Method

FlyGEM, a full transcriptome array platform for the *Drosophila* community

Rick Johnston[*], Bruce Wang[*], Rachel Nuttall[*], Michael Doctolero[*], Pamela Edwards[†], Jining Lü[†], Marina Vainer[*], Huibin Yue[*], Xinhao Wang[*], James Minor[*], Cathy Chan[*], Alex Lash[‡], Thomas Goralski[*], Michael Parisi[†], Brian Oliver[†] and Scott Eastman[*]

Addresses: [*]Incyte Genomics, Palo Alto, CA 94304, USA. [†]Laboratory of Developmental and Cellular Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, 50 South Drive, Room 3339, Bethesda, MD 20892, USA. [‡]Gene Expression Omnibus, National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20892, USA.

Correspondence: Brian Oliver. E-mail: oliver@helix.nih.gov

Published: 26 February 2004

Genome ***Biology*** 2004, **5:**R19

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/3/R19

Received: 17 October 2003
Revised: 16 January 2004
Accepted: 27 January 2004

Abstract

We have constructed a DNA microarray to monitor expression of predicted genes in *Drosophila*. By using homotypic hybridizations, we show that the array performs reproducibly, that dye effects are minimal, and that array results agree with systematic northern blotting. The array gene list has been extensively annotated and linked-out to other databases. Incyte and the NIH have made the platform available to the community via academic microarray facilities selected by an NIH committee.

Background

Several technologies, such as SAGE, ESTs, gene chips, and spotted arrays are in use to monitor global changes in mRNA expression patterns in a number of organisms including *Drosophila*. While all of these technologies are valuable tools, it is critically important to evaluate the accuracy, precision, and reliability of the data generated from each of them [1]. In a full-genome-scale analysis, errors of a few percent will generate hundreds of false readings. This may exceed the real biological changes one wishes to monitor. Understanding system performance allows the researcher to make provisions for suitable replication of the genomic assay in the experimental design. Similarly, understanding more pernicious artifacts that replicate, but do not accurately reflect, the underlying biology is critical for design of secondary screens.

We have constructed a fly gene expression microarray (FlyGEM) containing 94% of the release-1 predicted genes for *Drosophila melanogaster* [2], and 75% of the release-3 predicted genes [3]. We show that many of the predicted genes that were 'retired' between release-1 and release-3 are in fact expressed in array experiments, highlighting the ongoing chore of genome annotation. The FlyGEM is a spotted array, but differs substantially from other spotted cDNA arrays. Rather than amplifying cDNAs, we generated the DNA fragments used in microarray fabrication by PCR with exon-specific primers and genomic DNA. The exon-specific design allowed us to use sophisticated bioinformatic algorithms [4] to ensure that we not only cover most of the *Drosophila* genome, but also that most elements uniquely monitor expression of only one transcript. Because the sequences of the primers are known and amplification of the expected target sequence is easy to verify, the sequence of each array element is defined. This makes it easy to update the platform to conform with annotation gold standards at FlyBase. In addition, newly discovered genes or alternative exons can be

appended to the platform simply by adding additional amplicons to the set.

As with any genomic gene-expression platform, there are multiple process and biological variables that could adversely influence the quality of the data generated on the FlyGEM. Broadly classified, these would fall into the areas of array design, fabrication, sample handling and preparation, protocols, and biological variance. We have extensively investigated all aspects of FlyGEM performance, and this report presents data that directly measure the accuracy, precision, and reliability of FlyGEM data. The experimental design includes replicates for array elements, dye efficiency, labeling reactions and biological sample preparation. We find that most of the variability in FlyGEM results is due to the hybridization and labeling reactions. We strongly recommend replicate hybridizations, even if the array data are used as preliminary screens for cherry-picking genes of interest. Cy3/Cy5 dye effects are pervasive in many array experiments [5], but we find that using calibrated prelabeled short oligos for the labeling reactions and calibrated scanners effectively eliminates this variable. The correlation between FlyGEM results and northern blotting is high, suggesting that the false-positive and false-negative rates are low. Certainly, one would perform multiple types of assays if the goal were to advance candidate genes for extensive classical molecular genetic analysis. However, with replicate FlyGEM hybridizations, there is little need for systematic validation when addressing many genome-scale questions, such as gene-expression clusters or neighborhoods. The results presented here give us a high degree of confidence in overall platform performance.

Finally, wide access to both affordable arrays and array data is essential for the *Drosophila* community. A limited number of aliquots of the FlyGEM primers have been made available to regional, national and international centers at no cost, so that arrays may be manufactured by, and distributed to, the worldwide *Drosophila* community on a cost-recovery basis. These primers have been delivered to the *Drosophila* Genomic Resources Center in Bloomington, Indiana [6], to representatives of the International *Drosophila* Array Consortium in Cambridge in the UK [7], the Canadian *Drosophila* Micorarray Centre in Toronto [8], and the Keck Microarray Resource in New Haven, Connecticut [9]. Each of these delivered sets of primers should be sufficient for tens of thousands of arrays. Mining preexisting array data should be made as easy as possible. Like scientific literature and DNA sequences, the true value of array data is not fully realized until data are available from stable public databases. These data should be stored in simple formats so that a wide variety of current and future programs can retrieve and analyze them. To this end, information on the FlyGEM is at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) repository under accession GPL20 [10,11].

Results and discussion

Array manufacture

With the draft release of the entire genomic sequence of *D. melanogaster* and predicted coding regions [2], we sought to develop a PCR-based microarray with elements for each of the predicted transcripts. Primer3 was used for the design as outlined below [4]. The objective was to array exons of each transcript that would allow unique identification of each message. We wanted to avoid cross-hybridization between elements from members of any of a number of gene families, or cross-hybridization due to low sequence complexity [12]. To do this, we chose primers from gene regions with low homology to genes elsewhere in the genome. Some researchers may choose to label cDNA using oligo(dT), so we biased amplicons to the predicted 3'-ends. Another consideration in the design is ease of amplification. The amplicons used to build the array ranged in length from 150 to 600 base pairs (bp), with an average of 410 bp and a standard deviation of 100 bp. The primer-selection process was iterative. We first selected amplicon candidates so that there would be minimal cross-homology between elements on the array and the fly genome. In addition to the exclusion of common protein-coding motifs, this also excluded repetitive and low-complexity elements found in the draft sequence. Although there may be collapsed repeats in the assembly, those amplicons would be expected to fail in the PCR tests because of the absence of a product, in the case of a long length of unanticipated repetitive DNA, or an unexpected fragment size. The second algorithm determined the primers that would work best for amplifying each of the selected fragments.

As early experience with cDNA arrays has made abundantly clear [13,14], array element tracking is essential. Two methods ensure that the identity of each element is known through the entire array manufacture, and indeed to the end-user's bench (Figure 1). First, the primers are stored in left and right plates, such that amplification only occurs when the appropriate plates are joined together for PCR. More importantly a PCR amplicon was designed to a nontranscribed region of the *Drosophila* genome. Primers producing this amplicon species were placed at three locations in each of the 96-well plates as a unique identifier or 'barcoding element' for each plate (Figure 1b). The length of the amplicon (>700 bp) allows discrimination of the barcoding element by gel electrophoresis following amplification (Figure 1c). Furthermore, hybridization to this element provides in-process quality control at multiple steps. For example, hybridization to this element confirms that plates were loaded in the correct order during printing (Figure 1e). Additionally, because there are 900 of these elements on the array, they are also convenient for determining background hybridization.

Before arraying, several in-process quality-control measures were employed (Figure 1a). We used agarose gel electrophoresis as a qualitative tool to identify amplification failures due to multiple PCR products or incorrect length (Figure 1c).

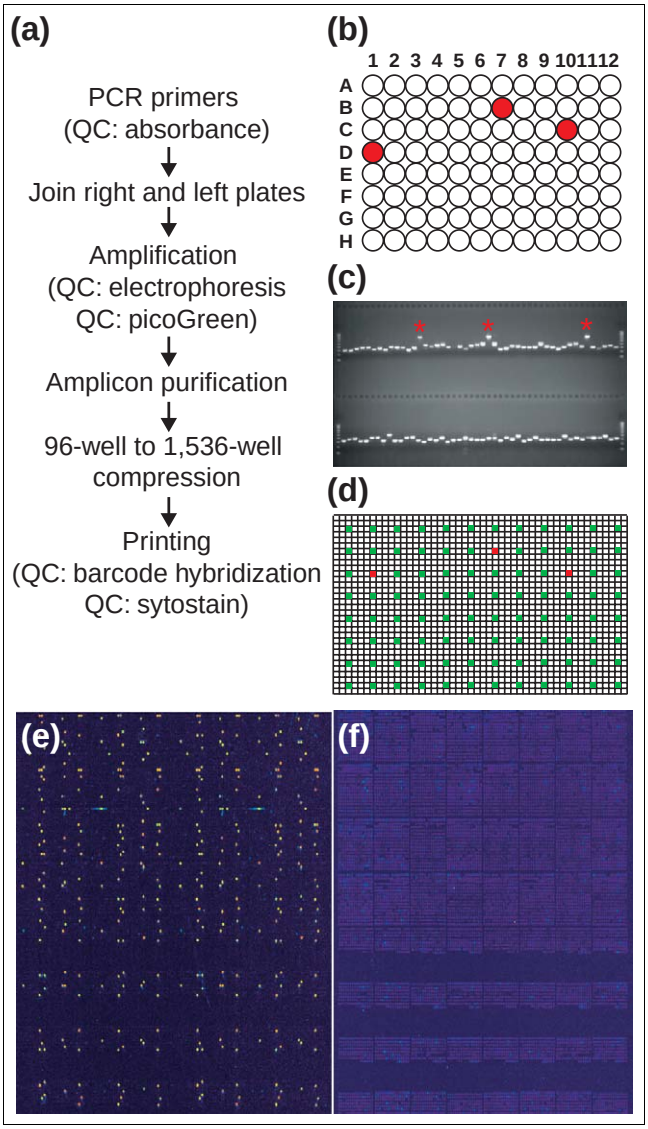

Figure 1
Process flow and in-process quality control for manufacturing the *Drosophila* FlyGEM. **(a)** The process flow for generation of PCR product for arraying and in-process quality-control measures (QC) are represented. **(b)** Barcode primers are located at unique positions in each plate (red wells in plate cartoon). **(c)** A typical agarose gel following electrophoresis of amplicons. The barcode amplicons migrate more slowly and allow for tracking after PCR (red asterisk). **(d)** 96-well plates of purified PCR product were collapsed into ten 1,536-well plates for printing. An individual plate barcode position (red) and all other barcodes (green) are shown. Post-hybridization QCs included **(e)** oligo hybridization to the barcoding elements and **(f)** syto-staining. Note the unprinted area available for adding new elements to the platform.

We failed PCRs if the products were more than 80% or less than 140% of the predicted size, or if there were multiple product lengths. A fluorescent PicoGreen assay was used to quantify the PCR product. This is superior to absorbance measurements, which are confounded by the presence of unincorporated nucleotides. We failed PCRs where amplicon concentrations were less than 75 ng/μl. For those PCRs that consistently failed, we designed a new set of primers. These were synthesized and appended to the plate set (PCR failures were simply annotated as such to prevent inclusion in the data analysis, but entire plate sets of amplicons including failed reactions were arrayed). The absence of amplicons, wrong lengths or multiple products accounted for the majority of element failures (4.8%), and are likely to represent issues of primer design or genome assembly of the draft sequence. Liquid handling and primer synthesis showed 0.9% and 0.6% failure rates. On the basis of these broadly favorable results, the amplicons were compressed into 1,536-well plates and were printed as microarrays (Figure 1d). Final quality-control measures were the hybridization with barcode element sequences and Syto-staining of printed arrays (Figure 1e,f).

The annotation of the array elements has been ongoing and is available at GEO [10] under accession GPL20 [11,15]. In addition to critical information such as the primer sequence, genome location and array element position, the current annotation includes web-based link-outs to the *Drosophila* genome database, FlyBase [16,17], gene models at NCBI LocusLink [18,19], gene functions at Gene Ontology (GO) [20,21], GenBank [22] accessions and, of course, the associated data. Because we reliably detect transcription from many release-1 predicted genes that were deleted from release-3, and because the validity of many gene-model joins are untested biologically (where one or more release-1 genes are combined into a single gene model in release 3.1), we include both sets of gene-model identifiers in the GEO platform description [2,3]. We have also included the following six flag categories that may be of interest to *Drosophila* researchers.

First, 'PCR failure' signifies data from elements where the PCR failed is suspect (see preceding paragraph). Second, we note when a sequence maps to an unassembled region of the genome (heterochromatic regions are less likely to be fully assembled). These elements might be revisited as heterochromatic regions become better assembled and annotated. The third category comprises possible secondary amplicons, based on relaxed criteria for amplification. Even if a PCR reaction passes the quality-control tests, there are cases where background amplification is more likely. The fourth category consists of multiple or secondary BLAST alignments between predicted amplicons and the *Drosophila* genome; some potential for cross-hybridizing species is inevitable. The fifth category includes amplicons mapping to a single genome location but to multiple genome features (as a result of overlapping genes, for example). The sixth category comprises problematic annotations in release 3.1 according to FlyBase [16].

Homotypic responses

To estimate the accuracy and precision of expression data generated by the FlyGEMs, we embarked on a series of

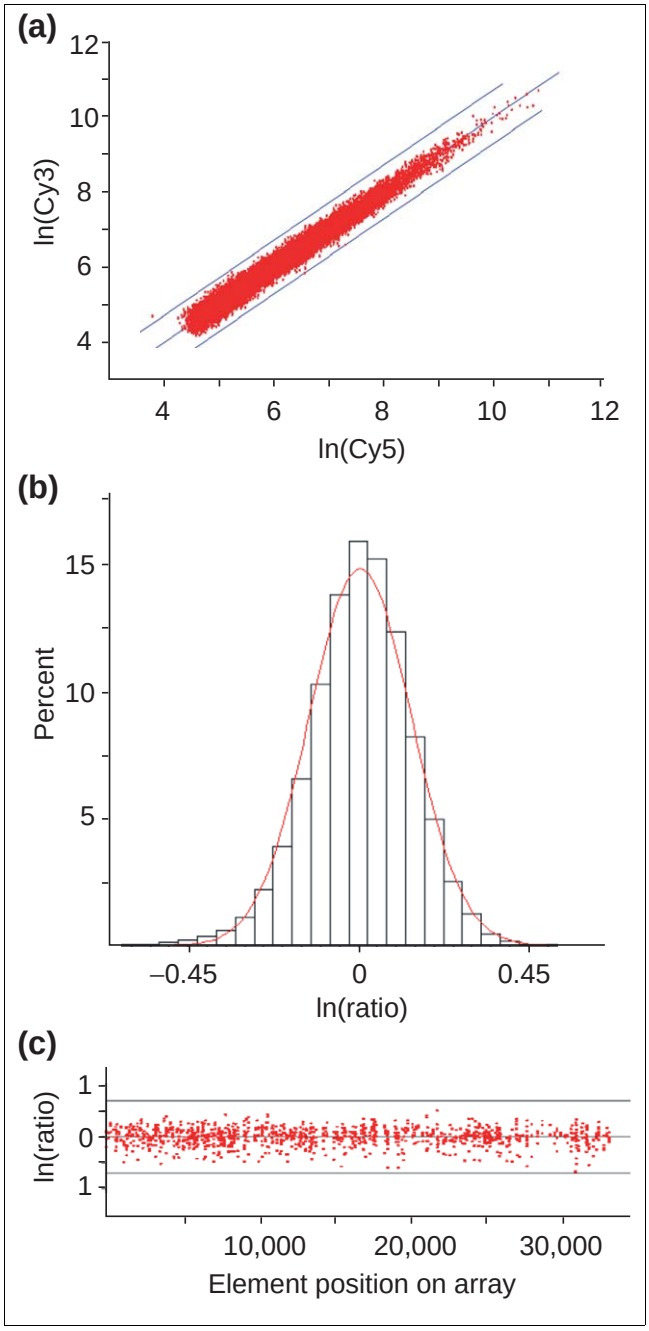

Figure 2
Homotypic hybridizations. **(a)** A typical hybridization where a single RNA pool was split and labeled with Cy3 and Cy5. For each element, intensities in each channel are plotted against each other. The central diagonal line represents equivalent intensities and the flanking lines twofold differences in intensity. **(b)** Data points from four such homotypic hybridizations were used to construct the histogram, which shows the distribution or 'bandwidth' of gene elements (as a percentage of the total) around the natural logarithm of the expected ratio of 1.0. As relative fluorescence can vary with laser power, spectral line and bandwidth and other detector parameters, it is more useful to express results as ratios, a unit-less term. **(c)** The ratio plotted against position of the element on the array. The parallel lines are at equivalent intensities and at twofold differences in intensity.

replicate experiments using various self versus self, or homotypic, hybridizations [23]. For example, a competitive hybridization of fluorescently labeled Cy3 cDNA and Cy5 cDNA, both prepared from the same mRNA sample, should theoretically give a fluorescence ratio of 1.00 for all 29,222 (14,611 transcript elements in duplicate) arrayed elements. By performing four replicate labeling reactions and hybridizations, we evaluated the overall precision of the data using statistical parameters, and estimated its accuracy on the basis of deviation(s) observed from the 1.00 expected theoretical value (0 in log space). Indeed, virtually all gene elements lie very close to the line corresponding to the expected differential expression ratio of 1.00 in these experiments, as shown by intensity scatter plots (Figure 2a), histograms of intensity ratios (Figure 2b), or intensity ratios versus array element position (Figure 2c). In three quadruplicated experiments, the average calculated relative fluorescence ratios for all elements were 1.0051, 1.0057 and 1.0052. These values are in good agreement not only with themselves, but also with the expected value of 1.00. Overall system response is linear over about three orders of magnitude. The coefficient of variation, or relative standard deviation, provides a useful estimate of the precision of measurement. The average coefficient of variation for 'differential expression' of any element in these homotypic experiments is 14% over the entire signal range. Four other homotypic hybridizations were performed with mRNA from adults of two other genotypes (for a total of 12 hybridizations) and identical coefficients of variation were observed (data not shown).

From the homotypic data, we can calculate what change in relative fluorescence ratio is required before that change has significance. Mathematically, this can be determined in terms of the two-sided statistical tolerance interval for the differential expression of non-differentiated elements. A statistical tolerance interval is one that contains a specified portion, P, of the entire sampled population with a specified degree of confidence, 100 (1-q)%. Table 1 shows the 99.5% tolerance intervals for the elements from each genotype tested - all observed values fall between ± 1.4 to 1.5. Thus as a first approximation, differences in relative fluorescence ratios of ± 1.5 or greater (lesser) are deemed to have significance in terms of measurement. The amount of a particular species of mRNA in a sample will depend on controlled and uncontrolled variables. Implicit in this analysis, however, is the advantage of concordance of replicate hybridizations. Any false negatives or false positives observed in a single hybridization do not replicate if it is a random event due to the measurement itself.

We used analysis of variance (ANOVA, restricted maximum likelihood) to estimate the contribution of specific potential sources of variance to the overall variance measured (Table 2). A random-effect model was used to estimate six general sources of variation in the ln differential expression ratios: (top/bottom) position in the sandwich hybridization, micro-

Table 1

Tolerance intervals (99.5%) for homotypic hybridizations

Source	Tolerance interval (fold difference)
Wild type:wild type	(-1.454, 1.454)
ap:ap	(-1.423, 1.423)
Antp:Antp	(-1.434, 1.434)
All combined	(-1.433, 1.433)

Table 2

Variance component (VC) estimation for homotypic hybridizations

Variation source	Estimated VC contribution
Top/bottom position	0.0%
Microarray print batch	0.0%
Sample genotype	0.0%
Hybridization	14 %
Replicate elements	9%
Gene sequence	0%
Total	16.7%

array printing batch, sample source (biological source tissues for the homotypic hybridization: wild-type, *apterous* and *Antennapedia*), array-to-array hybridization variance (including sample preparation/labeling), replicate elements within the array and gene sequence variance. Table 2 lists the estimated contribution of these potential sources of variation to the overall variance measured. The two sources contributing the most to the overall error are hybridization (14%; the variable we call hybridization includes all steps from labeling of the RNA to scanning) and variations in the array elements (9%). This points out the need to replicate hybridizations in considering the design of array experiments. The contribution of the array elements to the variance in homotypic experiments suggests that individual array elements have different and perhaps unique noise characteristics within the 1.5-fold confidence bands for overall performance. Examination of duplicate elements showing a difference in intensity usually reveals signal due to dust, scratches or other processing defects, and highlights the utility of having duplicate spots for flagging purposes (although we did not flag elements in this study). Interestingly, in contrast to the cDNA arrays which show nearly 10% contribution of gene sequence to variance [23], the differences in sequences from gene to gene (variation source, gene sequence) was not a major contributor to variation. Thus, there are many fewer elements that are inherently noisy, indicating that the approach of using an array with gene-specific primers may be superior to cDNA clones.

Differential expression

In a typical array experiment, one is looking for differences in gene expression between two samples, rather than remeasuring the same sample. We have extensively used the FlyGEM to analyze the gene-expression profiles in female versus male *Drosophila* and in tissues where there are very substantial differences between the two samples (greater than 30% of transcripts showing sex-bias) [15]. However, because even a very small misscall rate can swamp the genes of interest when there are only a few differentially expressed genes, we carefully examined gene-expression profiles between adult *Drosophila* bearing mutations resulting in visible phenotypes. Different genotypes should show some differences in gene expression [5,24], but given that all three genotypes give rise to adult flies, we expected that most expressed genes would be at equal levels - yielding a few differentially expressed genes deviating from 1.00, with most showing similar expression and thus clustering near a value of 1.00.

Six sets of experimental conditions to measure pairwise differential expression between adult strains with wild-type (Oregon-R), Antennapedia (*In(3R)Antp^{76B}/TM3*), and apterous (*ap^{56f}*) phenotypes were evaluated in quadruplicate. Although we did measure system precision and detection limits in these experiments, it is not possible to address accuracy because the expected ratio for any gene expressed in the two genotypes is unknown. Many of the elements, as predicted, are observed to fall on or very close to the 45° line representing equivalent hybridization and thus equivalent expression (Figure 3a). However, in contrast to results with homotypic experiments, elements are also observed to fall outside the twofold differential expression interval that we demonstrated is statistically significant in the homotypic experiments (Figure 2). From six such replicate experiments in this set, we calculated a coefficient of variance for each of the elements and plotted it against the dynamic signal range (Figure 3b). The average coefficient of variance was 12-15% across the entire range, although it was clear that there was slightly greater variation at the low end of the signal range. The coefficient of variance for array elements representing differentially expressed and non-differentially expressed genes were similar (Figure 3c).

Dye-flip experiments

Two-channel competitive hybridizations are a valuable way to minimize the problem of the unique signal/noise characteristics of each array element. Expressing data as a ratio cancels these inherent problems. Theoretically, it should not matter whether sample cDNAs from a given tissue are prepared with either Cy3- or Cy5-labeled primers. However, any differences in labeling or scanning efficiency,

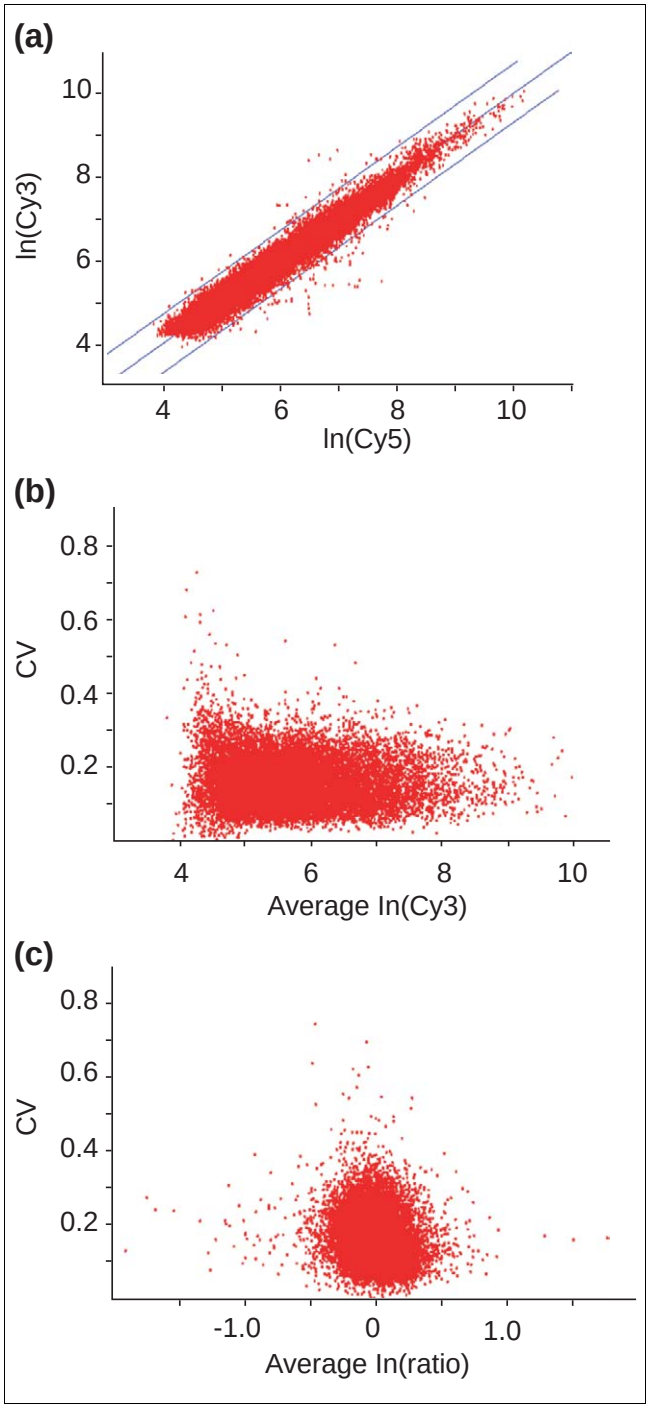

Figure 3
Heterotypic hybridizations. **(a)** A plot of a Cy3-labeled RNA from *Antp*^{76B}/*TM3* adults competitively hybridized to the array with Cy5-labeled RNA prepared from wild-type adults. **(b)** The coefficients of variance (CV) for all gene elements in (a) are plotted out as a function of Cy3 signal intensity. **(c)** The CV plotted as a function of differential expression.

photostability or other unequal behavior between channels will bias the results. We labeled samples using random nonamers, prelabeled with Cy3 or Cy5, rather than incorporating the dyes directly or utilizing conjugation following cDNA synthesis. This allows for greater control of specific activity within an experiment and between experimental series. We carried out a series of experiments specifically designed to test for dye effects.

We compared data from four replicates of the wild-type versus *ap*^{56f} hybridization to data from four replicates of the same mRNAs reciprocally labeled (Figure 4). We obtained similar results from 16 additional FlyGEMs, where reciprocal labeling experiments were performed with mRNAs from a different genotype (data not shown). For each element we can measure dye effects by averaging the two ratios (Cy3 sample A/Cy5 sample B and Cy3 sample B/Cy5 sample A) to obtain an axial symmetry of reflection (ASR). Calculated ASR values of 1.0004 (Figure 4a), 1.0005 and 1.0004 were obtained from the wild-type versus *ap*^{56f}, wild-type versus *Antp*^{76B}*/TM3*, and *Antp*^{76B}*/TM3* versus *ap*^{56f} datasets respectively, in good agreement with the theoretical value of 1.00. These are very similar to the histogram observed for non-differentiated elements (Figure 2b), and have the same standard deviation. The absence of widespread dye effects is also evident in plots of channel ratios versus position on the array. Dye-flip results are essentially mirror images (Figure 4b). These data indicate that any variation observed in a biologically relevant experiment is likely to result from real variations in experimental mRNA levels, not a byproduct of the labeling system.

Northern blots versus FlyGEM

We have shown that the FlyGEMs perform reproducibly. While reproducibility is clearly essential, every assay will suffer from different repeatable biases. We therefore asked how the FlyGEM expression measurements compare with expression measurements using a well-established procedure. A subset of 96 elements were chosen as probes for northern blotting and comparison to array data. We chose northerns as there is no reverse transcriptase step, which might introduce biases due to template preference, for example. For these experiments we chose samples with dramatically differing gene-expression profiles - adult females versus adult males [15]. The array elements chosen for this test covered the full range of hybridization intensities and the full range of differential expression in the array experiments. The relative measure of differential expression for the two platforms showed very good correlation (Figure 5). The data points fall along the diagonal ($y = 0.975x + 0.093$) with a calculated r^2 of 0.714. In no case did a gene switch from high in one sex to high in the other as a function of measurement method, indicating that reversed calls from array experiments are rare. The slope of the regression line and the intercept show that the ratios obtained from the array results did not under- or overestimate the differential expression across the full range of ratios.

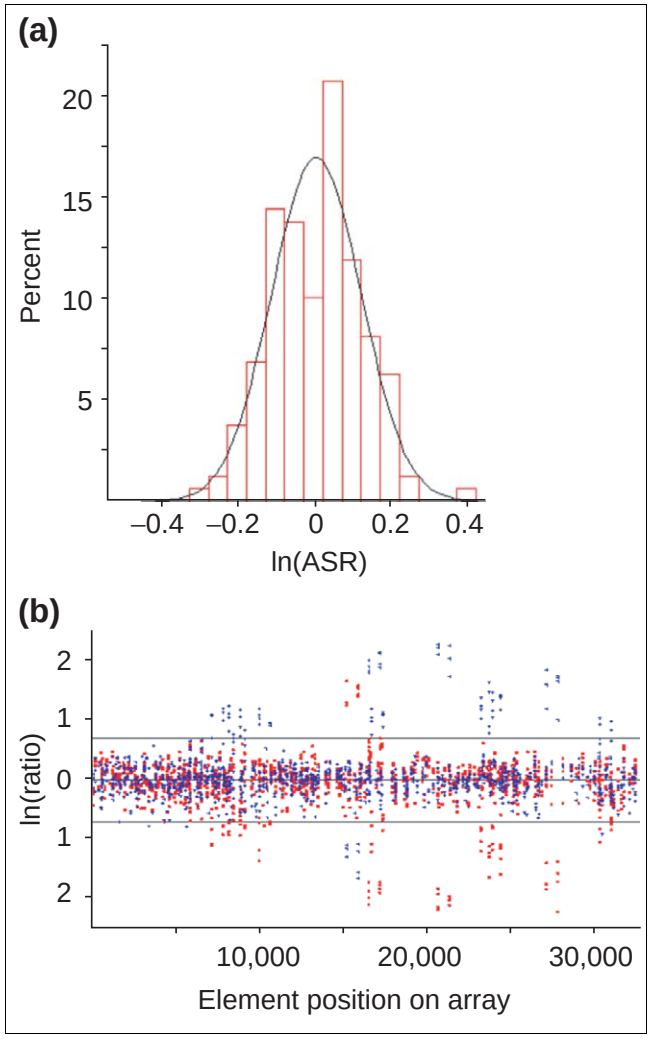

Figure 4
Dye-flip hybridizations. **(a)** A histogram showing the distribution of all elements (as a percent of the total) as a function of axial symmetry of reflection (ASR, a ratio vs ratio plot, see text). This is a measure of the contribution of dye effects to the results of a heterotypic hybridization. **(b)** The ratio plotted against position of the element on the array. The dye-reversed hybridizations are coded blue or red. Note the symmetrical patterns of differential expression as well as the bulk non-differential expression at a ratio of 1 (0 in log space). Experiments were performed in quadruplicate.

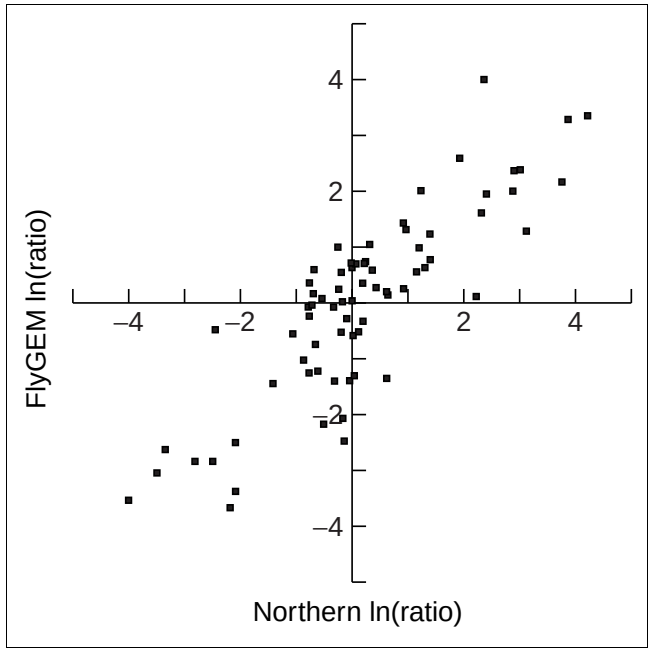

Figure 5
Comparison of FlyGEM and northern expression analysis. Differential expression values for whole adult females versus males determined from northern analysis or microarray analysis are plotted.

Conclusions

Collectively, the results presented in this report show the robust performance of the FlyGEM platform, which consists of evaluating a competitive hybridization between two differentially labeled cDNAs to a series of target sequences bound to glass. These labeled cDNAs are readily prepared from purified mRNAs using reverse transcriptase and labeled nonameric primers, and are readily applicable to a wide range of biological source materials. This platform should provide the high-quality data needed to establish accurate and reliable expression databases of great potential utility to *Drosophila* researchers. FlyGEM primers have been delivered to the *Drosophila* Genomic Resources Center in Bloomington, Indiana [6], to representatives of the International *Drosophila* Array Consortium in Cambridge, UK [7], to the Canadian *Drosophila* Microarray Centre in Toronto [8], and to the Keck Microarray Resource in New Haven, Connecticut [9]. Printed arrays should be available from those centers in the near future.

Materials and methods

Exon selection and primer design

The objective in building this array was to maximize the coverage of the entire *Drosophila* genome while minimizing cross-hybridization due to gene-family members, repeat sequences and low sequence complexity. For this, the Celera/Berkeley database (release 1) downloadable from NCBI [25] was used to identify 14,220 transcripts from the *Drosophila* genome and 13 from the *Drosophila* mitochondrion genome. For genes and putative coding regions, primers were designed primarily to DNA segments that shared no homology with other genes (<70% identity over 100 bp). DNA segments with homology between 70% and 100% identity over 100 bp were used to design primers as a last resort. Optimal PCR fragment size was 400-600 bp. If we identified no suitable segment of this length, segments of 300-400 bp were selected. If this also failed, we selected 200-300 bp segments. Finally, if no DNA segments could be picked using these criteria, two nearby

exons with an intron of less than 100 bp between them were used. DNA segments closest to the 3' end of the transcript were given preference. Primer3 [26] was used to design primers with the following settings: length 20-28 nucleotides, optimally 24 nucleotides; T_m 56-68°C, optimally 60°C; GC content 35-65%. To avoid any false priming during PCR reactions, we BLASTed [27,28] all primer sequences against the *Drosophila* genome - primers with 15 bp aligning with non-target genomic sequences were not used. PCR amplicons were assigned to locations in 96-well plates on a random basis. A single PCR amplicon was designed to a noncoding portion of the *Drosophila* genome. This amplicon was placed in a unique location in each primer plate pair to serve as a unique barcode for each plate. The size of the barcoding element (700 bp) versus the reporting elements is easily resolved by agarose gel analysis. In addition, it served to control for background (both substrate and on-spot DNA background), plate tracking, probe sensitivity, spike-in ratios, process gradients, and DNA contamination. The array platform includes 14,611 experimental elements printed in duplicate (29,222 total) in different regions of the array.

Reannotation of array elements

The GPL20 microarray platform annotation, available from the GEO on the NCBI website [10,11], was assembled using amplicons that represented exons from approximately 93% of the genes predicted by release 1 of the *D. melanogaster* genome annotation. The *Drosophila* genome sequence and annotations have undergone significant changes since the initial release [2,3] and as a reflection of this, based on release 3.1, the microarray now only represents approximately 75% of the predicted genes. The combination of an increase in the number of predicted genes, the dropping of some previous predictions and the merging of others into a single gene model contributed to this decrease.

To update the array annotation, we realigned the primer sequences to the most current release of the whole genome sequence, annotations, heterochromatin scaffolds and genome map, and release-1 genome sequence. These were downloaded from FlyBase [17,29]. The mitochondrial genome (for controls) and LocusLink information were downloaded from NCBI [25]. The primers were aligned to the release-3 genome sequence using BLAST [27,28] with the BLASTn program option, no masking, and a word size of 20. The word size used was the size of the smallest primer sequence in the set. From the BLAST output, valid potential ranges for amplicons were obtained by requiring that the primers in a given pair align with 100% identity to opposite strands in an orientation that could produce an amplicon of length less than 5 kb. Multiple amplicons were allowed per query. The primer pairs for approximately 30 queries failed to predict amplicons using the given BLAST arguments and the release-3 sequence database. To obtain amplicon predictions for these queries, the 5' ends of the primers were allowed to mismatch, and heterochromatin (WGS3) scaffolds from Celera and release-1 genome sequences were included in the BLAST database.

Next, the predicted amplicons were aligned to the release-3 genome sequence and heterochromatin WGS scaffolds using MegaBLAST with no masking and a word size of 50. The ranges obtained from the amplicon alignments were then mapped to the feature ranges from the release-3.1 annotation and transcript ranges from the genome map (gnomap file). For each query, the BLAST hit with the highest raw score that mapped to a feature and a transcript, if available, was selected to represent the ID in the platform table. If a BLAST hit mapped to the transcripts of multiple genes, the length of the overlapping region was considered. However, if a 'tie' was still present, then preference was given, in the following order, to features not flagged problematic; not located in a heterochromatin region; not RNAs (CRxxx features); not transposable elements (TExxx features), and sequences where annotations have GO terms.

Several additional 'status' flags have been included in the updated GPL20 platform. Along with whether the PCR failed, it is noted whether a given primer pair had multiple predicted amplicons, whether the representative amplicon had multiple BLAST hits, and whether the representative BLAST hit overlapped multiple annotation features. In addition, we noted whether FlyBase has flagged the mapped feature annotation as problematic. Overall, approximately 19% of the queries do not have an 'OK' status. However, data generated from these elements should not automatically be considered unreliable. Amplicons and BLAST hits that were computationally allowed and resulted in an element's flag may not have biological significance. For example, the potential for secondary amplicons was not always born out by an actual failed PCR when the amplicons were generated. As another example, for those elements with multiple feature mappings, usually only one of the candidate features has a transcript with an exon whose range overlaps that of the amplicon. These elements might well detect a single type of transcript despite the flag. Briefly, the flags should be viewed as a warning. If a scientist has a particular interest in the gene in question, a review of the evidence that led to the element's identity may be useful.

PCR product generation

Master solutions of PCR primers (Operon Technologies, Alameda, CA) are approximately 50 μM and were kept in left and right 96-well plates. A working stock of 7.5 μM was prepared by dilution of aliquots of the master plate. Primer concentrations for each plate were verified by absorbance readings (260 and 280 nm) of a dilution of the working stock. Primer concentrations less than 10% the expected were failed. *Drosophila* genomic DNA (Clontech, Palo Alto, CA) was quantified by absorbance and PicoGreen fluorometry (Molecular Probes, Eugene, OR) [23]. PCR was performed by adding 100 ng *Drosophila* DNA to 75 μl reaction buffer, containing 10 mM Tris-Cl pH 8.3, 1.5 mM $MgCl_2$, 50 mM KCl, 0.2 mM

each dNTP, 0.5 μM each primer, and 2 units Taq polymerase. Primers were added to the reaction mixture from the individual left and right working stocks. The mixture was incubated for 2 min at 95°C, and 40 cycles of PCR were performed at 94°C for 30 sec, 55°C for 30 sec, and 72°C for 120 sec. A final incubation for 5 min at 72°C was followed by reduction of the temperature to 4°C to terminate the reaction. Duplicate PCR reactions were pooled and purified with multiscreen filter plates (Millipore, Bedford, MA) and resuspended in 110 μl water. We concentrated amplicons by desiccation. Amplicons were resuspended in 12.5 μl 2x SSC and solubilized by 45 cycles of heating to 85°C for 30 sec and cooling to 20°C for 30 sec. A 1/10 dilution of each of the amplicons was used for qualification by agarose gel analysis. PCR products were failed if no bands appeared, if multiple bands appeared, or if the observed size was not between 80-140% of the expected size. Furthermore, PCR products were quantified by a PicoGreen fluorescent assay [23]. Yields below 20 ng/μl of the 1/10 dilution were failed.

Arraying

Plates (96-well) containing the qualified amplicons were condensed to ten 1,536-well plates robotically with a V-prep liquid-handling robot (Velocity 11, Palo Alto, CA). Arraying was performed with a DotBot, a prototype arrayer (Velocity 11). The arrayer uses 16-pen printing with 170 μm spacing, a 500-slide platen with automated slide placement, ultrasonic and 90°C active water-pen washing, a pen test fire station, environmental control and a cooled peltier plate holder to minimize evaporation. Amplicons were arrayed in duplicate on a slide. The number of spots printed necessitated the use of two slides per full array. Each pen prints a 13 × 13 subarray. The six quadrants on a slide are composed of 16 subarrays. Amplicons were arrayed on amino-modified glass slides [30]. DNA adhesion to the glass was achieved by UV irradiation using a Stratalinker Model 2400 UV Illuminator (Stratagene, San Diego, CA) with light at 254 nm and an energy output of 120,000 μJ/cm². The microarrays were washed for 2 min in 0.2% SDS (Life Technologies, Rockville, MD), followed by three rinses in water for 1 min each, then treated with 0.2% (w/v) I-block (Tropix, Bedford, MA) in PBS for 30 min at 60°C. Finally, they were washed again for 2 min in 0.2% SDS, rinsed three times in water for 1 min each before drying by a brief centrifugation.

A random sampling of arrays was stained with Syto61 (Molecular Probes) to quantify the amount of DNA deposited to the slide and identify dropouts [23]. In addition, a sample of slides was hybridized with a Cy3-labeled oligonucleotide probe to specifically detect the barcoding element (see below). This allowed qualification of the arraying process.

Array hybridization and scanning

Living cultures of wild-type, *In(3R)Antp^{76B}/TM3* and *ap^{56f}* flies were obtained from Carolina Biological Supply Co. (Burlington, NC). Male and female *y w^{67C}* flies were grown at 25 ± 0.5°C on GIF or PB media (KD Scientific, Columbia, MD) and aged for 5-7 days before use. Briefly, mRNA from the indicated flies was isolated by a single round of poly(A) selection using Oligotex resin (Qiagen, Valencia, CA). The purified mRNA was quantified using RiboGreen dye (Molecular Probes) in a fluorescent assay as previously described [23]. Briefly, RiboGreen dye was diluted 1:200 (v/v final) and mixed with Millennium RNA size ladder (Ambion, Austin, TX) in known RNA concentrations to generate standard curves. Unknown samples were diluted as necessary. Fluorescence was measured in 96-well plates with a FLUOstar fluorometer (BMG Lab Technologies, Germany) fitted with 485 nm (excitation) and 520 nm (emission) filters. mRNAs (25-100 ng) were separated on an Agilent 2100 Bioanalyzer, a high-resolution electrophoresis system (Agilent Technologies, Palo Alto, CA), to examine the mRNA size distribution.

Purified mRNA (600 ng) was converted to either Cy3- or Cy5-labeled cDNA probes using a custom labeling kit (Incyte Genomics, Fremont, CA). Each reaction contains 50 mM Tris-HCl pH 8.3, 75 mM KCl, 15 mM $MgCl_2$, 4 mM DTT, 2 mM dNTPs (0.5 mM each), 6 μg Cy3 or Cy5 random 9-mer (Trilink, San Diego, CA), 60 U RNase inhibitor (Ambion, Austin, TX), 600 U MMLV RT RNase (H-) (Promega, Madison, WI) in 75 μl. Labeled Cy3 or Cy5 cDNA products were combined and subsequently de-pooled into three aliquots and purified with ChromaSpin+ TE-30 gel-filtration spin column (Clontech). The probes were then re-pooled, concentrated by ethanol precipitation and resuspended in hybridization buffer.

Hybridization was performed in custom-made chambers allowing simultaneous exposure of the probe solution to both slides representing the entire *Drosophila* transcriptome. Spots (1 μl each) of a 40% suspension (v/v) of 30 μm ceramic microspheres in water were placed at four locations along each side of one slide and allowed to dry. The second slide was placed over the first slide such that the spotted parts of the slides were facing each other and the beads maintained proper spacing between the slides. Hybridization solution was applied at one end and covered the array surfaces by capillary action. Hybridization of labeled cDNA probes was performed in 50 μl 5x SSC, 0.1% SDS, and 1 mM DTT at 60°C for 6 h. Hybridization with a Cy3-labeled oligonucleotide (5' TTTGACACGTGCATACCAACTTGCAACGGTTTTATTTTCAC TTTTTTTGGACATGTGAA-3') (Operon Technologies) specific for the barcoding element was performed at 10 ng/μl in 5x SSC, 0.1% SDS, 1 mM DTT at 60°C for 1 h. The microarrays were washed after hybridization in 1x SSC, 0.1% SDS, 1 mM DTT at 45°C for 10 min, and then in 0.1x SSC, 0.2% SDS, 1 mM DTT at room temperature for 3 min. After drying by centrifugation, microarrays were scanned with an Axon GenePix 4000A fluorescence reader at 535 nm for Cy3 and 625 nm for Cy5 and GenePix software was used for image capture (Axon, Palo Alto, CA). An image-analysis algorithm in GEMTools software (Incyte Genomics) was used to quantify signal and

background intensity for each element. Intensity scales are arbitrary. Intensities similar to those obtained using GenPix are approximated by multiplying by 1,000. The ratio of the two corrected signal intensities was calculated and used as the differential expression ratio for this specific gene in the two mRNA samples.

The Axon scanner was calibrated using a primary standard and a secondary standard to account for the differences in scanner performance (laser and photomultiplier tube (PMT)) between the Cy3 and Cy5 channels. For the primary standard, hundreds of probe samples were prepared which were fluorescently balanced in Cy3 and Cy5 channels as determined by a Fluorolog3 fluorescence spectrophotometer (Instruments S.A., Edison, NJ). These probes were hybridized to microarrays and the scanner PMTs were adjusted to give balanced fluorescence and the greatest dynamic range. Using these PMT values, a fluorescent plastic slide was scanned to obtain corresponding fluorescent values. This secondary standard was used to calibrate other scanners on a daily basis.

Data acquisition and analysis

An image analysis algorithm in GEMTools software was used to quantify signal and background intensity for each target element. Two low-frequency data-correction algorithms were applied to compensate for systematic variations in data quality. The first procedure, a gradient-correction algorithm, modeled the signal-response surfaces of each channel. On a 10,000-element microarray, the signal response of Cy3 and Cy5 should be random as a result of the random physical location of the target elements. The signal-response surfaces were first examined for nonrandom patterns. If nonrandom patterns were detected, a second-order response model was applied to model the gene signal responses according to their positions on the surface. The nonrandomness was then corrected using the fitted model. The second procedure, a signal-correction algorithm, corrected for differential rates of the incorporation of the Cy3 and Cy5 dyes. In an idealized homotypic hybridization, a scatter plot of log Cy3 signal versus log Cy5 signal should show a signal distribution along a line with a slope of 1. If the center line of the signals does not have a slope of 1, there may be different rates of the incorporation of Cy3 and Cy5 dyes. The signal-correction algorithm tested whether the regression line slope of log Cy3 signal versus log Cy5 was 1, and applied a regression model to rotate the regression line to a slope of 1 if necessary.

ANOVA was used to estimate the contribution of specific potential sources of variance to the overall variance measured. Analyses were performed using the method of restricted maximum likelihood (REML) under SAS 8.2 for Windows version 8.02 procedure PROC MIXED [31]. Three variance components listed as 0.0%. The actuarial variance may not be 0.0%. They are estimated to be 0.0% by the REML algorithm. The two sources contributing most significantly to the overall variation were hybridization variance and sequence variance. Microarray batches and source tissue were not significant sources of variance.

Array data reported here is available under GEO sample accessions: GSM11026; GSM11080; GSM11081; GSM11088; GSM11104-GSM11111; GSM11113; GSM11115; GSM11128-GSM11132; GSM11134-GSM11136; GSM11352-GSM11363; GSM11365 and GSM11367.

Northern analysis

Ninety-six elements were chosen as probes for northern blotting on Hybond-N+ membranes (Amersham, Piscataway, NJ). Probes were selected to cover the full range of absolute intensities and male/female differential expression revealed in array experiments [15]. Blotted mRNAs were from flies wild type with respect to sex ($y\ w^{67C}$). These same genotypes were used for labeling reactions in previously reported array experiments [15]. Amplicon probes were made using the same primer pairs used in array construction and were labeled using Redi-prime II (Amersham). Northerns were hybridized at 42°C in UltraHyb (Ambion) in 15-ml conical tubes in a bacterial shaker. Blots were imaged on a Storm 860 phosphorimager and quantified using ImageQuant (Molecular Dynamics, Sunnyvale, CA). Signal within each lane was background corrected using inter-lane intensity. When multiple transcripts were detected, the summed intensities of those bands was recorded. Similarly, in cases of smearing indicative of message-specific degradation (all northerns were prepared from the same mRNA sample) all in-lane signal was recorded. Seventy-three northerns were successful (passing visual inspection and showing bands above background).

Acknowledgements

This article is dedicated to the memory of Jeff Seilhamer. We thank our many colleagues at Incyte and NIH for valuable discussions and for helping to make this collaboration possible, in particular Vaijayanti Gupta, Jeff Seilhamer, Virgina Ozer, Michael Edwards and Linda Schilling.

References

1. Churchill GA: **Fundamentals of experimental design for cDNA microarrays.** *Nat Genet* 2002, **32 Suppl:**490-495.
2. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
3. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, *et al.*: **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review.** *Genome Biol* 2002, **3:**research0083.1-0083.22.
4. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132:**365-386.
5. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G: **The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*.** *Nat Genet* 2001, **29:**389-395.
6. ***Drosophila*** **Genomics Resource Center** [http://dgrc.cgb.indiana.edu]
7. **International *Drosophila* Array Consortium** [http://

www.indac.net]
8. **Canadian *Drosophila* Microarray Centre** [http://www.flyarrays.com/index.html]
9. **HHMI Biopolymer/Keck Foundation Biotechnology Resource Laboratory** [http://keck.med.yale.edu/dna_arrays.htm]
10. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30:**207-210.
11. **GEO: Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo]
12. Evertsz EM, Au-Young J, Ruvolo MV, Lim AC, Reynolds MA: **Hybridization cross-reactivity within homologous gene families on glass cDNA microarrays.** *Biotechniques* 2001, **31:**1182-1186.
13. Kargul GJ, Dudekula DB, Qian Y, Lim MK, Jaradat SA, Tanaka TS, Carter MG, Ko MS: **Verification and initial annotation of the NIA mouse 15K cDNA clone set.** *Nat Genet* 2001, **28:**17-18.
14. Halgren RG, Fielden MR, Fong CJ, Zacharewski TR: **Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones.** *Nucleic Acids Res* 2001, **29:**582-588.
15. Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B: **Paucity of genes on the *Drosophila* X chromosome showing male-biased expression.** *Science* 2003, **299:**697-700.
16. **FlyBase, a database of the *Drosophila* genome** [http://flybase.bio.indiana.edu]
17. FlyBase Consortium: **The FlyBase database of the *Drosophila* genome projects and community literature.** *Nucleic Acids Res* 2003, **31:**172-175.
18. **LocusLink** [http://www.ncbi.nlm.nih.gov/LocusLink]
19. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
20. Gene Ontology Consortium: **Creating the gene ontology resource: design and implementation.** *Genome Res* 2001, **11:**1425-1433.
21. **Gene Ontology Consortium** [http://www.geneontology.org]
22. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2003, **31:**23-27.
23. Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, *et al.*: **An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression.** *Nucleic Acids Res* 2001, **29:**E41.
24. Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL: **Sex-dependent gene expression and evolution of the *Drosophila* transcriptome.** *Science* 2003, **300:**1742-1745.
25. **NCBI FTP site** [http://www.ncbi.nlm.nih.gov/Ftp/index.html]
26. **Primer3 software distribution** [http://www.broad.mit.edu/genome_software/other/primer3.html]
27. **NCBI BLAST** [http://www.ncbi.nlm.nih.gov/BLAST]
28. Altschul SF, Gish W: **Local alignment statistics.** *Methods Enzymol* 1996, **266:**460-480.
29. **FlyBase Reference Manual D: bulk FlyBase data retrieval** [http://flybase.bio.indiana.edu/docs/lk/refman/refman-D.html]
30. Lee PH, Sawan SP, Modrusan Z, Arnold LJ, Reynolds MA: **An efficient binding chemistry for glass polynucleotide microarrays.** *Bioconjug Chem* 2002, **13:**97-103.
31. Littel RC, Milliken GA, Stroup WW, Wolfinger RD: *SAS System for Mixed Models* Cary, NC: SAS Insitute; 1996.