Minireview

Impressive expressions: developing a systematic database of gene-expression patterns in *Drosophila* embryogenesis

Haiqiong Montalta-He and Heinrich Reichert

Address: Institute of Zoology, Biocenter/Pharmacenter, Klingelbergstrasse 50, University of Basel, CH-4056 Basel, Switzerland.

Correspondence: Haiqiong Montalta-He. E-mail: Haiqiong.he@stud.unibas.ch

Published: 28 January 2003

Genome ***Biology*** 2003, **4**:205

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/2/205

Abstract

The establishment of a database of gene-expression patterns derived from systematic high-throughput *in situ* hybridization studies on whole-mount *Drosophila* embryos, together with new information on the reannotated *Drosophila* genome and several recent microarray-based genomic analyses of *Drosophila* development, vastly increase the breadth and depth that can be reached by developmental genetics.

These are exciting days for developmental genetics; the rapid advance of functional genomic analyses of key model systems is creating possibilities that were only scientific fantasies a few years ago. One of these fantasies was to know not only all of the genes in your favorite model organism, but also the expression patterns, in time and space, of all the genes during development. Imagine, for example, what you could do if you were a developmental biologist interested in the formation of midline structures who had access to a database that revealed the identity of all of the fly genes expressed at the midline during embryogenesis. Instead of spending time and money establishing a subtractive library or doing differential display to look for midline-specific genes, you could simply go to the database and query it for all the genes that are expressed at the midline. Then, knowing all the potential genes of interest, you could proceed directly to the functional analysis of these genes and the genetic networks in which they are involved. A recent report in *Genome Biology* by Tomancak *et al.* [1], on the systematic determination of patterns of gene expression during *Drosophila* embryogenesis, shows that this fantasy is rapidly becoming a reality. The systematic establishment of a gene-expression database, together with a flurry of new information on the annotated *Drosophila* genome [2-10] and several other recent microarray-based, functional genomic analyses of *Drosophila* development [11-14], allow us to study development at a new, more detailed level of resolution.

Driven by advances in DNA-sequencing technology, early genomic projects consisted principally of the large-scale sequencing of whole genomes. Currently, approximately 100 genomes have been completely sequenced, including around 90 prokaryotic and 8 eukaryotic genomes, and the complete genome sequences of an increasing number of model organisms are now becoming available [15]. A first annotated version of the *Drosophila melanogaster* genome was released in March 2000 [16-18], and this was the first metazoan genome to be successfully sequenced by the whole-genome shotgun method [19,20]. Since then, the *Drosophila* genome has been reannotated twice, and the most recent of these reannotations, called Release 3, has now been finished and is available online in FlyBase [3,21]. Established by human curators with the help of sophisticated new software and significantly increased amounts of experimentally derived data from cDNAs and expressed sequence tags (ESTs), Release 3 provides a euchromatic sequence that is virtually free of gaps and is highly accurate [3,4,6-8]. The number of genes has not changed much compared with the previous annotation, Release 2, but Release 3 contains more exons and more transcripts and, importantly, has changes in over 40% of the predicted protein sequences [7]. It is believed that this new release of the *Drosophila* genome sequence is now a reliable resource for molecular and genetic experiments as well as for computational analysis.

With the rapid progress of genome sequencing projects, microarrays have become powerful and popular tools with which to investigate biological questions at a genome-wide level. The adoption of microarray technology for the study of the development of *Drosophila* was initially rather slow, but its use has accelerated markedly, especially in the past year [22]. One of the first microarray-based analyses of *Drosophila* development focused on the process of metamorphosis, using microarrays containing cDNAs corresponding to several thousand gene sequences; this was carried out before sequence information on the entire genome became available [23]. Similar microarrays were combined with automated embryo sorting by Furlong and colleagues [11] to identify the targets of the transcription factor *twist,* which plays a key role in mesoderm development. More recently, a systematic study of gene expression throughout *Drosophila* development using microarrays has been carried out, in which approximately one third of all genes were surveyed at different stages of development - embryos, larvae, pupae and adults [12]. In two further recent investigations [13,14], whole-genome oligonucleotide microarrays representing the entire protein-coding capacity of the *Drosophila* genome (over 13,500 genes) have been used to study specific aspects of embryogenesis in the fly. Stathopoulos and colleagues [13] focused on dorsal-ventral patterning in the *Drosophila* embryo and used whole-genome microarrays to identify targets of the transcription factor *dorsal*; their work identified over 40 novel *dorsal* target genes as well as several new tissue-specific enhancers of *dorsal* targets. We and our colleagues [14] have studied gliogenesis in *Drosophila* embryos by using whole-genome microarrays to identify downstream targets of the *glial cells missing* gene, which controls the determination of glial versus neuronal cell fate.

Although these microarray experiments have each provided a quantitative overview of changes in gene-expression levels over developmental time or between different experimental conditions [11-14], they still suffer from several limitations that have been discussed in similar studies of other organisms. Transcripts of low abundance, which are often involved in regulatory processes and thus may be of great interest for understanding development, are typically underrepresented in RNA probe pools and are therefore hard to detect in microarray experiments [24]. Moreover, in multicellular organisms, cell division and differentiation leads to an increase in tissue complexity throughout development, but whole-animal microarray analysis cannot document this spatial information. One can try to isolate mRNA from every tissue at each developmental stage and then define gene-expression information in different tissues at different times, but this is a formidable task and requires the establishment of reliable methods for tissue-specific mRNA isolation and probe preparation. Furthermore, false-positive results can arise from technical problems such as cross-hybridization of target-probe pairs or incorrect annotation of genome sequences leading to false gene-model predictions [24,25]. For all these reasons, validation of microarray data with histological methods such as RNA *in situ* hybridization is both important and necessary. Indeed, all of the recent whole-genome microarray studies of *Drosophila* development incorporate selected *in situ* hybridization experiments to confirm and localize expression for a subset of the studied genes [11-14,26]. Given the massive quantitative expression dataset that is coming out of whole-genome microarray experiments, it now becomes important to have access to equally massive amounts of *in situ* hybridization data. Ideally, one would like to have access to the expression patterns of all genes in the genome in all major embryonic tissues at all embryonic stages. This is the goal of the online *in situ* gene-expression atlas that Tomancak and colleagues are assembling [1].

To achieve this formidable task, the authors [1] have devised a high-throughput whole-mount *in situ* hybridization protocol in which RNA probes are generated from the set of cDNA clones that comprise the *Drosophila* gene collections [3,27,28] and are then hybridized to *Drosophila* embryos in 96-well plates. Gene-expression patterns are documented by assembling digital photographs of individual embryos that are ordered according to developmental stage, in order to visualize time-dependent expression changes. To facilitate subsequent analysis, the expression patterns of all genes are annotated by a single human curator using a controlled vocabulary that describes the developmental and spatial relationships between embryonic tissues. Hierarchical clustering [29] is then used to group together genes with similar expression patterns, as well as to group embryonic tissues with similar sets of expressed genes. All these data - digital images as well as annotations - are stored in a relational database and presented in a searchable form on the web [30], allowing any interested researcher to query the database rapidly and to compare results in a rigorous manner. In addition, quantitative expression levels determined by whole-genome microarrays have been obtained for each gene and each developmental stage studied, and these data are also presented along with the images and annotations of *in situ* expression patterns in the database, thus making a direct comparison of the two complementing data sets possible. Figure 1 shows the pipeline used in the construction of the database.

Currently, over 2,000 genes, or about one sixth of all *Drosophila* genes, have been examined by *in situ* hybridization in embryos, and over 25,000 digital photographs of gene-expression patterns have been taken, annotated and stored in the database [1]. On the basis of current production rates, the authors estimate that a first pass through the existing cDNA collections, which represent about 70% of the *Drosophila* genes, should be finished within a year; probes for genes that lack a suitable cDNA clone but that show significant expression by microarray analysis will be generated

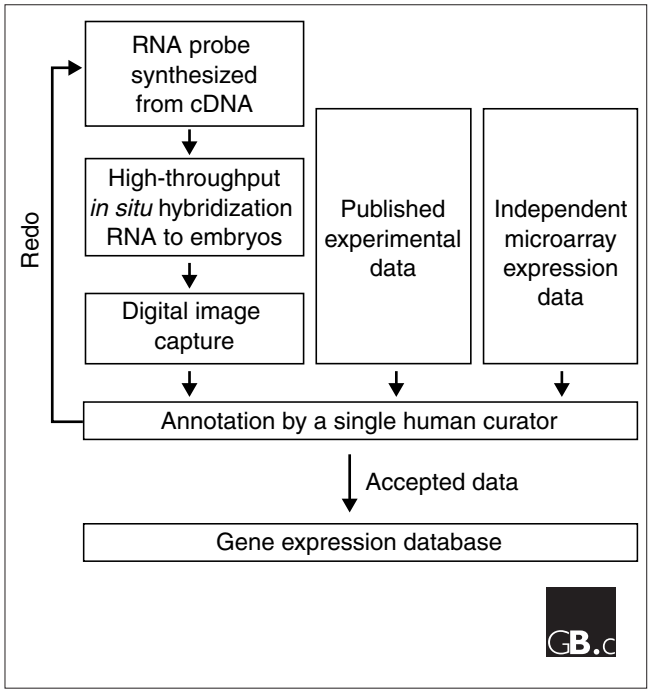

Figure 1
An overview of the pipeline used for the construction of the gene-expression database by Tomancak *et al.* [1].

by genomic PCR so that expression patterns for these genes can also be determined [1]. This will represent a major step towards the overall goal of determining the expression patterns of all genes in the fly genome and creating an integrated public resource of image-oriented gene-expression data analogous to the repositories of DNA sequences. The project will not stop there, however, but will continue to be refined as more accurate information on gene sequences, coding regions, and cDNAs becomes available. Release 3 of the fly genome has already presented marked improvements in all of these areas, and regular updates of the fly genome and the *in situ* expression database are planned [1-10].

Systematic high-throughput *in situ* hybridization of whole-mount embryos as used by Tomancak and colleagues [1] provides a powerful method for the global survey of gene expression in embryos [31]. Combined with data obtained by microarray analysis, this method makes it possible to investigate gene-expression profiles in both a quantitative and a qualitative manner. Analysis of this kind of gene-expression dataset will provide a rich source of developmental-genetic information and should also make it possible to identify genes involved in developmental processes that have been missed by traditional, mutagenesis-based genetic analysis. According to published estimates for flies and other animals, less than one-third of genes lead to obvious phenotypes when mutated [32-34], so a lot remains to be discovered. The exciting days of developmental genetics have only just begun.

References

1. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM: **Systematic determination of patterns of gene expression during *Drosophila* embryogenesis.** *Genome Biol* 2002, **3:**research0088.1-0088.14.
2. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E, *et al*: **Finishing a whole-genome shotgun: Release 3 of the *Drosophila* euchromatic genome sequence.** *Genome Biol* 2002, **3:**research0079.1-0079.14.
3. Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, *et al.*: **A *Drosophila* full-length cDNA resource.** *Genome Biol* 2002, **3:**research0080.1-0080.8.
4. Mungall CJ, Misra S, Berman BP, Carlson J, Frise E, Harris NL, Marshall B, Shu S, Kaminker JS, Prochnik SE, *et al.*: **An integrated computational pipeline and database to support whole-genome sequence annotation.** *Genome Biol* 2002, **3:**research0081.1-0081.11.
5. Lewis SE, Searle SMJ, Harris NL, Gibson M, Iyer VR, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, *et al.*: **Apollo: A sequence annotation editor.** *Genome Biol* 2002, **3:**research0082.1-0082.14.
6. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, *et al.*: **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review.** *Genome Biol* 2002, **3:**research0083.1-0083.22.
7. Kaminker JS, Bergman C, Kronmiller B, Carlson J, Svirskas R, Patel NH, Frise E, Wheeler DL, Lewis SE, Rubin GM, *et al.*: **The transposable elements of the *Drosophila melanogaster* euchromatin - a genomics perspective.** *Genome Biol* 2002, **3:**research0084.1-0084.20.
8. Hoskins RA, Smith CD, Carlson JW, Carvalho AB, Halpern A, Kaminker JS, Kennedy C, Mungall CJ, Sullivan BA, Sutton GG, *et al.*: **Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly.** *Genome Biol* 2002, **3:**research0085.1-0085.16.
9. Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, *et al.*: **Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome.** *Genome Biol* 2002, **3:**research0086.1-0086.20.
10. Ohler U, Liao G-C, Niemann H, Rubin GM: **Computational analysis of core promoters in the *Drosophila* genome.** *Genome Biol* 2002, **3:**research0087.1-0087.12.
11. Furlong EE, Andersen EC, Null B, White KP, Scott MP: **Patterns of gene expression during *Drosophila* mesoderm development.** *Science* 2001, **293:**1629-1633.
12. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: **Gene expression during the life cycle of *Drosophila melanogaster*.** *Science* 2002, **297:**2270-2275.
13. Stathopoulos A, Van Drenth M, Erives A, Markstein M, Levine M: **Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo.** *Cell* 2002, **111:**687-701.
14. Egger B, Leemans R, Loop T, Kammermeier L, Fan Y, Radimerski T, Strahm MC, Certa U, Reichert H: **Gliogenesis in *Drosophila*: genome-wide analysis of downstream genes of glial cells missing in the embryonic nervous system.** *Development* 2002, **129:**3295-3309.
15. **Entrez Genome**
[http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/org.htm]
16. 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.
17. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, *et al.*: **A whole-genome assembly of *Drosophila*.** *Science* 2000, **287:**2196-2204.
18. Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE: **Genome annotation assessment in *Drosophila melanogaster*.** *Genome Res* 2000, **10:**483-501.
19. Loder N: **Celera's shotgun approach puts *Drosophila* in the bag.** *Nature* 2000, **403:**817.
20. Hartl DL: **Fly meets shotgun: shotgun wins.** *Nat Genet* 2000, **24:**327-328.

21. **FlyBase** [http://flybase.bio.indiana.edu/]
22. Livesey R: **Have microarrays failed to deliver for developmental biology?** *Genome Biol* 2002, **3:**comment2009.1-2009.5.
23. White KP, Rifkin SA, Hurban P, Hogness DS: **Microarray analysis of *Drosophila* development during metamorphosis.** *Science* 1999, **286:**2179-2184.
24. Chudin E, Walker R, Kosaka A, Wu SX, Rabert D, Chang TK, Kreder DE: **Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays.** *Genome Biol* 2002, **3:**research0005.1-0005.10.
25. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
26. Klebes A, Biehs B, Cifuentes F, Kornberg TB: **Expression profiling of *Drosophila* imaginal discs.** *Genome Biol* 2002, **3:**research0038.1-0038.16.
27. Rubin GM, Hong L, Brokstein P, Evans-Holm M, Frise E, Stapleton M, Harvey DA: **A *Drosophila* complementary DNA resource.** *Science* 2000, **287:**2222-2224.
28. Stapleton M, Liao G, Brokstein P, Hong L, Carninci P, Shiraki T, Hayashizaki Y, Champe M, Pacleb J, Wan K, *et al.*: **The *Drosophila* gene collection: identification of putative full-length cDNAs for 70% of *D. melanogaster* genes.** *Genome Res* 2002, **12:**1294-1300.
29. **Annotation clustering** [http://www.fruitfly.org/ex/annotation_clustering.html]
30. **Patterns of gene expression in *Drosophila* embryogenesis** [http://toy.lbl.gov:8888/cgi-bin/ex/insitu.pl]
31. Simin K, Scuderi A, Reamey J, Dunn D, Weiss R, Metherall JE, Letsou A: **Profiling patterned transcripts in *Drosophila* embryos.** *Genome Res* 2002, **12:**1040-1047.
32. *Caenorhabditis elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
33. Thatcher JW, Shaw JM, Dickinson WJ: **Marginal fitness contributions of nonessential genes in yeast.** *Proc Natl Acad Sci USA* 1998, **95:**253-257.
34. Ashburner M, Misra S, Roote J, Lewis SE, Blazej R, Davis T, Doyle C, Galle R, George R, Harris N, *et al.*: **An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region.** *Genetics* 1999, **153:**179-219.