Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 360 (2005) 78 – 82

GENE
SECTION
FUNCTIONAL GENOMICS

www.elsevier.com/locate/gene

List of lists-annotated (LOLA): A database for annotation and comparison of published microarray gene lists

Patrick Cahan [a,1], Amera M. Ahmad [a], Harry Burke [a], Sidney Fu [a], Yinglei Lai [b], Liliana Florea [c], Nachiket Dharker [a], Todd Kobrinski [a], Prachee Kale [a], Timothy A. McCaffrey [a,*]

[a] *The George Washington University Medical Center, Department of Biochemistry and Molecular Biology, 2300 I Street NW. Ross Hall 541, Washington, D.C. 20037, United States*

[b] *Department of Statistics, United States*

[c] *Department of Computer Sciences and The Catherine Birch McCormick Genomics Center, United States*

Received 11 May 2005; received in revised form 7 July 2005; accepted 11 July 2005
Available online 2 September 2005
Received by A.J. van Wijnen

Abstract

Microarray profiling of RNA expression is a powerful tool that generates large lists of transcripts that are potentially relevant to a disease or treatment. However, because the lists of changed transcripts are embedded in figures and tables, they are typically inaccessible for search engines. Due to differences in gene nomenclatures, the lists are difficult to compare between studies. LOLA (Lists of Lists Annotated) is an internet-based database for comparing gene lists from microarray studies or other genomic-scale methods. It serves as a common platform to compare and reannotate heterogeneous gene lists from different microarray platforms or different genomic methodologies such as serial analysis of gene expression (SAGE) or proteomics. LOLA (www.lola.gwu.edu) provides researchers with a means to store, annotate, and compare gene lists produced from different studies or different analyses of the same study. It is especially useful in identifying potentially "high interest" genes which are reported as significant across multiple studies and species. Its application to the fields of stem cell, cancer, and aging research is demonstrated by comparing published papers.

Keywords: Microarray; Transcript profiling; Expression profiling; Bioinformatics; Gene expression; Gene annotation; Aging

1. Introduction

Microarray analysis of gene expression is capable of screening the entire transcriptome of a cell type or organ to find genes which change in association with a disease or drug treatment. With the exponential growth in gene expression data, biologists are faced with the challenge of extracting useful information from this raw data, and subsequently developing a coherent storage and retrieval system for the analysis of processed data. Several powerful software tools and databases have been built to facilitate the management and curation of raw microarray data. Among them are OncoMine (Rhodes et al., 2004), SMD (Gollub et al., 2003), Longhorn (Killion et al., 2003), Array Express (Brazma et al., 2003), GEO (Edgar et al., 2002), DAVID (Dennis et al., 2003), and SOURCE (Diehn et al., 2003). Some serve as repositories of primary gene expression data from a specific platform, while others provide functional annotations of genes or gene lists under study (Table 1). Despite the availability of these tools for managing raw data, the published results remain surprisingly inaccessible for further comparison.

Abbreviations: AML, acute myeloid leukemia; *ApoD*, apolipoprotein D; ESC, embryonic stem cells; LOLA, list of lists annotated; NPC, neural progenitor cells; RDBMS, relational database management system; RPC, retinal progenitor cells; SAGE, serial analysis of gene expression.

* Corresponding author. Tel.: +1 202 994 8919; fax: +1 202 994 8924.
E-mail address: mcc@gwu.edu (T.A. McCaffrey).
[1] Current address: Division of Biology and Biomedical Research, Washington University, St. Louis, MO 63110, USA.

0378-1119/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2005.07.008

Table 1
Comparison of features of microarray databases, PubMed, and LOLA

	Raw data repositories	PubMed	LOLA
Stores microarray images	Y	N	N
Stores raw data	Y	N	N
Ability to process raw data	Y	N	N
Stores results (gene lists)	Y/N	Y	Y
Compares results from different studies	Y/N	N	Y
Cross platform comparison of results	N	N	Y
Cross species comparison of results	N	N	Y
Searchable access to published lists	N	N	Y

Microarray studies typically generate lists of 100–500 transcripts which are changed by a given disease or treatment. There is no uniform and easily accessible method of comparing lists of genes thought to be significantly linked to a specific disease or treatment. Enormous resources have been invested in microarray screening of aging, cardiovascular diseases, cancer, and dozens of other diseases, and yet, the resulting data is not amenable to electronic access, comparison, or updating of the annotations. The comparison of gene lists, arising from microarray, serial analysis of gene expression (SAGE), or differential display, is severely hindered by different platforms and different gene nomenclatures, and the fact that the resulting gene lists are inaccessible by conventional PubMed searches. The results of microarray studies, which identify potentially important lists of genes, could, and should be compared in order to:

- Identify genes that are robust across different signal generation, normalization and analysis methods within the same microarray study.
- Identify genes that are observed in common between different, independent studies of the same disease or treatment.
- Identify sets of genes that may be modulated in common between different disease states or drug treatments.

To address these needs, LOLA was developed as a web-enabled database for comparison and annotation of microarray-generated lists of genes from divergent platforms and studies.

LOLA allows users to upload their gene lists in order to identify genes that are observed in common between their list and other gene lists. LOLA provides a method to compare and analyze thousands of genes in multiple lists simultaneously and provides a uniform measure of gene list similarity. LOLA circumvents many of the problems associated with integrating and comparing raw microarray data produced in different studies by allowing users to simply compare results of prior analyses (Fig. 1).

Recently, The Tumor Analysis Best Practices Working Group recommended that Affymetrix microarray analyses utilize at least two probe set algorithms in order to identify

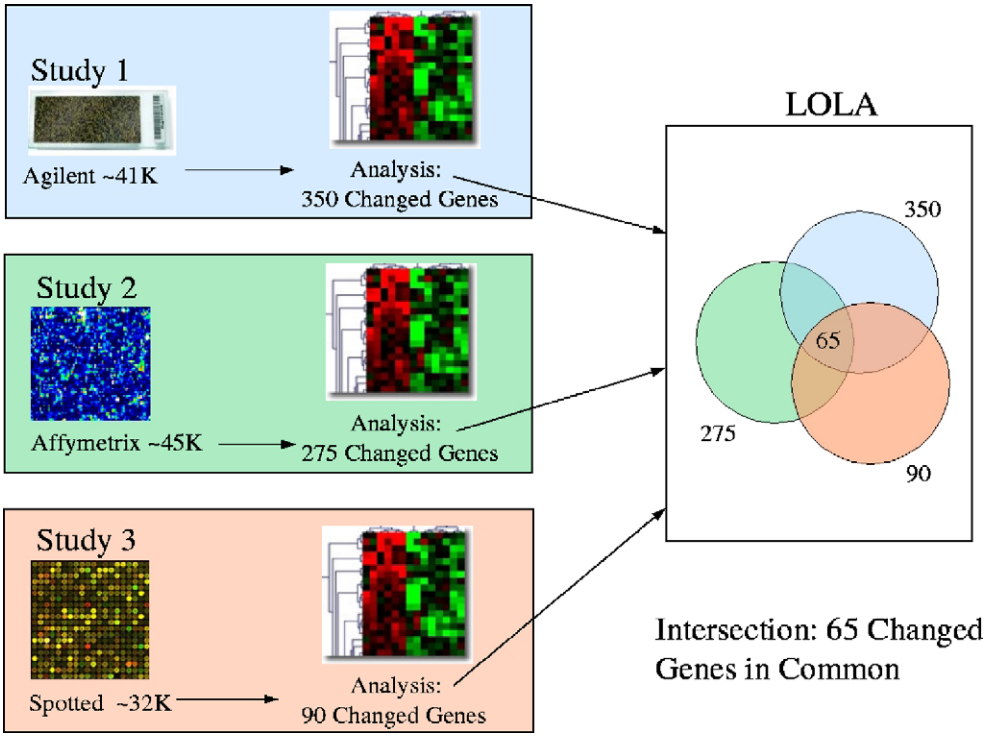

Fig. 1. Comparing the results of microarray studies from different platforms using LOLA. Different microarray platforms, such as Agilent, Affymetrix, and spotted arrays can be processed through a variety of different analytical methods, such as SAM, *t*-tests, or clustering, to produce smaller gene lists that are thought to be related to a particular disease or drug treatment. Regardless of the platform involved, the resulting lists can be compared for similarity/dissimilarity using LOLA.

robustly significant genes (Hoffman et al., 2004). LOLA is an ideal tool to compare lists produced using various expression measures. Additionally, LOLA provides links to Entrez Gene (Wheeler et al., 2004), GeneCards (Rebhan et al., 1997), NetAffx, and PubMed in one interface. Presently, LOLA archives data on human, mouse and rat genomes.

2. Construction and content

2.1. Database structure

LOLA employs a three-tier web application architecture. In tier 1, data is stored in the relational database management system (RDBMS) MySQL. The tier 2 application layer is implemented in PHP (www.php.net), which is executed via calls from the tier 3 Apache web server running on the Linux operating system. In addition to requirements for data consistency and the ability to update easily, the data model was designed to allow for flexible and efficient gene list comparisons and searches.

Gene annotations are stored and updated by automatically parsing annotation files produced using BioConductor's AnnAffy and MetaData packages (Ihaka and Gentleman, 1996). Each row of an annotation file contains an Affymetrix probe set identifier, LocusLink/Gene identifier, the gene name and gene symbol, when known. The number of genes currently archived in LOLA is illustrated in Fig. 2.

2.2. Statistics

The concordance between any two gene lists is calculated as the number of genes in common divided by the number of genes in both of the lists (the intersection divided by the union). The significance of this concordance is evaluated in two different ways. First, the variance of the concordance is calculated as $A(B+C)/(A+B+C)^3$ where A is the number of common genes, while B and C are the numbers of unique genes on their respective lists (Burke and Hoang, in preparation). The variance is used to compute the 95% confidence interval (CI) by standard methods. If the CI does not include the value 0, which indicates no concordance, then the concordance is significant, with 95% confidence. Secondly, the *P*-value of a given intersection size is the probability of observing an intersection size that occurs by chance and is larger than the given one. It is calculated as: $\Pr(X > c | m_1, m_2, n_1, n_2, s) = \sum_{x>c} \Pr(x \mid m_1, m_2, n_1, n_2, s)$, where c is the number of genes common to both lists, m_1 and m_2 are the sizes of two lists, n_1 and n_2 are the sizes of two arrays, and s is the total number of probes from which the two arrays were sampled. Evaluating the statistical significance of the concordance, however, is limited by a number of potential biases that could systematically favor the presence of gene subsets on a list. These biases include the selection bias of named genes for inclusion on microarrays, and the abundance of some transcripts, which would favor their detection experimentally.

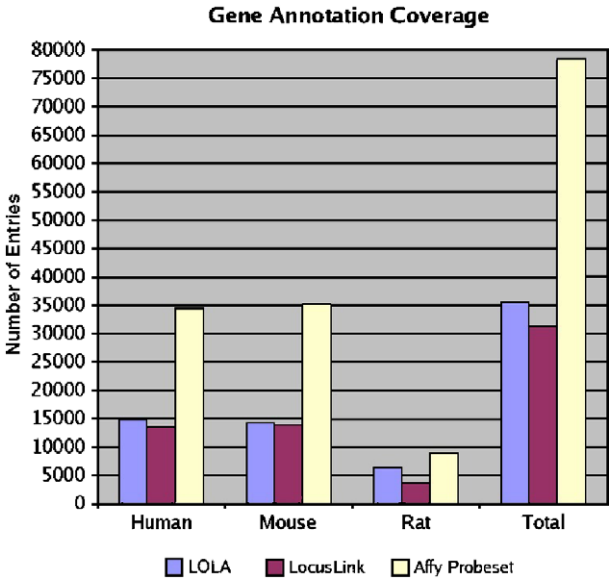

Fig. 2. The number of searchable elements referenced in LOLA. The number of possible probes encompassed in LOLA is graphed as a function of the species and probe/gene identifier type.

3. Results and discussion

3.1. Data access

LOLA provides a simple interface for viewing, storing, re-annotating, and comparing gene lists. Users create a new list by uploading a tab-delimited text file of specified format or by pasting a gene list into a form on the LOLA site. The user must select the identifier type that is used to identify each gene. Currently, LOLA supports both Gene and Affymetrix probe set identifiers. If a matching gene identifier is not found, that gene will not be stored in the database, and the submitter notified. A single gene is allowed to occur multiple times in the same list although it will only be used once in comparisons. Reference information, such as the PubMed citation, study design, sample types, analysis method and criteria used in generating the list, can be saved along with the gene list. After a list is submitted, LOLA will display the annotated gene list for confirmation by the user prior to saving the list in the database. Users can create new folders and move lists among folders to assist in organization. The list is available for all users to view and compare after the curator has verified it.

Gene list summaries, which include list title, description, analysis type, and the number of genes, are displayed in a folder hierarchy. Published gene lists are hyperlinked to their associated PubMed abstracts. The gene names and descriptions in a list are viewed by clicking on the list title.

Each gene is linked to GeneCards, Entrez Gene and, when applicable, NetAffx through their Affymetrix probe set ID (Liu et al., 2003). Fold changes or log ratios appear color-coded in the gene list display.

Gene lists are compared by selecting their check boxes and pressing the 'compare lists' button. LOLA performs a pair wise comparison between each of the selected lists and produces a table that reports the number of genes in common between the lists, the concordance calculation, confidence interval, and *P*-value. The names of the genes in common are retrieved by clicking on the numerical entry in the Intersection column. When two lists derive from different species, a gene is counted as common to both if its homolog is found in the other list, as reported from NCBI's HomoloGene. Users can also compare one list of interest to all accessible lists by selecting the 'compare to all' function.

3.2. Validation of stem cell microarray papers

Lists from a published paper on stem cells were tested on LOLA in order to demonstrate its usefulness and validate its results. Fortunel et al. (Fortunel et al., 2003) carried out gene expression profiling of three stem cell types: embryonic stem cells (ESC), neural progenitor cells (NPC) and retinal progenitor cells (RPC) with their differentiated progenies. The intersection of these lists defined a list of 385 'stemness' genes expressed by all three stem cells. The same lists of genes were uploaded into LOLA and compared. LOLA found the same 385 genes enriched in all three stem cells as was identified in the original analysis. LOLA confirmed a significant concordance (95% CI) between ESC and NPC of 0.19 (±0.014), between NPC and RPC of 0.29 (±0.016), and between ESC and RPC of 0.28 (±0.016), with all *P*-values less than 10^{-200}.

3.3. Comparison of cancer microarray studies

It can be quite difficult to compare the results of microarray studies from different laboratories studying the same disease. Two recent reports by Bullinger et al. (Bullinger et al., 2004) and Valk et al. (Valk et al., 2004) used microarrays to identify genes that may have value in classifying acute myeloid leukemia (AML). The accompanying Perspectives (Lui and Karuturi, 2004) and Editorial (Grimwade and Haferlach, 2004) emphasized a robust similarity in the findings of the 2 studies, but did not specify the number of similar genes that were identified. Comparing the 133 genes identified by Bullinger (122 found in LOLA) and the 182 genes identified by Valk, LOLA identified only 9 genes (Table 2) that were in common (concordance = 0.040 (+0.026), $p = 2.66 \times 10^{-7}$). While it suggests that the 2 studies did not observe highly similar results (<5% concordance), the lower end of the 95% confidence interval falls slightly above 0, indicating a significant concordance, and the odds of such an overlap occurring randomly are quite small. In this case, LOLA generates a testable hypothesis that the 9 genes that are reproducible between the 2 studies may have greater predictive value than genes that did not reproduce in both studies.

Table 2
Genes identified by 2 different studies of acute myeloleucytic leukemia (AML)

D2S448	melanoma associated gene
FLJ23058	hypothetical protein FLJ23058
HOXA4	homeo box A4
HOXB2	homeo box B2
HOXB5	homeo box B5
LCN2	lipocalin 2 (oncogene 24p3)
LOC57228	hypothetical protein from clone 643
MSLN	mesothelin
PBX3	pre-B-cell leukemia transcription factor 3 (occurs twice)

3.4. Interspecies comparison of microarray analysis of the aging brain

One method of evaluating microarray results would be to determine which genes reproducibly change in different studies employing a similar design. Aging studies should be reasonably similar between different species, and so we conducted an in-depth literature search of microarray studies involving aging. A total of 25 separate gene lists, from 20 papers, were uploaded into LOLA for comparison. To illustrate the cross-species function of LOLA, 3 different microarray studies of brain aging, in 3 different species, were identified and compared using LOLA. Human post-mortem frontal cortex specimens from subjects ranging in age from 26 to 106 were analyzed by Affymetrix GeneChips to identify genes increased or decreased in subjects over 40 (Lu et al., 2004). Likewise, aging-related genes in rat hippocampus were identified by microarray analysis (Blalock et al., 2003), as were genes associated with aging of the neocortex in mice (Prolla, 2002). Using the cross-species features of LOLA, a 3-way comparison revealed that any 2 lists had significant concordance of 6–8 genes in common from a total of 134–145 genes per list. However, only 1 gene was common to all 3 lists: apolipoprotein D (*ApoD*). *ApoD* has previously been associated with cellular senescence (Provost et al., 1991) and with aging of the human brain (Kalman et al., 2000), possibly as a function of the increased presence of reactive astrocytes in aged brain (del Valle et al., 2003). Thus, LOLA allows the identification of genes that change reproducibly in different studies, even when the studies are in different species.

3.5. Future directions

LOLA's coverage of gene expression data will continue to expand as more genomes become available and existing genome annotations are augmented. An algorithm that will facilitate comparisons based on Gene Ontologies (GO) (Ashburner et al., 2000) is under development. In this way,

lists that share closely related genes, expressed as detailed GO terms, will still correlate well. TIGR's microarray analysis software, MeV (Saeed et al., 2003), has been adapted in such a way that it can upload new lists directly to LOLA. There are several other features that are currently under development:

- Sorting lists and genes according to name, symbol, fold changes, etc.
- An option for automatic email notification when a newly uploaded list meets a threshold similarity to an author's existing list.
- Expanded gene and probe ID associations to allow for upload and comparison of probes which do not have an Entrez Gene or Affymetrix ID.
- Improved searching for genes and gene ontologies in lists.
- Downloadable version for local installation.

3.6. Availability and requirements

LOLA is freely available for browsing, uploading of lists, and comparing publicly available gene lists. Registration is required to upload gene lists. See: http://www.lola.gwu.edu/. The source code is available upon request for academic users.

3.7. Conclusions

LOLA stores analyzed microarray gene lists, not raw data, and measures the concordance among gene lists. LOLA can be used to answer several types of questions, including: (1) What genes are changed in all microarray studies of a particular biological phenomena? (2) What genes are shared between an investigator's list and other analyzed lists, which focus on his area of interest? (3) Within a microarray experiment, which genes remain significant across different signal summarization, normalization, and analysis methods? While useful for microarray data comparison, LOLA is not limited to microarray results, and is equally capable of comparing results from differential display, SAGE, or proteomic analyses.

Acknowledgements

The authors are grateful for the financial support of the NIH/National Institutes on Aging (AG12712), and The Catherine Birch McCormick Genomic Center.

References

Ashburner, M., et al., 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. 25, 25–29.

Blalock, E.M., et al., 2003. Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment. J. Neurosci. 23 (9), 3807–3819.

Brazma, A., et al., 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 31, 68–71.

Bullinger, L., et al., 2004. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. N. Engl. J. Med. 350 (16), 1605–1616.

Burke, H., Hoang, A., in preparation. A new measure of gene expression concordance.

del Valle, E., Navarro, A., Astudillo, A., Tolivia, J., 2003. Apolipoprotein D expression in human brain reactive astrocytes. J. Histochem. Cytochem. 51 (10), 1285–1290.

Dennis, G., et al., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 4 (5), P3.

Diehn, M., et al., 2003. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res. 31, 219–223.

Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210.

Fortunel, N.O., Otu, H.H., Ng, H.H., Chen, J., 2003. Comment on "'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature" (I). Science 302, 393b.

Gollub, J., et al., 2003. The Stanford microarray database: data access and quality assessment tools. Nucleic Acids Res. 31, 94–96.

Grimwade, D., Haferlach, T., 2004. Gene-expression profiling in acute myeloid leukemia ([Editorial]). N. Engl. J. Med. 350 (16), 1676–1680.

Hoffman, E., et al., 2004. Expression profiling—best practices for data generation and interpretation in clinical trials. Tumor Analysis Best Practices Working Group. Nat. Rev., Genet. 5 (3), 229–237.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comp. Graph. Stat. 5, 299–314.

Kalman, J., McConathy, W., Araoz, C., Kasa, P., Lacko, A.G., 2000. Apolipoprotein D in the aging brain and in Alzheimer's dementia. Neurol. Res. 22 (4), 330–336.

Killion, P., Sherlock, G., Iyer, V., 2003. The Longhorn Array Database (LAD): an open source, MIAME compliant implementation of the Stanford Microarray Database (SMD). BMC Bioinformatics 4 (1), 32.

Liu, G., et al., 2003. NetAffx: affymetrix probesets and annotations. Nucleic Acids Res. 31, 82–86.

Lu, T., et al., 2004. Gene regulation and DNA damage in the ageing human brain. Nature 429 (6994), 883–891.

Lui, E., Karuturi, K., 2004. Microarrays and clinical investigations. [Perspective]. N. Engl. J. Med. 350 (16), 1595–1597.

Prolla, T.A., 2002. DNA microarray analysis of the aging brain. Chem. Senses 27 (3), 299–306.

Provost, P.R., Marcel, Y.L., Milne, R.W., Weech, P.K., Rassart, E., 1991. Apolipoprotein D transcription occurs specifically in nonproliferating quiescent and senescent fibroblast cultures. FEBS Lett. 290 (1–2), 139–141.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D., 1997. GeneCards: Encyclopedia for Genes, Proteins and Diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center, Rehovot, Israel. http://bioinformatics.weizmann.ac.il/cards.

Rhodes, D.R., et al., 2004. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 6 (1), 1–6.

Saeed, A., et al., 2003. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34, 374–378.

Valk, P.J., et al., 2004. Prognostically useful gene-expression profiles in acute myeloid leukemia. N. Engl. J. Med. 350 (16), 1617–1628.

Wheeler, D.L., et al., 2004. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res. 32, D35–D40.