe level views in GeneAnnot (grey)

algorithm searches for genes that match the probeset with more than four probes, and selects those genes that have the maximal number of matching probes. For example, if a probeset is linked to two genes with ten probes, and to three other genes with fewer than ten probes, we will only take the first two genes. This cutoff was first applied to probesets with annotation quality 1. For probesets that remained with no annotation, the same cutoff procedure was then applied for annotation quality 2.

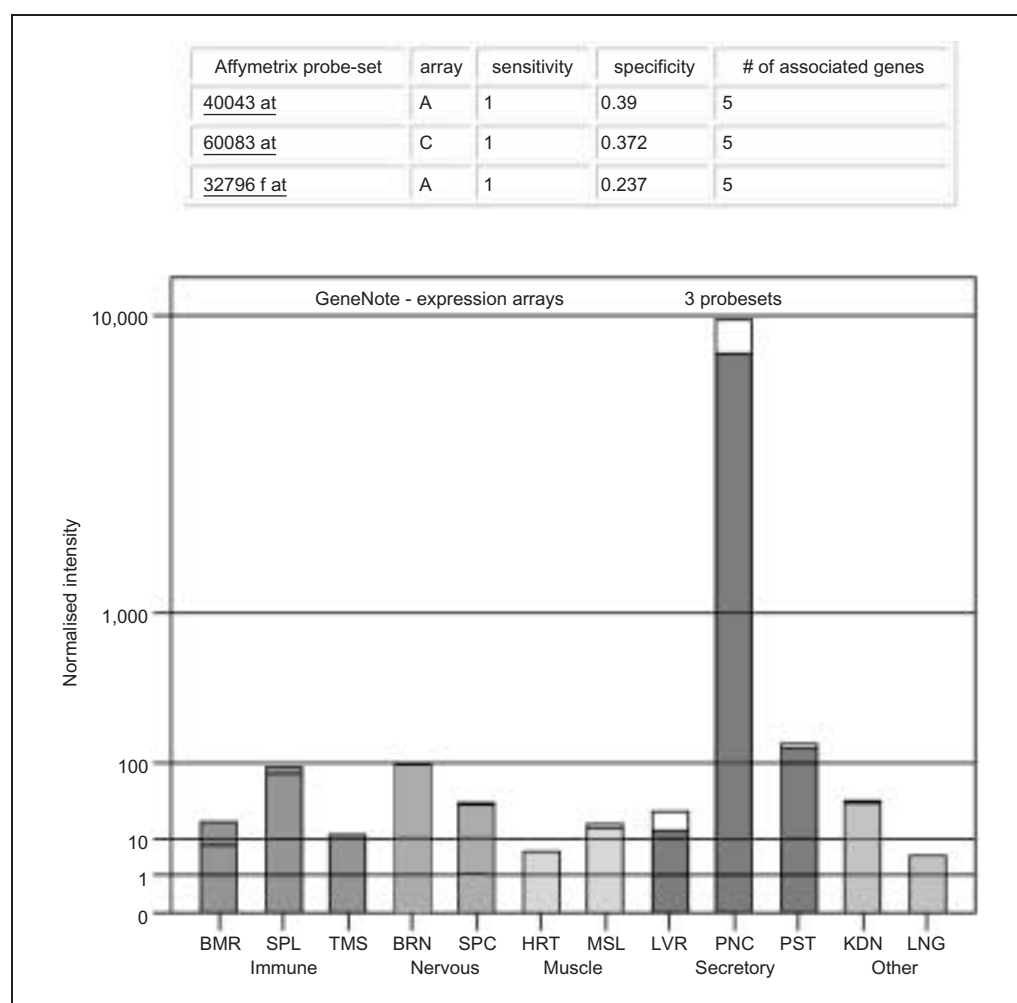
To summarise, GeneAnnot provides the user with detailed information on the probeset-to-gene relationships and probeset-to-mRNA relationships at the probe level. In addition, it provides summarised scores for the sensitivity and specificity of each probeset-to-gene pair,

and applies cutoffs for assigning probesets to genes based on sensitivity scores. As shown in Figure 5, the mutual linking between GeneCards and GeneAnnot websites enables searching GeneAnnot results from two perspectives:

- From the array point of view: one may enter a list of probesets (and in the future, upload Affymetrix result files) to GeneAnnot, get a summarised annotation table, and follow the links to GeneCards.
- From the gene point of view: one may search GeneCards by a keyword (eg gene symbol, gene alternative name, protein domain, disease/phenotype name, subcellular organelle, accession number from external database,

**GeneAnnot data may be accessed by probeset names and by gene-related keywords, through GeneAnnot and GeneCards search engines, respectively**

**Figure 6:** GeneAnnot and GeneNote results display in GeneCards. Data are from the PRSS3 (protease, serine, 3 (mesotrypsin)) GeneCard. The top table shows the probesets linked to the gene, with their Sn (sensitivity), Sp (specificity) and Ng (no. of associated genes) scores. The bar graph at the bottom shows the averaged expression profiles of these probesets. Tissue abbreviations: BMR, bone marrow; SPL, spleen; TMS, thymus; BRN, brain; SPC, spinal cord; HRT, heart; MSL, skeletal muscle; LVR, liver; PNC, pancreas; PST, prostate; KDN, kidney; LNG, lung



chromosomal position, article title, etc), access the cards of the relevant genes and follow the links to GeneAnnot.

Moreover, from the gene entries both in GeneCards and in GeneAnnot it is possible to go to GeneNote for detailed expression data, and to GeneLoc for genomic positional information.

**GeneCards displays an aggregate expression profile graph per gene, as well as details on the associated probesets**

### GENEANNOT GENERATES NOVEL GENECARDS EXPRESSION DISPLAY

Expression data are displayed in the GeneCards category 'Expression in Human Tissues', part of which is shown in Figure 6. The bar graph shows GeneNote in-house experimental expression data, where the white rectangles above the bars show the range for the duplicate measurements of each tissue. Tissues are colour-coded according to their types. Data were normalised and root scale computed as described.<sup>3,19</sup> The table above the graph shows the probesets associated with the gene, after applying the graded cutoff in the probeset annotation procedure described above.

Each probeset is annotated with its array affiliation and its Sn, Sp and Ng scores. Clicking on the probeset name in the table leads to detailed information about the genes that match this probeset, at the probe level (Figure 7).

A further cutoff is applied for determining which probesets to average in order to produce the aggregate expression profile bar graph. Only probesets that have a normalised expression value higher than 10 in at least one of the tissues are included; however, if all probesets associated with the gene have expression values lower than 10 in all tissues, they will be averaged and included in the graph.

In a real-world search aimed at retrieving gene expression information for all genes related, for example, to a given disease or a protein domain, one can search GeneCards with the relevant keyword, and for each resulting gene, retrieve the following information:

- Gene expression in normal human tissues, according to linked GeneNote database<sup>19</sup> (and see below).

**Table 3:** GeneAnnot search results. Upon entering a probeset list to GeneAnnot website, a table is retrieved with summarised information for each probeset and its associated gene(s). Only genes that passed the graded sensitivity cutoff (see text) are displayed. Note that probeset 32796\_f\_at matched a total of five genes, as indicated by the Ng score, of which three passed the cutoff

Probe set	Array	Annotation quality (1 = best)	Resource	Accession in resource	Gene symbol	Gene description	Chromosomal location	Sn score	Sp score	Nr genes
<u>34894_r_at</u>	HG-U95A	1	<a href="#">GeneCards</a>	<a href="#">GC16M002923</a>	<b>PRSS22</b>	protease, serine, 22	<a href="#">Chr 16 (-) 2936370 – 2941782 bp</a>	1	1	1
<u>63329_at</u>	HG-U95C	1	<a href="#">GeneCards</a>	<a href="#">GC21M0039410</a>	<b>TMPRSS2</b>	transmembrane protease, serine, 2	<a href="#">Chr 21 (-) 39493446 – 39537043 bp</a>	0.625	1	1
<u>32796_f_at</u>	HG-U95A	1	<a href="#">GeneCards</a>	<a href="#">GC09P033919</a>	<b>PRSS3</b>	protease, serine, 3 (mesotrypsin)	<a href="#">Chr 9 (+) 33919958 – 33968672 bp</a>	1	0.237	5
				<a href="#">GC07P140731</a>	<b>PRSS2</b>	protease, serine, 2 (trypsin 2)	<a href="#">Chr 7 (+) 140730711 – 140776170 bp</a>	1	0.237	5
				<a href="#">GC07P140730</a>	<b>PRSSI</b>	protease, serine, 1 (trypsin 1)	<a href="#">Chr 7 (+) 140730711 – 140776170 bp</a>	0.938	0.222	5

**Information for probe set 32796\_f\_at (HG-U95A)**

Probe to GeneCards ID Match

	Gene symbol	Probes																Scores	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Sn	Sp
1	<a href="#">PRSS2</a>	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.000	0.238
2	<a href="#">PRSS3</a>	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.000	0.238
3	<a href="#">PRSS1</a>	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	0.938	0.222
4	<a href="#">TRBVOR0@</a>				+	+	+	+	+		+	+	+	+	+	+	0.750	0.159	
5	<a href="#">TRY6</a>				+	+	+	+		+	+	+	+	+	+	+	0.688	0.144	

Gene Details

Gene symbol	GeneCards ID	Title	Chr	Strand	Start	End
<a href="#">PRSS2</a>	<a href="#">GC07P140731</a>	protease, serine, 2 (trypsin 2)	7	+	<a href="#">140730711</a>	<a href="#">140776170</a>
<a href="#">PRSS3</a>	<a href="#">GC09P033919</a>	protease, serine, 3 (mesotrypsin)	9	+	<a href="#">33919958</a>	<a href="#">33968672</a>
<a href="#">PRSS1</a>	<a href="#">GC07P140730</a>	protease, serine, 1 (trypsin 1)	7	+	<a href="#">140730711</a>	<a href="#">140776170</a>
<a href="#">TRBVOR9@</a>	<a href="#">GC09U990073</a>	T cell receptor beta variable orphans on chromosome 9	9			
<a href="#">TRY6</a>	<a href="#">GC07P140752</a>	trypsinogen C	7	+	<a href="#">140752139</a>	<a href="#">140755798</a>

**Figure 7:** GeneAnnot display of the GeneCards genes associated with a given probeset, at the probe level. This example shows all five genes linked with probeset 32796\_f\_at. For each gene, a plus sign marks those probes that matched at least one of the mRNA sequences associated with the gene

**GeneCards provides links to individual probeset expression and annotation data in GeneNote and GeneAnnot, respectively**

- Information on variability in the expression of the probesets related to this gene. If the correlation between the probesets is low, the user may wish to further view the expression of each probeset individually, by exploring GeneNote. This is done by clicking on the expression profile bar graph (see more below).
- Relationship between the probesets related to the gene and known splice variants, through GeneAnnot.
- Information on the quality of the probesets associated with the gene. Probesets with value 1 for all of the scores (sensitivity, specificity and number of associated genes) are best. If

such probesets are associated with the gene, the user may wish to disregard additional probesets associated with that gene if they have lower specificity scores, unless they represent a distinct, interesting splice variant.

- The names of the best probesets associated with the gene may be used to search other human gene expression data sets that utilised the same Affymetrix arrays.

In addition, GeneCards provides E-northern data, computed for the same set of 12 human tissues included in GeneNote, obtained by mining the UniGene database for information about the number of unique clones per gene per

**GeneAnnot website provides details on the mRNAs and genes that matched each probe in the probeset**

tissue. Also, links are provided to additional gene expression results, in normal and the respective abnormal human tissues, via the link to the SOURCE database at Stanford.<sup>22,23</sup>

### THE GENEANNOT WEBSITE

A GeneAnnot website search yields a table containing the relevant probesets, the genes associated with each probeset, gene descriptive and genomic positional information, the annotation quality rank, and the Sn, Sp and Ng scores (Table 3). Note that the table only contains probeset/gene pairs that passed the graded sensitivity cutoff described above. By pressing on the probeset name, one receives a table that relates the individual probes in the probeset to all matched genes (without cutoffs) (Figure 7). Matched probe/gene pairs are marked with a '+' sign. Descriptive gene

information, with links to GeneCards, Ensembl, LocusLink and GeneLoc are provided in a separate table.

Below the gene details table, there are links to view probe matches at the mRNA level (not shown). These are divided into two pages, 'probe to RefSeq/Ensembl matches' and 'Probe to GenBank mRNAs', whose format is similar to that of the probeset to gene relationship page (Figure 8). Note that each of the RefSeq and Ensembl sequence collections aims at producing a non-redundant list of human transcripts. In case there are several transcripts per gene, these are supposed to be splice variants. On the other hand, when multiple transcripts are available per gene in GenBank, this may be also due to multiple, independent submissions of cDNAs of the gene to GenBank. While GenBank contains submitted cDNA sequences, RefSeq and Ensembl also

**Figure 8:** GeneAnnot display of the GenBank mRNAs associated with a given probeset, at the probe level. On the website there is an additional table with mRNA descriptive information, including the respective GeneCards gene for each mRNA (not shown). Note that only Sn scores are indicated. Sp scores are meaningless here, owing to the redundancy of mRNA representation in GenBank. A similar web page shows the RefSeq/Ensembl mRNAs associated with the probeset (not shown)

	GeneBank ID	Probes																Sn score
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	<a href="#">M27602</a>	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.000
2	<a href="#">BC030260</a>	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	1.000
3	<a href="#">M22612</a>	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		0.938
4	<a href="#">X71345</a>	+	+	+	+	+	+	+			+	+	+	+	+	+	+	0.875
5	<a href="#">X72781</a>	+	+	+	+	+	+	+	+			+	+	+	+	+		0.875
6	<a href="#">BC030238</a>		+	+	+	+	+	+			+	+	+	+	+	+	+	0.812
7	<a href="#">AF009664</a>				+	+	+	+	+	+	+	+	+	+	+	+	+	0.812
8	<a href="#">X15505</a>	+	+	+	+	+	+	+			+	+	+	+	+	+	+	0.812
9	<a href="#">U66061</a>				+	+	+	+	+	+	+	+	+	+	+	+	+	0.812
10	<a href="#">D45417</a>		+	+	+	+	+	+	+			+	+	+	+	+		0.812
11	<a href="#">AF029308</a>				+	+	+	+			+	+	+	+	+	+	+	0.688
12	<a href="#">U70137</a>				+	+	+	+	+	+								0.438
13	<a href="#">AF315310</a>				+	+	+	+	+	+								0.438
14	<a href="#">AF315309</a>				+	+	+	+	+	+								0.438

contain sequences of predicted genes, based on the genomic sequence.

## THE GENENOTE DATABASE

GeneNote may be accessed either through GeneCards, or through its own website, using field-specific searches. Current search options include gene symbol, GeneCards ID, Ensembl ID, LocusLink ID, probeset ID and GenBank accession. Searches by expression level and by tissue-specific thresholds are planned for future versions. When the GeneNote site is searched directly, the user may choose to see the expression values in one of the following formats: MAS 5.0 raw data, MAS 5.0 normalised data and, in an upcoming version, data analysed by robust multi-array analysis (RMA<sup>24</sup>). GeneNote's search retrieves aggregated expression information for the respective gene. The table at the bottom includes summarised GeneAnnot information, and by pressing on probeset names, the expression profiles of the individual probesets, both in numbers and as a bar graph, are displayed.

### Acknowledgments

This work was supported by grants from the Abraham and Judy Goldwasser Fund and from the Crown Human Genome Center.

### References

1. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1998), 'GeneCards: A novel functional genomics compendium with automated data mining and query reformulation support', *Bioinformatics*, Vol. 14(8), pp. 656–664.
2. Safran, M., Solomon, I., Shmueli, O. *et al.* (2002), 'GeneCards 2002: Towards a complete, object-oriented, human gene compendium', *Bioinformatics*, Vol. 18(11), pp. 1542–1543.
3. Safran, M., Chalifa-Caspi, V., Shmueli, O. *et al.* (2003), 'Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE', *Nucleic Acids Res.*, Vol. 31(1), pp. 142–146.
4. URL: <http://bioinfo.weizmann.ac.il/genecards>
5. Rosen, N., Chalifa-Caspi, V., Shmueli, O. *et al.* (2003), 'GeneLoc: Exon-based integration of human genome maps', *Bioinformatics*, Vol. 19, Suppl. 1, pp. I222–I224.
6. URL: <http://genecards.weizmann.ac.il/geneloc/>
7. Wheeler, D. L., Church, D. M., Federhen, S. *et al.* (2003), 'Database resources of the National Center for Biotechnology', *Nucleic Acids Res.*, Vol. 31(1), pp. 28–33.
8. Clamp, M., Andrews, D., Barker, D. *et al.* (2003), 'Ensembl 2002: Accommodating comparative genomics', *Nucleic Acids Res.*, Vol. 31(1), pp. 38–42.
9. Liu, G., Loraine, A. E., Shigeta, R. *et al.* (2003), 'NetAffx: Affymetrix probesets and annotations', *Nucleic Acids Res.*, Vol. 31(1), pp. 82–86.
10. Guffanti, A., Reid, J. F., Alcalay, M. and Simon, G. (2002), 'The meaning of it all: Web-based resources for large-scale functional annotation and visualization of DNA microarray data', *Trends Genetics*, Vol. 18(11), pp. 589–592.
11. Zhong, S., Li, C. and Wong, W. H. (2003), 'ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis', *Nucleic Acids Res.*, Vol. 31(13), pp. 3483–3486.
12. Draghici, S., Khatri, P., Martins, R. P. *et al.* (2003), 'Global functional profiling of gene expression', *Genomics*, Vol. 81(2), pp. 98–104.
13. Draghici, S., Khatri, P., Bhavsar, P. *et al.* (2003), 'Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate', *Nucleic Acids Res.*, Vol. 31(13), pp. 3775–3781.
14. Knudsen, S., Workman, C., Sicheritz-Ponten, T. and Friis, C. (2003), 'GenePublisher: Automated analysis of DNA microarray data', *Nucleic Acids Res.*, Vol. 31(13), pp. 3471–3476.
15. Dennis, G., Jr, Sherman, B. T., Hosack, D. A. *et al.* (2003), 'DAVID: Database for Annotation, Visualization, and Integrated Discovery', *Genome Biol.*, Vol. 4(5), p. P3.
16. Zhang, J., Carey, V. and Gentleman, R. (2003), 'An extensible application for assembling annotation for genomic data', *Bioinformatics*, Vol. 19(1), pp. 155–156.
17. Kulkarni, A. V., Williams, N. S., Lian, Y. *et al.* (2002), 'ARROGANT: An application to manipulate large gene collections', *Bioinformatics*, Vol. 18(11), pp. 1410–1417.
18. Chalifa-Caspi, V., Yanai, I., Ophir, R. *et al.*, 'GeneAnnot: Comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes'. Unpublished.
19. Shmueli, O., Horn-Saban, S., Chalifa-Caspi, V. *et al.* (2003), 'GeneNote: Whole genome expression profiles in normal human tissues', *Proc. French Acad. Sci.*, in press.

**GeneNote website displays expression profiles for the individual probesets, for both raw and normalised data**

20. Kent, W. J. (2002), 'BLAT – the BLAST-like alignment tool', *Genome Res.*, Vol. 12(4), pp. 656–664.
21. Karolchik, D., Baertsch, R., Diekhans, M. *et al.* (2003), 'The UCSC Genome Browser Database', *Nucleic Acids Res.*, Vol. 31(1), pp. 51–54.
22. Diehn, M., Sherlock, G., Binkley, G. *et al.* (2003), 'SOURCE: A unified genomic resource of functional annotations, ontologies, and gene expression data', *Nucleic Acids Res.*, Vol. 31(1), pp. 219–223.
23. URL: <http://source.stanford.org>
24. Irizarry, R. A., Bolstad, B. M., Collin, F. *et al.* (2003), 'Summaries of Affymetrix GeneChip probe level data', *Nucleic Acids Res.*, Vol. 31(4), p. e15.