GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways

DNA microarrays are used to measure simultaneously the expression levels of thousands of genes. New tools are needed to relate the large amounts of microarray data generated to known models of cell biology and biochemistry. We have developed a free stand-alone computer program called GenMAPP (Gene Microarray Pathway Profiler), designed for viewing and analyzing gene expression data in the context of biological pathways. GenMAPP displays gene expression data on pathways by color-coding the genes based on data and criteria provided by the investigator. GenMAPP also has graphics tools for constructing and modifying pathways. In addition, it provides access to annotation for genes and a connection with pathway experts. This program thus provides a new format to organize, publish and share DNA microarray data that will increase our understanding of biology in the genomics era.

Current tools for analyzing microarray data include statistical filters and pattern-finding algorithms such as hierarchical clustering[1,2]. The power of such tools is that they evaluate the gene expression data without the bias of prior knowledge of gene function, providing a valuable first step in analysis by generating lists of interesting genes. The next logical step is to analyze the gene expression changes in the context of known biological pathways[3]. Several public and commercial pathway resources currently exist, including the Alliance for Cellular Signaling, BioCarta, EcoCyc[4] and MetaCyc[5], the Kyoto Encyclopedia of Genes and Genomes[6] (KEGG) and PathDB. These databases contain large amounts of curated information, and some (EcoCyc, MetaCyc and KEGG) can also be used to view simple gene expression data sets in the context of preexisting pathways[7–9]. GenMAPP extends the capabilities of these pathway resources by allowing users to modify pathways for their own use, to design new pathways and to apply complex criteria for viewing gene expression data on those pathways. GenMAPP also complements the current pathway databases by providing a means of freely exchanging pathway-related data among investigators.

GenMAPP represents biological pathways in a special file format called ‘MAPPs’. MAPPs are independent of the gene expression data and can be used to group genes by any organizing principle, for example: metabolic pathways (such as fatty-acid degradation; see Fig. 1 and URLs A–C), signal transduction cascades (see URL D), gene families (see URL E), subcellular components (see Web Fig. F online) or custom MAPPs for hypothesis testing (including the most interesting genes in a microarray experiment). Investigators can construct custom MAPPs with the graphics tools provided by the program, assigning each gene an identification (ID) from GenBank, SWISS-PROT[10] or a user-defined ID system. The gene ID is the link between the gene object on the MAPP, the gene expression data and the annotation for that gene contained in an underlying GenMAPP database. These annotations, the data and the hyperlinks to the public databases can be accessed by simply clicking on each gene.

In addition to constructing custom MAPPs, investigators can download existing MAPP files from the growing archive at our website. Archived MAPPs derive from two sources. First, MAPPs for a few sample pathways have been drawn based on textbooks, review articles and public pathway databases. At the time of this writing, this category includes approximately 50 MAPPs each for mouse and human (see Fig. 1 and URLs A–D). Second, MAPPs containing lists of functionally related genes have been generated from the public database maintained by the Gene Ontology Project[11,12]. At the time of this writing, this category includes 958 MAPPs for mouse (see URLs E–G), 1,856 MAPPs for human, 339 MAPPs for rat and 631 MAPPs for yeast. This starter set of pathways is in no way comprehensive and will require the contributions of additional pathways from GenMAPP users in their fields of expertise. MAPP files are small enough to be shared easily with colleagues by e-mail and can also be exported to HTML for display on websites. Moreover, contact information for the author is provided right on each MAPP. Thus, Gen-

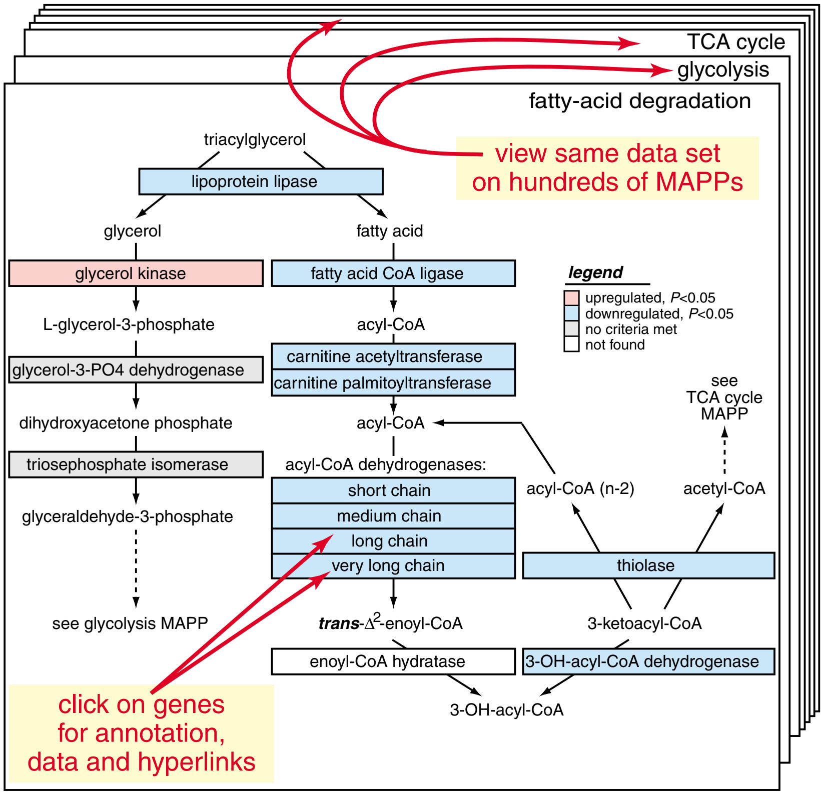

Fig. 1 Fatty-acid degradation pathway adapted from a view in GenMAPP. Each box represents a gene. For the interactive output of this MAPP and other MAPPs from GenMAPP, see URLs A–G. This MAPP is color-coded with gene expression data from a mouse model of dilated cardiomyopathy[14]. Red indicates a fold change greater than 1.2 in the disease group (nine mice) versus a control group (eight mice) with a significant *P* value of less than 0.05. Blue indicates a fold change of less than –1.2 and *P*<0.05. Gray indicates that neither of the above criteria were met. White indicates that the gene is not found on the array. TCA, tricarboxylic acid.

MAPP can promote the efficient exchange of data and ideas among investigators.

To foster this exchange, we have formed GenMAPP.org, a nonprofit, academically based organization that develops and supports GenMAPP. To assure the growth in the numbers of MAPPs freely available to the community, GenMAPP.org will provide a distribution center for MAPPs at our website and will support the establishment of other MAPP distribution sites. As the collection of MAPPs continues to grow, we are creating navigation and search tools that will allow users to easily find the MAPPs of greatest interest. In addition, the GenMAPP database will be expanded to include more gene cataloging systems and additional species. Currently, human, mouse, rat and budding yeast[13] are supported.

To view gene expression data on MAPPs, the user imports the data in a comma-separated values format. GenMAPP has the flexibility to accept numeric and character data types, calculated values (such as *P* values), data from several experiments and data from both custom-spotted and commercial microarrays. GenMAPP converts the expression data into a data set that can then be viewed on any MAPP with any number of color-coding criteria sets. GenMAPP comes with a sample data set, an extensive help file and an online tutorial that shows how to import data and establish color-coding criteria. The sample gene expression data set is from a mouse model of dilated cardiomyopathy that has many statistically significant changes in gene expression[14]. Hierarchical clustering suggested that the expression of genes encoding enzymes involved in fatty-acid degradation was coordinately downregulated. When we viewed these data using GenMAPP, we saw the changes in gene expression in a biological context not attainable by viewing one gene at a time (Fig. 1). Thus, GenMAPP complements and extends the statistical and clustering algorithms already available.

The GenMAPP program is a powerful tool for graphically viewing microarray data in an intuitive way familiar to biologists—in the context of pathways. Moreover, GenMAPP connects investigators who carry out DNA microarray experiments with those interested in organizing genomic information into interpretive maps. It is our hope that GenMAPP will contribute to the next task in genomics: describing all of the pathways that make up a cell and how they interact as a system in the overall physiology of an organism.

URLs. For supplementary figures (URLs) A–G, see http://www.genmapp.org/Supplementary/Web%20Figure%20A.htm (for figures B–G, substitute the appropriate letter for 'A'). The GenMAPP program (Windows) and MAPP files are available free-of-charge from http://www.GenMAPP.org; Gene Ontology MAPP files were derived from http://www.geneontology.org.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments

We thank S. Doniger, N. Fidelman, G. Howard, R. Mahley, S. Ordway, C. Redfern, M. Schambelan, B. Taylor, J. Warrington, K. Weisgraber, S. Young and M. Zeiger for valuable discussions and their contributions to the preparation of this manuscript. Support for this work included funding from the J. David Gladstone Institutes, the San Francisco General Hospital General Clinical Research Center, the National Heart, Lung, and Blood Institute and the NHLBI Programs for Genomic Applications (BayGenomics).

Competing interests statement

The authors declare that they have no competing financial interests.

Kam D. Dahlquist[1,2], Nathan Salomonis[1], Karen Vranizan[1], Steven C. Lawlor[1] & Bruce R. Conklin[1–3]

[1]Gladstone Institute of Cardiovascular Disease, [2]Cardiovascular Research Institute and [3]Departments of Medicine and Molecular and Cellular Pharmacology, University of California, San Francisco, California 94141-9100, USA. Correspondence should be addressed to B.R.C. (e-mail: bconklin@gladstone.ucsf.edu).

1. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
2. Tomayo, P. *et al. Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
3. DeRisi, J.L., Iyer, V.R. & Brown, P.O. *Science* **278**, 680–686 (1997).
4. Karp, P.D. *et al. Nucleic Acids Res.* **30**, 56–58 (2002).
5. Karp, P.D. *et al. Nucleic Acids Res.* **30**, 59–61 (2002).
6. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. *Nucleic Acids Res.* **30**, 42–46 (2002).
7. Karp, P.D., Krummenacker, M., Paley, S. & Wagg, J. *Trends Biotechnol.* **17**, 275–281 (1999).
8. Karp, P.D. *Science* **293**, 2040–2044 (2001).
9. Nakao, M. *et al. Genome Inform. Ser Workshop Genome Inform.* **10**, 94–103 (1999).
10. Bairoch, A. & Apweiler, R. *Nucleic Acids Res.* **28**, 45–48 (2000).
11. Ashburner, M. *et al. Nature Genet.* **25**, 25–29 (2000).
12. Ashburner, M. *et al. Genome Res.* **11**, 1425–1433 (2001).
13. Dwight, S.S. *et al. Nucleic Acids Res.* **30**, 69–72 (2002).
14. Redfern, C.H. *et al. Proc. Natl Acad. Sci. USA* **97**, 4826–4831 (2000).