LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-JRNL-219041

Array2BIO: A Comprehensive Suite of Utilities for the Analysis of Microarray Data

G. G. Loots, P. S. G. Chain, S. Mabery, A. Rasley, E. Garcia, I. Ovcharenko

February 16, 2006

BMC Bioinformatics

Array2BIO: A Comprehensive Suite of Utilities for the Analysis of Microarray Data.

Gabriela G. Loots[1], Patrick S. G. Chain[1], Shalini Mabery[1], Amy Rasley[1], Emilio Garcia[1], and Ivan Ovcharenko[2*]

[1] Biosciences Directorate, [2] Computational Directorate, Lawrence Livermore National Laboratory, Livermore, California 94550, USA

* Phone: 1 (925) 422 5035; Fax: 1 (925) 422 2099; Email: ovcharenko1@llnl.gov

ABSTRACT.

We have developed an integrative and automated toolkit for the analysis of Affymetrix microarray data, named Array2BIO. It identifies groups of coexpressed genes using two complementary approaches – comparative analysis of signal versus control microarrays and clustering analysis of gene expression across different conditions. The identified genes are assigned to functional categories based on the Gene Ontology classification, and a detection of corresponding KEGG protein interaction pathways. Array2BIO reliably handles low-expressor genes and provides a set of statistical methods to quantify the odds of observations, including the Benjamini-Hochberg and Bonferroni multiple testing corrections. Automated interface with the ECR Browser provides evolutionary conservation analysis of identified gene loci while the interconnection with Crème allows high-throughput analysis of human promoter regions and prediction of gene regulatory

elements that underlie the observed expression patterns. Array2BIO is publicly available at http://array2bio.dcode.org.

INTRODUCTION

Microarray experiments provide rapid and direct access to whole genome gene expression profiling. Recently, this experimental approach has become routine for the *en masse* identification of genes associated with different biological processes. We have developed a multifunctional, user-friendly, web-interactive microarray analysis tool, Array2BIO, that identifies and functionally characterizes coexpressed genes. It also integrates other genomic, transcriptional and gene regulatory tools available at http://www.dcode.org to further explore in greater detail the underlying mechanisms of gene co-regulation. Array2BIO permits the users to functionally characterize clusters of co-expressed genes, identify putative biological activities, build interaction networks, as well as predict modules of transcription factors regulating eukaryotic gene expression in different tissues and under different conditions. This tool is freely available at http://array2bio.dcode.org or through the http://www.dcode.org web portal.

IMPLEMENTATION

Figure 1 summarizes the schematics behind the Array2BIO analysis. Users are required to submit only the input .CEL files – (ie. the standard output data derived from Affymetrix microarray experiments). Array2BIO performs multi-step data analysis and filtering, including background correction, exclusion of non-specific hybridizing probes, normalization and logarithmic transformation of raw intensities. Individual probes are automatically mapped to Affymetrix tags and subsequently to UCSC 'known genes' (1). In contrast to other available microarray analysis software, Array2BIO analysis also incorporates a balanced analysis of low- and high-expressor genes thus providing a

reliable method for handling low-expressors that otherwise lead to false positive predictions.

Two complementary methods of microarray data analysis are incorporated into the Array2BIO software: 1) comparative and 2) clustering analyses. Comparative analysis identifies genes that are differentially regulated in reference to a control sample (for example gene expression in transgenic animals compared with non-transgenic, wild-type littermates). Clustering analysis identifies groups of genes that are co-expressed under several different conditions (e.g. when analyzing timecourse experiments).

The automated functional classification of co-expressed genes is based on the Gene Ontology (2) database and allows for the identification of 'enriched' or 'depleted' categories in assigned biological processes, molecular functions, and cellular components. Automated KEGG (3) classification of gene interaction identifies the major biochemical processes that underlie any observed differences in gene expression and groups them into five main categories - (1) metabolism, (2) genetic information processing, (3) environmental information processing, (4) cellular processes, and (5) human diseases.

Every group of differentially expressed genes identified using Array2BIO is dynamically linked to the Evolutionary Conserved Region (ECR) Browser (http://ecrbrowser.dcode.org; (4)) and to the Cis-REgulatory Module Explorer (Crème 2.0; http://creme.dcode.org; (5)) tools, as well as to the NCBI database

(http://www.ncbi.nlm.nih.gov/). The ECR Browser provides multi-species evolutionary conservation information for individual genes, and the NCBI database prompts in detailed information about mRNA sequences and related proteins, while the Crème 2.0 tool allows the users to perform an additional step to further functionally annotate the group of human genes by analyzing the promoter elements of these genes and identifying shared clusters of evolutionary conserved transcription factor binding sites within these promoters. Combined, these tools provide a wealth of information regarding the gene(s) in question, their conservation, their transcripts, as well as potential transcriptional networks underlying the observed transcriptional response provided by the microarray data.

APPLICATION

Plague (commonly known as the Black Death that devastated much of the known world in the 14th century) is a primarily a disease in rodents caused by the bacterium *Yersinia pestis*. Plague, which is transmitted to humans through the bite of infected fleas, has swept across the world, killing up to 200 million people during three major pandemics. To address the host-pathogen interactions and the molecular mechanisms underlying the extreme virulence of this pathogen during human infection, we have analyzed microarray data of temporal gene expression patterns of human dendritic cells subjected to *Y. pestis* infection using Array2BIO. HG-U133A microarray expression data of human dendritic cells at 4 hours after exposure to *Y. pestis* was compared to mock-exposed cells. We observed significant increases and decreases in expression (as measured using the Welch's t-test analysis with Benjamini-Hochberg correction for multiple testing) for 139

and 81 human genes, respectively. Gene Ontology (GO) analysis identified 31 'enriched' biological processes and 5 molecular functions corresponding to the upregulated genes; while none were found for downregulated genes. Expectedly, the majority of these categories were related to the human immune response, including the "response to pest, pathogen or parasite" (Table 1). The chemokine (cytokines with chemotactic activities) category was ~20-fold 'enriched' when compared to the expected values due to chance alone. Eighteen percent of all human chemokines (primarily CXC chemokines) are activated in response to *Y. pestis* invasion. KEGG analysis of the corresponding gene interactions identified a family of upregulated CXC cytokines acting upstream of the IL8RB receptor, and several other receptor genes (Figure 2). These pathways are likely to reflect the core response of the human dendritic cells to this infection The KEGG analysis of overpopulated cellular processes provides some clues to the mechanisms of *Y. pestis* infection. There are two related subcategories that are identified by Array2BIO: (1) apoptosis ($p<0.001$) and (2) cell growth and death ($p<0.002$). Six genes are shared between these two subcategories and may be key players in plague's disease etiology.

We performed Crème 2.0 analysis on 25 genes identified in this study that are related to the "response to pest, pathogen or parasite" GO category. Crème 2.0 predicted transcription factors that potentially act as key regulators of these genes and are likely to upregulate their expression during *Y. pestis* infection. Several transcription factors binding sites conserved between human and rodents were significantly enriched in the promoters of these genes, including several members of the STAT and NFKB families, as well as TATA transcription factor. While the TATA transcription factor plays a basal

role in the TATA-box recognition, the two other identified transcription factor families are known to be involved in regulating the immune system. STAT and NFKB proteins respond to cytokines, are associated with inflammatory disease and can lead to inappropriate immune cell development (6,7).

METHODS

Microarray data analysis.

Background correction. Array2BIO follows the original Affymetrix procedure of background correction. An array of probes is separated into 16 zones (4x4 grid). Raw intensities for each zone are ranked and the background level is selected as a 2% lowest intensity for the zone. Distances from each probe to different zone centers are used to estimate the background level at the probe location, which is then subtracted from the raw probe intensity.

Filtering out non-specific hybridization. Each probe intensity is measured in duplicate - a perfect match (PM) intensity and a mismatch (MM) intensity, where the MM intensity estimates the cross-reactivity with other genes. Array2BIO excludes all the probes with the PM intensity less than 1.25*MM intensity. It also calculates the ratio of probes with specific hybridization that survive this filtering. MM intensity is substracted from the PM intensity for the surviving probes, so the raw intensity is measured as the relative(PM-MM) intensity.

Normalization and Log2 transformation. Median (PM-MM) array intensity $\tilde{I}$ is calculated for the remaining probes after the filtering step. Individual (PM-MM) probe intensities I_i undergo normalization and a base 2 logarithmic transformation:

$EP_i = \log_s(I_i / \tilde{I})$.

Probe to tag mapping. Affymetrix *.CDF* files are used to map individual probe intensities EP_i onto different Affymetrix gene tags GP_j. Usually each tag accumulates ~10 good probes that span the corresponding gene transcript.

Averaging experiment replicas. Several experimental replicas can be averaged in comparative analysis to reliably estimate signal and background gene expression levels.

Filtering out the outliers. It is common to observe that the expression level of several gene probes differs significantly from the median level of the transcript gene expression $\tilde{GP}_j$. To filter out the outliers, Array2BIO excludes transcript probes that differ from $\tilde{GP}_j$ in expression by the given number of standard deviations σ_j (as specified by the user). By default, a strict filtering ($1*\sigma_j$) and a medium stringency filtering ($2*\sigma_j$) are selected for comparative and clustering analyses, respectively.

Statistical methods (comparative analysis).

Handling low-expressors. The significance of fold-difference in intensity values (ie. expression) varies dramatically for low- vs. high-expressor genes simply because for low-expressors, the division of a small number by another small number can result in a non-significantly large number. Array2BIO utilizes local mean normalization and local variance correction across intensities to reliably handle low- and high-expressors and define differential fold-difference thresholds for different level intensities. In Array2BIO, the approach is very similar to the previously described SNOMAD method (8). Briefly, fold-expressions of every Affymetrix tag are ordered by the average expression of signal

and control tags (the latter provides an average level for the tag expression). Then the tags are divided into 100 groups by their average expression levels and a distribution of fold-expressions is calculated for each group. Z-value (based on the average) and standard deviation of fold-expressions in the group is assigned to each tag. Tags with Z-values greater than 2 are selected for further analysis (Figure 3).

Welch's t-test of differential expression significance. Signal and control tag expressions that survive the balance analysis of low- and high-expressors are then subjected to statistical testing using the Welch's t-test method. Statistical testing is performed on the average signal and contol tag expression using standard deviations of their probe expression distributions. A p-value is assigned to every differentially expressed tag and the tags with p-values of less than 0.05 are selected for multiple testing correction analyses.

Mapping Affymetrix tags onto UCSC known genes. Array2BIO first identifies a set of unique (non-overlapping) genes in a genome matching the original *.CEL* files by using the 'known genes' annotation provided by the UCSC Genome Browser database. Then the Affymetrix tags are mapped onto (and are grouped by) corresponding 'known genes'. Accession numbers of corresponding mRNA sequences and their genomic locations are retrieved for each gene during the mapping process. These are then used to dynamically link genes to the NCBI database and to the ECR Browser.

Gene Ontology (GO) and KEGG analyses of biological functions and gene interactions. Array2BIO utilizes a locally installed version of the GO and KEGG databases to contrast the distribution of differentially expressed functional categories of

genes to the genome average population of these categories. Observed and expected category population values are compared and the statistical 'enrichment' (or 'depletion') of a category is quantified by using hypergeometric distribution statistics. Functional categories with p-values smaller than 0.05 are selected for subsequent multiple testing correction analyses. The Gene Ontology database provides biological classification of gene function represented by membership to a functional category that relates to a particular biological process, to a molecular function, or to a cellular component. The KEGG database combines information on gene interactions that are grouped into (1) metabolism, (2) genetic information processing, (3) environmental information processing, (4) cellular processes, and (5) human diseases categories.

Correction for multiple testing. Array2BIO performs a correction for multiple testing to exclude false positive predictions associated with the fact that the statistical testing of differential tag expression or enrichment/depletion in GO and KEGG categories is performed multiple times. Array2BIO provides two statistical methods to correct for multiple testing and also allows omitting multiple testing if the user does not want to apply this function. The default method used by Array2BIO is the medium stringency Benjamini-Hochberg correction. Alternatively, the Bonferroni correction method can be applied. The latter is one of the most stringent multiple testing correction methods and can be used to select for the most outstanding overexpressor genes or enriched/depleted functional categories.

Clustering analysis.

Microarray data clustering. Array2BIO utilizes the Unix version of the Cluster tool (9). The hierarchical cluster analysis is implemented into Array2BIO, which allows clustering of genes and/or conditions; provides 9 distance measures and 4 methods. Due to Cluster limitations, Array2BIO restricts the maximum number of clustered transcripts to under 2500 genes. The genes are ranked by their deviation in expression across different conditions and those with the largest deviation are selected for clustering.

Interactive tree visualization. Array2BIO provides an interactive web utility for the visualization of clustering results. Clustered gene expression across multiple conditions is visualized in a matrix format. The tree of clustering relationships is given to the left of the gene expression image (Figure 4A). Mouse click on a tree branch generates a 'zoom in' image of the branch and gives a detailed description of related genes (including gene names, accession numbers, corresponding Affymetix tags, and genomic locations) (Figure 4B).

Interconnection with external tools.

ECR Browser - evolutionary conservation analysis. ECR Browser is a dynamic whole-genome navigation tool for visualizing and studying evolutionary relationships between genomes. Evolutionary Conserved Regions (ECRs) are mapped to alignments and are graphically visualized in relation to known genes that have been annotated in the reference genome.

Creme 2.0 - identification of clusters of transcription factor binding sites in promoters. Crème 2.0 relies on a database of putative transcription factor binding sites that have been carefully annotated across the human genome using evolutionary

conservation with the mouse and rat genomes. An efficient search algorithm is applied to this data set to identify combinations of transcription factors, whose binding sites tend to co-occur in close proximity to the promoter regions of the input gene set. These combinations are statistically evaluated, and significant combinations are reported and visualized.

NCBI - detailed sequence information. Detailed mRNA transcript information including: nucleotide and protein sequences, related publications, gene annotation, etc. are provided through the dynamic interconnection to the NCBI database.

SUMMARY

In summary, Array2BIO is a significant addition to the Dcode.org collection of tools (10) that permits the efficient and unique integration of Dcode.org utilities for multi-functional analysis of gene expression data and perhaps more importantly, Array2BIO represents the first web-based tool/utility for sophisticated analysis of microarray expression data. A "single-click" implementation of the variety of biological characterizations into a single tool permits the standardized, prompt identification of co-expressed genes, their functional annotation, the identification of related interaction pathways, and prediction of key transcription factors underlying observed gene expression responses.

This work was performed under the auspices of the U. S. Department of Energy by University of California, Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

Table 1. Five of the most overpopulated GO biological processes and molecular functions that corresponding to *Yersinia pestis* infection.

Biological processes	***Enrichment***	***p-value***	***Observed / expected***	***Total category count***
response to biotic stimulus	4.16	1.85e-14	37 / 8.9	873
response to stress	4.05	6.37e-13	34 / 8.4	824
immune response	4.36	3.33e-12	30 / 6.9	676
response to pest, pathogen or parasite	5.30	4.48e-12	25 / 4.7	463
defense response	3.99	1.27e-11	31 / 7.8	762

Molecular functions	***Enrichment***	***p-value***	***Observed / expected***	***Total category count***
chemokine activity	17.07	1.72e-8	8 / 0.5	46
chemokine receptor binding	17.07	1.72e-8	8 / 0.5	46
G-protein-coupled receptor binding	15.40	4.01e-8	8 / 0.5	51
cytokine activity	5.27	7.35e-6	11 / 2.1	205
transcription factor activity	2.48	2.92e-4	18 / 7.3	712

FIGURE LEGENDS

Figure 1. Schematic flowchart of the Array2BIO analysis.

Figure 2. Snapshot of cytokine-cytokine receptor interactions related to the *Y. pestis* infection.

Figure 3. SNOMAD local Z-test for handling low-expressors. Signal versus control fold different in expression is plotted against the median signal and control expression. Orange dots represent the selection of over- and under-expressors.

Figure 4. Clustering analysis visualization. A full clustering tree across 5 control (cN) and 5 signal (sN) conditions (A) and a zoom in into two genes (B) selected by a mouse click on the full clustering tree as depicted by the orange rectangle.

Figure 1.

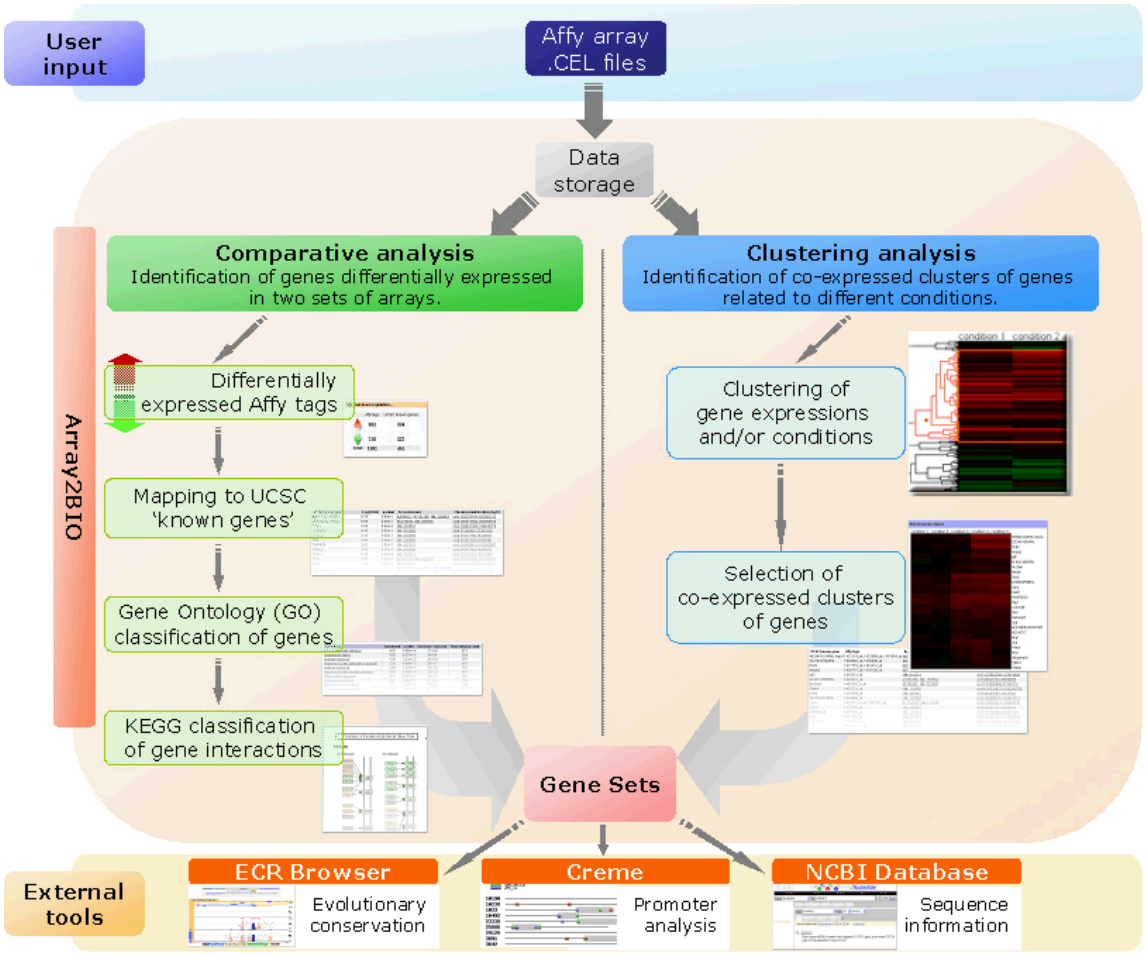

User input
Affy array .CEL files
Data storage
Array2BIO
Comparative analysis
Identification of genes differentially expressed in two sets of arrays.
Clustering analysis
Identification of co-expressed clusters of genes related to different conditions.
Differentially expressed Affy tags
Mapping to UCSC 'known genes'
Gene Ontology (GO) classification of genes
KEGG classification of gene interactions
Clustering of gene expressions and/or conditions
Selection of co-expressed clusters of genes
Gene Sets
External tools
ECR Browser
Evolutionary conservation
Creme
Promoter analysis
NCBI Database
Sequence information

Figure 2.

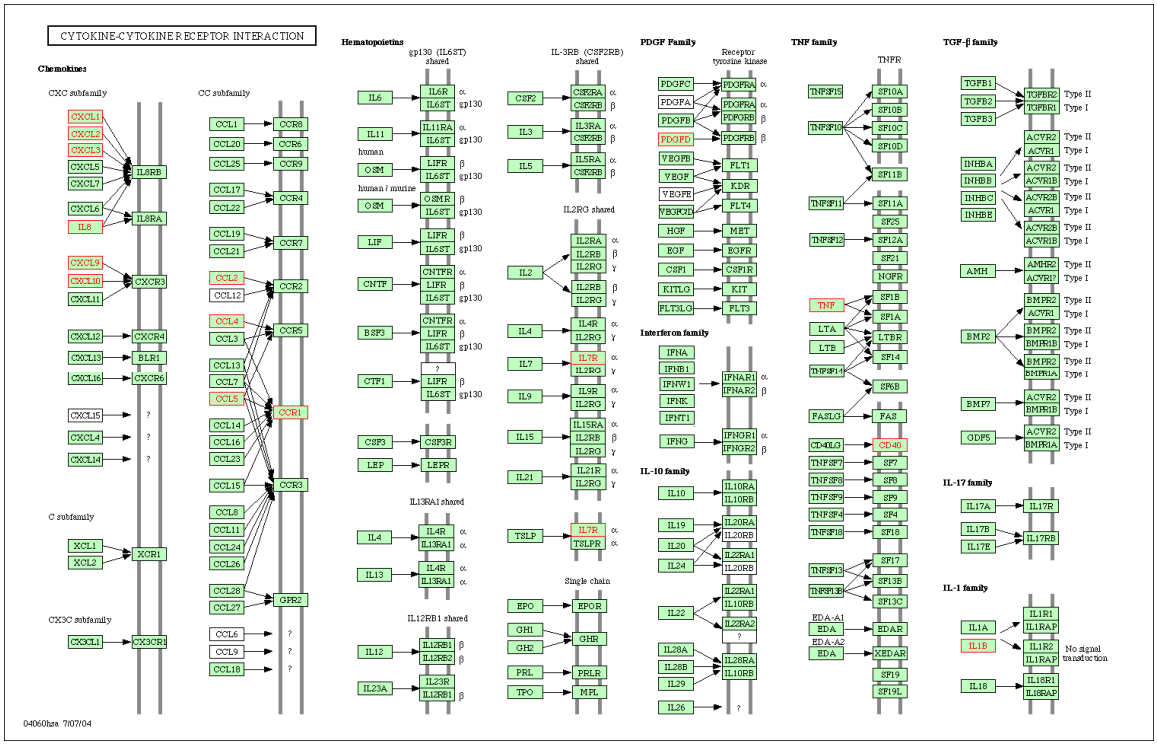

CYTOKINE-CYTOKINE RECEPTOR INTERACTION
Chemokines
CXC subfamily
CC subfamily
C subfamily
CX3C subfamily
Hematopoietins
gp130 (IL6ST) shared
IL13RA1 shared
IL12RB1 shared
IL-3RB (CSF2RB) shared
IL2RG shared
Single chain
PDGF Family
Receptor tyrosine kinase
Interferon family
IL-10 family
TNF family
TNFR
TGF-β family
IL-17 family
IL-1 family
04060hsa 7/07/04

FIGURE 3.

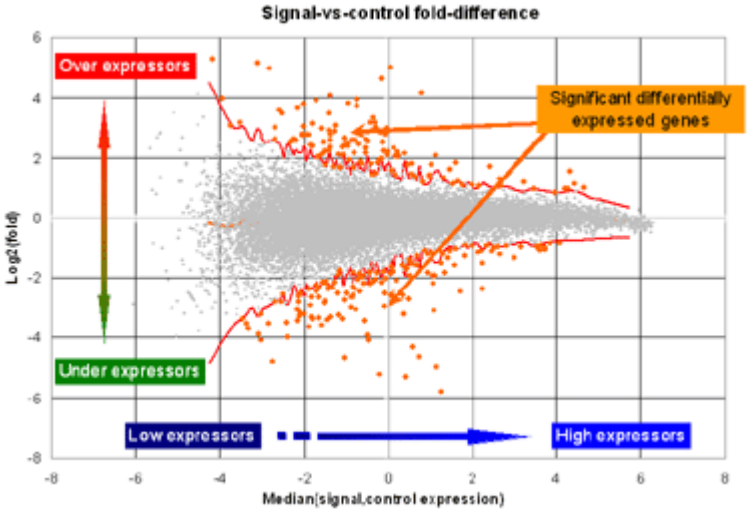
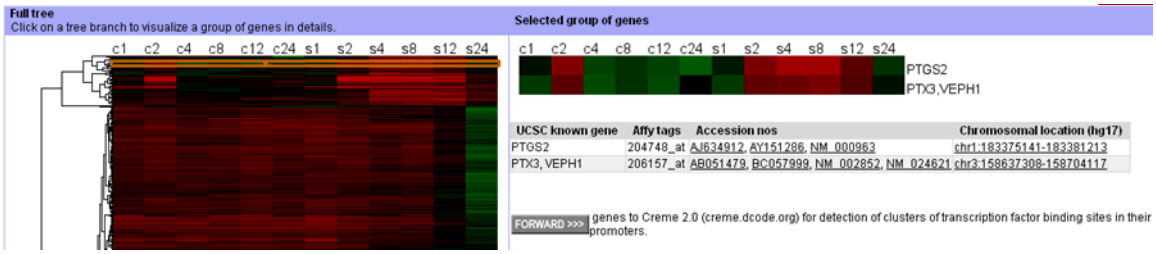

Signal-vs-control fold-difference
Over expressors
Significant differentially expressed genes
Under expressors
Low expressors
High expressors
Log2(fold)
Median(signal,control expression)
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6
8

FIGURE 4.

A) **B)**


Full tree
Click on a tree branch to visualize a group of genes in details.
c1 c2 c4 c8 c12 c24 s1 s2 s4 s8 s12 s24
Selected group of genes
c1 c2 c4 c8 c12 c24 s1 s2 s4 s8 s12 s24
PTGS2
PTX3,VEPH1
UCSC known gene
Affy tags
Accession nos
Chromosomal location (hg17)
PTGS2
204748_at
AJ634912, AY151286, NM_000963
chr1:183375141-183381213
PTX3, VEPH1
206157_at
AB051479, BC057999, NM_002852, NM_024621
chr3:158637308-158704117
FORWARD >>>
genes to Creme 2.0 (creme.dcode.org) for detection of clusters of transcription factor binding sites in their promoters.

REFERENCES

1. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*, 31, 51-54.
2. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32, D258-261.
3. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 27, 29-34.
4. Ovcharenko, I., Nobrega, M.A., Loots, G.G. and Stubbs, L. (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res*, 32, W280-286.
5. Sharan, R., Ben-Hur, A., Loots, G.G. and Ovcharenko, I. (2004) CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res*, 32, W253-256.
6. Hirayama, T., Dai, S., Abbas, S., Yamanaka, Y. and Abu-Amer, Y. (2005) Inhibition of inflammatory bone erosion by constitutively active STAT-6 through blockade of JNK and NF-kappaB activation. *Arthritis Rheum*, 52, 2719-2729.
7. O'Shea, J.J., Park, H., Pesu, M., Borie, D. and Changelian, P. (2005) New strategies for immunosuppression: interfering with cytokines by targeting the Jak/Stat pathway. *Curr Opin Rheumatol*, 17, 305-311.
8. Colantuoni, C., Henry, G., Zeger, S. and Pevsner, J. (2002) SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis. *Bioinformatics*, 18, 1540-1541.
9. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95, 14863-14868.
10. Loots, G.G. and Ovcharenko, I. (2005) Dcode.org anthology of comparative genomic tools. *Nucleic Acids Res*, 33, W56-64.