

Research news

## Co-regulation of mouse genes predicts function

Jonathan B Weitzman

Published: 6 December 2004

*Journal of Biology* 2004, **3**:19

The electronic version of this article is the complete one and can be found online at <http://jbiol.com/content/3/5/19>

© 2004 BioMed Central Ltd

### Large-scale microarray analyses reveal that transcriptional co-regulation patterns can be remarkably helpful in predicting the function of novel mouse genes.

Every eukaryotic genome-sequencing project to date has revealed the presence of thousands of novel predicted genes. Researchers interested in functional genomics now face some formidable challenges: defining how many unknown genes are yet to be discovered and working out what they do. Now, in *Journal of Biology* [1], Timothy Hughes and colleagues show that techniques that were first applied to yeast can be used to predict gene function in mice (see 'The bottom line' box for a summary of the work).

Hughes became something of a microarray aficionado during his postdoc at Rosetta Inpharmatics, LLC in Seattle, USA. He and his colleagues there demonstrated that a careful combination of genome-wide microarray analysis of gene expression patterns and sophisticated statistical methods could be used to predict gene function. Specifically, they showed that patterns of transcriptional co-regulation could effectively predict the biological function of novel genes [2]. But those impressive studies were performed in a unicellular yeast, which has around 6,000 genes in total. It wasn't clear how well the approach would fare with larger mammalian genomes and the

complexity of multicellular organisms. When Hughes moved to the University of Toronto, Canada, he was eager to give it a try. Mark Gerstein of Yale University says that the Hughes study has tackled an important problem in functional genomics: "That is, translating ideas that were found applicable in

simple unicellular organisms to more complicated mammalian systems."

#### A mountain of microarray data

Hughes' first concern was which genes to spot onto his microarray slides (see the 'Background' box). Researchers are

#### The bottom line

- Genome-wide studies of gene expression in yeast, using microarrays, showed that patterns of transcriptional co-regulation can predict the biological function of novel genes.
- Microarrays have also been used to analyze the expression of 40,000 known or predicted mouse mRNAs across a range of 55 tissues.
- Sophisticated machine-learning algorithms (support vector machines) can assign genes to transcriptional co-regulation groups, and these can be matched to predicted functional categories, using Gene Ontology, to predict gene function.
- The results challenge the conventional wisdom that tissue-specific expression is indicative of gene function in mammals.
- The enormous gene-expression dataset generated during the study will be an important open resource for future functional studies in mice.

## Background

- High-density **microarrays** (often referred to as 'DNA chips') are powerful tools for analyzing the expression profiles of all transcripts under multiple conditions. Microarrays contain thousands of spots of either cDNA fragments corresponding to each gene or short synthetic oligonucleotide sequences. By hybridizing labeled mRNA or cDNA from a sample to the microarray, transcripts from all expressed genes can be assayed simultaneously; one microarray experiment can give as much information as thousands of northern blots.
- The National Center for Biotechnology Information (**NCBI**) has created many resources for genome **annotation**, the process of identifying all genes and ascribing functions to the proteins that they encode. The **XM sequence database** contains about 40,000 known and predicted mRNA sequences generated automatically by an *ab initio* gene-identification computer algorithm.
- NCBI's **RefSeq** project provides a curated database of non-redundant DNA, RNA and protein sequences for major model organisms. RefSeq sequences are substantially based on sequence records from GenBank, which in turn comprises original data from gene- and genome-sequencing projects. An independent database is maintained by the **RIKEN** Institute in Japan and contains the sequences of over 60,000 full-insert mouse cDNAs. A third source of annotated sequences is the Wellcome Trust's **Ensembl** project, which automatically annotates metazoan genome sequences.
- A **support vector machine (SVM)** is a supervised learning algorithm (or computer program). The algorithm addresses the general problem of learning to discriminate between positive and negative members of a given class of n-dimensional vectors. The SVM works by mapping a given training set into a multi-dimensional space and attempting to locate in that space a plane that distinguishes between different groups.
- The **Gene Ontology** is a controlled vocabulary consisting of three structured networks of defined terms that are used to describe the attributes of gene products in terms of Molecular Function, Biological Process and Cellular Component.

still undecided about how many genes make up a mouse. "There is no 'gold standard' cDNA database for mouse genes," explains Hughes. His team chose to start with a single source, the **XM sequences** from NCBI (see Table 1 for a list of the resources mentioned in this article). "We downloaded the XM collection from the NCBI. It's almost

certainly not perfect, as it's all done using draft genome sequence, but it seems to contain a large majority of the known genes and a bunch of predicted genes, many of which were detectable on the arrays," says Hughes. "The collection contains about 75% of the current **RefSeq** sequences, it contains the majority of **Ensembl** genes,

but it's missing a lot of the **RIKEN** clones." The team then made a single 60-residue oligonucleotide for each of the potential genes.

The Hughes team next got hold of as many different sources of mouse mRNA as they could and hybridized them to the microarrays carrying over 40,000 spots. They found that 21,622 transcripts were expressed in at least one of the 55 tissues examined. "We didn't really expect everything to be expressed," comments Hughes (see the 'Behind the scenes' box for more of the rationale for the work). "We mostly looked at adult tissues and we tried not to look at stress responses." He notes, however, that the latest estimates for the number of mouse genes are somewhere around the 20,000 to 25,000 mark.

Mining the resulting data mountain required a sophisticated bioinformatic approach. "You have to know what you are looking for and be able to formulate questions mathematically and execute them on a computer," notes Hughes. Hughes teamed up with computational colleagues in Brendan Frey's team and applied some fancy statistical tricks, such as 'variance stabilizing normalization', to allow comparison across the tissues, and implemented a learning algorithm called a **support vector machine (SVM)** [3]. "If you have a bunch of points in two- or three-dimensional space, an SVM looks for ways to distinguish between the ones that have a given feature and the ones that don't. No one had used SVMs before on this scale. If we have 55 tissues, then we are looking at 21,000 objects in a 55-dimensional space and trying to separate the ones that have a function from those that don't."

The statistical analysis revealed that quantitative co-expression could identify groups of genes with related functions; the functions were determined as similar because **annotation** designated the genes as belonging to the same functional category within the **Gene Ontology** (see Figure 1). In

fact, the SVM method was so effective that it could be used to predict functions for hundreds of genes of unknown function; indeed, the SVM was a much better predictor of gene

function than were the simple tissue-specific gene-expression patterns.

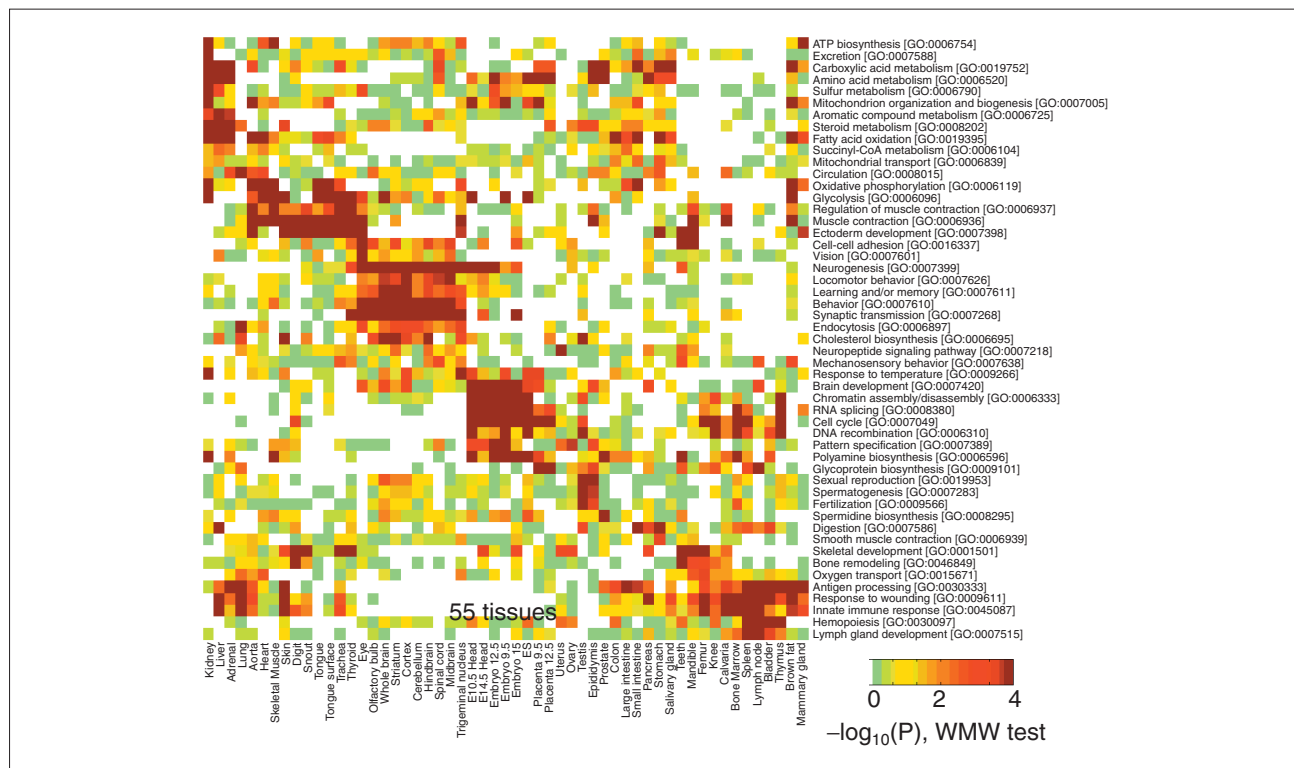
The Canadian group is not the first to carry out such large-scale analyses of mammalian gene expression [4-6].

“But what I like about this paper is that it’s really rock solid,” says Stuart Kim of the Stanford University Medical Center, USA. “This is really believable stuff. It is really well grounded in the

**Table 1**

**The online genome-annotation and gene-listing resources described in this article**

Resource	URL	Contents
NCBI XM sequences (from the non-redundant (NR) database)	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein	Non-redundant protein entries from a variety of sources, including translations from annotated coding regions in GenBank and RefSeq
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/	A comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major model organisms
Ensembl	http://www.ensembl.org/	Annotated metazoan genomes
RIKEN FANTOM cDNA database	http://fantom.gsc.riken.jp/	Functional annotation of mouse full-length cDNA clones
Gene Ontology	http://www.geneontology.org	Genes annotated according to three structured networks of defined terms



**Figure 1**  
Correspondence between gene expression patterns and GO annotations. Significance values resulting from applying a statistical test to each correlation of a Gene Ontology functional category with expression in the indicated tissues shown with colors. See [1] for further details.

## Behind the scenes

*Journal of Biology* asked Timothy Hughes about the background and outlook for his ambitious project to map the functional landscape of the mouse genome.

### What motivated you to embark on the mouse microarray project?

My group has mostly worked on yeast in the past. We had a lot of success using microarrays to look at how gene expression can be used to predict gene functions and to find transcriptional regulatory pathways. So, when the mouse genome came out we thought that this was a reasonable thing to try, assuming that if it worked in mouse it would probably work in humans. There was the added bonus that we could use the microarrays to validate the expression of predicted genes and contribute to the big goal of finding all mammalian genes.

### How long did the experiments take and what were the steps that ensured success?

It took about two years: one to get the data and another year to do the analysis. We did several things differently from other groups looking at expression in different tissues: first, we think it was a good choice to use NCBI's XM gene collection. Then, we tested all the tissues from labs in the Toronto area and I hired a medical student, Richard Chang, to dissect mice over the summer to obtain the tissues we were still missing. Our bioinformatics collaborators, Brendan Frey and his postdoc Quaid Morris, were also indispensable; without them our paper would look an awful lot like many other microarray papers.

### What was your initial reaction to the results and how were they received by others?

We were happy to see a good correspondence between gene expression patterns and functional categories. It's not trivial to figure out how all the genes are regulated or how to use that data to figure out their functions, but it's worth doing. I think that it's helpful to look at the co-regulation patterns in an arbitrary sense rather than getting hung up on exactly what tissue a pattern corresponds to. That's the aspect that people are most surprised about, and some members of the mammalian research community are skeptical about whether it's right. But we saw the same thing in yeast, that genes are co-regulated in functional groups.

### What are the next steps?

We are collaborating with local labs that do gene-trap mutagenesis and make knockout mice. We plan to test several dozen of our functional predictions, but these experiments literally take years. An important point here is that showing that it works once or twice from a biological standpoint is actually not as rigorous, from a statistics standpoint, as doing the full cross-validation test which we did in the paper. Also, we will probably work on computational approaches to find possible *cis*-regulatory sites.

statistics, avoiding simplistic non-mathematical concepts like 'on and off' or 'two-fold up and two-fold down'. They did fairly sophisticated statistical analyses to make sure that the trends they were seeing were really valid. It's important to get better and better datasets published." John Hogenesch of Novartis Research Foundation Genomics Institute in San Diego, California, notes that "[Hughes'] application of SVMs and Gene Ontology to provide preliminary functional annotation for thousands of genes of unknown function is a major advance." The Hogenesch group is also creating an atlas of mammalian genes [5]. "This approach had been used in yeast and worms, but it hadn't yet been applied to mammalian gene expression. Hughes' paper now provides testable hypotheses for the roles of thousands of genes in the genome."

### An open resource at the click of a mouse

Hughes' analysis revealed that the results from the extensive mouse tissue-specific dataset correlates very well with the results of studies from other laboratories. One notable feature of the Hughes dataset is that it has been made openly accessible to the research community [1,7]. The additional data with the published article, and the Hughes lab website, provide information about the microarray oligonucleotide sequences, the SVM predictions, gene annotation, and so on, all of which can be downloaded without restriction and free of charge.

Kim points out that this is really important. "I think that every person that works on mice should now go to this study and type in the name of their favorite gene(s) and see where it is expressed in 55 tissues. It will cost nothing and then you will know where it is expressed strongly. You can make sure there are no hidden surprises [in your experiments] or find out what the hidden surprises are." Hogenesch concurs: "Most users will use the

database to see where their gene of interest is expressed and what pathway it might participate in. Others will use the dataset itself to ask questions using other methodologies (tissue-specific gene expression, regulatory-element analysis, functional classification, and so on). The types of things you can do with a dataset like this are numerous, which is why it's important that the data are available."

Kim's group is building large genetic networks based on microarray datasets [8]. "We use more than just tissue specificity to build our networks – we use everything that we can grab. So, we will go and grab these data and fold them into ours. Our next paper will include 1,700 mouse microarrays folded into the human-yeast-fly-worm networks. In worms, many labs have used our resource and published some pretty awesome papers based on the genetic network." Kim thinks that the networks will be even more powerful in accelerating the pace of research in mammalian systems, where classical experimental approaches are slow and expensive. Mark Gerstein agrees: "This is an important advance in helping to unravel the functions of the tens of thousands of human genes using functional genomics approaches."

Hughes has enjoyed the transition from studying yeast to working on mice, and is eager to collaborate with mouse geneticists to test some of the predictions that come out of the current study. And he wants to understand more about the correlation between co-regulation patterns and gene function. "As a yeast researcher the thing that blows my mind is how many things animal cells do. I learned a lot just looking at all the functional categories and Gene Ontology," admits Hughes. "The correlation between transcriptional co-regulation and function is very strong. It's much, much higher than you would get if genes were just expressed at random. But it's not absolute either. So, annotating function is a hard problem to crack and that gives us plenty to work on."

## References

1. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, Laurin N, Eftekharpour E, Sat E, Grigull J, Pan Q, Peng WT, Krogan N, Greenblatt J, Fehlings M, van der Kooy D, Aubin J, Bruneau BG, Rossant J, Blencowe BJ, Frey BJ, Hughes TR: **The functional landscape of mouse gene expression.** *J Biol* 2004, **3**:21.
2. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ:

**Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31**:255-265.

3. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97**:262-267.
4. Bono H, Yagi K, Kasukawa T, Nikaido I, Tominaga N, Miki R, Mizuno Y, Tomaru Y, Goto H, Nitanda H, et al.: **Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays.** *Genome Res* 2003, **13**:1318-1323.
5. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al.: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101**:6062-6067.
6. Schadt EE, Edwards SW, GuhaThakurta D, Holder D, Ying L, Svetnik V, Leonardson A, Hart KW, Russell A, Li G, et al.: **A comprehensive transcript index of the human genome generated using microarrays and computational approaches.** *Genome Biol* 2004, **5**:R73.
7. **The functional landscape of mouse gene expression** [<http://hugheslab.med.utoronto.ca/Zhang>]
8. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules.** *Science* 2003, **302**:249-255.

Jonathan B Weitzman is a scientist and science writer based in Paris, France.

E-mail: [jonathanweitzman@hotmail.com](mailto:jonathanweitzman@hotmail.com)