J. Biol. Journal of Biology

BioMed Central

Research news

Transcriptional territories in the genome

Jonathan B Weitzman

Published: 25 June 2002

Journal of Biology 2002, **1**:2

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/1/1/2

An analysis of numerous *Drosophila* microarray experiments reveals that the genome has many large groups of adjacent genes that are expressed similarly but are not functionally related.

Newly completed genome sequences are emerging at ever-faster rates, and **microarrays** (‘**DNA chips**’) are now routine tools for exploring genome-wide changes in mRNA levels (see the ‘Background’ box). Most journals are bursting at the seams with genome maps and brightly colored gene-**expression profiling** data. But few studies have sought to explore the relationship between the organization of the genome and the **transcriptome**.

In the maiden publication of the *Journal of Biology*, Paul Spellman and Gerald Rubin, of the Howard Hughes Medical Institute and the University of California, Berkeley, describe how they analyzed microarray data from 88 different experimental conditions according to the chromosomal location of each gene within the *Drosophila* genome [1] (see ‘The bottom line’ box for a summary of their work). They come up with the remarkable observation that the genome contains many large groups of adjacent genes that are expressed similarly but are not functionally related to one another. These results challenge the way we think about the mechanisms of gene regulation and the influence of local ‘territories’ within chromosomes.

A dynamic duo

Spellman and Rubin form an ideal partnership for this ‘flies and chips’ project. Rubin is one of the leading scientists in the *Drosophila* field. Using carefully crafted genetic screens, his laboratory has helped to make the fruit fly *Drosophila melanogaster* the star model of developmental genetics. Rubin also leads the Berkeley *Drosophila* Genome Project, a unique collaboration between academic laboratories and the industrial sequencing giant

The bottom line

- Spellman and Rubin took hundreds of microarray profiles acquired under 88 experimental conditions and mapped the profile for each gene to the gene’s position along the *Drosophila* chromosomes.
- They found that the *Drosophila* genome contains over 200 groups of adjacent genes that are expressed together.
- Each of these groups contains 10-30 genes that are not related to one another in sequence or function, and each group spans hundreds of kilobases.
- Spellman and Rubin propose that local changes in chromatin structure might define chromosomal domains that in turn control the expression of large groups of genes; perhaps the regulation of large groups reflects an ‘open’ or ‘closed‘ chromatin state around a gene whose expression it is important to turn ‘on’ or ‘off’.
- Analyzing other genomes, both for the presence of similar gene groups and for conserved ordering of grouped genes, will help in assessing the functional importance of co-regulated gene domains.

Celera Genomics that generated the first whole-genome shotgun sequence of a eukaryote genome in record time [2,3].

Spellman is a graduate of Patrick Brown's laboratory at Stanford University, where he mastered chip technology and helped develop computational algorithms for analyzing microarray results. Pioneering work from the Brown and Botstein laboratories exploited their home-made chips to demonstrate the power of microarray technology, analyzing genome-wide changes in gene-expression levels in the yeast *Saccharomyces cerevisiae* under different experimental conditions [4]. Spellman is now a postdoctoral researcher in Rubin's laboratory where he is chipping away at the fly transcriptome.

Microarray manipulation

The Rubin laboratory has been using a DNA chip called the GeneChip Drosophila Genome Array, containing nearly 200,000 spots that represent the approximately 13,500 predicted fly genes; the chip was created by Affymetrix Inc., a leading manufacturer of high-density oligonucleotide microarrays. Spellman and Rubin pooled microarray data from 88 different experiments (mostly unpublished) using *Drosophila*, corresponding to 267 separate hybridizations.

Many of those who have plunged into the world of transcriptome analysis have found that the biggest challenge lies in picking out the jewels from the mountain of hybridization results. The authors of early microarray papers contented themselves with providing a list of genes whose expression increased ('turned on') or decreased ('turned off') under certain conditions. The underlying assumption was that changes in the expression of an individual gene are of biological relevance.

In 1998 Spellman co-authored a paper with Michael Eisen, then in David Botstein's laboratory at Stanford, describing a method for analyzing microarray data by 'hierarchically clustering' genes with similar expression profiles [5]. Today, authors of most microarray papers apply bioinformatics tools such as the 'Eisen clustering algorithms' and associated **TreeView** software to make sense of their data. These turn hybridization data into '**ratiograms**', a mass of green and red bars (see Figure 1) that are easier on the eye than a mass of raw data presented numerically (although color-blind researchers are at a distinct disadvantage).

The assumption has been that groups of co-regulated genes have potential biological significance; they may represent subsets of genes required for a particular transcriptional program or physiological process. This idea has been reinforced by the observation that expression profiling can identify groups of genes that effectively distinguish between different forms of cancer or even predict clinical outcomes [6].

Charting chromosomal territories

Now, in their *Journal of Biology* paper, Spellman and Rubin have taken a different approach, investigating genes according to their position on the chromosome and relating this to similarities in their expression patterns. When they analyzed data from 267 hybridizations from adult flies and embryos, they found that groups of physically adjacent genes had strikingly similar expression profiles; one fifth of all genes lie within about 200 such groups, spread throughout the *Drosophila* genome.

"I was stunned when I saw the first results" recalls Spellman. "We really hadn't predicted this." (See the 'Behind the scenes' box for more of the background to the work.) The groups have an average size of 12-15

Background

- High-density **microarrays** (often referred to as '**DNA chips**') are powerful tools for analyzing the **expression profiles** of all transcripts ('the **transcriptome**') 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**.**

- **TreeView** is a microarray analysis program that defines groups of genes with similar expression patterns by clustering them hierarchically. Expression profiles are most often depicted as a **ratiogram**, a grid of red (high relative expression) and green (low relative expression), in which individual genes are represented by rows in the grid and individual experimental conditions by columns.

- **Gene Ontology** (GO) is a genome annotation tool that attempts to define a unified vocabulary that relates primary DNA sequence to gene function in terms of biochemical and cellular activity within biological processes. Using GO terms allows computational analysis of whether genes have related functions.

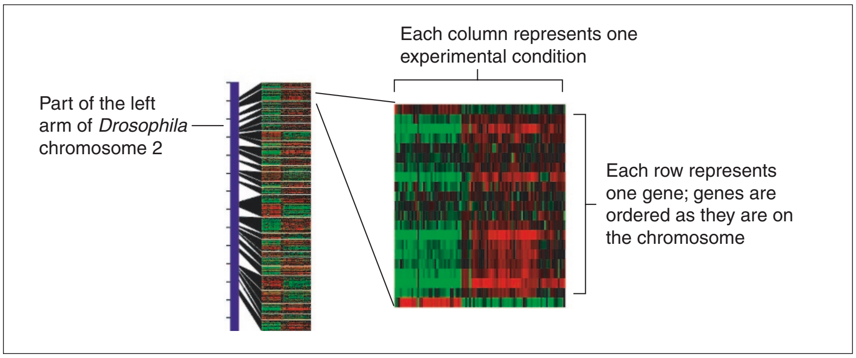

Figure 1
An example of a group of adjacent genes that are similarly expressed (adapted from [1]). For each square on the grid, red denotes relative expression higher than the average for a gene in an experiment, green denotes lower relative expression and black indicates that the expression is equal to the average. There are over 200 such groups within the *Drosophila* genome.

genes, with individual groups spanning up to 450 kilobases. Spellman carried out a series of rigorous checks to ensure that the results were real. The gene groups are not related to the banding patterns of *Drosophila* polytene chromosomes nor to known chromosomal structures such as scaffold-attachment sites. Computational analysis of the groups also demonstrated that they could not be explained by any detectable similarity between the genes within similarly expressed groups – not gene homology, **gene ontology**, or related function.

The idea of co-regulation of adjacent clusters of genes has been around for a while. More than 50 years ago François Jacob and Jacques Monod discovered operons — groups of genes that are expressed from a single promoter in the form of a polycistronic mRNA. Operons are common in bacteria and archaea and usually encode proteins that function together. But there are few examples of polycistronic mRNA in eukaryotes. Some examples of co-regulated clusters have been extensively characterized in mammals, but these predominantly contain related genes and have been considered to be rare cases, for example the developmentally regulated Hox genes and β-globin locus, or genes within the major histocompatibility complex.

Some studies have hinted that eukaryotic genes may be organized in distinct domains that are coordinately expressed. "These provocative results [from Spellman and Rubin] are reminiscent of what's been seen in yeast," says George Church (Harvard Medical School, Boston) referring to a study from his lab which showed that pairs or triplets of yeast genes that are highly expressed are often adjacent on the chromosome [7].

Spellman speculates that similar domains probably exist in most animal genomes. Indeed, a study from Rogier Versteeg's group (University of Amsterdam) used data from SAGE (serial analysis of gene expression) experiments in a range of human tissues and cancer cells to show that highly expressed genes were often grouped together in chromosomal domains [8]. "These are probably just the extremes of a continuum," says Versteeg; "the whole genome might be divided into domains of weak, high or intermediate gene expression." Very recent work from Laurence Hurst's lab (Bath University, UK) has also shown that genes that are expressed in most tissues (which they call 'housekeeping' genes) are found in groups within the human genome [9].

Unlike SAGE data, microarray results are usually expressed as expression ratios rather than as absolute expression levels. Spellman has tried turning the fly chip data into expression intensities and found that highly expressed genes were also in groups. Interestingly, Spellman notes that "there is only a partial correlation between domains defined by expression profile and those of high expression intensity."

How and why?

The intriguing observations by Spellman and Rubin pose a number of challenges about how chromosomal domains are created and maintained, why the genome contains such large clusters of similarly regulated genes, and the nature of transcriptional control. "It raises a lot of questions," says microarray aficionado Brian Oliver (NIH, Bethesda), referring to the Spellman and Rubin paper as a "call to exploration" and predicting a flood of papers exploring these domains. "Is control at the level of individual genes or whole domains?" asks Versteeg. "That's the most important question, but it's too early to say and it might take a long time to answer."

Mapping the transcriptome back onto the genome may help to link what is known about the fine-detail and large-scale regulation of transcription. In the good old days (before genome sequences and chips) the detection of quantitative changes in the expression of an individual gene (usually by northern blot analysis) was followed by a systematic and laborious characterization of its promoter and nearby enhancer sequences that act as a switch to determine whether a gene is on or off. This led to exquisite models of transcriptional regulation controlled by a precise network of sequence motifs and *cis*-regulatory modules.

Behind the scenes

Journal of Biology asked Paul Spellman to comment on why he and Gerry Rubin began their analysis of expression clusters in the *Drosophila* genome.

What was the question you wanted to address when you embarked on this study?
I had hierarchically clustered the data for a number of different experiments and noticed that most genes were preferentially expressed in either adults or in embryos. I presented this result at a meeting last fall, where Michael Ashburner commented that it had once been hypothesized that there were separate ‘genomes’ or gene complements for adults and for embryos. This led us to ask the question whether our gene-expression data segregated into ‘adult’ and ‘embryo’ by genome location.

What was your initial reaction to the results, and how were they received by others?
I was surprised and extremely excited, since we had no real expectation that there would be any correlation. We shared the results with a number of people prior to submission and they were all very interested. It doesn’t directly challenge any of the central tenets of biology, but it suggests that the mechanisms for controlling gene expression are more complicated than many had suspected.

How long did the project take?
We already had the data so it took a week or so of coding to show that there was a very strong preference for genes with similar expression patterns to be near one another. And it took us another few weeks to work out metrics that we could use to determine significance. The only real concern was how big expression domains are.

What are the next steps?
We know basically nothing about these domains. We want to determine how important they are to gene function, map the boundaries accurately, isolate potential boundary sequence elements, and determine if the domains are conserved in other species. There’s a ton of experiments to do.

More recently, enhancers have been found capable of regulating genes from quite substantial distances. Additional complexity has been revealed by studies of chromatin structure: different conformations of chromatin can regulate transcription and the accessibility to transcription factors by creating physical domains that are effectively ‘open’ or ‘closed’ for protein-DNA interactions.

Spellman and Rubin found that there is often a predominant gene within each chromosomal territory that is most strongly expressed or repressed and they suggest that the behavior of neighboring genes might reflect a general ‘sloppiness’ in transcriptional control. “We don’t have a mechanism,” admits Spellman, “but I think the most likely explanation is regulation at the level of chromatin.” Open chromatin conformations may be created to drive the expression of a certain key gene in the domain with the rest of the nearby genes “in effect being carried along for a ride” [1]. As long as their expression is not harmful to the cell, the changes in transcription of most genes may not be too important. “The regulation of transcription may be precise when it is needed and sloppy when it is not important,” write Spellman and Rubin [1]. Versteeg strongly rejects the notion of sloppiness in gene control, however, citing the catastrophic consequences of trisomy in humans.

Biologists will be keen to understand how the territories are established. “My gut feeling is that it’s driven by boundary elements” says Spellman. Church agrees that defining the nature of the domain boundaries is an important challenge. “If we have enough examples it might be possible to search using motif alignment tools,” he says, but he predicts that this will be harder than it was for promoter motifs.

It might be some time before the mechanisms involved and the biological consequences are clearly understood. “It’s possible that the expression domains are regulated by the three-dimensional structure of the nucleus and the ‘nuclear address’ of specific chromosome regions,” speculates Oliver. Versteeg proposes that “genes that are highly expressed might be clustered together to facilitate post-transcriptional functions such as splicing and RNA processing,” citing the existence of nuclear speckles — sites of splicing and RNA metabolism. Experiments with directed transgene insertions may help to address some of these issues. Comparison with similar studies in other organisms, and correlations with regions of conserved synteny within the genome, are likely to provide insights. And evolution may give us hints about what’s going on and about biological relevance.

The paper from Spellman and Rubin [1] represents a delicious taste of what’s to come in the post-genomic era, as extensive genome and transcriptome datasets become available. One result of the work is that microarray analysts might henceforth choose to map their gene-expression profiles to the relevant genomic location, before they construct elaborate theories about specific transcriptional programs on the basis of which genes are turned on and off. We clearly have a lot to learn about chromosomal territories and boundaries within the fly genome and perhaps in the genomes of the worm, the weed, mouse and man.

References

1. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the *Drosophila* genome.** *Journal of Biology* 2002, **1:**5.
2. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
3. **FlyBase** [http://www.fruitfly.org]
4. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
5. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
6. Ring BZ, Ross DT: **Microarrays and molecular markers for tumor classification.** *Genome Biol* 2002, **3:**comment2005.1-2005.6.
7. Cohen BA, Mitra RD, Hughes JD, Church GM**: A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nat Genet* 2000, **26:**183-186.
8. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, *et al.:* **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 2001, **291:**1289-1292.
9. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome.** *Nat Genet* 2002, **31:**180-183.

Jonathan B Weitzman is a science writer based in Paris, France.
E-mail: jonathanweitzman@hotmail.com