Journal of Biology

BioMed Central

Research article

Evidence for large domains of similarly expressed genes in the *Drosophila* genome

Paul T Spellman and Gerald M Rubin

Address: Howard Hughes Medical Institute and Department of Molecular and Cell Biology, University of California, Berkeley CA 94720-3400, USA.

Correspondence: Paul T Spellman. E-mail: spellman@bdgp.lbl.gov

Published: 18 June 2002

Journal of Biology 2002, **1**:5

The electronic version of this article is the complete one and can be found online at http://jbiol.com/content/1/1/5

ISSN 1475-4924

Received: 28 March 2002
Revised: 7 May 2002
Accepted: 17 May 2002

Abstract

Background: Transcriptional regulation in eukaryotes generally operates at the level of individual genes. Regulation of sets of adjacent genes by mechanisms operating at the level of chromosomal domains has been demonstrated in a number of cases, but the fraction of genes in the genome subject to regulation at this level is unknown.

Results: *Drosophila* gene-expression profiles that were determined from over 80 experimental conditions using high-density oligonucleotide microarrays were searched for groups of adjacent genes that show similar expression profiles. We found about 200 groups of adjacent and similarly expressed genes, each having between 10 and 30 members; together these groups account for over 20% of assayed genes. Each group covers between 20 and 200 kilobase pairs of genomic sequence, with a mean group size of about 100 kilobase pairs. Groups do not appear to show any correlation with polytene banding patterns or other known chromosomal structures, nor were genes within groups functionally related to one another.

Conclusions: Groups of adjacent and co-regulated genes that are not otherwise functionally related in any obvious way can be identified by expression profiling in *Drosophila*. The mechanism underlying this phenomenon is not yet known.

Background

The regulation of gene expression is a fundamental process within every cell that often allows exquisite control over a gene's activity (for review see [1]). Altering transcription rates is an effective strategy for regulating gene activity. It is well established that transcription of a given gene is dependent upon a promoter sequence located within a few hundred base pairs of the transcriptional start site. Promoter activity is modulated by sequence-specific transcription factors that physically interact either with the protein complexes that make up the core transcriptional machinery or with the promoter sequence itself.

In eukaryotes, the activity of a promoter can be modified by transcription factors binding to DNA sequences (frequently termed *cis*-regulatory modules or enhancers) that are located from hundreds to hundreds of thousands of base pairs away from the promoter. These regulatory modules can either increase or decrease the rate of transcription for a target gene, depending on the cellular state and the activities of the bound transcription factors. There are several mechanisms by which transcription factors bound to regulatory modules exert their effects. First, many transcription factors interact directly with the core transcriptional machinery by recruiting the latter's protein complexes to the promoter. Second, transcription factors may bend or twist the DNA, altering the way in which other transcription factors interact with the DNA. Finally, transcription factors can alter local chromatin structure by modifying histones (typically through methylation, acetylation, and substitution of histone subunits) to permit or restrict access to the DNA. Modifications of chromosome structure also occur at much larger scales. Most eukaryotes exhibit distinct chromosomal regions that are usually either transcriptionally active (euchromatin) or inactive (heterochromatin). In animals, heterochromatin is typically found near centromeres and other regions of low sequence complexity.

Less clear are the mechanisms by which the regulation provided by a *cis*-regulatory module is restricted to specific target genes. Several examples of insulators - sequences that prevent neighboring modules from affecting transcription - have been identified (reviewed in [2]). Insulators seem to function not by deactivating *cis*-regulatory modules but by preventing their influence from being propagated along the chromosome. It is not known how common insulators are in the *Drosophila* (or any other) genome. Some insulator-binding proteins localize to a few hundred chromosomal positions, and these positions coincide with genomic sequences that are not heavily compacted by chromatin structure (the 'interbands' of polytene chromosomes) [3]. There is substantial evidence that, although gene expression can be tightly controlled, neighboring genes or chromatin regions are important for the expression of individual genes. For example, otherwise identical transgenes inserted into different chromosomal sites show varying levels of expression [4].

Two recent observations lend credence to the idea that genomes may be divided into domains important for controlling the expression of groups of adjacent genes. First, there is evidence from budding yeast that some genes are found in pairs or triplets of adjacent genes that display similar expression patterns [5]. Second, about 50 much larger regions of the human genome show a strong clustering of highly expressed genes [6], which is caused by clustering of genes that are expressed in nearly all tissues [7]. We have examined the fraction of genes in the *Drosophila* genome that are subject to regulation that reflects large domains, using data from high-density oligonucleotide microarrays that reflect over 80 experimental conditions, and have found more than 20% of the genes clustered into co-regulated groups of 10-30 genes.

Results

Many neighboring genes show similar expression patterns

We collected relative gene-expression profiles covering 88 distinct experimental conditions from 267 Affymetrix GeneChip Drosophila Genome Arrays (see Materials and methods section). When the genes in this dataset were organized according to their positions along the chromosome, we observed numerous groups of physically adjacent genes that shared strikingly similar expression profiles. We sought to measure the magnitude of this effect by identifying all groups of physically adjacent genes that showed pair-wise correlations between their expression profiles that were higher than expected by chance.

Visual inspection of the entire dataset using TreeView software [8] revealed that groups of adjacent genes with similar expression patterns appeared frequently in our real dataset but rarely in a randomized dataset. The size of these groups varied, but appeared to average about 10 genes. In order to systematically identify groups of adjacent, similarly expressed genes, we calculated the average pair-wise Pearson correlation of gene expression for genes in a sliding ten-gene window across the genome. The Pearson correlation is a commonly used metric for determining the similarity between two gene expression profiles [8], and the average pair-wise correlation is the average of the Pearson correlations of all 45 possible pairs of genes within the ten-gene set. We estimated the probability of the average correlation scores by randomly sampling one million times from the dataset and calculating the average pair-wise correlation for windows of ten genes. We also created a random dataset of the same size, by randomly shuffling the associations from genes to expression profiles, and used this to illustrate the significance of our results. Our analyses show that groups of physically adjacent genes with similar expression are common; nearly 1,100 such groups are significant at a p value of 10^{-2} (Table 1). In more conservative analyses (requiring an uncorrected p value of 10^{-4}), where we expect to observe only one group by chance, in fact we observed 124 groups (Table 1).

To ensure that ten-gene windows were appropriate, we repeated the analysis using windows of various sizes. As

Table 1

The number of ten-gene groups of adjacent, similarly expressed genes that are found in ordered and randomized datasets, or are expected to be found in a randomized dataset

Significance (*p* value)	Ordered dataset	Randomized dataset	Expected
10^{-4}	124	0	1
10^{-3}	352	6	13
10^{-2}	1,077	106	130

The 'Expected' column gives an approximate number.

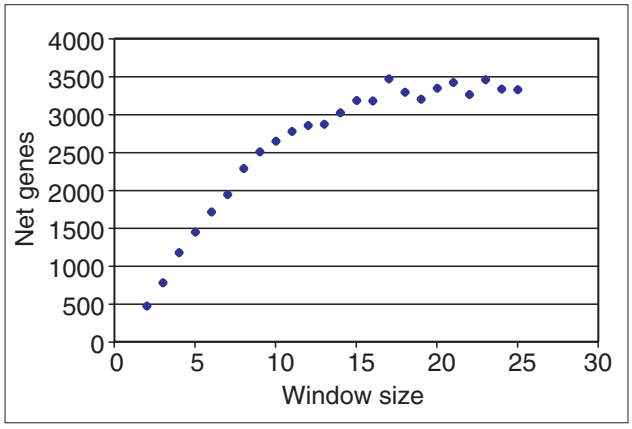

Figure 1
The number of genes identified as being in groups when different window sizes are used. In order to identify groups of adjacent, similarly expressed genes, the average pair-wise correlation of gene expression was calculated for genes in a sliding window across the genome, and this process was repeated for windows of different sizes. The net number of genes (that is, the number of genes in groups in the ordered dataset minus the number of genes in groups from the random dataset) is plotted against window size.

the window size increases from two to eight genes, the net number of genes in groups (that is, the genes in groups in the ordered dataset minus genes in groups from the random dataset) increases linearly. At a window size of about ten genes, the net number of genes begins to plateau (Figure 1). This suggests that most groups include about ten genes, so we used a window size of ten for the remainder of our analysis. There are no qualitative differences in the nature of groups identified by larger window sizes.

Many of the ten-gene groups that have high average pair-wise correlations of gene expression represent physically overlapping stretches of genes (that is, genes n through $n + 9$ make up one group and genes $n + 1$ through $n + 10$ form another). For all further analyses, therefore, we collapsed all groups that bordered one another into a single group. This substantially reduced the number of groups, showing that the effect on expression extends well beyond ten genes (Table 2). Nearly 1,100 ten-gene groups are significant at $p < 10^{-2}$, but these collapse into only 211 groups with an average group size of greater than 15 genes. As the p values decrease the average group size also decreases, but even at $p < 10^{-4}$ there are, on average, 12 genes in each group (553 genes in 46 groups; see Table 2).

The 44 groups (681 genes in total) that map to the left arm of chromosome two and have a p value of less than 10^{-2} are shown, using a ratiogram [8] aligned to the chromosome arm, in Figure 2. The distribution of groups along the chromosomes appears random and there is little bias for genes in a group to be on the same strand. The length of genomic sequence occupied by similarly expressed gene groups is highly variable. The average group size is nearly 125 kilobase pairs (kbp) in length, with a standard deviation of about 90 kbp, while the smallest group is 22 kbp and the largest is over 450 kbp. As might be expected, there is a relationship between the number of genes in a group and the length of genomic DNA covered by each group (Pearson correlation 0.59).

Gene groups are not explained by gene function or homology

Many genes that are related by function share similar expression patterns, and it is plausible that the same is true for homologous genes, particularly those that arose from recent duplications. In *Drosophila* there are 2,207 genes for which there is a homolog within the genome and the two homologs are separated by less than 10 genes. To determine whether homologs account for our observations, we repeated our analysis on a dataset from which homologs that are physically near one another were removed. This dataset is just under 12,000 genes, and although there is a significant decrease in the numbers of

Table 2

The number of groups of genes, and total numbers of genes in groups, that are identified at various levels of significance (*p* values)

	Groups		Genes	
Significance (*p* value)	Ordered dataset	Randomized dataset	Ordered dataset	Randomized dataset
10^{-4}	46	0	553	0
10^{-3}	93	5	1,219	51
10^{-2}	211	53	3,228	586

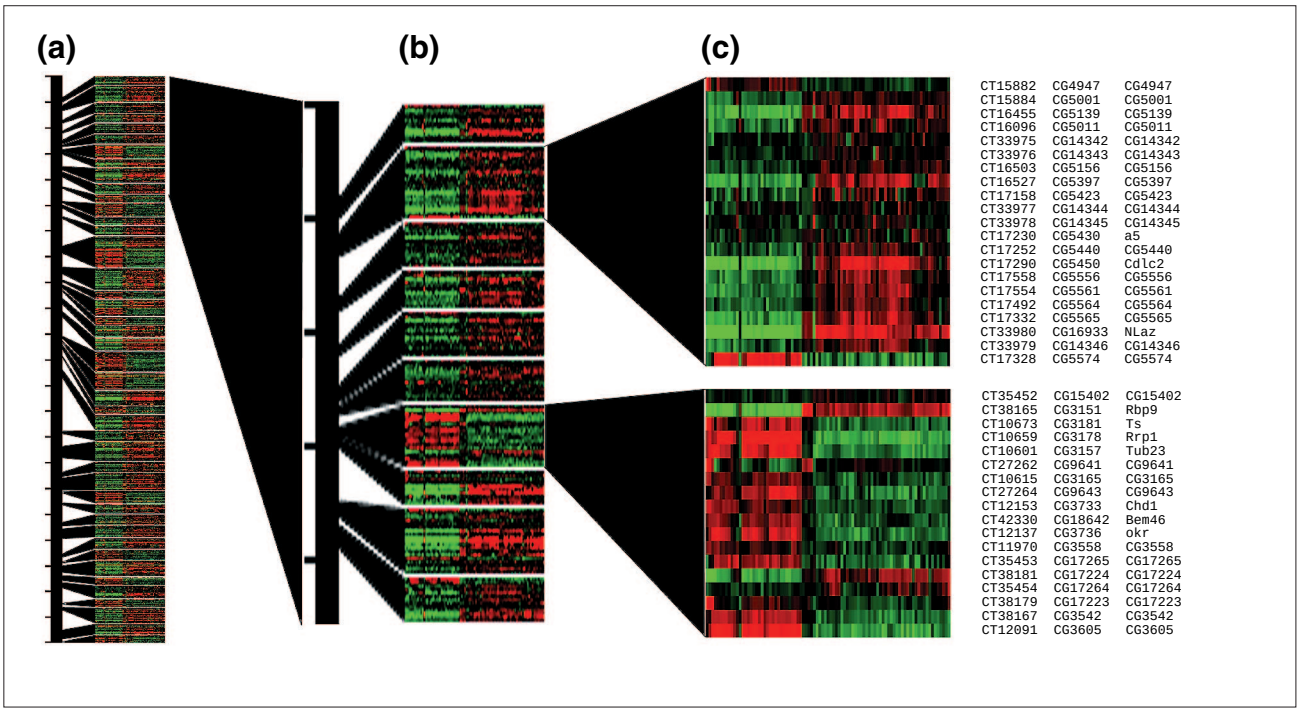

Figure 2
Similarly expressed adjacent genes on the left arm of *Drosophila* chromosome 2 (2L). **(a)** Ratiograms show the relative expression of all gene groups on 2L that are significant at $p < 10^{-2}$. In each ratiogram, columns represent individual experimental conditions and rows represent individual genes. For each square on the resulting 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. The black bar represents the chromosome, and the ticks along its left side mark 1 megabase (Mb) distances. The black shapes link the positions of groups on 2L to the expanded views of certain groups that are shown in **(b,c)**. (b) An expanded view of about 5 Mb. (c) The genes in two groups are shown in detail. The CT (computed transcript identifier), CG (computed gene identifier), and gene name are shown for each of the genes in these two groups. Each of the two expanded sections represents one group.

genes found to be in groups in this dataset (Table 3), 176 groups remain, containing about 2,500 genes.

We considered an extreme model to account for our observations - that evolutionary selection has organized gene groups according to the biological processes the genes are involved in, so that their expression can be coordinately regulated. We sought to test this model using the Gene Ontology (GO) database [9,10] as a source of annotations of biological processes. We first used the hypergeometric distribution to calculate the probability of observing each GO term as enriched in each group, on the basis of the number of genes in the group, the number of genes in that group that are annotated with that GO term, and the number of genes in the genome that are annotated with that GO term. We then selected all GO 'process' terms associated with a group at $p < 0.05$ where at least two genes had the selected GO term. Of the 211 groups identified in our full dataset and the 176 groups from the 'homologs-removed' dataset, 43 and 11 GO terms, respectively, have associations to groups that meet the above criteria. These numbers are modestly higher than would be expected by placing a random selection of genes into groups, where we would expect 7 ± 2 from the full dataset and 4 ± 2 from the homologs-removed dataset. The observed enrichment is clearly dependent on homologs, however, given the nearly four-fold decrease in observed associations when homologs are excluded from the analysis. Thus, with the present level of functional annotation, the vast majority of gene groups we observe are not composed of genes with similar biological processes, and the extreme model is not supported.

Similarly expressed gene groups can be identified from smaller datasets

Our dataset is derived from RNA samples taken from embryos or adults (primarily males). The groups in our dataset show a pattern of gene expression that mirrors this

Table 3

The number of groups of genes, and total numbers of genes in groups, from a dataset containing no physically close homologs

	Groups		Genes	
Significance (*p* value)	Ordered dataset	Randomized dataset	Ordered dataset	Randomized dataset
10^{-4}	18	2	200	21
10^{-3}	62	7	767	80
10^{-2}	176	49	2,561	576

bifurcation: most genes are expressed at higher levels in either adults or embryos. We wished to determine whether our observations of groups reflect this division, so we divided our dataset in two, creating one dataset of 'embryo' experiments and one of 'adult' experiments. It should be noted that four of the adult experiments contained RNA from males and from females, which contain a substantial number of oocytes, whereas the rest of the dataset was only from males. We calculated the average pair-wise correlations for all groups of genes in each of the two new datasets; Table 4 summarizes the number of genes in groups for the embryo and adult datasets (both randomized and ordered). The gene numbers are remarkably similar to those found for the entire dataset, as are the numbers of groups (see the Additional data files with this article online).

We wished to know if there was a correlation between the gene groups identified in the adult, embryo, and combined datasets. To do this we tabulated all genes identified in each dataset at each of three *p* values (10^{-2}, 10^{-3} and 10^{-4}) and calculated the Pearson correlation between each pair of datasets at each *p* value (Table 5). The average correlation between either the embryo or the adult dataset and the combined dataset is about 0.35, while the average correlation between the adult and embryo datasets is lower (about 0.23). The number of genes involved makes little difference, because the correlations are similar at each *p* value, despite the vastly different numbers of genes identified at different *p* values. In all, 890 genes are present in a group defined by one of the three datasets at $p < 10^{-4}$. After correcting for genes expected to be found in groups by chance, about 2,250 genes are identified in one of the three datasets at a *p* value of 10^{-3} and about 4,000 genes are identified at 10^{-2}.

Correlations with known chromosome structures

We attempted to determine whether the locations of similarly expressed gene groups correlate with known chromosome structures. Polytene chromosomes show a distinct, reproducible pattern of extended and compacted regions. The compacted regions contain the vast majority of the DNA, although the amount of DNA in each band can vary by more than one order of magnitude. The mean DNA content of each band is approximately 25 kbp [11,12] as compared with approximately 125 kbp for each group of co-expressed genes. We calculated the number of bands that overlap (or are contained in) each group and compared this with the number of bands that overlap (or are contained in) a randomly placed group matched for size. There was very little difference in the average number of bands overlapping each co-expressed group or each randomly placed group (5.9 versus 6.6).

It has been proposed that *Drosophila* chromosomes are attached to a nuclear scaffold at precise locations [13], but there is very limited mapping data on the position of these attachments. Mirkovitch *et al.* [13] mapped four attachment sites in a 320 kbp region near the *rosy* gene on chromosome 3R, dividing the region into a number of discrete domains of average size 50 kbp, each containing many genes. We wished to determine whether the groups we identified might correspond to distinct regions between attachment sites, as several of our groups fall in the region

Table 4

The number of genes within groups identified in either 'adult' or 'embryo' experiments

	Embryo		Adult	
Significance (*p* value)	Ordered dataset	Randomized dataset	Ordered dataset	Randomized dataset
10^{-4}	285	0	371	0
10^{-3}	1,159	52	1,139	114
10^{-2}	3,108	686	3,144	938

Table 5

The correlation between sets of genes identified in the adult, embryo and combined datasets

Significance (*p* value)	Combined: adult	Combined: embryo	Adult: embryo
10^{-4}	0.33	0.41	0.24
10^{-3}	0.34	0.34	0.23
10^{-2}	0.38	0.28	0.22

studied by Mirkovitch *et al.* [13]. We attempted to align these regions but there are no clear overlaps; the sizes and positions of the domains identified between attachment sites did not correspond to the groups we found.

Discussion

We have found that over 20% of the genes in the *Drosophila* genome appear to fall into groups of 10-30 genes such that the genes within each group are expressed similarly across a wide range of experimental conditions. Our data do not reveal the mechanism(s) responsible for the observed similarities in expression of adjacent genes but we believe the findings are most consistent with regulation at the level of chromatin structure, for the following reasons. First, the regions showing similarities in expression are quite large, containing on average 15 genes, with each gene presumably having its own core promoter. Second, it is frequently the case that one or two genes in a group display a high level of differential expression (see Figure 2c). If the chromatin in a region of the chromosome that contained many genes was 'opened' so that a single target gene could be expressed, it might increase the accessibility of the promoters and enhancers of other genes to the transcriptional machinery, leading to modest parallel increases in their expression. Such an effect could account for the observations we have made.

Discussions of transcriptional regulation often emphasize the belief that the process is tightly controlled and essentially error-free. We believe that the degree of precision, at least at a quantitative level, may be less than is generally assumed. For example, only a few genes show an obvious phenotype when heterozygous, and heterozygosity generally results in a two-fold reduction in expression level [14]. Moreover, there are numerous examples in the literature of genes that, when misexpressed either temporally or spatially, do not generate a phenotype. Although it is difficult to prove that individuals carrying such traits are as fit as their normal relatives, it is likely that the precise regulation of many genes is allowed to vary considerably. If we presume that the groups we have observed arise because of selection on the regulation of a small subset of genes in each group, then the vast majority of genes are in effect being 'carried along for a ride'. The regulation of transcription may be precise when it is needed and sloppy when it is not important.

If coordinated gene expression is unimportant, there should be no selection that drives the groups of co-regulated genes we observed to be evolutionarily conserved. It will be possible to test this when the *D. pseudoobscura* sequence is completed. If the groups of genes we identify here are found to be more syntenic in the *D. melanogaster* and *D. pseudoobscura* genomes than expected, that would support the idea that the observed coordinated expression is advantageous.

Although we have assayed a relatively large number of biological samples, we cannot infer the profiles of unique cellular states. As further experiments are carried out it may be that our observation of similarly regulated groups will grow to include all genes - that is, the entire euchromatic genome may be structured in such domains.

Materials and methods

Data collection

We collected a dataset composed of 88 experimental conditions hybridized to a total of 267 GeneChip Drosophila Genome Arrays (Affymetrix, Santa Clara, CA, USA) [15]. This dataset came from six independent investigations that will be described in detail elsewhere (A. Bailey, personal communication; M. Brodsky, personal communication; [16]; E. De Gregorio personal communication; A. Tang, personal communication; and P. Tomancak, personal communication), which study five different experimental questions - aging, DNA-damage response, immune response, resistance to DDT, and embryonic development. Supplemental data including software used in this study and the underlying expression dataset is available at our website [17] and from the ArrayExpress database [18] with the accession id E-RUBN-1.

Data processing

Genes are represented on the GeneChip Drosophila Genome Array by one or more transcripts, which in turn are represented by a probe set. Each probe set has 14 pairs of perfect match (PM) and mismatch (MM) oligonucleotides. Data were collected at the level of the transcript, but for ease in the text, the data are referred to by gene. Intensity data for each feature on the array were calculated from the images generated by the GeneChip scanner, using the GeneChip Microarray Suite. These intensity data were loaded into a MySQL database where information on each of the features was also stored. The difference between the PM and MM oligonucleotides (probe pair) was calculated, and the mean PM-MM intensity for each array was set to a constant value by linearly scaling array values. The mean intensity of individual probe pairs was calculated across all arrays, and the $\log_2$ ratio of each value to this mean was stored. Next, all log ratios for each probe pair set (transcript) were averaged, creating one measurement for each transcript on each array. The final dataset was generated by averaging data

for each transcript on replicate arrays and subtracting the average log ratio of each gene in the dataset.

Definition of homologs

BLAST scores based on predicted protein sequence were obtained from Gadfly (Release 2) [19]. We used these scores to define a homolog pair as those gene pairs for which BLAST *E* values are less than 10^{-7}.

Identification of adjacent similarly regulated genes

We calculated the average pair-wise correlation of gene-expression profiles for all genes that were within *n* genes (an *n*-gene window) of one another using the Pearson correlation. Significance (*p* values) was estimated by sampling random sets of *n* genes 1 million times to determine the likelihood distribution for the dataset. We also calculated the average pair-wise correlation for a random dataset in which the associations between genes and expression profiles were shuffled. We have calculated the number of genes in groups at each of the three *p* values, namely 10^{-2}, 10^{-3}, and 10^{-4}, for window sizes ranging from 2 to 25 genes.

Next, we set out to show that homologs did not account for the increase in the number of gene groups with higher than expected average correlations. We searched for cases in which homologs (as defined above) were near each other in the genome by scanning the set of genes for each chromosome from one end to another. If a gene showed homology to another gene that appeared less than 10 genes ahead, it was removed from the dataset, although no break in gene order was created. For example, in a set of 11 genes where the third and fourth were homologs, gene 3 would be removed, and a ten-gene group would consist of genes 1, 2 and 4 through 11. In total, 1,369 genes were removed from the dataset. This 'homologs removed' dataset was subjected to the average pair-wise algorithm, as was a randomized version of it.

We also constructed two non-overlapping subsets of the total data matrix. All hybridizations were divided into either the 'embryo' or 'adult' dataset on the basis of the source of the RNA used in that hybridization. In total, 35 experiments remained in the embryo dataset and 53 experiments remained in the adult dataset. The random pair-wise correlation algorithm was applied independently to each of these datasets as well as to randomized versions of each dataset.

Significantly enriched GO terms among gene groups

GO terms for all genes were obtained from the GO database [10]. Using the hypergeometric distribution, the probability of observing each GO term with each group was calculated. Briefly, the probability *p* that a GO term is significantly enriched among a specified set of genes can be calculated with the following formula:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{A}{i}\binom{G-A}{n-1}}{\binom{G}{n}}$$

where *k* is the number of genes in the group, *G* is the total number of genes, *n* is the number of genes in the group with a given annotation and *A* is the total number of genes with a given annotation. Because many sets of GO terms (> 1,000) were tested on many groups of genes (> 200), there is a problem of multiple testing. All GO terms significantly associated with a group of similarly expressed genes at a *p* value of less than 5×10^{-4} were recorded.

Correlation of groups with known chromosomal structures

We determined the number of polytene bands present in each group of similarly expressed genes. The coordinates of each group were determined by using the transcription start sites (from GadFly Release 2) [19] of the genes at each end of a group. We then determined how many bands overlapped each group based on the positions reported [11,12]. We also calculated the number of bands that overlap randomly placed groups (with the same sizes as the real groups).

Additional data files

The following are provided as supplemental materials; a tab-delimited text file of the underlying expression data; the perl scripts used to process the data; and a text file used to generate Figure 2. All expression data are reported as log base 2 and are mean centered (the mean expression value for each gene in all experiments is zero). The first column of each expression data file is the CT identifier of each transcript. The second column is a description field, which includes the CT identifier, CG identifier, gene name, and brief Gene Ontology annotations. The remainder of the columns contain expression data, classified by the column header (either adult or embryo). The data used to generate Figure 2 can be loaded into the TreeView software [8] to visualize individual groups (null data rows indicate boundaries between groups). The software and underlying expression dataset are also available at our website [17] and from the ArrayExpress database [18] with the accession ID E-RUBN-1.

Acknowledgements

We thank Adina Bailey, Michael Brodsky, Amy Tang, and Pavel Tomancak for sharing data prior to publication. P.T.S. was a recipient of an NSF Biocomputing postdoctoral fellowship. G.M.R. is an investigator of the Howard Hughes Medical Institute.

References

1. Emerson BM: **Specifity of gene regulation.** *Cell* 2002, **109:**267-270.
2. Bell AC, West AG, Felsenfeld G: **Insulators and boundaries: versatile regulatory elements in the eukaryotic genome.** *Science* 2001, **291:**447-450.
3. Zhao K, Hart CM, Laemmli UK: **Visualization of chromosomal domains with boundary element-associated factor BEAF-32.** *Cell* 1995, **81:**879-889.
4. Spradling AC, Rubin GM: **The effect of chromosomal position on the expression of the *Drosophila* xanthine dehydrogenase gene.** *Cell* 1983, **34:**47-57.
5. 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.
6. Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus M-C, van Asperen R, Boon K, Vouöte PA, *et al.*: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains.** *Science* 2001, **291:**1289-1292.
7. 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.
8. 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.
9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
10. **The Gene Ontology Consortium** [http://www.geneontology.org/]
11. Ashburner M, de Grey A: **Cytological table used to infer a genetic map position from a published cytogenetic map position** [http://fly.ebi.ac.uk:7081/maps/lk/cytotable.txt]
12. Saura AO, Saura AJ, Sorsa V: **Electron micrograph maps of *Drosophila melanogaster* polytene chromosomes.** [http://www.helsinki.fi/~saura/EM/index.html]
13. Mirkovitch J, Spierer P, Laemmli UK: **Genes and loops in 320,000 base-pairs of the *Drosophila melanogaster* chromosome.** *J Mol Biol* 1986, **190:**255-258.
14. Lindsley DL, Sandler L, Baker BS, Carpenter AT, Denell RE, Hall JC, Jacobs PA, Miklos GL, Davis BK, Gethmann RC, *et al.*: **Segmental aneuploidy and the genetic gross structure of the *Drosophila genome*.** *Genetics* 1972, **71:**157-184.
15. **Affymetrix GeneChip Drosophila Genome Array** [http://www.affymetrix.com/products/arrays/specific/fly.affx]
16. De Gregorio E, Spellman PT, Rubin GM, Lemaitre B: **Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays.** *Proc Natl Acad Sci USA* 2001, **98:**12590-12595.
17. Spellman PT, Rubin GM: **Web supplement to "Identification of adjacent gene groups showing similar expression"** [http://www.fruitfly.org/expression/dse/]
18. **ArrayExpress** [http://www.ebi.ac.uk/microarray/ArrayExpress/arrayexpress.html]
19. **Gadfly** [http://www.fruitfly.org/annot/index.html]