Molecular Systems Biology 3; Article number 72; doi:10.1038/msb4100112
**Citation: *Molecular Systems Biology* 3:72**

www.molecularsystemsbiology.com

molecular systems biology

# Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis

**Sean D Hooper[1,5], Stephanie Boué[1,5], Roland Krause[2,3,5], Lars J Jensen[1], Christopher E Mason[4], Murad Ghanim[4], Kevin P White[4], Eileen EM Furlong[1,*] and Peer Bork[1,*]**

[1] Structural and Computational Biology Unit, EMBL, Heidelberg, Germany, [2] Department Vingron, Max-Planck-Institute for Molecular Genetics, Berlin, Germany, [3] Department Zychlinsky, Max-Planck-Institute for Infection Biology, Berlin, Germany and [4] Department of Genetics, Yale University School of Medicine, New Haven, CT, USA
[5] These authors contributed equally to this work
* Corresponding authors. E Furlong, Gene Expression Unit, EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany. Tel.: +49 6221 387 8416; E-mail: furlong@embl.de or P Bork, Structural and Computational Biology Unit, EMBL, Meyerhofstrasse 1, Heidelberg 69117, Germany. Tel.: +49 622 1387 8526; Fax: +49 622 1387 8517; E-mail: bork@embl.de

Received 12.12.05; accepted 3.11.06

**Time-series analysis of whole-genome expression data during *Drosophila melanogaster* development indicates that up to 86% of its genes change their relative transcript level during embryogenesis. By applying conservative filtering criteria and requiring 'sharp' transcript changes, we identified 1534 maternal genes, 792 transient zygotic genes, and 1053 genes whose transcript levels increase during embryogenesis. Each of these three categories is dominated by groups of genes where all transcript levels increase and/or decrease at similar times, suggesting a common mode of regulation. For example, 34% of the transiently expressed genes fall into three groups, with increased transcript levels between 2.5–12, 11–20, and 15–20 h of development, respectively. We highlight common and distinctive functional features of these expression groups and identify a coupling between downregulation of transcript levels and targeted protein degradation. By mapping the groups to the protein network, we also predict and experimentally confirm new functional associations.**
*Molecular Systems Biology* 16 January 2007; doi:10.1038/msb4100112
*Subject Categories:* stimulation and data analysis; development
*Keywords: Drosophila* embryogenesis; Notch pathway; supervised clustering; transient expression

## Introduction

For the purpose of tracking relative transcript levels during development, microarray analysis has proven to be invaluable. The partial transcriptomes of two major model organisms have already been analyzed, the nematode *Caenorhabditis elegans* (Baugh *et al*, 2003) during embryogenesis, and an extensive developmental time series in *Drosophila melanogaster* of approximately 30% of all genes covering the entire lifespan, from the first minutes of development to aging adults (Li and White, 2003). The latter study gave the first insights into global changes of regulation, such as the prominent biphasic expression of many genes in two major stages, either in the embryo and pupa, or in the larva and adult, revealing the molecular similarities in these stages of the lifecycle.

Recently, genome-wide expression in fruitfly was measured at the exon level, providing enough resolution to identify alternative splicing in 40% of predicted genes and to identify at least 15% as developmentally regulated (Stolc *et al*, 2004). However, as only two time points were considered during early and late embryo development, many transient and tightly regulated processes during embryogenesis would not have been detected.

Here, we perform an extensive analysis of the fly transcriptome, comprising 12 868 genes (FlyBase 4.0 release; Drysdale and Crosby, 2005), during 30 time points, covering the entire 24-h period in which the fertilized egg develops into a larva. This enabled us to identify transition points of sharp changes in transcript levels, as well as groups of genes with similar expression profiles during embryo development. Several common functional features were identified among the proteins encoded by genes with similar temporal co-expression patterns. For example, a significant enrichment of physical interactions implies the presence of entire programs of 'effector' genes involved in a common developmental process. Furthermore, we show on a more global level that tight regulation of transcript levels is often accompanied by targeted protein degradation. As a result, we increase our understanding of several vital developmental pathways, and we suggest new, as yet unknown, members of these pathways, some of which we validate experimentally.

## Results

### Generation of the developmental time-series data

Thirty-one-hour time points, which spanned all stages of embryogenesis, were collected as described previously (Arbeitman *et al*, 2002). To capture the rapid developmental

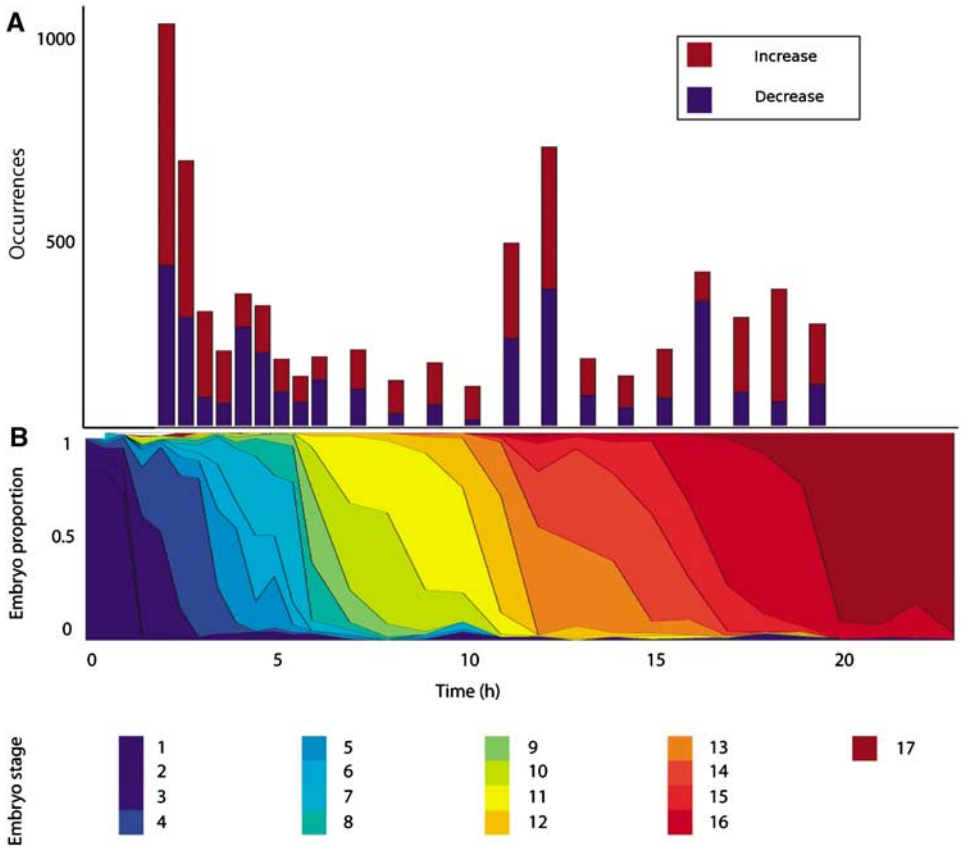


**Figure 1** (**A**) Increase and decrease of fly gene transcript levels during embryogenesis. Red bars indicate points of sharp expression changes from low to high and blue bars signify changes from high to low expression. (**B**) Distribution of embryo stages at sampling times. For instance, at 12 h, a majority of embryos have reached stages 12–13. Samples are taken every half an hour at the start of the study.

changes that occur during the first half of embryogenesis, overlapping 1-h time points were obtained for the first 6.5 h of development. The stages of all samples were verified and only tightly staged embryo collections were used for RNA isolation and microarray analysis in order to minimize the overlap between measurements (see Figure 1 for the distribution of stages at each time point).

Three independent embryo collections were used for each stage of development. Details of the microarray hybridizations are found in Arbeitman *et al* (2002). All samples were hybridized together with a common reference sample, which was made from pooled samples for each transcript from all stages of the *Drosophila* lifecycle. Thus, the expression level of each gene in the sample can be compared relative to its corresponding reference expression level. The microarrays used for this study consist of PCR fragments of one exon of every predicted *Drosophila* gene (Li and White, 2003). The data were normalized using the intensity-dependent Qspline method (Workman *et al*, 2002) and subsequently corrected for spatial biases (see Materials and methods for details).

## Estimating expression changes of genes during embryogenesis

Two different statistical methods were used for identifying genes that change in expression during embryonic development. First, the widely used analysis of variance (ANOVA) was applied to identify significant changes in the general expression level. We found significant changes in transcript levels for 86% ($P<0.05$) of all genes. This compares to *C. elegans* (Baugh *et al*, 2003) and an earlier analysis of *D. melanogaster* (Arbeitman *et al*, 2002), where 68 and 95% respectively ($P<0.05$) of the genes were found to change expression during embryogenesis. However, the actual implementation of ANOVA may differ slightly. Second, to explicitly analyze the temporal dependency of expression levels in individual genes, a runs test was used; it suggests temporal changes in transcript levels for 65% of genes during embryogenesis.

These estimates give a global overview of the amount of change occurring during the entire embryonic development; however, they do not pinpoint when transitions in gene expression programs occur. To get a more exact time measure, we searched for changes in expression levels using local convolution methods (Supplementary Figure S1). More specifically, we required four points of low expression and four subsequent points of high expression (or vice versa) even if the amplitude change was relatively low (see Materials and methods). This type of convolution not only requires a sharp increase or decrease of expression, but also that the change in transcript level is consistent over a period of time, thereby reducing the rate of false positives owing to individual outliers.

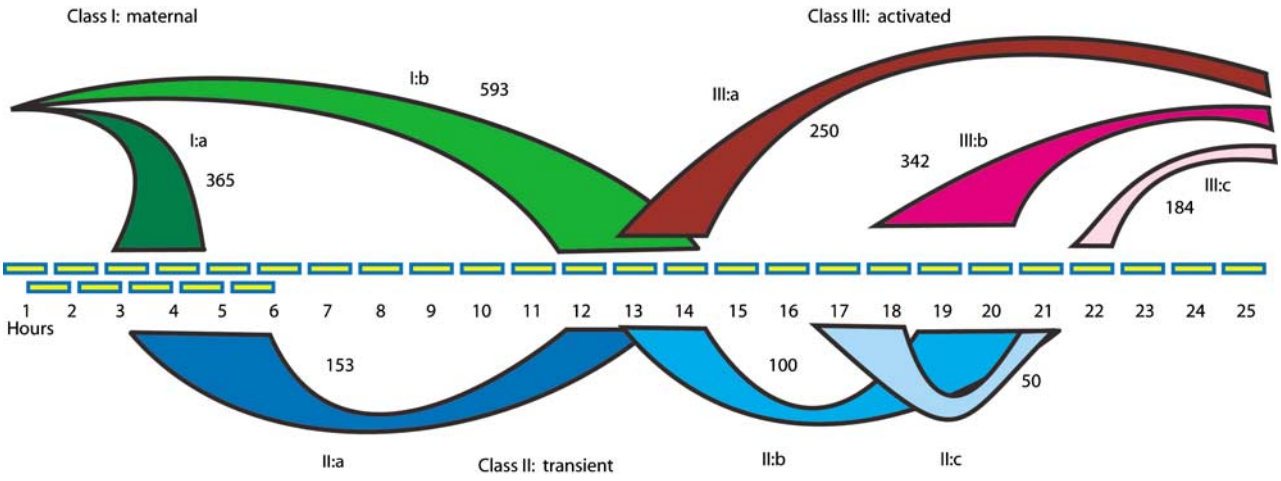


**Figure 2** Major classes of transcript levels, as determined by global convolution. Arcs represent the dominant subgroups within class **I**, **II,** and **III** transcripts. For instance, class **I** is dominated by two main subgroups **I:a** and **I:b**, represented by the pink and red arcs, respectively. Time is in hours, and yellow rectangles signify measurement points. The time of increase and decrease of the transcript groups coincides with those derived by local convolution (Supplementary Figure S1). Note the interplay between groups as a decrease of one transcript group is followed by an increase of another.

Using this approach, we mapped 6233 'sharp' changes in transcript levels (2808 increases and 3425 decreases) to time points and developmental stages (Figure 1). As indicated already in the study of Arbeitman *et al* (2002), several developmental stages show an increased frequency of transcript level changes during embryogenesis, which can now be confirmed genome-wide. The local convolution analysis also revealed that the increase in the transcript level of one group of genes often coincides with the decrease of transcript expression of another group of genes and vice versa, indicating coordinated waves of expression (Figure 2).

A flurry of expression changes was observed at 2–3 h (embryo stage ~5), representing the initiation of zygotic transcription and the parallel decay of some maternal transcripts. This first stage of dramatic change coincides with events just after cellularization—the process during which each nucleus is enclosed to form a cell by invagination of the plasma membrane—for example, the major morphogenetic changes leading to germ layer formation that occur during that time period in the cellularized embryo. It might also reflect the embryo patterning that begins along both the anterior–posterior and dorsal–ventral axis (Wolpert *et al*, 2002).

A second major period of transcript expression change (both increase and decrease) was observed at roughly 12 h, when most embryos had reached stages 12–14, corresponding to the end of the dorsal closure, the terminal differentiation of many tissues, and to the invagination of the epithelial cells that will become the imaginal discs.

A third period of gene expression change is observed at 16 h (stages 14–16) when a discrete set of transcripts decrease their expression levels, followed by an intense increase of transcript levels of another set of genes (17–19 h). This could possibly be in preparation for the transition to the larval stage. Generally, sharp decreases of mRNAs seem to be more confined to particular time points.

The correlation between times of increase and decrease in transcript levels suggests the existence of coregulated groups of genes that drive major developmental events during embryogenesis.

## Classification of gene expression behavior

To be able to group genes whose transcripts follow similar patterns over the full 30 time points, and to correlate this with the periods of rapid expression changes, we used global convolution (see Materials and methods and Supplementary Figure S1) to assign genes to general expression classes. These are characterized by distinct plateaus of low and high expression during embryogenesis: class **I** (maternal) genes encoding transcripts that start with a high relative transcript level, which subsequently decreases; class **II** (transient) genes whose transcripts levels first increase and later decrease, and thus do not seem to be maternally deposited and, Finally, class **III** (activated) genes encoding transcripts for which we only observe an increase in expression.

Class **III** gene transcripts are most likely not present at high levels during the entire *Drosophila* life cycle, as most of the corresponding genes return to low levels at various time points beyond embryogenesis (Arbeitman *et al*, 2002). Although the transcripts from the transiently expressed genes (class **II**) may be present in later stages in the larva, pupa, or the adult fly, we expect only a few genes with multiple expression peaks during embryogenesis (Arbeitman *et al*, 2002), also indicated by the runs test above (data not shown).

Using global convolution (see Materials and methods), we found strong and significant expression correlation coefficients ($r > 0.8$, $P < 10^{-4}$, $t$-test) between transcript levels and global convolution profiles for 26% (3379) of the transcripts present on the array. Of these, we classified 1534 as class **I**, 792 as class **II**, and 1053 as class **III**. Many genes do not fit to any of these three classes, for instance genes that are constitutively transcribed (or not transcribed at all) during embryogenesis. Furthermore, the requirements for assignment were rather stringent; the entire expression profile must fit the categorization to a high degree (30 time points), whereas in the local

convolution, only eight time points are considered (corresponding to $P<10^{-4}$, *t*-test). This leads to a low rate of false positives at the cost of sensitivity. For instance, we are likely to miss many genes with very short periods of transcriptional activity.

Within each of the three global classes, groups of genes can now be readily identified with common times of increase and/or decrease of relative transcript levels (Figure 1A). More than 62% of the class **I** (maternal) genes can be classified into two major groups: class **I:a** and **b**, whose transcript levels decrease at 3–5 and 12–14 h, respectively. For the class **II** genes, even though there are 276 possible combinations of time points of increase and decrease of transcript levels (see Materials and methods), as many as 38% of the 792 class **II** genes fall into only three groups: **II:a** (2.5–12 h), **II:b** (11–20 h), and **II:c** (15–20 h) containing 153, 100, and 50 genes, respectively. Of the genes whose transcript levels increase but are not observed to decrease during embryogenesis (class **III**), more than 73% can be classified into three main groups; class **III:a**, **b**, and **c** (times of increase at 13–14, 18–20, and 22 h of development).

In order to identify biological principles underlying these different coexpression groups, and also as a general quality control, we studied these groups using various sources of biological information (Table Ia–c and Materials and methods). For example, proteins encoded by genes assigned to each of the three major classes have more interactions with each other than expected from a random group of the same size ($P<0.01$; Table Ia–c), showing that the gene products are not only coexpressed but also tend to interact physically to perform related cellular functions.

## Class I: maternal genes

Class **I** contains 1534 genes, and thus represents the most populated of the three major expression classes. It is long been known that a large number of transcripts are deposited in the oocyte during gametogenesis. Among other vital functions, they have been shown to be responsible for establishing the major body axes and for the initiation of zygotic transcription (Luschnig *et al*, 2004). The importance of these genes is reflected by a high proportion of lethal genes (observed: 10%; expected: 5%; $P<10^{-3}$; $\chi^2$-test) and a higher than average fraction of orthologs in this group shared with *Anopheles gambiae* (Table Ia), indicating functional conservation. Furthermore, analysis of Flybase GO annotation (Drysdale and Crosby, 2005) revealed that there is a significant overrepresentation of genes involved in 'nuclear organization and biogenesis', 'nuclear mRNA splicing', and 'DNA metabolism' in the class **I** group. These genes facilitate the organization of the chromosomes and nuclei during the very rapid cell divisions in the precellular blastoderm embryo. Proteins encoded by the class **I** genes also display a strong physical interconnectivity: 1097 connections within 1534 proteins (expected connections: 517; $P<10^{-3}$; $\chi^2$-test). Furthermore, they are enriched in transcription factors ($P<0.05$). A total of 365 out of the 1534 genes change transcript levels at 1.5–3 h (group class **I:a;** Figure 2). As expected, the functionally characterized genes of the class **I:a** group are mostly involved in early development and in the cell cycle according to the interactive fly database (Brody, 1999). The class **I:b** group consists of 593 genes, which encode transcripts with decreased levels of expression by 10–11 h. These genes are annotated in the interactive fly database with functions involved in the cell cycle, chromatin organization, and DNA replication. Class **I:b** is also enriched in lethal genes ($P<0.01$; $\chi^2$-test). In developmental terms, the decrease of the transcripts in class **I:a** coincides with the start of gastrulation, and **I:b** with the initiation of germ band retraction and dorsal closure.

## Class II: transient genes

Despite the stringent requirements of class **II** genes to have both a sharp increase and decrease in transcript levels, 792 fly

**Table Ia** Database analysis of class I transcripts

| Evidence | **I** | **I:a** | **I:b** |
|---|---|---|---|
| Orth | + ($P<10^{-9}$) | NS | + ($P<10^{-6}$) |
| Lethal | + | | |
| Transcription factors | + | | |
| Interaction | + ($P<10^{-3}$) | NS | + ($P<10^{-3}$) |
| PEST | + ($P<10^{-20}$) | + ($P<10^{-2}$) | + ($P<10^{-10}$) |
| Pathway1 | + ($P<0.03$) | NS | + ($P<0.02$) |
| PathwayX | + ($P<0.01$) | NS | + ($P<0.01$) |
| GO—molecular function | RNA and DNA metabolism, nuclear division, cell cycle | Catalytic activity | RNA metabolism, cell cycle |
| GO—biological process | | dvpt (pattern specification, reproduction), cell growth, protein metabolism | |
| GO—cellular compartment | Intracellular | Polar granule | |
| *In situ* | NS | Maternal ($P<0.05$) | Central nervous system ($P<0.05$) |
| Sum | 1534 | 365 | 593 |
| Start | | 1 | 1 |
| Stop | | 4–7 | 17–18 |

Orth: proportion of genes with orthology to *A. gambiae*. Lethal: proportion of genes annotated as 'Phenotypic class: Lethal'. Transcription factors: proportion of genes annotated as 'transcription factor'. Interaction: proportion of genes coding for proteins with known or predicted protein interactions in STRING (see Materials and methods). Pathway1: proportion of genes (counted once) involved in known pathways. PathwayX: as Pathway1, but genes may be counted any number of times, if it appears in several pathways. GO: annotation for the categories *molecular function*, *biological process,* and *cellular compartment.* The dominant categories are listed, and *P*-values are given if significant. *In situ*: major *in situ* annotation of the class of transcripts. *P*-values are given if significant. Sum: total number of transcripts in the class. Start: general time of increase. Stop: general time of decrease. For the categories Lethal and Transcription factors, a '+' denotes an overrepresentation and '−' an underrepresentation. The distributions are significant at $P<10^{-3}$ ($\chi^2$-test). NS, nonsignificant.

**Table Ib** Database analysis of class II transcripts

| Evidence | **II** | **II:a** | **II:b** | **II:c** |
|---|---|---|---|---|
| Orth | NS | NS | NS | NS |
| Lethal | + | | | |
| Transcription factors | + | | | |
| Interaction | + ($P<10^{-4}$) | + ($P<10^{-4}$) | NS | NS |
| PEST | + ($P<10^{-7}$) | + ($P<10^{-4}$) | NA | + ($P<0.02$) |
| Pathway1 | NS | NS | NS | NS |
| PathwayX | + ($P<0.02$) | + ($P<0.02$) | NS | NS |
| GO—molecular function | Transcriptional regulation, antioxidant activity | Transcription regulation, binding | Oxidoreductase activity | Structural constituent of cuticle |
| GO—biological process | Development, cell communication, transcription | Transcription, Notch signaling pathway, cell differentiation, dvpt | | |
| GO—cellular compartment | Nucleus, plasma mbn | Nucleus | | |
| *In situ* | NS | Mesectoderm and derivatives ($P<0.05$) | Dorsal ectoderm and derivatives ($P<0.05$) | NS |
| Sum | 792 | 153 | 100 | 50 |
| Start | | 6–10 | 18–19 | 22–23 |
| Stop | | 17–18 | 24–26 | 26 |

See Table Ia for details.

**Table Ic** Database analysis of class III transcripts

| Evidence | **III** | **III:a** | **III:b** | **III:c** |
|---|---|---|---|---|
| Orth | NS | NS | −($P<0.02$) | −($P<0.01$) |
| Lethal | — | | | |
| Transcription factors | — | | | |
| Interaction | NS | NS | NS | NS |
| PEST | − ($P<10^{-9}$) | NS | −($P<10^{-3}$) | −($P<10^{-9}$) |
| Pathway1 | NS | NS | NS | NS |
| PathwayX | NS | NS | NS | NS |
| GO—molecular function | Structure(cuticule), catalytic activity | Monovalent inorganic cation transporter activity ($P<10^{-4}$) | Structural, enzyme inhibitor, transporter activity | Catalytic activity ($P<10^{-6}$) |
| GO—biological process | Muscle contraction, metabolism | Muscle contraction ($P<10^{-5}$) | Metabolism | Catabolism ($P<10^{-6}$) |
| GO—cellular compartment | Muscle fiber, vacuole, extracellular | Muscle fiber ($P<10^{-7}$) | Vacuolar membrane | |
| *In situ* | NS | NS | NS | NS |
| Sum | 1053 | 250 | 342 | 184 |
| Start | | 18–19 | 23–25 | 27 |
| Stop | | 30 + | 30 + | 30 + |

See Table Ia for details.

genes were classified in this class. This is consistent with earlier reports in *D. melanogaster* (Arbeitman *et al*, 2002) and also *C. elegans* (Baugh *et al*, 2003). As shown in Table Ib, the class **II** genes are enriched in transcription factors and lethal phenotype genes (hereafter referred to as *lethals*; see Materials and methods) and not surprisingly, this class shows an over-representation of genes with functions involved in development. More specifically, many of these genes are annotated as encoding proteins involved in 'histogenesis', 'organogenesis', 'ectoderm development', 'cell differentiation', and 'cell fate commitment' (Table Ib; GO analysis). The class **II** group is also enriched in genes that encode well-characterized transcription factors ($P<0.05$, $t^2$-test).

Of the class **II** genes, 303 (38%) are found in only three groups; **a**, **b**, and **c**, each defined by specific times of increase and decrease of transcript levels (Figure 2). In addition to the common features above, the groups also differ from each other. For example, more genes from class **II:a** (with a plateau of increased transcript levels starting at 3–6 h and decreasing at 12–13 h) have been functionally characterized than average for the genome ($P<0.01$, $t^2$-test). Conversely, genes in class **II:b** (13–14 to 19–21 h) and **II:c** (17–18 to 21 h) are poorly characterized. Class **II:a** corresponds roughly to the time of cellularization and gastrulation, up to the point of germ band retraction. Class **II:b** and **II:c** range between germ band retraction and late embryonic stage.

The class **II:a** group (153 genes, of which 109 are annotated) contains 11 genes (of 29 on the array) that are a part of the Notch pathway (see below). Some other Notch members are found in class **I**, as many are maternally inherited. The drop in their transcript levels do however tend to coincide with **II:a**. The remaining 142 genes are often annotated as being implicated in 'neuroblast cell fate determination' suggesting that this group is highly enriched for both the regulators

(members of the Notch pathway) and their downstream effector molecules (genes involved in neurogenesis).

Genes involved in dorso-ventral patterning are also significantly overrepresented in the class **II** groups ($P<0.01$) and in particular in the class **II:a** category ($P<0.05$). Genes from the class **II:a** group are expressed in the procephalic ectoderm, ventral ectoderm, sensory complex, and central brain neurons (*in situ* data, see Materials and methods). Again, this strongly suggests a role for class **II:a** genes in nervous tissue and brain development.

Despite the low proportion of functionally characterized genes in the class **II:b** group (51 of 100 genes), there is an enrichment in oxidoreductase and peroxidase functions (GO classification). Moreover, defense and immune response are common functional classifications in this group. For instance, eight (of 20 known) members of the Osiris cluster (Dorer *et al*, 2003) are present in class **II:b**. This cluster is part of the largest region of synteny between *D. melanogaster* and *A. gambiae*, and encodes one of the largest gene families in fruitfly (Zdobnov *et al*, 2002). The genes of the Osiris cluster are still poorly characterized, but they are known to be under strong selection pressure both on protein sequence and expression level (Dorer *et al*, 2003). The transcript levels of Osiris genes 3, 7, 9, 17, 18, and 20 are known to be high during embryogenesis from stages 13 to 16 (Dorer *et al*, 2003), compatible with our findings.

The class **II:c** group encompasses 50 genes. Only 19 of these are annotated, and most are involved in cuticle formation. The fly embryo secretes a hard proteinaceous material, which forms a thick protective cuticle surrounding the larvae. Furthermore, class **II:b** (see above) is linked via *in situ* data to the dorsal ectoderm, suggesting that also many genes in class **II:b** may contribute to cuticle formation (Ostrowski *et al*, 2002). Therefore, groups **II:b** and **II:c** may be of interest when designing insecticides or planning experiments that target the cuticle.

### Class **III**: activated genes

Of the 1053 transcripts with sharp increases in expression levels but without a subsequent decrease during embryogenesis (class **III**), 250 are activated at 11–12 h (**III:a**), 342 at 16–18 h (**III:b**), and 184 at 20 h (**III:c**). These genes are the most species specific of the three categories: we found significantly fewer orthologs to predicted genes from *A. gambiae* than for the genes represented on the microarray as a whole. Furthermore, known transcription factors are significantly underrepresented in this group (Table Ic), which implies that the mRNA expression of transcriptional regulators is likely to be under tight regulation. **III**:**a** starts roughly at the same time as the dorsal closure, and **III:b** coincides with the late embryonic stage. The initiation of **III:c** cannot be correlated to any distinct developmental event.

## Coordinated regulation of transcripts and protein products

The decrease in transcript levels during embryogenesis in the class **I** and **II** genes suggests that it is important to reduce the levels of the respective protein in a temporally controlled manner. Transcriptional regulation alone is not sufficient to ensure a rapid reduction in protein levels, as the protein degradation may take hours. We thus hypothesize that the protein products of most transcriptionally repressed genes are inactivated, for example through targeted degradation controlled by PEST regions. This mechanism has previously been suggested to be responsible for the degradation of maternal proteins in *C. elegans* (Baugh *et al*, 2003) as well as for proteins that are periodically expressed during the mitotic cell cycle in several eukaryotes (de Lichtenberg *et al*, 2005; Jensen *et al*, 2006).

To test our hypothesis, we used a computational method to systematically predict PEST regions in the *D. melanogaster* proteome and compared the percentage of PEST-containing proteins encoded by the genome to that of the genes in each expression class. The highest percentages of PEST-containing proteins are encoded by class **I** (39%) and class **II** genes (37%). Both groups are significantly enriched in PEST regions compared to the proteome-wide content (31%) with *P*-values of $10^{-11}$ and $10^{-3}$, respectively. The difference between class **I** and **II** is not statistically significant. In contrast, PEST regions are found in only 21% of the proteins encoded by class **III** genes whose RNA levels remain high, which is significantly less than expected at $P<10^{-9}$. These observations are consistent with the hypothesis of a coordinated downregulation of a gene's expression in *Drosophila* at the level of both their RNA and protein products during embryogenesis, as suggested previously for other organisms (Baugh *et al*, 2003; de Lichtenberg *et al*, 2005; Jensen *et al*, 2006).

## Mapping coordinately expressed groups of genes to protein interactions

As suggested above, transcripts in the individual expression groups are tightly coexpressed at the same stages of development and are enriched in particular groups of functional (GO) categories. Consequently, one would expect that the protein products preferentially interact with each other to perform common functions. To test this hypothesis, we analyzed the integrated *D. melanogaster* protein interaction network from the STRING database, which contains large-scale interaction data from fruitfly yeast two-hybrid screens (Giot *et al*, 2003), small-scale interactions stored in dedicated databases and extracted from the literature and inferred interactions from different species, all embedded into a unified scoring scheme (von Mering *et al*, 2005). Indeed, proteins encoded by genes in all groups except class **I**:**a** and class **III:a** have more interactions between themselves than with proteins outside that group ($P<0.01$). Certain classes of proteins and pathways contribute strongly to this enrichment, including proteins involved in energy production, transcription factors (such as *bicoid*, FBgn0000166), or members of the Notch pathway (Figure 3). The latter also indicates that functions are not only performed within a particular coexpression group but also that some larger developmental pathways require a concerted action of genes with different expression profiles.

## Exploration of coexpressed genes involved in common pathways: Notch as an example

The Notch signaling pathway regulates cell fate determination, and consists of 61 proneural and neurogenic genes (Brody,

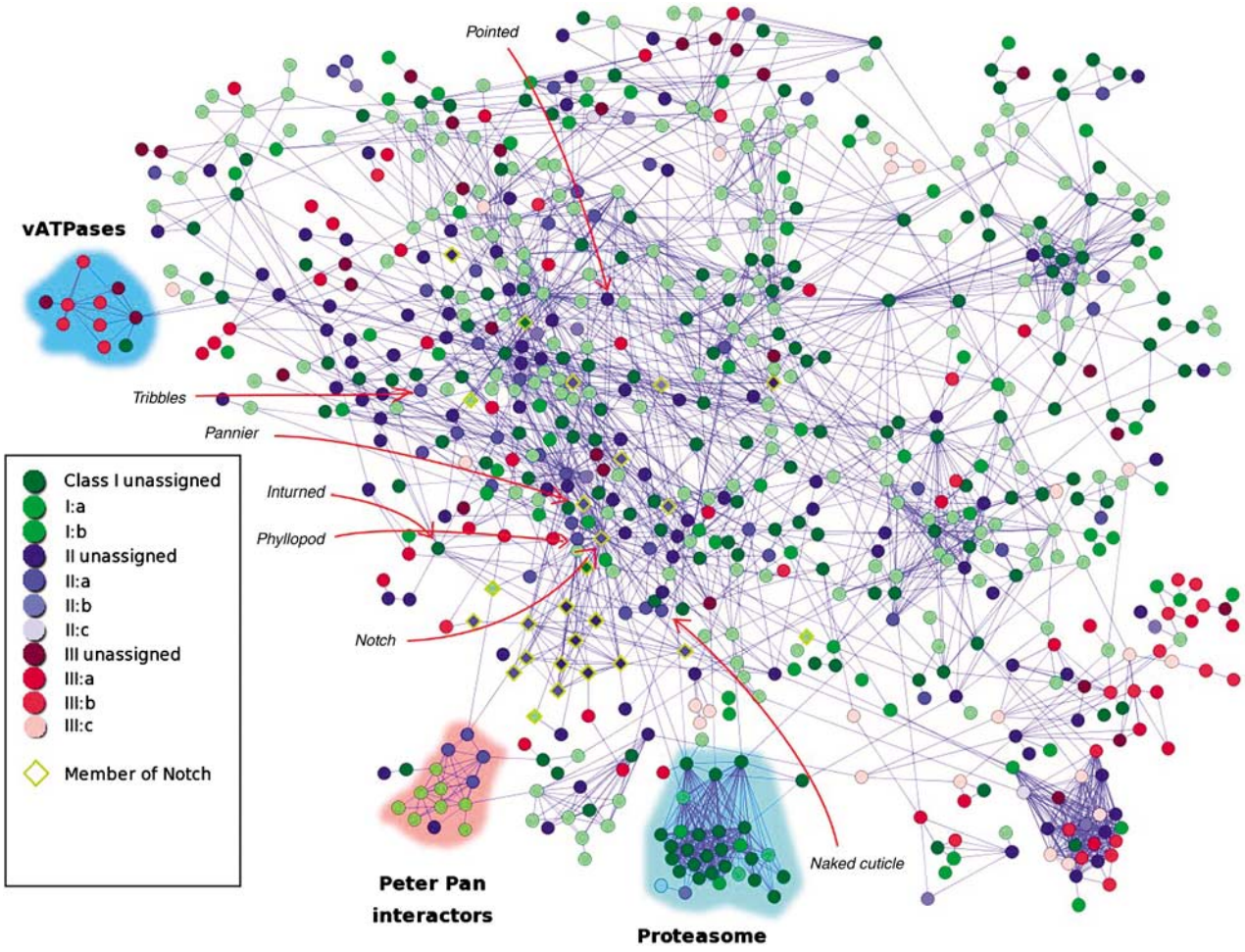


**Figure 3** Major component of a literature-derived protein interaction subnetwork, obtained from the String database at a reliability score of at least 0.3 (von Mering *et al*, 2005). It reveals that several well-known interacting proteins also show similar expression profiles. Examples are the highlighted vacuolar ATPases, the proteasome, or interactors of Peter Pan. *Notch* appears as the central node of the network and contributes to the high interconnectivity of the (transient) class **II:a** group. Some genes with expression profiles very similar to Notch (labeled by red arrows) are currently only loosely associated with the pathway (see text), but might share more functionality with Notch than previously thought. *Notch* is labeled for reference purposes. Note that *unassigned* in the legend means that the respective genes belong to the class but not to any of the major subclasses. To explore this complex network in full detail, see the interactive figure and data files in Supplementary information.

1999). Of these 61 genes, 52 were assayed on the array and 20 fit significantly to the category of class **II** genes. Thus, >32% of the Notch pathway genes have highly coordinated and transient gene expression, as compared to an expected 6.2% (based on 792 of 12 868 transcripts in class **II;** $P<10^{-4}$, *t*-test). This reflects the transient role of this pathway in cell fate specification. Roughly half of these genes are found in **II:a**, along with *Notch* itself. Also, we find four pathway members (e.g. *deltex*: FBgn0000524) in the maternal class **I:b**, whose transcripts decrease at the same time as class **II:a**.

Some members of the Notch pathway in class **II** include *big brain* (FBgn0000180; an ion channel concentrated at apical adherence junctions), *neuralized* (FBgn0002561; a gene involved in ubiquitination), and a key transcription factor acting in parallel to the proneural genes, *soxneuro* (FBgn0029123; Overton *et al*, 2002).

*Notch* itself (FBgn0004647) and many of its pathway members are highly expressed up to the 12th hour of development (stages 12–14 in this developmental time series), but then have an abrupt decrease in expression, dropping to very low transcript levels after this point. This sharp decrease of *Notch* and some other transcripts suggests that the pathway is not needed anymore and should be suppressed, indicating that the major specification events have been completed. This is further indicated by the presence of a PEST sequence in *Notch*, and the fact that the prolonged transcript expression of *Notch* gives rise to disease (Joutel and Tournier-Lasserve, 1998).

## Coexpressed classes of genes show coordinated expression *in vivo*

As outlined above, the tight temporal patterns of expression observed by >30% of the Notch family members suggest a coordinated regulation in expression. This observation makes a number of predictions about the expression patterns of the other ~100 class **II:b** genes, the majority of which are poorly characterized. Firstly, class **II** genes should have transient

expression, initiating at stages 6–7 and decreasing by stage 13. Second, as these genes have coordinated expression with Notch family members, they are highly likely to be colocalized in the same cells. To test these hypotheses, we selected five candidates from the **II:a** group, for costaining with *Delta* (FBgn0000463; the ligand of the Notch receptor) by double *in situ* hybridization. The candidates were selected based on their strength of correlation to the **II:a** profile, their fold change, and their gene annotation.

Four of the five genes gave specific patterns of expression. We were not able to obtain an *in situ* hybridization probe for the fifth gene (*CG1316*). For all four genes tested, their expression initiates early in development (~stage 7), and is dramatically reduced or absent by stage 13 (Figure 4). There is residual expression in small groups of cells at stage 13 for three genes, for example in the brain (*worniu* and *CG13333*), in the foregut (*pdm2*), and segmentally repeated groups of cells in the ectoderm (*CG13333*). No expression was observed for *CG4440* at stage 13, indicating that this gene is no longer transcribed. Therefore, the temporal window and transient nature of expression of all four genes map to the prediction for class **II:a** genes.

We next examined the spatial colocalization of these four genes with the Notch pathway, using *Delta* as a marker. Delta is a membrane-bound ligand for Notch, and is therefore expressed on the 'Notch signal-sending' cell. While Notch itself is a membrane bound receptor on the 'Notch-receiving cell'. As Delta is tethered to the membrane, in contrast to other signaling pathways, the 'Notch signal-sending' cells and 'Notch-receiving' cells are usually adjacent to each other or within the same field of cells. A colocalization of tissue expression in the Delta-expressing cell or the neighboring Notch-expressing cell would suggest that the transcript may play a role either in the Notch–Delta pathway or in a tightly coordinated parallel pathway. The transcription factors *pdm2* and *worniu* are essential for brain and neural development, but are not known to be linked to either Notch or Delta specifically.

Figure 5 shows *in situ* hybridization of stage 11–12 embryos, which have peak expression for class **II:a** genes expression. As the Notch pathway is highly active during this time of embryogenesis, *Delta* has a very broad expression pattern making it problematic to discern specific colocalization. Given this potential difficulty, we could observe colocalization of *worniu* and *pdm2* in the ventral nerve cord and brain. These genes are colocalized in a subset of neuroblasts indicating specific colocalization in these cells at this stage of development. *CG13333*, a gene of unknown function, has a broad expression in a number of tissues including the developing foregut, hindgut, and trachea. Again, we observed specific colocalization with Delta in the brain. Interestingly, both *CG13333* and the second uncharacterized gene *CG4440* are expressed in ectodermal strips which are directly adjacent to the Delta-expressing ectodermal strips. This neighboring expression may represent *CG13333* and *CG4440* expression in Notch-receiving cells, rather than the Delta-sending cells, which would implicate these genes in Notch-regulated processes in development.

## Discussion

Our approach focused on sharp changes in transcript levels in order to identify genes that may be subject to tight regulation.

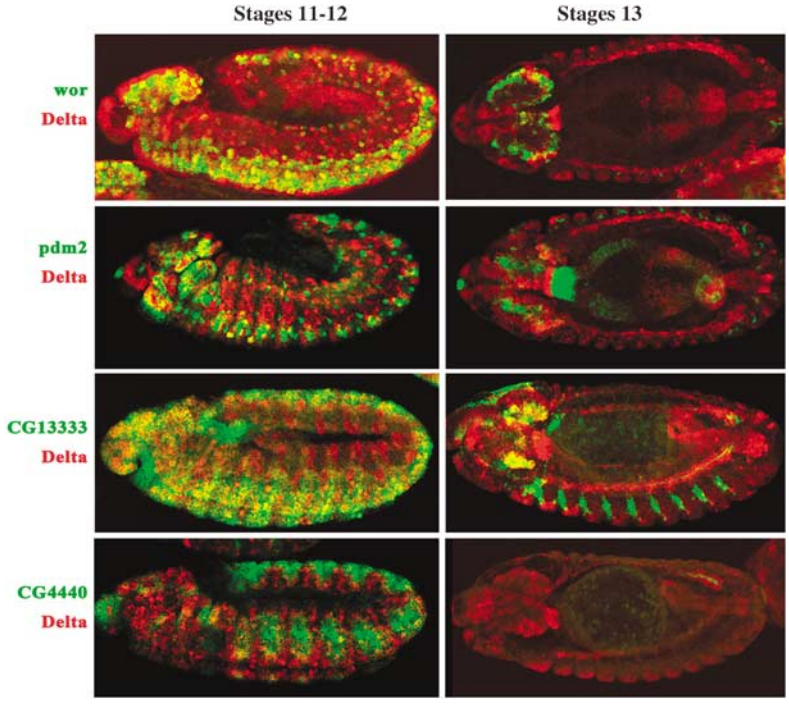


**Figure 4** Decline of transcript levels of four class **II:a** genes as predicted by the array analysis. Transcript levels are high until stage 12, but decline rapidly after stage 13.

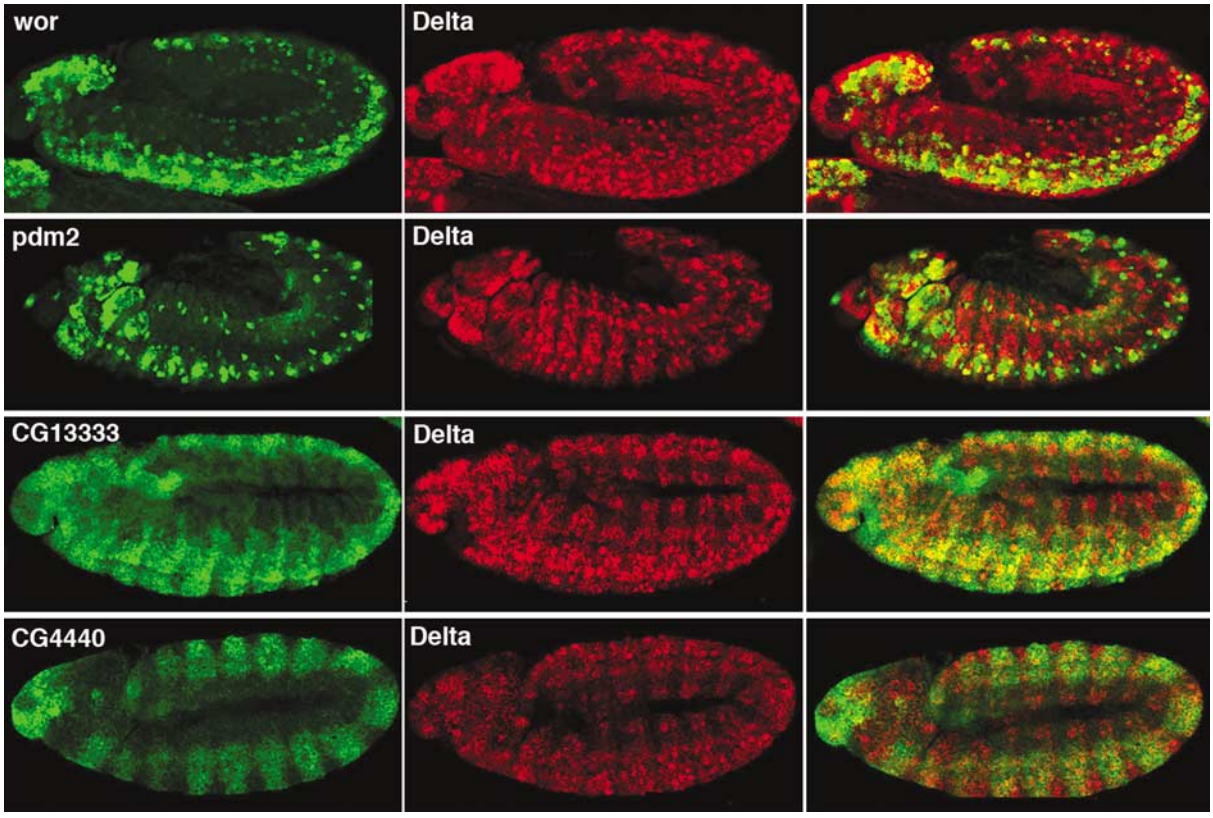


**Figure 5** Spatial colocalization of worniu, pdm2, CG13333 and CG4440 with Delta in embryos stage 11–12. Columns 1 and 2 show stainings individually and column 3 shows colocalization. CG4440 exhibits an anti-correlation, suggesting colocalization with Notch rather than Delta.

The functional signals (i.e. orthology, degree of interaction, etc.) that are shared among expression groups and also the signals that distinguish groups from each other lend support to our analysis strategy, despite inclusion of data of varying quality. Although these signals come at the cost of low sensitivity (groups are likely to have more members than this study can identify), the expression categories and groups are a starting point for exploring different functional features. For instance, we note that the transient class **II** genes are likely to play vital roles in development, based on the behavior of their transcripts. Overall, we find a strong consistency between the global clustering method, which is conceptually based on time-dependent data, and various sources of purely biological information. This further underlines the advantages of time-series arrays as opposed to non-temporal studies.

Owing to the low number of genes in some of the groups and to their limited annotation, we were unable to uncover the functionality behind the simultaneous decrease of one expression group and the increase of another (e.g. 12–14 h; class **II:a** and **I:b** transcript levels decrease, whereas **II:b** and **III:a** increase; Figure 2). However, our analyses clearly reveal the existence of these transitions between groups. The most dramatic of these expression changes appears to occur in the developmental stage 8 (12–14 h; Figure 2). This stage tends to be either the point of increase or decrease of transcript levels for many of the expression groups described above. The expression group class **II:a** decreases sharply at this point and contains a high number of genes encoding transcription factors, which most likely induce expression of various downstream pathways. Together with class **I:b**, whose transcript levels also decrease at that time, class **II:a** is best annotated. This group contains the highest fraction of orthologs and is enriched in lethal genes, suggesting that a coordinated decrease of transcript levels at this stage is essential for embryogenesis. The decrease of class **I:b** and class **II:a** transcript levels is followed by a burst of sharp transcript level increases from other groups of genes. There are many possible biological explanations for the distinct transition time between class **I:b**/**II:a** and class **III:a**/**II:b**. One major developmental event at this time is the end of the cell fate determination phase, coordinated by the Notch pathway. Not only do transcript levels decrease but also the proteins they encode, as we observe an enrichment of PEST motifs in the respective protein sequences indicative of a controlled degradation upon phosphorylation, whereas PEST motifs are underrepresented in proteins from class **III**. In this study, we have identified roughly 100 genes in class **II** that may be involved in cell determination, for instance in association with Notch or Delta. For a few of these, we have shown experimentally that this prediction is valid. A more detailed study of *in situ* patterns may provide insights into many more of these uncharacterized genes.

Taken together, our initial analysis of embryonic gene expression in *D. melanogaster* not only confirmed a number of known expression patterns but also revealed a surprisingly low number of well-defined expression groups that are sharply

activated and suppressed together. Some of them (class **I:b**, class **II:a**) contain many well-characterized genes, whereas others (class **II:b**, class **II:c**) reveal common temporal aspects of poorly annotated genes that should help in initiating more targeted functional studies. The coupling of sharp gene suppression and targeted protein degradation suggested here should enable network studies that combine temporal regulation and physical protein interaction.

# Materials and methods

## Microarrays and sample preparation

In this work, we used the same samples as described in detail by Arbeitman *et al* (2002). A brief summary is as follows. Canton S wild-type embryos were collected at 30 time points over a 24-h time period, with overlapping 1-h time points during the first 6.5 h, followed by hourly sampling. The stages of all samples were verified and only tightly staged embryo collections were used for RNA isolation and microarray analysis. The distribution of stages within each time points is shown in Figure 1. All samples were hybridized to a common reference sample. The reference sample, described by Arbeitman *et al* (2002), was made from pooled samples from all stages of the *Drosophila* lifecycle and therefore should represent a median level of expression for all genes in the genome. This serves as a constant denominator to which the relative levels of expression of each gene in the experimental samples can be prepared. The microarrays used for this study consist of PCR fragments of one exon of every predicted *Drosophila* gene from release one of the genome (for more details, see supplemental material of Li and White, 2003).

Array data are available online at Gene Expression Omnibus with the accession number GSE6186 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6186).

## Normalization of microarrays

The spot intensities of the two channels (Cy3 and Cy5) on each microarray were individually normalized using the Qspline method (Workman *et al*, 2002) with a log–normal distribution as target ($M$=ln 1000, $S$=ln 1000). The channels were further normalized to correct for spatial biases using a Gaussian smoother with sigma=0.8 (Workman *et al*, 2002). After adding a regularization background intensity of 100 to the normalized intensities, a log-ratio was calculated for each gene on each spotted array. This value was semi-empirically chosen to make the spread of log-ratios independent of the spot intensities.

## Identification of significantly regulated genes

We performed an ANOVA on each gene in order to determine significant changes in expression as has been carried out in earlier studies of time series (Arbeitman *et al*, 2002; Baugh *et al*, 2003). The resulting number of regulated genes was high (86% at $P<0.05$, 70% at $P<10^{-3}$), although Arbeitman *et al* (2002) reported even higher numbers (95% at $P<0.05$ and 86% at $P<10^{-3}$). The high number of genes with a significant change in expression level could suggest that an ANOVA is not sufficiently specific. We therefore also performed a runs test, which unlike ANOVA takes the temporal ordering of the data points into account and hence fits better with our subsequent analyses. Of 336 known transcription factors (not exclusively involved in embryogenesis), the runs test found 71% of them while suggesting 46% of all genes to be regulated ($P<0.05$). In comparison, ANOVA found 88% of the transcription factors, but suggested 86% of all genes to be regulated. In this case, ANOVA performs only barely better than random selection.

## Local convolution

In order to specify the times of activation and suppression, we convolved array data with vectors of eight integers, for instance $x$=[0 0 0 1 1 1 1]. We selected those matches with correlation coefficients exceeding 0.9, corresponding to $P<10^{-3}$ ($t = r/\sqrt{(1-r^2)/(N-2)}$, $t$-test). Throughout testing, trends remained when both the length of the x-vector and the correlation lower limit were either strengthened or weakened.

## Global convolution—supervised clustering

The goal of our clustering approach was to find transcripts that would be biologically easy to explain. In particular, we were interested in tying clusters of transcripts to specific stages of development, such as germ-band elongation or neurogenesis. Our strategy was therefore to cluster transcripts that had very consistent expression patterns. Consistency in this case would mean an unbroken state of, for example, high expression followed by an unbroken state of low expression, such as for the maternally inherited transcripts. The actual fold change of the transcript levels was not as important as the consistency requirement, meaning that we could study genes with less dramatic transcript levels than traditional clustering. The disadvantages of traditional clustering include (a) random transcript spikes leading to predictions that were difficult to assess biologically and (b) effects of noisy data points leading to false positives. Using a supervised clustering technique would classify these cases as true negatives.

When searching for plateaus of expression, we convolved the expression profiles according to $S_{i,j}=c(e, x_{i,j})$, where $e$ is the expression profile vector and

$$x_{i,j} = \begin{cases} 1 & \text{for } i, \ldots, j \\ 0 & \text{otherwise} \end{cases}, \; j>i$$

Here, $c$ is the correlation coefficient function. The expression profile is considered to be active from $i, \ldots, j$ if the maximum of $S$ exceeds 0.8 ($P<10^{-4}$, $t$-test). For example, Supplementary Figure S1 illustrates a high correlation between an expression profile and the filter vector $x_{i,j}$, where $i$=18 and $j$=25. The maximum value of $S$ in this case is 0.95. This approach is conceptually similar to the method employed by Šášik *et al* (2002), although here we actively look for steady plateaus followed by sharp declines in expression. Furthermore, an advantage of using correlation coefficients is that $P$-values are given, removing the need for random sampling. This method of clustering was chosen as it does not assume Euclidean distance between genes. A Euclidean distance implies that there is no time dependency, which is not consistent with our expectations. Finally, as mentioned above, the resulting clusters are more readily explained biologically as contiguous phases of up- and downregulation.

For class **II**, in order to distinguish transcript expression profiles from those that are maternally deposited, we set the first six data points to low expression and varied the remaining 24. Hence, the number of combinations is 276 and not 435. For class **III**, we required that expression be still high at the end of the time series.

## Database resources

### *In situ* data

The *in situ* data were retrieved from the Berkeley database (Tomancak *et al*, 2002). This database contains *in situ* data annotation for 2152 genes with 211 anatomical terms that are based on pictures taken at five different developmental times. Data are available for 253 out of our 842 significantly correlated genes.

### Pathways

A total of 1309 genes grouped in 31 pathways were retrieved from the Interactive Fly database (http://flybase.bio.indiana.edu/allied-data/lk/interactive-fly/aimain/1aahome.htm). Out of these, 86% are spotted on the microarray.

### Orthology

Orthologs in *A. gambiae* were retrieved from the STRING database (von Mering *et al*, 2005).

### Lethality
We consider all genes as lethal that were annotated as 'Phenotypic class: lethal' in Flybase (Drysdale and Crosby, 2005) unless the time of manifestation was given and stated manifestation of the phenotype only in the larvae or later stages. In total, 711 genes of the 12 868 genes (5.5%) tested were considered as lethal.

### Transcription factors
A total of 336 genes annotated as 'transcription factor' were extracted from FlyBase (Drysdale and Crosby, 2005) and manually curated (Tobias Doerks, personal communication).

### GO annotation
Overrepresentation of GO categories was analyzed using the GOSSIP program (Bluthgen *et al*, 2005), correcting for multiple testing using FDR and FWER.

### Protein–protein interactions
The interaction network was obtained from release 6.2 of the STRING database (von Mering *et al*, 2005), which includes curated data from the yeast two-hybrid screen by Giot *et al* (2003) and many individually reported interactions. All interaction data for *D. melanogaster* were used with the exception of links inferred from mRNA coexpression data (Arbeitman *et al*, 2002). The cutoff for STRING scores was set at 0.3. The network figure was created using Cytoscape 2.1 (Shannon *et al*, 2003) and Medusa (Hooper and Bork, 2005).

### PEST degradation signals
The PEST-find program (Rechsteiner and Rogers, 1996) was used to perform proteome-wide computational search for PEST regions. The number of PEST-containing proteins was counted among all significantly regulated genes as well as within each 'expression class'. The statistical significance of PEST overrepresentation was calculated for each 'expression class' compared to all significantly regulated genes using the exact hypergeometric test.

### Supplementary information
Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

## Acknowledgements
We thank Tobias Doerks for his transcription factor expertise, as well as Florian Raible and members of the Bork group for helpful discussions. This work was partly supported by EU grants LSHG-CT-2003503329 and QLRT-2001-02062. SD Hooper was supported by the Knut and Alice Wallenberg Foundation.

## References
Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297:** 2270–2275

Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP (2003) Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development* **130:** 889–900

Bluthgen N, Kielbasa SM, Herzel H (2005) Inferring combinatorial regulation of transcription *in silico*. *Nucleic Acids Res* **33:** 272–279

Brody T (1999) The Interactive Fly: gene networks, development and the Internet. *Trends Genet* **15:** 333–334

Dorer DR, Rudnick JA, Moriyama EN, Christensen AC (2003) A family of genes clustered at the Triplo-lethal locus of *Drosophila melanogaster* has an unusual evolutionary history and significant synteny with *Anopheles gambiae*. *Genetics* **165:** 613–621

Drysdale RA, Crosby MA (2005) FlyBase: genes and gene models. *Nucleic Acids Res* **33:** D390–D395

Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamodar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley Jr RL, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM (2003) A protein interaction map of *Drosophila melanogaster*. *Science* **302:** 1727–1736

Hooper SD, Bork P (2005) Medusa: a simple tool for interaction graph analysis. *Bioinformatics* **17:** 656–657

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* **443:** 594–597

Joutel A, Tournier-Lasserve E (1998) Notch signalling pathway and human diseases. *Semin Cell Dev Biol* **9:** 619–625

Li TR, White KP (2003) Tissue-specific gene expression and ecdysone-regulated genomic networks in *Drosophila*. *Dev Cell* **5:** 59–72

de Lichtenberg U, Jensen LJ, Brunak S, Bork P (2005) Dynamic complex formation during the yeast cell cycle. *Science* **307:** 724–727

Luschnig S, Moussian B, Krauss J, Desjeux I, Perkovic J, Nusslein-Volhard C (2004) An F1 genetic screen for maternal-effect mutations affecting embryonic pattern formation in *Drosophila melanogaster*. *Genetics* **167:** 325–342

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33:** D433–D437

Ostrowski S, Dierick HA, Bejsovec A (2002) Genetic control of cuticle formation during embryonic development of *Drosophila melanogaster*. *Genetics* **161:** 171–182

Overton PM, Meadows LA, Urban J, Russell S (2002) Evidence for differential and redundant function of the Sox genes Dichaete and SoxN during CNS development in *Drosophila*. *Development* **18:** 4219–4228

Rechsteiner M, Rogers SW (1996) PEST sequences and regulation by proteolysis. *Trends Biochem Sci* **21:** 267–271

Šášik R, Iranfar N, Hwa T, Loomis WF (2002) Extracting transcriptional events from temporal gene expression patterns during *Dictyostelium* development. *Bioinformatics* **18:** 61–66

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, Bussemaker HJ, White KP (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306:** 655–660

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, Rubin GM (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol* **3**, RESEARCH0088

Wolpert L, Beddington R, Jessell T (2002) *Principles of Development*. New York: Oxford University Press

Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, Saxild HH, Nielsen C, Brunak S, Knudsen S (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol* **3**, RESEARCH0048

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Lai Z, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, Yandell M, Collins FH, Ribeiro J, Gelbart WM, Kafatos FC, Bork P (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298:** 149–159