BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 9 2004, pages 1457–1458
DOI: 10.1093/bioinformatics/bth081

GeneAnnot: comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes

Vered Chalifa-Caspi[1], Itai Yanai[2], Ron Ophir[1], Naomi Rosen[2], Michael Shmoish[2], Hila Benjamin-Rodrig[3], Maxim Shklar, Tsippi Iny Stein[2], Orit Shmueli[2], Marilyn Safran[1] and Doron Lancet[2,*]

[1]Department of Biological Services, [2]Department of Molecular Genetics and [3]Department of Physics of Complex Systems, the Weizmann Institute of Science, 76100 Rehovot, Israel

Received on July 12, 2003; revised on September 2, 2003; accepted on December 19, 2003
Advance Access publication February 12, 2004

ABSTRACT

Motivation: High density oligonucleotide arrays are usually annotated in a one-to-one fashion, with each probeset assigned to one gene. However, in reality, subsets of oligonucleotides in a probeset may match sequences within more than one gene, potentially leading to misinterpretations. Moreover, a gene is often represented by more than one probeset, and analyzing probe matches at the mRNA level can help one deduce whether these probesets are derived from the same or different splice variants.

Results: The GeneAnnot system comprehensively documents the many-to-many relationship between oligonucleotide array probesets and annotated genes in GeneCards™. It performs pairwise alignments between the probe sequences and gene transcripts, and assigns sensitivity and specificity scores to each probeset/gene pair.

Availability: http://genecards.weizmann.ac.il/geneannot/

Contact: geneannot@weizmann.ac.il

Supplementary information: Program description and statistics http://genecards.weizmann.ac.il/geneannot/DOC/index.html

INTRODUCTION

Affymetrix GeneChip® expression array sets are designed to represent nearly the entire gene complement of an organism. Each array contains sets of oligonucleotide probes ('probesets') derived from cDNAs, predicted genes and expressed sequence tags (ESTs; 'representative sequences') available at the time of manufacture. To keep the probeset annotation up-to-date, the company continuously 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 (Liu *et al*., 2003). This typically results in the assignment of no more than one gene per probeset, even in cases where the probe sequences are actually shared by several genes. Further, this procedure does not provide quality assessment for each probeset annotation. The GeneAnnot system explores the many-to-many relationship between probesets and genes, by directly comparing the individual probe sequences with publicly available cDNAs and predicted genes from GenBank, RefSeq and Ensembl. The transcript sequences are further identified as GeneCards genes (Safran *et al*., 2002) using the GeneLoc system (Rosen *et al*., 2003), which merges LocusLink and Ensembl gene indices on the basis of their genomic position. The linking to GeneCards enables attachment of annotation from GeneCards' nearly 40 mined resources, including expression results in normal human tissues (GeneNote: Shmueli *et al*., 2003).

ALGORITHM AND RESULTS

GeneAnnot is implemented for the human HG-U95 array set, comprising 62 839 probesets on five arrays (A–E). However, it is applicable to any oligonucleotide set. The algorithm was executed as a three-tier procedure:

1. *Probe-to-transcript mapping.* Probes were mapped to full length transcripts or ESTs as follows: all 25mer probe sequences from the array set (typically 16 per probeset) were downloaded from the Affymetrix web site and compared, using the BLAT program (Kent, 2002), to all transcript sequences from the following resources: (a) human non-genomic sequences from GenBank's 'primate' division, (b) NCBI RefSeq sequences and (c) Ensembl transcripts. Probe/transcript matches were accepted if the probe alignment was in the mRNA orientation, and had no more than one mismatch. For probesets with no matching transcripts, the EST accessions of their representative sequences were stored.

2. *Transcript-to-gene mapping.* Transcripts were mapped to GeneCards genes if possible, or otherwise to UniGene

*To whom correspondence should be addressed.

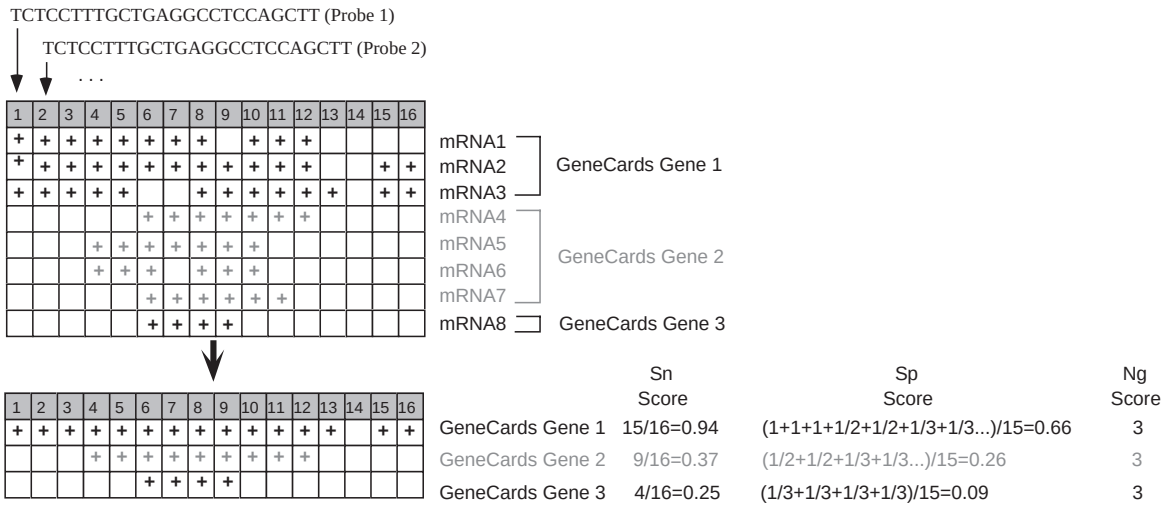

Fig. 1. Sequence-based matching of a probeset to its associated genes and sensitivity (Sn) and specificity (Sp) scores for each gene. The Ng score indicates the total number of genes matched to the probeset. Data were managed in a MySQL database using Perl.

clusters, as follows: (1) RefSeq, Ensembl and about half of the GenBank transcripts were mapped to their corresponding LocusLink/Ensembl genes, and these were further associated with GeneCards genes according to GeneLoc. (2) GenBank entries for which there was no information on the corresponding LocusLink gene were annotated as follows: their genomic coordinates were retrieved from UCSC, and GeneLoc was used to generate a link to a GeneCards entry, whenever at least one GeneLoc-recorded exon overlapped with the UCSC coordinates. (3) ESTs were mapped to their associated UniGene cluster.

3. *Summarized probeset-to-GeneCards annotation.* As shown in Figure 1, probeset annotation is recorded at both the transcript (top) and gene (bottom) levels. A probe is marked as associated with a GeneCards gene if it matches at least one of the transcripts related to that gene. Each probeset-to-gene pair may be 'connected' via 1–16 probes, a property denoted as the pairing 'sensitivity' (Sn score). The 'specificity' property, on the other hand, denotes how many other genes match this probeset and with how many probes (Ng and Sp scores). Such genes, usually paralogs, may contribute to the expression value observed for the probeset. GeneAnnot v0.2, 35 136 probesets were matched to 25 336 GeneCards genes. Among these, 85% matched one GeneCard, 10% matched two, 2% matched three, 1% matched four, and 3% matched five or more.

APPLICABILITY AND FUTURE DIRECTIONS

GeneAnnot constitutes an essential tool for interpreting expression array results in ways that relate explicitly to the rich annotation available for many genes. Interpretation may be fine tuned based on new knowledge about groups of genes linked to a given probeset, and by investigating the identity or discrepancy found in the patterns of expression seen for multiple probesets that represent the same gene. It is linked to GeneCards, GeneLoc, GeneNote and external databases. GeneAnnot results are also displayed in GeneCards, which enables their retrieval through GeneCards' free text search. Currently, efforts are underway to extend GeneCards to include more of the 'terra incognita' of the human genome, represented by probesets without links to GeneCards entries. GeneAnnot's flexible platform extendable to other species and to newer generations of arrays.

ACKNOWLEDGEMENTS

This work was supported by grants from the Abraham and Judith Goldwasser Fund, the Crown Human Genome Center and the Yeda fund. D.L. holds the Ralph and Lois Silver Chair in Human Genomics.

REFERENCES

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.

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.

Safran,M., Solomon,I., Shmueli,O., Lapidot,M., Shen-Orr,S., Adato,A., Ben-Dor,U., Esterman,N., Rosen,N., Peter,I. *et al.* (2002) GeneCards(TM) 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542–1543.

Shmueli,O., Horn-Saban,S., Chalifa-Caspi,V., Shmoish,M., Ophir,R., Benjamin-Rodrig,H., Safran,M., Domany,E. and Lancet,D. (2003) GeneNote: whole genome expression profiles in normal human tissues. *C. R. Biol.*, **326**, 1067–1072.