BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 12 2014, pages 1780–1781
doi:10.1093/bioinformatics/btu109

Gene expression

Advance Access publication February 25, 2014

MGDB: crossing the marker genes of a user microarray with a database of public-microarrays marker genes

Mario Huerta[1], Marc Munyi[2], David Expósito[2], Enric Querol[1] and Juan Cedano[3,*]

[1]Institut de Biotecnologia i Biomedicina, [2]Escola Tècnica Superior de Ingenieria, Universitat Autònoma de Barcelona; Bellaterra, Barcelona 08193, Spain and [3]Laboratorio de Inmunología, Universidad de la República Regional Norte-Salto, Rivera 1350, CP 50000 Salto, Uruguay

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: The microarrays performed by scientific teams grow exponentially. These microarray data could be useful for researchers around the world, but unfortunately they are underused. To fully exploit these data, it is necessary (i) to extract these data from a repository of the high-throughput gene expression data like Gene Expression Omnibus (GEO) and (ii) to make the data from different microarrays comparable with tools easy to use for scientists. We have developed these two solutions in our server, implementing a database of microarray marker genes (Marker Genes Data Base). This database contains the marker genes of all GEO microarray datasets and it is updated monthly with the new microarrays from GEO. Thus, researchers can see whether the marker genes of their microarray are marker genes in other microarrays in the database, expanding the analysis of their microarray to the rest of the public microarrays. This solution helps not only to corroborate the conclusions regarding a researcher's microarray but also to identify the phenotype of different subsets of individuals under investigation, to frame the results with microarray experiments from other species, pathologies or tissues, to search for drugs that promote the transition between the studied phenotypes, to detect undesirable side effects of the treatment applied, etc. Thus, the researcher can quickly add relevant information to his/her studies from all of the previous analyses performed in other studies as long as they have been deposited in public repositories.

Availability: Marker-gene database tool: http://ibb.uab.es/mgdb

Contact: jcedano@unorte.edu.uy

Received on February 18, 2013; revised on December 24, 2013; accepted on January 26, 2014

1 INTRODUCTION

Since 2002, the Nature journals, among others, have announced that authors were thereafter required to deposit microarray data in public repositories like Gene Expression Omnibus (GEO) (Barrett *et al.*, 2011) or ArrayExpress (Parkinson *et al.*, 2007) so that anyone could freely access and critically evaluate the data discussed in manuscripts (Nature, 2002). But how could all of these data help your particular research? How could all of this information enrich your microarray data analysis to make conclusions, to formulate hypotheses or even to construct new models from them?

To date, GEO archives ~20 000 studies comprising 500 000 samples, 33 billion individual measurements for >1300 organisms, submitted by 8000 laboratories from around the world, and supporting data for >10 000 published manuscripts (Barrett *et al.*, 2011).

Samples within GEO datasets are further grouped and classified into subsets according to the experimental variables under examination in each study, for instance 'tissue' or 'strain'. So, any gene of the microarray will be a marker gene of the microarray if its expression displays a significant effect in relation to subsets, that is, if the expression values pass a threshold of statistical difference between any experimental-variable subset and another (Barrett *et al.*, 2011).

The experimental variables are based on the sample origin, such as species, specimen, strain, individual, tissue, development stage, cell type, cell line, on individual features like age or gender, on pharmacological experimentation such as agent, dose, protocol, on the disease genesis such as genotype/variation, disease state, infection, shock, stress or on other experimental conditions like temperature or time.

How can we establish a correspondence between these sample subsets perfectly classified by experimental variable and the sample clusters obtained by statistical methods from our microarray under study? We can do this by verifying that marker genes are the same in both microarrays, similar to cMAP (Lamb *et al.*, 2006), because it can imply that the subsets of both microarrays describe the same phenotypes.

So, although the experimental-variable subsets of GEO microarrays are defined by microarray developers, and the experimental variables are subjected to the hypothesis that the researchers wanted to investigate, these micrroarray data can be reused for investigations around the world even when their hypotheses are completely different from the originals.

Our tool can be used for different purposes:

1. To study the role of marker genes of the user's microarray in other microarrays.
2. To assign biological meaning to the sample clusters of the user's microarray. This can include the following:
   - 2.1. To compare the user's experiments with experiments for the same pathology but in different tissues, in different species or directly different pathologies.
   - 2.2. To search for drugs whose effect causes the transition between the phenotypes studied in the user's microarray.

*To whom correspondence should be addressed.

2.3. To study the phenotype of different subsets of individuals under investigation.

2.4. To search undesirable side effects of a treatment studied in the user's microarray.

2 METHODOLOGY

Database of microarray marker genes: The samples of GEO microarray datasets are classified by experimental variable such as treatment, protocol, disease state, patients' condition, tissue. So, our database of microarray marker genes contains the genes with statistically significant differences in their expression between any experimental variable subset and another for each GEO microarray. The system obtains these marker genes directly from GEO. The database is updated monthly with the marker genes of the new GEO datasets.

Marker-gene search in the user's microarray: The definition of sample clusters and the marker-gene search in the user's microarray is totally versatile. Our system provides the sample clusters calculated by common clustering methods (HC, SOM, SOTA) or allows the researcher to define them based on his/her hypothesis. The marker genes can be searched for by having some of the clusters being upregulated or downregulated with respect to the basal value, or by being overexpressed or underexpressed with respect to other clusters. All possible combinations are allowed and different combinations will supply different sets of marker genes (Huerta *et al*., 2009). This procedure permits the researcher to limit the search of matching microarrays to a specific phenotypic change.

Crossing the user's microarray marker genes with the database of microarray marker genes: By comparing the marker genes of the user's sample clusters with the marker genes of the database, the system returns all microarrays with common marker genes with respect to the specific search for the user's microarray. Then, the correspondence between the sample-clusters of the user's microarray and the matching microarrays can be elucidated using the graphical interface (Fig. 1). A correspondence can be established between two clusters if the common marker genes overexpressed in one cluster are overexpressed in the other.

3 RESULTS

When our application crosses the marker genes of the user's microarray sample clusters with the database of microarray marker genes, two types of lists are provided:

List-of-microarrays view: The list of matching microarrays is ordered by the number of marker genes in common between the user's microarray specific search and the matching microarrays. The list of microarrays can be filtered by the experimental variable, like 'agent', 'dose' or 'time', to limit the search to a concrete scope, or by keywords like 'breast cancer'. The common marker genes between the user's microarray and each matching microarray can be analyzed in the *list-of-marker-genes view*.

List-of-marker-genes view: The common marker genes between the two microarrays are listed. The cluster distribution along the gene expression is shown for each marker gene and both microarrays. In this way, the researcher can quickly establish the correspondence between the clusters of his/her own microarray and the public microarray. The more microarrays satisfactorily compared, the more attributes that can be assigned to the user's microarray sample clusters, which could have been calculated by statistical methods and, thus, their biological significance be unknown.

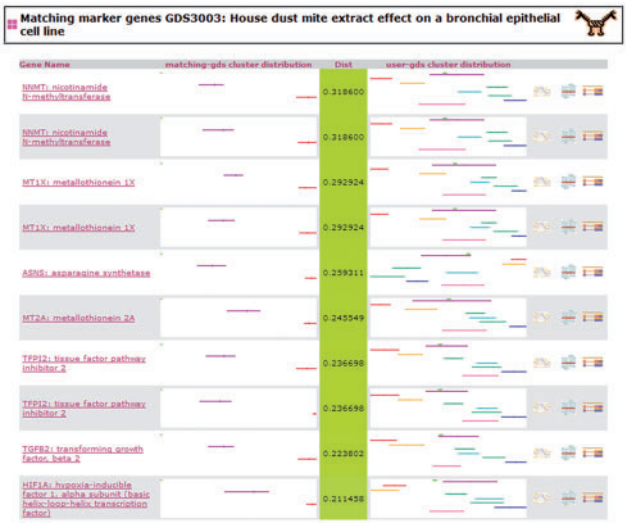

Fig. 1. The list of marker-genes view. The common marker genes between a matching microarray from GEO and the user's microarray are listed. The distribution of the clusters along the gene expression is shown for each gene and the two microarrays. Comparing the two bar charts of each marker gene the researcher can establish the correspondences between the clusters of his/her microarray and the subsets of the GEO microarray. Dist value shows the difference in expression among the clusters specified in the marker-gene search

As a result, the user can enrich his/her microarray data analysis and sample clusters, improve his/her future experimental design and check the hypotheses generated from the data in the ways cited in the Section 1.

Funding: MCYT (BFU2010-22209-C02-01); the Centre de Referència de R+D de Biotecnologia de la Generalitat de Catalunya; Comisión Coordinadora del Interior (Uruguay).

Conflict of Interest: none declared.

REFERENCES

Barrett,T. *et al.* (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.

Huerta,M. *et al.* (2009) PCOPGene-Net: holistic characterisation of cellular states from microarray data based on continuous and non-continuous analysis of gene-expression relationships. *BMC Bioinformatics*, **10**, 138.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Nature. (2002) Microarray standards at last. *Nature*, **419**, 323.

Parkinson,H. *et al.* (2007) ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.