Materials and Methods

1. Probe design

In a previous study, 36mer oligo arrays were shown to yield both high signal intensity and high sequence specificity when tested with a set of 96 Drosophila genes (Nuwaysir et al. 2002). 36mer probes do not suffer from the same degree of sequence-specific variability as traditional 25mer probes, yet identification of large numbers of unique probes is easier than with longer oligos such as 70mers. The 36-mer oligonucleotide probe sequences were obtained using an algorithm that minimizes the potential cross-hybridization of each probe to the rest of the genome. Sequence-dependent factors such as length, extent of complementarity and the overall base composition were used to optimize probe selection. The first generation of 36-mer oligonucleotide probes that were identified for the *Drosophila* genome contained probes that have unique subsequences of 18-mers that are found only once within the complete genome sequence. We used the NASA Oligo Probe Selection Algorithm (NOPSA) to scan the genome with single base intervals and created a database of the frequency of every 18-mer in the genome using hash algorithm, using chaining to resolve collisions (Weiss 1993). Then the genome was re-scanned again and the average frequency of a 36-mer, for every 36-mer in the genome was calculated from frequencies of each subsequence 18-mer within a 36-mer and its reverse complement. Thus, 36-mer oligonucleotides that have frequency equal to one were selected. For some regions in the genome there is no unique probe. The probe selection module of NOPSA was run a second time for those regions without unique probes. 36-mer oligonucleotides in those regions that have frequency equal to 2 were selected to cover the possible duplicated-gene regions. If probes with the above criteria could not be found in the regions, the probes with the least frequency were used to cover those regions in the genome. NOPSA code is freely available upon request. Contact Viktor Stolc [vstolc@mail.arc.nasa.gov].

2. Array Synthesis

Arrays were synthesized according to previously published procedures (Singh-Gasson 1999; Nuwaysir et al. 2002). Briefly, Standard DNA synthesis reagents (Glen Research, Sterling, VA), (Proligo, Boulder, CO), (Amersham Pharmacia, Piscataway, NJ), or (Applied Biosystems, Foster City, CA) were used on Expedite DNA synthesizers (Applied Biosystems). The photolabile phosophoramidites (NPPOC- dAdenosine (N6-tac) -Cyanoethylphosphoramidite, NPPOC-dCytidine (N4-Isobutyryl) -Cyanoethylphosphoramidite NPPOCdGuanosine
(N2-ipac) -Cyanoethylphosphoramidite, NPPOC-dThymidine--Cyanoethylphosphoramidite) were from Proligo. The MAS units (NimbleGen Systems, Madison, WI) were connected to the Expedites to manufacture the custom arrays. Arrays were designed with ArrayScribe™ software (NimbleGen Systems). After synthesis on the MAS was completed, the base protecting groups were removed in a solution of ethylenediamine:ethanol (1:1 v/v) (Aldrich, St. Louis, MO) for two hours. The arrays were rinsed with water, dried and stored desiccated until use.

3. Correcting for Probe Sequence Bias
To correct for probe sequence bias, we fit a position-dependent model to the subset of NEPs having a G+C content between 3 and 8 nucleotides out of 36 (GC3-8 probes). The linear model allows different segments along the probe to make independent contributions to non-specific binding. Fit parameters for the model were selected using a forward selection algorithm. Fitting was done separately for each channel (i.e. each unique combination of stage, array, and dye) and the resulting parameters were to correct the log-intensity of all probes (EP, NEP, SJP and NCP). These corrected values were used in sections 4 and 5.

4. Detecting probes expressed above background (PEAB)
For detecting significantly expressed probes, we took into account the fact that there was considerable variation in signal intensity between arrays, even when the same developmental stage was being assayed. We therefore computed p-values separately for each channel.

Negative control probes (NCPs) were defined as the subset of NEPs with a G+C content of 10 or 11 nucleotides out of 36 (GC10-11). For Table S4, GC3-8 probes were used as NCPs, yielding a less conservative analysis.

For each EP and NEP probe, a p-value was derived reflecting the likelihood that its intensity belong to the NCP distribution (SJPs were compared to the WJP distribution). Each probe therefore had 24 associated p-values, one for each combination of array and dye. These were combined into a single p-value using the Fisher method (Fisher 1950). The false discovery rate (FDR) formalism (Benjamini and Hochberg 1995) was applied to the resulting pooled p-values, using an FDR threshold of 5%. A more complete explanation of our analysis can be found at: http://bussemaker.bio.columbia.edu/papers/Science2004/

5. Detecting differential expression using ANOVA
To identify differentially expressed probes, we adapted the formalism of Kerr and Churchill (2001). For each probe, we set the gene-specific (G)model parameter to zero and calculated the variety-specific contribution to the probe's expression (VG).
We derived a distribution of VG estimates for each of the six stages using bootstrap sampling (5,000 iterations). This distribution was compared with the average of the six observed VG values to yield variety-specific p-values for each probe. In analogy with the procedure for determining significantly expressed probes, these p-values were combined using a Fisher test and a false discovery rate procedure was applied using an FDR threshold of 5%.

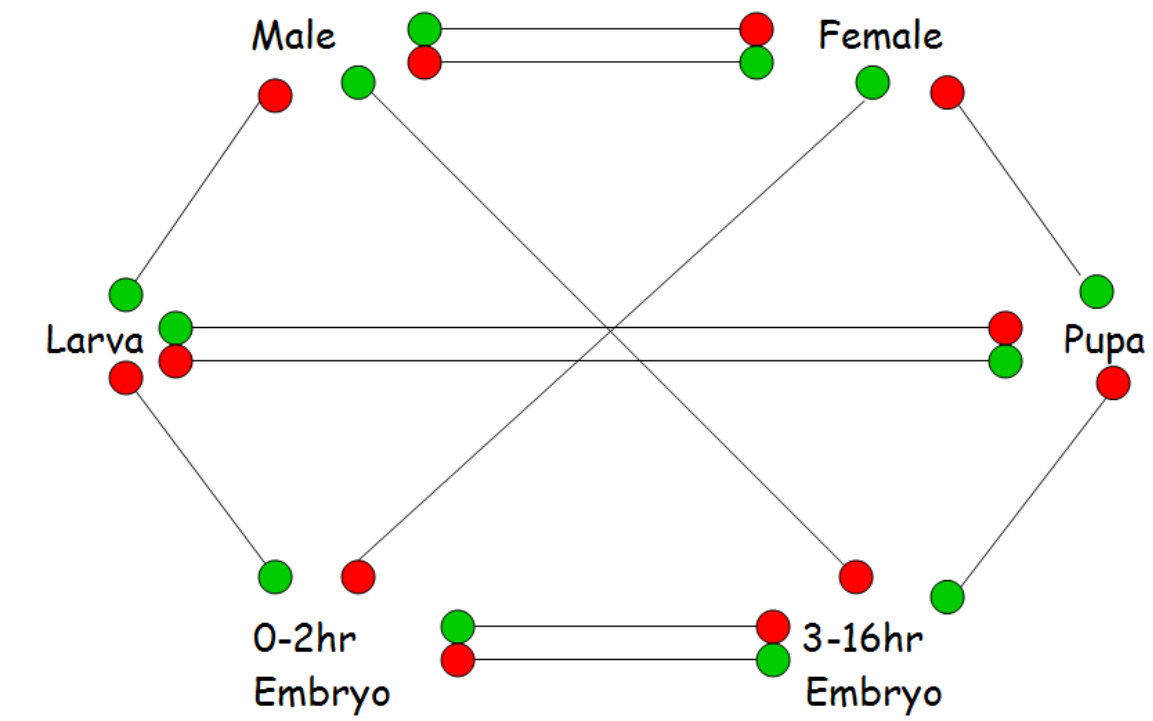
6. Pattern Separation Algorithm
Given the set of expression data from the exons of a gene, we developed a method to identify sets of exons sharing a common pattern of expression. Our

technique makes use of LLE (locally linear embedding) of the normalized data (Roweis et al. 2000). The six time-point data of each exon is normalized to other exons by considering the values minus their mean, divided by the sum of the squares, then graphed. The graph is analyzed to search for its components (or "clusters") that exhibit similar patterns (clusters), our sub-patterns.

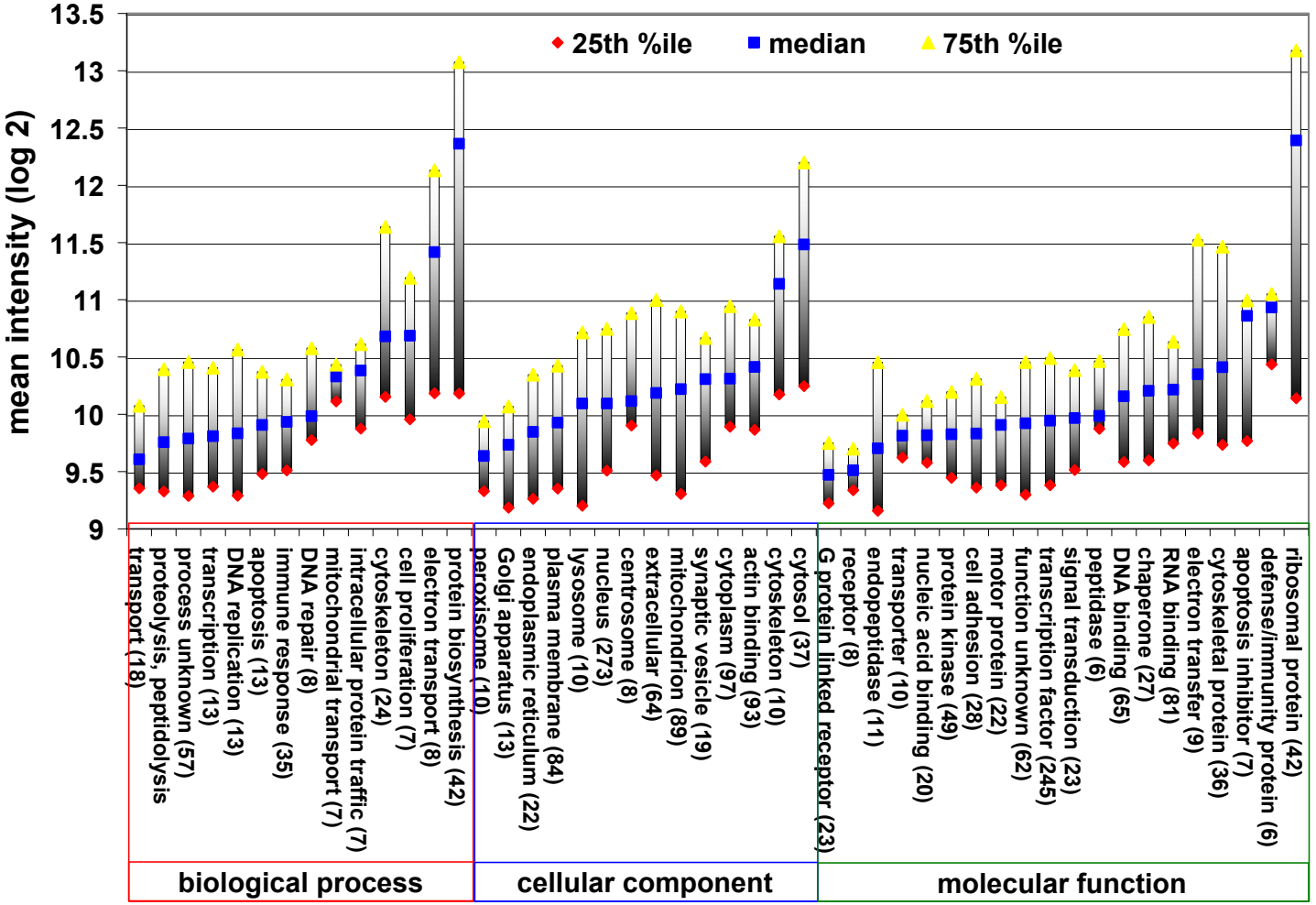
References

Y. Benjamini, Y. Hochberg, Journal of the Royal Statistical Society B, vol 75, No.1 pp 289-300 (1995)
R. A. Fisher, in Statistical Methods for Research Workers. (1950), vol. Section 21.1, pp. 99.
M. K. Kerr, G. A. Churchill, Proc Natl Acad Sci U S A 98, 8961 (Jul 31, 2001).
E. F. Nuwaysir et al., Genome Res 12, 1749 (Nov, 2002).
Roweis ST, Saul LK., Nonlinear dimensionality reduction by locally linear embedding. Science. 290(5500):2268-9 (Dec 22, 2000).
S. Singh-Gasson et al., Nat Biotechnol 17, 974 (Oct, 1999).
M. A. Weiss, "Data Structures and Algorithm Analysis in C", Chapter 5 Hashing, page 152-157 Benjamin/Cummings Publishing (1993)

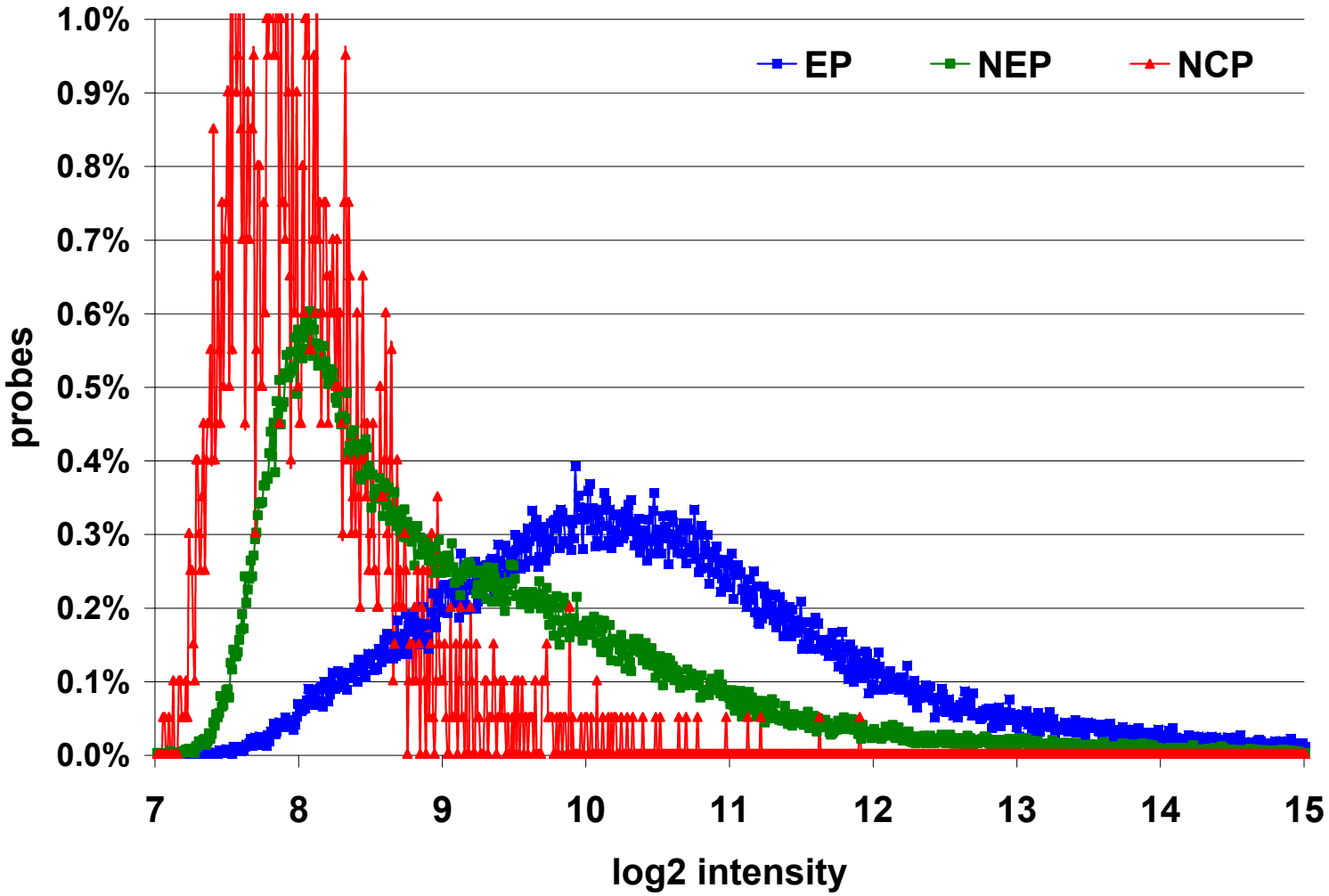
Supplemental 1: Experimental Design



Supplemental 2: Expression Levels by Gene Ontology (GO) category



Supplemental 3: Signal Intensity Distributions

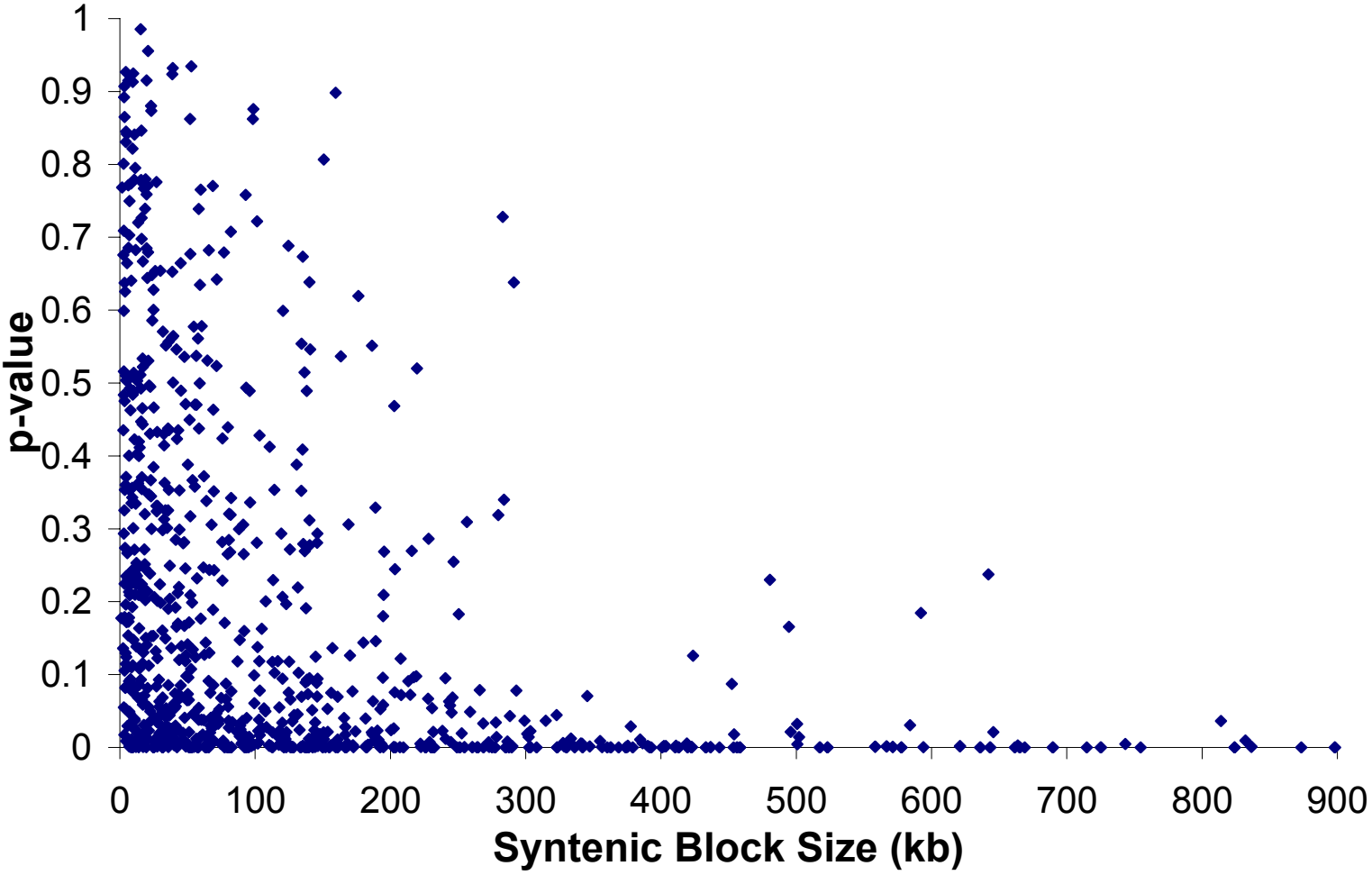


Supplemental 4: Genome-wide Statistics for Expressed Probes

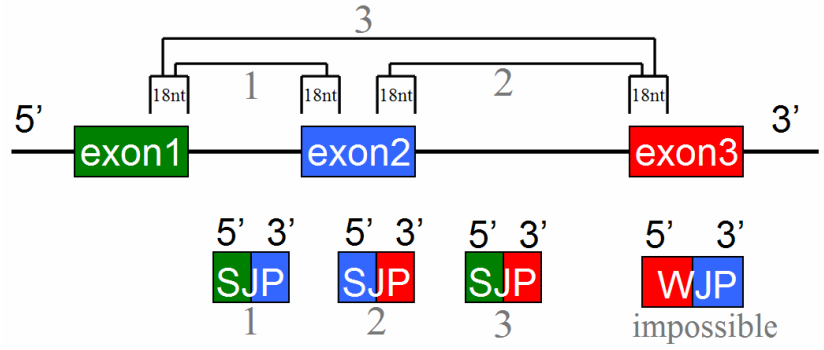
			Significant Probes (FDR=0.05)		
		Total Probes	**Probe Expression Above Background (PEAB)**	**Analysis of Variance (ANOVA)**	**Overlap: ANOVA & PEAB**
Probe Type	**NEPs**	87,814	48,241	6,789	5,953 (88%)
	EPs	61,371	53,381	27,176	25,554 (94%)

Lenient Background Model

Supplemental 5: Distribution of Gene Correlations in Syntenic Blocks



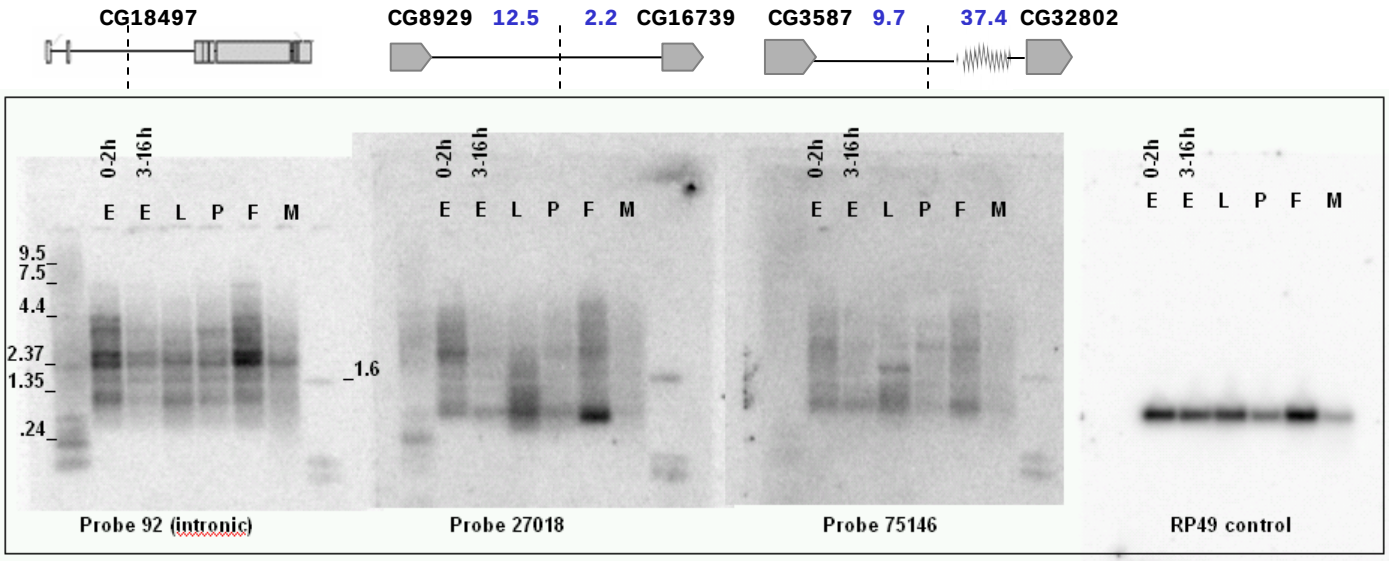
Supplemental 6: WJPs (Wrong Junction Probes) - Negative Controls for SJPs



Supplemental 7: Table of EP Activity Correlations to NEP Activity, by Pairwise Comparison

	Stage comparison														
	m-f	m-0	m-3	m-l	m-p	f-0	f-3	f-l	f-p	0-3	0-l	0-p	3-l	3-p	l-p
closest noncoding to an exon	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	7.3E-09	0.0E+00	5.6E-01	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00
closest noncoding 3' of an exon	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	2.9E-11	0.0E+00	1.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00	0.0E+00
closest noncoding 5' of an exon	2.3E-05	3.5E-10	1.9E-05	9.0E-06	1.2E-07	2.1E-07	2.5E-10	9.9E-01	5.9E-03	9.9E-01	0.0E+00	0.0E+00	0.0E+00	2.6E-11	6.4E-10
closest exon 5' of a noncoding	5.3E-03	7.2E-13	0.0E+00	2.5E-10	2.7E-12	1.8E-10	2.4E-03	1.7E-02	3.6E-06	1.0E+00	2.8E-04	3.9E-03	5.4E-03	1.5E-07	8.9E-08
closest exon 3' of a noncoding	1.0E+00	9.6E-01	1.2E-01	1.0E+00	9.9E-01	9.3E-01	1.0E+00	1.0E+00	9.8E-01	1.0E+00	8.4E-01	4.5E-03	1.0E+00	9.2E-01	1.0E+00
closest exon to a noncoding	1.4E-02	6.7E-11	2.5E-13	1.1E-08	2.4E-11	4.9E-08	3.1E-03	7.3E-03	3.6E-05	1.0E+00	2.2E-05	2.1E-02	2.5E-03	7.2E-06	4.7E-08

Supplemental 8: Verification of noncoding, expressed sequences



Supplementary Table 9: Expressed Sequences within cis-Regulatory Modules (CRMs)

CRM	CRM Strand	Array Probe ID	Chrom.	Position	Probe Strand
PD enhancer	+	66722	3R	26664683	+
ventral imaginal disc enhancer	-	56641	3R	12593982	+
late element 2	+	19650	2R	5042091	+
late 7-stripe element	+	93187	X	20361835	+
h stripe 5+1	+	36048	3L	8628474	+
boundary enhancer (vgBE)	+	21655	2R	7954168	+
dll 215 enhancer	+	29814	2R	19845396	+
prd stripe P enhancer	-	8198	2L	12077017	+
AD+PD enhancer	+	66723	3R	26666183	+
dll 304 early element	+	29815	2R	19846896	+
visceral mesoderm enhancer	-	56617	3R	12563002	+
run stripe 3+7	+	93183	X	20356921	+
ventral repression element (VRE)	-	49267	3R	2581374	+
run stripe 1+7	+	93179	X	20352520	+
stripe 5	+	19654	2R	5048097	+
AD2	+	30087	2R	20269660	+
ABX enhancer	-	56559	3R	12508280	+
epidermal autoregulatory enhancer	+	49288	3R	2612142	+
stripe 3+7	+	19647	2R	5036204	+
CD1	+	30085	2R	20266651	+
posterior enhancer	-	75472	X	2188331	+
late element 1	+	19646	2R	5034704	+
BXD enhancer	-	56628	3R	12575784	+
dll 208 enhancer	+	29816	2R	19848396	+
h stripe 3+4	+	36044	3L	8623437	+
tracheal enhancer	-	7706	2L	11438599	+
gsb late enhancer	+	29967	2R	20104111	+