Software

GeneHopper: a web-based search engine to link gene-expression platforms through GenBank accession numbers

B Anders T Svensson, Arja J Kreeft, Gert-Jan B van Ommen, Johan T den Dunnen and Judith M Boer

Address: Center for Human and Clinical Genetics/Leiden Genome Technology Center, Leiden University Medical Center, Wassenaarseweg 72, 2333 AL Leiden, The Netherlands.

Correspondence: Judith M Boer. E-mail: j.m.boer@lumc.nl

Published: 25 April 2003

Genome **Biology** 2003, **4**:R35

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/5/R35

Received: 12 November 2002
Revised: 13 March 2003
Accepted: 19 March 2003

Abstract

Global gene-expression analysis is carried out using different technologies that are either array- or sequence-tag-based. To compare experiments that are performed on these different platforms, array probes and sequence tags need to be linked. An additional challenge is cross-referencing between species, to compare human profiles with those obtained in a mouse model, for example. We have developed the web-based search engine GeneHopper to link different expression resources based on UniGene clusters and HomoloGene orthologs databases of the National Center for Biotechnology Information (NCBI).

Rationale

Genome-wide analysis of gene expression provides insight into the transcriptional state of a cell or tissue sample, measuring RNA levels for thousands of genes in parallel. Among the most commonly applied technologies are photolithographically synthesized oligonucleotide chips [1], printed microarrays using cDNA probes or oligonucleotides [2-5], and serial analysis of gene expression (SAGE) [6]. Array-based technologies measure hybridization signal intensities in one or two channels for each array feature or probe, resulting in absolute or relative information about the expression levels of the corresponding transcripts in the samples. Array probes are produced by polymerase chain reaction (PCR) amplification of selected cDNAs or designed on the basis of cDNA or gene sequence information. SAGE results in quantitative information connected to 10-14 base-pair (bp) sequence tags derived from the 3′ ends of transcripts.

Expressed sequence tag (EST) sequencing projects have generated millions of cDNA sequences for human, mouse and other organisms that are identified by an accession number in databases such as GenBank. To group these individual sequence reads into sets representing one transcript, several efforts to cluster the ESTs were developed [7-9]. UniGene, developed at the National Center for Biotechnology Information (NCBI) [10], automatically clusters ESTs derived from one organism on the basis of sequence homology, generating a nonredundant set of clusters representing (parts of) transcripts [7]. As GenBank is growing, the clustering is carried out regularly, resulting in new so-called UniGene builds. A number of commercial cDNA clone libraries and commercial chip formats are based on UniGene clusters (including [11,12]). In general, a sequence-specific identifier (GenBank accession number) serves as a reference to the array probe sequence. Likewise, SAGE tags, which are unique for each transcript, are mapped to UniGene clusters and GenBank accession numbers [13,14].

The existence of several technologies for measuring gene expression makes cross-technology comparison an important issue. Cross-referencing array probes allows comparison of studies carried out using different technologies, and facilitates validation of gene expression. In addition, gene-expression data need to be compared between species, signifying the need to link genes from different organisms. Because no unique gene nomenclature or sequence identifier is used for all platforms, a database linking the different identifiers belonging to one gene and its orthologs is required.

To address these issues, we have developed GeneHopper, which provides a web-based user interface to link gene-expression platforms within and between species in a batch-wise fashion. Currently, GeneHopper contains the most commonly used array resources for human, mouse and rat. In addition, the database can be used to annotate array probes with reference sequences and updated gene descriptions from UniGene. While this work was in progress, a microarray annotation tool with a different focus, RESOURCERER, was presented that allows cross-species and cross-platform comparisons [15,16].

The GeneHopper database system

The probe identifiers and the associated GenBank accession numbers representing the genes on widely used microarray platforms were parsed and loaded into a database system. These include cDNA clone libraries, Affymetrix GeneChips, and oligonucleotide sets for human, mouse and rat (Table 1). Linking of genes represented on expression platforms from a particular species is based on UniGene, of which new builds are downloaded monthly [10]. A GenBank accession number serves as a unique gene identifier for which the UniGene cluster from the currently used build is retrieved. One or more accession numbers in the target resource belonging to the same cluster are returned. For between-species searches, the UniGene cluster identifiers are subsequently used to retrieve a pair of ortholog clusters based on HomoloGene. [17]. Only UniGene clusters for reciprocal best BLAST hits and curated orthologs in the target species are retrieved and linked to the resource requested. Additional services that are offered by the GeneHopper database are updated annotation of accession numbers with NCBI's RefSeq [18], UniGene cluster, LocusLink identifier [19], Gene Ontology functional annotation [20,21], chromosomal localization, gene title, and so on.

On the GeneHopper search page, a service is selected from a dropdown menu, which contains the represented microarray resources for within-species and cross-species queries. Next, a list of user-defined accession numbers is uploaded to the database by submitting them as a plain-text file. The information retrieved is a tab-delimited file listing input accession number, UniGene cluster and build, gene symbol and title, and probe location (cDNA and oligo libraries) or probe identifier (Affymetrix). For an ortholog gene search, the homologous gene name, UniGene cluster, and the type of homology (calculated or curated) are also indicated.

It is possible that a submitted accession number yields several hits in the target dataset. The main reason for this redundancy is that array clones or sequences were selected on the basis of earlier builds of UniGene and that the clustering has changed over time. On Affymetrix chips several transcripts from one gene, including alternative splice forms, can be represented by different probe sets. Failure to identify a linking probe occurs when the input accession number is not

Table 1

Target data sets currently represented in GeneHopper

Species	Dataset*	Elements	Source
Human	Sequence-validated human cDNA library 40K	41,472	Research Genetics [11]
	Affymetrix HG-U133A	22,283	Affymetrix [12]
	Affymetrix HG-U133B	22,645	Affymetrix [12]
	Human OligoLibrary	18,885	Sigma-Genosys [31]
Mouse	NIA mouse 15K cDNA clone set	15,247	National Institute on Aging [32]
	MG-U74Av2	12,488	Affymetrix [12]
	MG-U74Bv2	12,477	Affymetrix [12]
	MG-U74Cv2	11,934	Affymetrix [12]
	Mouse OligoLibrary	7,524	Sigma-Genosys [31]
Rat	Sequence-verified rat clone collection	28,448	Research Genetics [11]
	RG-U34A	8,799	Affymetrix [12]
	RG-U34B	8,791	Affymetrix [12]
	RG-U34C	8,789	Affymetrix [12]
	Rat OligoLibrary	4,848	Sigma-Genosys [31]

*Not all Affymetrix arrays are presented in the table: most older versions and all subset arrays are available in GeneHopper as well.

used by UniGene, or when the UniGene cluster is not represented by a probe in the target dataset. For cross-species searches, the joined UniGene cluster may not be present in the HomoloGene data table. Special queries, such as finding all the accession numbers in a specific chromosomal region, or annotation of all genes in a set for protein identification, for example [22], can currently be carried out upon specific request to the database manager. When required, automated services will be implemented later.

Application examples

The GeneHopper database was developed to facilitate microarray research with cross-species and cross-platform questions. Three examples that demonstrate different types of applications are given below.

Within species: human

Commercial microarrays for gene-expression studies, such as Affymetrix GeneChips, have the advantage that they address a large number of genes, covering the whole genome for many organisms. Although giving very complete information about the transcriptional status of the cell or tissue, such screens are expensive and therefore not affordable when many samples need to be analyzed. After an initial analysis on genome-wide arrays, a researcher may desire to produce project-specific microarrays. Entering the accession numbers of genes that were selected as differentially expressed from the pilot study will generate the corresponding clones from an available cDNA clone collection, for example, the Human cDNA Library from Research Genetics (Figure 1a). From the test example, the majority of genes were represented as cDNAs in the library. Using the plate, row and column information for clone location in the database output file, a robot can be directed to rearray the desired clones into a sublibrary. PCR-amplified cDNA clones can be generated and used to produce a large number of the desired project-specific arrays at low cost.

Between species: human to mouse

Mouse and rat models are used to study human development, physiology and disease, generating the need to compare gene-expression profiles across species. In a study on Duchenne Muscular Dystrophy (DMD), gene regulation in mouse muscles lacking the DMD gene product had to be compared to that in muscle tissue from human DMD patients (Figure 1b). One hundred and thirty-one differentially expressed genes in patients, identified on the Affymetrix human HuGeneFL chips [23], were linked to the Affymetrix mouse U74Av2 chips that were used for a mouse model study [24]. For 64 out of these 131 human genes the curated or calculated best reciprocal hit mouse ortholog could be identified, represented by 82 different probe sets on the target chips.

Between technologies: mouse

Several expression-measurement platforms may be used in parallel for the same or similar samples, both in the same and different research groups. Expression analysis of the same samples on different platforms gives information on

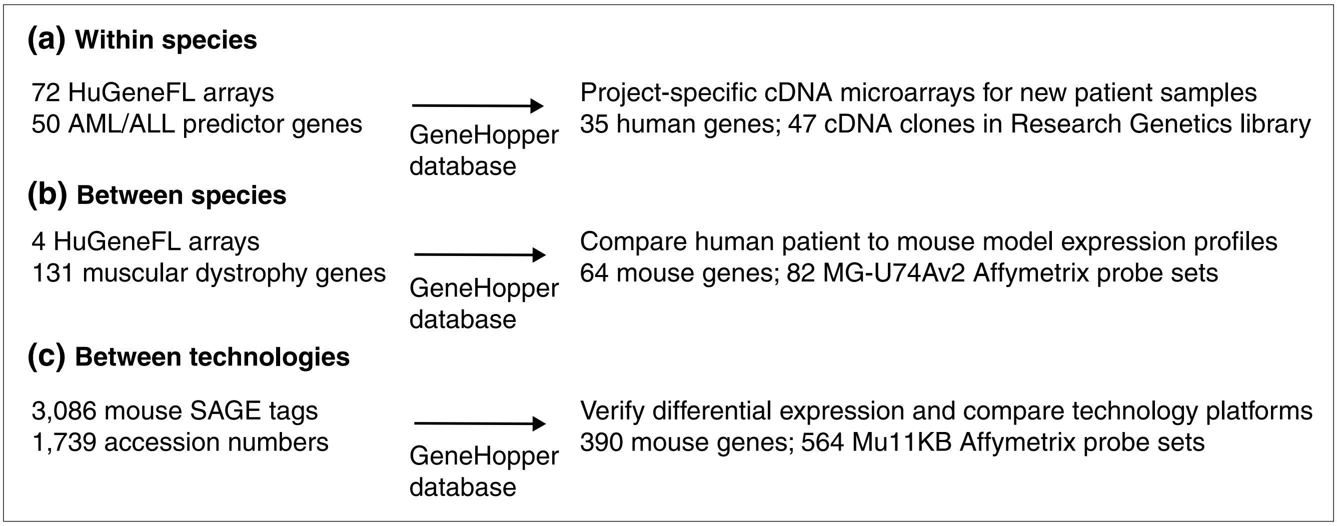
(a) Within species
72 HuGeneFL arrays
50 AML/ALL predictor genes
GeneHopper database
Project-specific cDNA microarrays for new patient samples
35 human genes; 47 cDNA clones in Research Genetics library
(b) Between species
4 HuGeneFL arrays
131 muscular dystrophy genes
GeneHopper database
Compare human patient to mouse model expression profiles
64 mouse genes; 82 MG-U74Av2 Affymetrix probe sets
(c) Between technologies
3,086 mouse SAGE tags
1,739 accession numbers
GeneHopper database
Verify differential expression and compare technology platforms
390 mouse genes; 564 Mu11KB Affymetrix probe sets

Figure 1

Application examples of the GeneHopper database. **(a)** A within-species search for cDNA clones from the Research Genetics human cDNA library corresponding to 50 marker genes to distinguish acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) that were identified by screening 72 leukemia patients on HuGeneFL Affymetrix arrays [33]. **(b)** A between-species search with genes deregulated in human muscular dystrophies, based on four HuGeneFL arrays [23], to compare to results of dystrophic mouse muscle analysis on MG-U74Av2 Affymetrix arrays [24]. **(c)** Comparison of results from two different technologies used to analyze expression in mouse livers. Accession numbers corresponding to SAGE tags were linked to probes on the Mu11KB Affymetrix array [27]. These searches were carried out using UniGene builds Hs.149 and Mm.105.

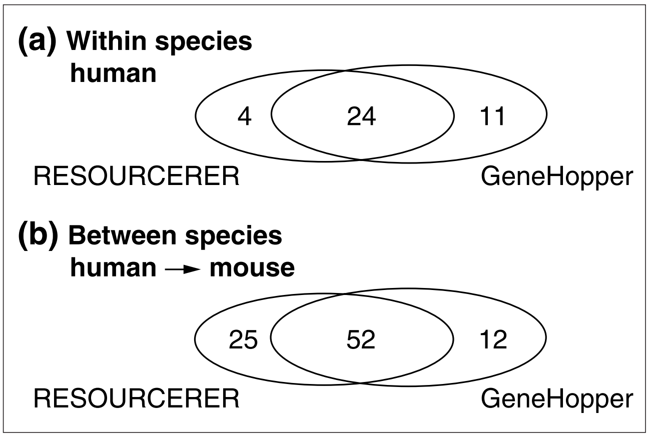

Figure 2
Comparison of search results between GeneHopper and RESOURCERER databases; queries of the latter were based on the TIGR Eukaryotic Gene Orthologs database [29]. **(a)** Within-species search for cDNA clones from the Research Genetics human library corresponding to 50 human genes. **(b)** Between-species comparison with 131 human genes to Affymetrix mouse MG-U74Av2.

the comparability of the technologies, thereby cross-validating the qualitative and quantitative measurements [25,26]. In addition, results from one technology can be verified and extended using another technology. Using the GeneHopper database, we linked accession numbers mapping to SAGE tags to the Affymetrix mouse Mu11KB chip in a study on arteriosclerosis in mice (Figure 1c and [27]). For about half of the SAGE tag accession numbers, a UniGene cluster could be found, and again for half of these a probe set was present on the Mu11KB Affymetrix chip.

Discussion

The GeneHopper database is a timesaving tool developed for microarray researchers to link and annotate different gene-expression analysis platforms across several species (Figure 1). A user-defined list of accession numbers is processed batchwise against a selected microarray resource, and the corresponding array probes are returned. The linking is carried out through residence in the same UniGene gene cluster. This approach has the advantage of identifying probes that are nonoverlapping in sequence. In addition, the database facilitates special searches, for instance to find accession numbers in a specific chromosomal region, or to annotate all genes in a set or array for chromosomal localization, protein identifier, and so on. These approaches may be useful for constructing a chromosomal region-specific or pathway-specific array.

Another high-throughput web-based database for the annotation and comparison of commonly available microarray resources, RESOURCERER, was developed at The Institute for Genomic Research (TIGR) [15,16]. This database allows comparisons between resources from the same species using either the TIGR Gene Indices [28] or UniGene, and between species using the TIGR EGO database [29]. The GeneHopper Database and RESOURCERER contain partly the same target data sets, but each also reflects the needs of local researchers both in available resources and output formats. GeneHopper is focused on uploading user-defined lists of accession numbers, which generally run faster than uploading a virtual set in RESOURCERER, while the latter very quickly combines predefined large array resources. Because the clustering algorithms underlying the TIGR and NCBI databases are different, the same queries give different, although largely overlapping, results (Figure 2). For the between-species comparison from human to mouse (Figure 2b), 4 of the 52 accession numbers that generated ortholog hits in both databases returned different genes. Although these examples are indicative of inherent differences between the search results in the two databases, the comparisons are too limited to draw general conclusions. However, it could be advisable to run queries against both databases, compare the results, and thereby increase the overall search quality.

Accessing the GeneHopper database

The database is freely accessible via the Leiden Genome Technology Center website [30], which includes information pages about the database and online help. The source code is available from the authors on request.

Acknowledgements

We thank Paul van der Elst for help with the server, Michel Villerius, Rolf Turk, Ellen Sterrenburg, Barbera van Schaik and Antoine van Kampen for fruitful discussions, and Tineke van der Pouw-Kraan for meeting B.A.T.S. on the train. This work was supported in part by a Biomolecular Informatics program grant from the Netherlands Organization for Scientific Research (NWO-BMI; B.A.T.S.) and the Centre of Biomedical Genetics, The Netherlands (J.M.B.).

References

1. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, *et al.*: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14:**1675-1680.
2. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: **Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.** *Nucleic Acids Res* 2000, **28:**4552-4557.
3. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
4. Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, Prokhorova A, Gieser L, Touma E, Lockner R, Tata M, *et al.*: **An assessment of Motorola CodeLink microarray performance for gene expression profiling applications.** *Nucleic Acids Res* 2002, **30:**e30.
5. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, *et al.*: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.** *Nat Biotechnol* 2001, **19:**342-347.
6. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270:**484-487.
7. Boguski MS, Schuler GD: **ESTablishing a human transcript map.** *Nat Genet* 1995, **10:**369-371.

8. Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA: **A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base.** *Genome Res* 1999, **9:**1143-1155.
9. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J: **The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species.** *Nucleic Acids Res* 2001, **29:**159-164.
10. **NCBI UniGene** [http://www.ncbi.nlm.nih.gov/UniGene]
11. **ResGen Invitrogen Corporation** [http://www.resgen.com]
12. **Affymetrix** [http://www.affymetrix.com]
13. **SAGEmap** [http://www.ncbi.nlm.nih.gov/SAGE]
14. **Serial Analysis of Gene Expression** [http://www.sagenet.org]
15. **RESOURCERER 6.0** [http://pga.tigr.org/tigr-scripts/magic/r1.pl]
16. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva S, Antonescu V, Cho J, Parvizi B, Cheung F, Quackenbush J: **RESOURCERER: a database for annotating and linking microarray resources within and across species.** *Genome Biol* 2001, **2:**software0002.1-0002.4.
17. **NCBI HomoloGene** [http://www.ncbi.nlm.nih.gov/HomoloGene]
18. **NCBI RefSeq** [http://www.ncbi.nih.gov/RefSeq]
19. **NCBI LocusLink** [http://www.ncbi.nlm.nih.gov/locuslink]
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
21. **Gene Ontology Consortium** [http://www.geneontology.org]
22. **SWISS-PROT and TrEMBL** [http://www.expasy.ch/sprot]
23. Chen YW, Zhao P, Borup R, Hoffman EP: **Expression profiling in the muscular dystrophies: identification of novel aspects of molecular pathophysiology.** *J Cell Biol* 2000, **151:**1321-1336.
24. Boer J, de Meijer EJ, Mank EM, van Ommen GJB, den Dunnen JT: **Expression profiling in stably regenerating skeletal muscle of dystrophin-deficient *mdx* mice.** *Neuro Musc Disord* 2002, **12:**S118-S124.
25. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18:**405-412.
26. Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ, Sealfon SC: **Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays.** *Nucleic Acids Res* 2002, **30:**e48.
27. Kreeft AJ, Moen CJA, Hofker MH, Frants RR, Vreugdenhil E, Gijbels MJJ, Havekes LM, Datson NA: **Identification of differentially regulated genes in mildly hyperlipidemic ApoE3-Leiden mice by use of serial analysis of gene expression.** *Arterioscler Thromb Vasc Biol* 2001, **21:**1984-1990.
28. **TIGR Gene Indices** [http://www.tigr.org/tdb/tgi]
29. Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, *et al.*: **Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA).** *Genome Res* 2002, **12:**493-502.
30. **GeneHopper** [http://www.lgtc.nl/GeneHopper]
31. **Sigma-Genosys** [http://www.sigma-genosys.com]
32. Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, *et al.*: **Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray.** *Proc Natl Acad Sci USA* 2000, **97:**9127-9132.
33. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.