Supplementary methods

Lines

The lines were lines 18, 21, 23, 42, 46, 49, 51, 54, 59, 65, 67, and 71, chosen randomly from more than twenty IVe-39 sublines still alive in 2003 after approximately 200 generations of mutation accumulation and described by Houle and Nuzhdin[1]. They were expanded on a 12:12 light:dark cycle. The lines were maintained during mutation accumulation by brother-sister mating. Further details can be found in ref. (1) .

Array design

The sequences on the arrays were selected to minimize intra-family hybridization. To pick the primers, we followed a sequence of decreasingly stringent tests using D. melanogaster sequence release 1.0. For each gene, we looked first for a 500b to 2kb sequence at the 3' end and, failing this, two 200b to 500b sequences: (1) in coding sequence, completely within an EST clot, and unduplicated (BLAST<=10-4); (2) in coding sequence, partially within an EST clot, unduplicated; (3) in coding sequence, unduplicated; (4) an exon, completely within an EST clot, unduplicated; (5) an exon, partially within an EST clot, unduplicated; (6) an exon, unduplicated; (7) a sequence less than 2kb at the 3' end. We printed poly-lysine coated arrays on a Gene Machines arrayer and post-processed the arrays using 1-2 dichloroethane, N-methylimidazol, and succinic anhydride[2].

In updating our annotation to *Drosophila melanogaster* release 4.0 (http://flybase.bio.indiana.edu/annot/dmel-release4-notes.html), we flagged spots according to the following procedure:

1) Model PCR amplification using isPcr[3] on the transcript, gene, and chromosome files from release 4.0[4-6] Sequence from release 3.2 was used for

heterochromatic regions not covered in release 4.0[7]. Since these are mRNA hybridizations, spot annotation is based primarily on the transcript results. For example, if a region amplified by a primer pair is part of a transcript for gene A and an intron for gene B (on the opposite stand), the spot is considered to measure gene A

2) If the amplified sequence falls within a gene but not completely within a transcript, it is considered to represent that gene. This does introduce some inconsistency with step (1) in how complete we regard the annotation, but it only affects a few probes.
3) If the amplified sequence overlaps a transcript or gene by more than half its length (with the other part running off the 5' or 3' end of the gene) it is accepted as a good probe.
4) If there are multiple predicted sequences for a given primer pair which do not fall within a single gene, flag the spot as bad.
5) If the primers amplified multiple bands in the actual PCR, only flag it as good if isPcr predicts multiple bands all within the same gene.
6) If the sequences for several spots fall within the same gene, determine whether they fall within the same transcripts and group them according to transcripts with the caveat that singletons are not grouped into larger transcript groups. For example, if a gene with five transcripts ($Tr_A...Tr_E$)is represented by five spots on the array with sequences ($S_1...S_5$), and the sequences fall within the following transcripts $\{S_1 \subset Tr_A, Tr_E; S_2 \subset Tr_B; S_3 \subset Tr_B, Tr_C; S_4 \subset Tr_C, Tr_D; S_5 \subset Tr_B\}$, then group S_3 and S_4 together and consider them as replicates in the following gene-specific analyses (although not in the global normalizations). Analyze S_1 on its own and group S_2 and S_5 together. This causes some complications when using gene-based functional annotations in later analyses, but since transcripts and not genes are biologically active

molecules, this avoids grouping some potentially functionally dissimilar mRNAs together.

After this processing, 1,292 genes were represented by multiple groups of spots on the arrays leaving us with 13,567 potentially good spots on the array distributed into 12,017 transcript groups, some of which were flagged for other reasons in the ensuing analysis (Supplementay data 2). In the rest of the paper, the term *gene* refers to a transcript group.

Experimental design

We designed the hybridizations to efficiently estimate stage-specific across-line and within-line variance, without using an excessive number of microarrays. We divided the twelve lines into two hexagons (different for each stage). Each line was hybridized to four other lines in its own hexagon, two lines in the other hexagon in the same stage, and twice to itself at the other stage (Supplementary figure 1). In all, each of the twelve lines was measured eight times at each of two stages with dye swapping to control for systematic labeling biases. The raw data for a spot for each dye consists of the median of the pixel intensities of the spot minus Spot's morphological opening local background estimate[8].

Preliminary analyses and global normalizations

After scanning the arrays with a confocal laser scanner (Axon), we analyzed the images using Spot[8], subtracting the morph background estimates giving us log base 2 measurements for median intensities for each channel in each spot on the array. All subsequent analyses were coded in python[9,10] with calls to the R statistical package[11] and PROC MIXED in SAS software v. 8.2 (SAS Institute, Cary, North Carolina[12-14]. On an array-by-array basis, we performed a series of normalizations on the log base 2 transformed data to remove technical bias and noise

First, we subtracted the mean for each dye.

$$\vec{R}_{ag}^{dye} = \vec{R}_{ag}^{raw} - R_{a\cdot}^{raw}; \vec{G}_{ag}^{dye} = \vec{G}_{ag}^{raw} - G_{a\cdot}^{raw} \quad \text{(S1)}$$

Second, we locally regressed the difference of the Cy5 and Cy3 data for a spot on their average intensity and the spot location in pixel coordinates (*x,y*) using the loess procedure in R with a span of 0.4[15]. For each spot, we subtracted half the value of the loess curve at that spot from the Cy5 value and added half to the Cy3 channel.

$$\begin{aligned} &\vec{R}_{ag}^{loess} = \vec{R}_{ag}^{dye} - \frac{1}{2}L_{a_g}; \vec{G}_{ag}^{loess} = \vec{G}_{ag}^{dye} + \frac{1}{2}L_{a_g}; \\ &L_a = loess(\vec{R}_{ag}^{dye} - \vec{G}_{ag}^{dye} \sim \frac{1}{2}(\vec{R}_{ag}^{dye} + \vec{G}_{ag}^{dye}), x_g, y_g, span = 0.4) \end{aligned} \quad \text{(S2)}$$

Third, we subtracted half the median difference between the log intensities from the Cy5 intensities and added half to the Cy3 intensities.

$$\begin{aligned} &\vec{R}_{ag}^{median} = \vec{R}_{ag}^{loess} - \frac{1}{2}median_a(\vec{R}_{ag}^{loess} - \vec{G}_{ag}^{loess}); \\ &\vec{G}_{ag}^{median} = \vec{G}_{ag}^{loess} + \frac{1}{2}median_a(\vec{R}_{ag}^{loess} - \vec{G}_{ag}^{loess}) \end{aligned} \quad \text{(S3)}$$

Finally, we accounted for scale differences among the arrays by dividing each intensity by the median absolute deviation of the log ratios on the array from the median log ratio on the array (which should be zero from the previous step) and then multiplied these by the geometric mean of these scaling factors across all the arrays in order to convert them back to a log base 2 scale[16].

$$\begin{aligned} &\vec{R}_{ag}^{MAD} = \frac{\vec{R}_{ag}^{median}}{MAD_a}; \vec{G}_{ag}^{MAD} = \frac{\vec{G}_{ag}^{median}}{MAD_a}; \\ &MAD_a = median_a\left(\left|\vec{R}_{ag}^{median} - \vec{G}_{ag}^{median}\right| - median_a(\vec{R}_{ag}^{median} - \vec{G}_{ag}^{median})\right) \\ &\vec{R}_{ag}^{geoMAD} = geoMAD \times \vec{R}_{ag}^{MAD}; \vec{G}_{ag}^{geoMAD} = geoMAD \times \vec{G}_{ag}^{MAD}; geoMAD = \sqrt[n]{\prod_{a=1}^{n} MAD_a} \end{aligned} \quad \text{(S4)}$$

These transformations account for global technical noise including (S1) additive translational differences in the distributions of intensities of dyes on a single array, (S2)

differences in the shapes of these distributions which are very according to intensity or spatial location on the array, (S3) relative intensities of the arrays, and (S4) multiplicative differences in scale between arrays of the log-ratios of the dyes. The final dataset consisted of this adjusted data for the eight replicates per lineage.

Mixed-model analyses

Starting from an identical level, the gene expression of particular genes in the lines, will diverge during mutation accumulation under the weight of mutations. Since gene expression levels vary during development, the levels as well as the amount of divergence may not be equal at the two stages. The full model accounts for differences in average levels in the two stages by means of a fixed effect stage-specific deviation from a grand mean level and for line-specific deviations with a random effect term accounting for the variance of the line means within a stage. The full model, allowing for stage-specific mean effects, sequence effects when dealing with spot groups, spot effects, stage-specific across line variances, and stage-specific residuals is:

$$\begin{aligned} y_{ijkq} &= \mu + Stage_i + Sequence_q + Array(Sequence)_{j(q)} + Line(Stage)_{k(i)} + \varepsilon_{ijkq} \\ Array(Sequence) &\sim N(0,\sigma_a^2); Line(Stage)_{(i)} \sim N(0,\sigma_i^2); \varepsilon \sim N(0,\sigma_{\varepsilon_i}^2) \end{aligned} \quad \text{(S5)}$$

where y_{gijkq} is the log base 2 measurement for gene g at stage i from line k from spot/sequence q on array j. Sequence was only included when two spots comprised a spot group. (SAS ran out of memory computing the model when more than two spots comprised a spot group, so for more than two spots we fit the data for each spot individually, averaged the ensuing likelihood contrasts (see below) and averaged the estimates.) This is a gene-based model, so the terms are gene-specific. The overall mean, stage, and sequence effects are fixed; the array, line, and error effects are random with the distributions above.

To test particular assumptions of the full model, we constructed a hierarchy of six alternative models (including the full one). Using a series of likelihood ratio tests, we determined which model best fit the data and used parameter estimates from that model for further analyses. Three complications arose: the six models were not all strictly nested, SAS was not able to fit all models for all genes, and the PROC MIXED KR degrees of freedom option discounted a model's degrees of freedom of by the number of parameter estimates equal to zero, leading to possible loss in the more complex models. We accounted for multiple testing in the likelihood ratio tests by a false discovery rate (FDR) procedure[17]. Our test sequence was (Supplementary figure 2):

0) geneG_models = {3r0, 2ar0, 2br0, 1r0, 0b, 0}
1) for each geneG:
    for test_2 in {3r0 vs. 2ar0, 3r0 vs 2br0}:
        p = p_value from likelihood ratio test_1
        if model 3r0 significantly better than model 2xr0 (p < FDR(0.05) cutoff):
            remove model 2xr0 from geneG_models
        else:
            remove model 3r0 from geneG_models
    if both models 2ar0 and 2br0 were eliminated, the gene fits model 3r0; eliminate all remaining models
2) for each geneG with model 2ar0 in geneG_models:
    p = p_value from likelihood ratio test 2ar0 vs. 1r0
    if model 2ar0 significantly better than model 1r0 (p < FDR(0.05) cutoff):
        remove model 1r0 from geneG_models
    else:
        remove model 2ar0 from geneG_models

if 2br0 and 1r0 were eliminated, the gene fits model 2ar0; eliminate all remaining models

3) for each geneG with model 2br0 in geneG_models:

for test_3 in {2br0 vs. 1r0, 2br0 vs 0b}:

p = p_value from likelihood ratio test_3

if model 2br0 significantly better than model 1r0/0b (p < FDR(0.05) cutoff):

remove model 1r0/0b from geneG_models

else:

remove model 2br0 from geneG_models

if models 2ar0, 1r0, and 0b were eliminated, the gene fits model 2br0; eliminate all remaining models

4) for each geneG with models 2ar0 and 0b in geneG_models:

p = p_value from likelihood ratio test 2ar0 vs. 0b

if model 2ar0 significantly better than model 0b (p < FDR(0.05) cutoff):

remove model 0b from geneG_models

else:

remove model 2ar0 from geneG_models

if 0b was eliminated, the gene fits model 2ar0; eliminate all remaining models

5) for each geneG with models 2ar0 and 2br0 in geneG_models:

p = p_value from likelihood ratio test 2ar0 vs. 2br0

if model 2ar0 significantly better than model 2br0 (p < FDR(0.05) cutoff):

remove model 2br0 from geneG_models

else:

remove model 2ar0 from geneG_models

the gene fits the model which wasn't eliminated; eliminate all remaining models

6) for each geneG with models 2ar0 and 2br0 in geneG_models:

p = p_value from likelihood ratio test 2ar0 vs. 2br0

if model 2ar0 significantly better than model 2br0 (p < FDR(0.05) cutoff):

remove model 2br0 from geneG_models

else:

remove model 2ar0 from geneG_models

the gene fits the model which wasn't eliminated; eliminate all remaining models

7) for each geneG with models 1r0 and 0b in geneG_models:

for test_7 in {1r0 vs. 0, 0b vs 0}:

p = p_value from likelihood ratio test_7

if model 1r0/0b significantly better than model 0 (p < FDR(0.05) cutoff):

remove model 0 from geneG_models

else:

remove model 1r0/0b from geneG_models

if models 0b and 0 were eliminated, the gene fits model 1r0

if models 1r0 and 0 were eliminated, the gene fits model 0b

8) for each geneG with models 1r0 and 0b in geneG_models:

p = p_value from likelihood ratio test 1r0 vs. 0b

if model 1r0 significantly better than model 0b (p < FDR(0.05) cutoff):

remove model 0 b from geneG_models

else:

remove model 1r0 from geneG_models

the gene fits the model which was not eliminated

All tests were two-tailed tests except for 2br0 vs. 0b (step 3b) and 1r0 vs. 0 (step 7a). In the cases where two models had the same degrees of freedom (making a conventional likelihood ratio test impossible) the more likely one was chosen. In cases where two models had the same degrees of freedom and likelihoods, the simpler one was chosen according to the following scale of model complexity: 3r0 > 2ar0 > 2br0 > 1r0 > 0b > 0. When SAS was not able to fit a particular model to the data, any tests involving that model were skipped. In some cases, this led to the likelihood ratio tests 3r0 vs. 1r0, 2ar0 vs. 0, and 2br0 vs. 0. We dealt with these on an individual basis with a significance cutoff of 0.05.

Comparative data update and extension

The core of the interspecific and *D. melanogaster* intraspecific data used in the comparative anlysis has been described before[18,19]. Rifkin *et al.*[18] used 4 arrays per line (except for *D. melanogaster* Samarkand which had 6 arrays). For the analyses of Gu *et al.*[19], enough arrays were added to bring these to 8 arrays per line. For the current paper, the original 4 *D. yakuba* arrays[18] were replaced by 4 new arrays. The array annotation was also updated to *D. melanogaster* annotation release 4.0 as described above. A new version of the results from Rifkin *et al.*[18] based on this new data can be downloaded from http://genome.med.yale.edu/Comparative/ and the raw data itself can be found in GEO using accession number GSE2642.

Assessing technical error

We collected large numbers of larvae, ground them up together, divided the homogenate into aliquots, and processed each sample separately as described in the Methods. We did the same for a collection of adults and performed 8 larvae-adult hybridizations (GEO accession GSE2641). After global normalizations as described above, we fit the data for each gene to the model:

$$y_{ijq} = \mu + Stage_i + Sequence_q + Array(Sequence)_{j(q)} + \varepsilon_{ijq}$$
$$Array(Sequence) \sim N(0, \sigma_a^2); \varepsilon \sim N(0, \sigma_{\varepsilon_i}^2) \quad \text{(S6)}$$

using PROC MIXED in SAS 8.2, consistent with the analysis above.

We assume that we have samples from two distributions. For the technical error data set each residual is a random sample from $N(0, \sigma_t^2)$. For the mutational variance data set each residual has the form $T + (1/30) \cdot M$ where *T* is the random variable for the technical error and *M* is the random variable for physiological variable. *T* has an $N(0, \sigma_t^2)$ distribution and we assume *M* has the distribution $N(0, \sigma_m^2)$. We will consider the hypothesis that the physiological variance is some α fold of the technical variance. Then, the mutational variance data set has the distribution $N(0, \sigma_m^2) = N(0, \sigma_t^2 + (1/30) \cdot \alpha\sigma_t^2)$. Now we consider the statistic:

$$\frac{S_t^2 / \sigma_t^2}{S_m^2 / [\sigma_t^2 (1 + \alpha/30)]} = (1 + \alpha/30) \frac{S_t^2}{S_m^2} \quad \text{(S7)}$$

where S_t^2 is the residual variance for the control experiment and S_m^2 is the residual variance for the mutational variance experiment. This is distributed as an *F*-distribution with 7 degrees of freedom for the technical error experiment and 7*12 = 84 for the mutational variance experiment.

We will be conservative and test the null hypothesis that $\alpha = 5$, i.e., that the physiological variance is at least 5 times bigger than technical variance. Thus we compute the right tail probability of $\frac{7}{6} \cdot \frac{S_t^2}{S_m^2}$ for the $F_{(7,84)}$ distribution. Using a Dunn-Sidak multiple-test correction to hold the experimentwise error rate to 5%, we reject the null hypothesis for 4% of the genes in either stage and conclude that technical error is not a large contributor to the residual variance in the mutational variance experiments.

Jackknifing

We fit models to 11,923 genes with at least 4 replicates per line at each stage. 4,046 of these had significant mutational variance in the blue gut stage and 3,686 had significant mutational variance at white prepupae with 5,008 overall. 4,599 showed a significant developmental change in expression between the two stages with 2,115 of these having significant mutational variance in either stage. To reduce bias in the estimates, we jackknifed the data from the 5,008 significant genes by fitting data from 11 lines at a time to the model chosen from the full dataset, ensuring that each of the 11 lines had at least 3 replicates per line in each stage. We computed the jackknife estimators for variances, BLUPs, and fixed effects by:

$$JS = \frac{\sum_{l}\left(12S_{f} - 11S_{-l}\right)}{12} \qquad \text{(S8)}$$

where JS is thejackknifed estimate, S_f is the estimate from the full dataset, and S_{-l} is the estimate from the dataset without line l[20-22]. SAS could not fit the model for all 12 jackknife datasets for 125 genes; we discarded these leaving us with 11,798 genes for further analysis. We use estimates from this jackknifed set for the analysis (Supplementary data 1).

GO molecular function tests

To test for significant patterns within the GO categories, we also used a Krukal-Wallis test followed by Wilcoxon two-sample tests for specific categories, again controlling multiple testing using the Dunn-Šidák method. Genes in categories at the top level of the GO molecular function classification hierarchy[23] significantly differed in their median h^2m (Kruskal-Wallis test, $p<.001$). Other results are reported in the main text.

Quality threshold clustering

We scaled BLUPs by the gene-specific mutational standard deviation at each stage and clustered them using a quality threshold (e.g. 24) of 0.576 absolute correlation (corresponding to p=0.05 for 12 lines). The 3,475 significantly varying genes at *wpp* sort into 281 clusters (254 for *bg*). The largest of these contains 229 (322) genes, and the sizes rapidly plummet before plateauing with many small clusters. We expect that the largest clusters reflect the effects of changes either in common, proximal *trans*-factors or earlier mutational effects which impact many genes. The smaller cluster sizes may reflect the consequences of mutations in *cis*-factors, epistatic effects of multiple *trans*-factors, or *trans*-effects at the early stages of their developmental impact. We propose that the number of clusters is an upper bound for the number of mutations affecting gene expression in these lines at a particular stage.

References

1. Houle, D. & Nuzhdin, S. V. Mutation accumulation and the effect of *copia* insertions in *Drosophila melanogaster*. *Genetical Research* **83**, 7-18 (2004).

2. Diehl, F., Grahlmann, S., Beier, M. & Hoheisel, J. D. Manufacturing DNA microarrays of high spot homogeneity and reduced background signal. *Nucleic Acids Res.* **29**, e38 (2001).

3. Kent, W. isPcr. 3.1. http://www.soe.ucsc.edu/~kent/src/isPrc31.zip (2004).

4. ftp://ftp.fruitfly.org/pub/download/compressed/na_euchromatin_genomic_dmel_RELEASE4.FASTA.gz (22 Dec 2004)

5. ftp://flybase.net/genomes/Drosophila_melanogaster/current/gff/dmel-[CHR]-r4.0.gff.gz (22 Dec 2004)

6. ftp://flybase.net/genomes/Drosophila_melanogaster/current/fasta/dmel-all-[transcript/gene]-r4.0.fasta.gz (22 Dec 2004)

7. ftp://flybase.net/genomes/Drosophila_melanogaster/current_hetchr/dmel-all-chromosome.hetr32b.fasta (28 Dec 2004)

8. Yang, Y. H., Buckley, M. J., Dudoit, S. & Speed, T. P. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* **11**, 108-136 (2002).

9. van Rossum, G. python. 2.3. http://www.python.org (2003).

10. Ascher, D., Dubois, P. F., Hinsen, K., Hugunin, J. & Oliphant, T. Numerical python. 23.0. http://numpy.sourceforge.net (2001).

11. Team, R. D. C. R: A language and environment for statistical computing. 1.8.1. http://www.R-project.org (2003).

12. Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. *SAS System for Mixed Models* (SAS Institute Inc., Cary, North Carolina, 1996).

13. Fry, J. D. in *Genetic analysis of complex traits with SAS* (ed. Saxton, A.) (SAS Institute Inc., Cary, North Carolina, 2004).

14. Hamlett, A., Ryan, L., Serrano-Trespalacios, P. & Wolfinger, R. Mixed models for assessing correlation in the presence of replication. *J Air Waste Manag Assoc* **53**, 442-50 (2003).

15. Yang, Y. H., Dudoit, S., Luu, P. & Speed, T. P. in *SPIE. Microarrays: Optical Technologies and Informatics* (eds. Bittner, M. L., Chen, Y., Dorsel, A. N. & Dougherty, E. R.) 141-152 (Society for Optical Engineering, San Jose, California, 2001).

16. Yang, Y. H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**, e15 (2002).

17. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289-300 (1995).

18. Rifkin, S. A., Kim, J. & White, K. P. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nature Genetics* **33**, 138-144 (2003).

19. Gu, Z. L., Rifkin, S. A., White, K. P. & Li, W. H. Duplicate genes increase gene expression diversity within and between species. *Nature Genetics* **36**, 577-579 (2004).

20. Quenoille, M. H. Approximate testes of correlation in time series. *Journal of the Royal Statistical Society, Series B* **11**, 68-84 (1949).

21. Tukey, J. Bias and confidence in not quite so large samples. *Annals of Mathematical Statistics* **29**, 614 (1958).

22. Sokal, R. R. & Rohlf, F. J. *Biometry* (W.H.Freeman and Company, New York, 1995).

23. Consortium, T. G. O. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29 (2000).

24. Heyer, L. J., Kruglyak, S. & Yooseph, S. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research* **9**, 1106-1115 (1999).

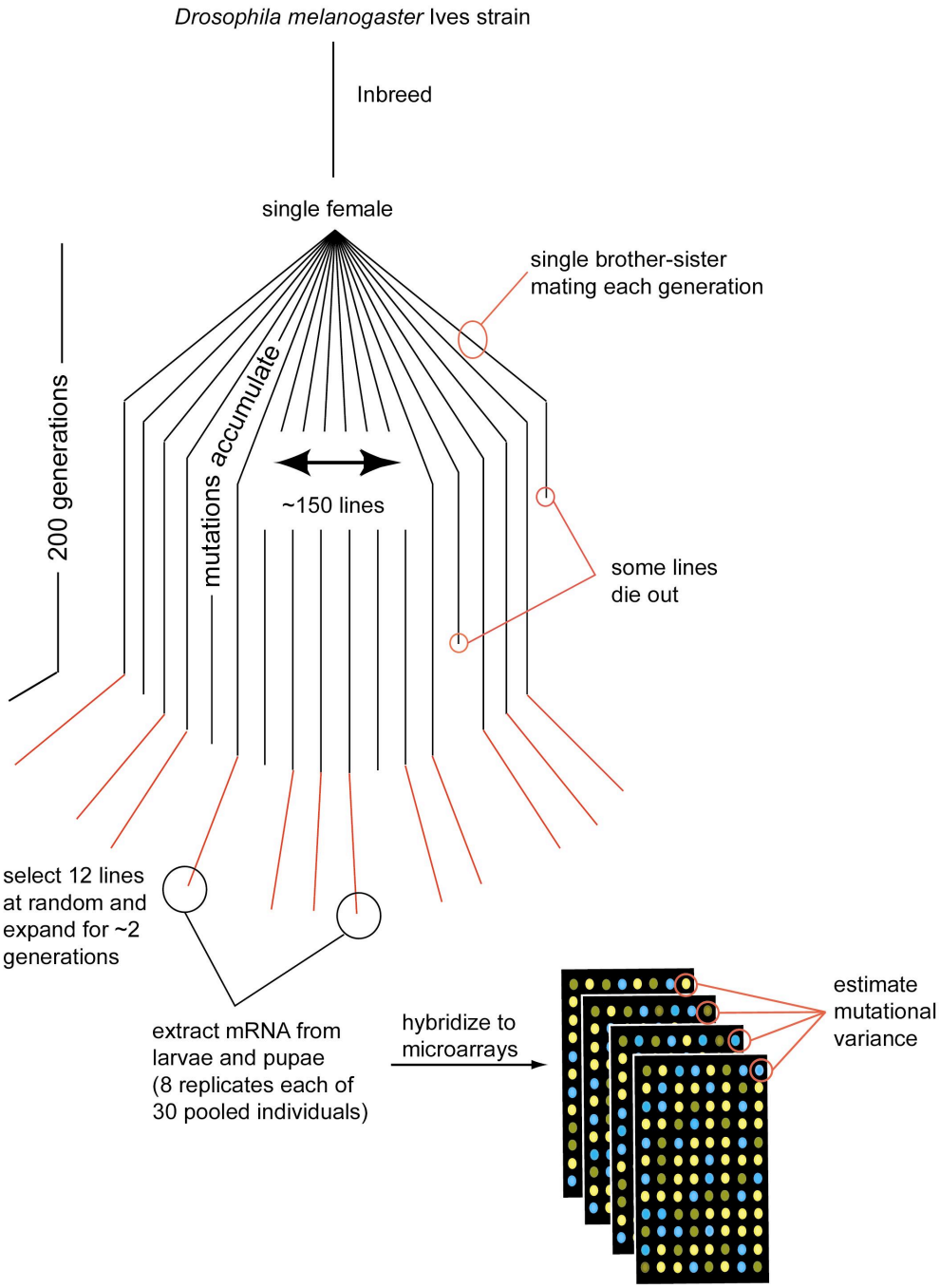
Supplementary data 2.
SAS PROC MIXED calls for the model hierarchy.

```
----------------------------------------
model 0 single spots: No between-line variance, single (spot-specific) within-
line variance.

proc mixed covtest CL IC ratio data = mv;
class slide stage line replicate;
model intensity = stage/solution ddfm = kr outp=resids0_;
random int/subject=slide;
ods output covparms = cp0_ convergencestatus=convs0_ iterhistory=ith0_
infocrit=info0_ solutionf=eblues0_ solutionr = eblups0_;
run;

----------------------------------------
model 0 multiple spots: No between-line variance, single (spot-specific) within-
line variance.

proc mixed covtest CL IC ratio data = mv;
class sequence slide stage line replicate;
model intensity = stage sequence/solution ddfm = kr outp=resids0_;
random sequence/subject=slide type=unr;
ods output covparms = cp0_ convergencestatus=convs0_ iterhistory=ith0_
infocrit=info0_ solutionf=eblues0_ solutionr = eblups0_;
run;

----------------------------------------
model 0b single spots: No between-line variance, stage-specific (spot-specific)
within-line variance.

proc mixed covtest CL IC ratio data = mv;
class slide stage line replicate;
model intensity = stage/solution ddfm = kr outp=resids0b_;
random int/subject=slide;
repeated stage/subject = replicate(line) type = unr(1);

ods output covparms = cp0b_ convergencestatus=convs0b_ iterhistory=ith0b_
infocrit=info0b_ solutionf=eblues0b_ solutionr = eblups0b_;
run;

----------------------------------------
model 0b multiple spots: No between-line variance, stage-specific (spot-
specific) within-line variance.

proc mixed covtest CL IC ratio data = mv;
class sequence slide stage line replicate;
model intensity = stage sequence/solution ddfm = kr outp=resids0b_;
random sequence/subject=slide type=unr;
repeated stage/subject = replicate(line) type = unr(1);
ods output covparms = cp0b_ convergencestatus=convs0b_ iterhistory=ith0b_
infocrit=info0b_ solutionf=eblues0b_ solutionr = eblups0b_;
run;
```

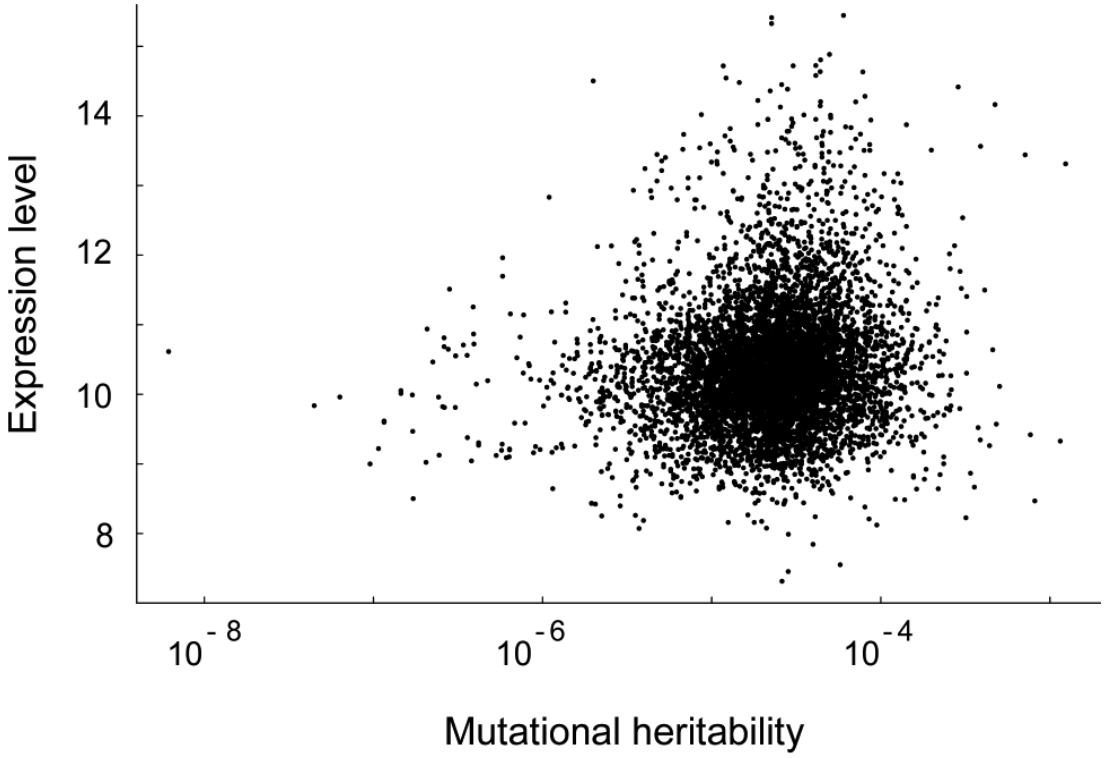
```
----------------------------------------
model 1r0 single spots: Single between-line variance, single (spot-specific)
within-line variance.

proc mixed covtest CL IC ratio data = mv;
class slide stage line replicate;
model intensity = stage/solution ddfm = kr outp=resids1r0_;
random stage/subject = line type = vc solution;
random int/subject=slide;
repeated stage/subject=replicate(line) type=vc;
ods output covparms = cp1r0_ convergencestatus=convs1r0_ iterhistory=ith1r0_
infocrit=info1r0_ solutionf=eblues1r0_ solutionr = eblups1r0_;
run;

----------------------------------------
model 1r0 multiple spots: Single between-line variance, single (spot-specific)
within-line variance.

proc mixed covtest CL IC ratio data = mv;
class sequence slide stage line replicate;
model intensity = stage sequence/solution ddfm = kr outp=resids1r0_;
random stage/subject = line type = vc solution;
random sequence/subject=slide type=unr;
repeated stage/subject=replicate(line) type=vc;
ods output covparms = cp1r0_ convergencestatus=convs1r0_ iterhistory=ith1r0_
infocrit=info1r0_ solutionf=eblues1r0_ solutionr = eblups1r0_;
run;

----------------------------------------
model 2br0 single spots: Single between-line variance, stage-specific (spot-
specific) within-line variance.

proc mixed covtest CL IC ratio data = mv;
class slide stage line replicate;
model intensity = stage/solution ddfm = kr outp=resids2br0_;
random stage/subject = line type = vc solution;
random int/subject=slide;
repeated stage/subject = replicate(line) type = unr(1);
ods output covparms = cp2br0_ convergencestatus=convs2br0_ iterhistory=ith2br0_
infocrit=info2br0_ solutionf=eblues2br0_ solutionr = eblups2br0_;
run;

----------------------------------------
model 2br0 multiple spots: Single between-line variance, stage-specific (spot-
specific) within-line variance.

proc mixed covtest CL IC ratio data = mv;
class sequence slide stage line replicate;
model intensity = stage sequence/solution ddfm = kr outp=resids2br0_;
random stage/subject = line type = vc solution;
random sequence/subject=slide type=unr;
repeated stage/subject = replicate(line) type = unr(1);
ods output covparms = cp2br0_ convergencestatus=convs2br0_ iterhistory=ith2br0_
infocrit=info2br0_ solutionf=eblues2br0_ solutionr = eblups2br0_;
```

```
run;

----------------------------------------
model 2ar0 single spots: Stage-specific between-line variance, single (spot-
specific) within-line variance.

proc mixed covtest CL IC ratio data = mv;
class slide stage line replicate;
model intensity = stage/solution ddfm = kr outp=resids2ar0_;
random stage/subject = line type = unr(1) solution;
random int/subject=slide;
ods output covparms = cp2ar0_ convergencestatus=convs2ar0_ iterhistory=ith2ar0_
infocrit=info2ar0_ solutionf=eblues2ar0_ solutionr = eblups2ar0_;
run;

----------------------------------------
model 2ar0 multiple spots: Stage-specific between-line variance, single (spot-
specific) within-line variance.

proc mixed covtest CL IC ratio data = mv;
class sequence slide stage line replicate;
model intensity = stage sequence/solution ddfm = kr outp=resids2ar0_;
random stage/subject = line type = unr(1) solution;
random sequence/subject=slide type=unr;
ods output covparms = cp2ar0_ convergencestatus=convs2ar0_ iterhistory=ith2ar0_
infocrit=info2ar0_ solutionf=eblues2ar0_ solutionr = eblups2ar0_;
run;

----------------------------------------
model 3r0 single spots: Stage-specific between-line variance, stage-specific
(spot-specific) within-line variance.

proc mixed covtest CL IC ratio data = mv;
class slide stage line replicate;
model intensity = stage/solution ddfm = kr outp=resids3r0_;
random stage/subject = line type = unr(1) solution;
random int/subject=slide;
repeated stage/subject = replicate(line) type = unr(1);
ods output covparms = cp3r0_ convergencestatus=convs3r0_ iterhistory=ith3r0_
infocrit=info3r0_ solutionf=eblues3r0_ solutionr = eblups3r0_;
run;

----------------------------------------
model 3r0 multiple spots: Stage-specific between-line variance, stage-specific
(spot-specific) within-line variance.

proc mixed covtest CL IC ratio data = mv;
class sequence slide stage line replicate;
model intensity = stage sequence/solution ddfm = kr outp=resids3r0_;
random stage/subject = line type = unr(1) solution;
random sequence/subject=slide type=unr;
repeated stage/subject = replicate(line) type = unr(1);
ods output covparms = cp3r0_ convergencestatus=convs3r0_ iterhistory=ith3r0_
infocrit=info3r0_ solutionf=eblues3r0_ solutionr = eblups3r0_;
run;
```

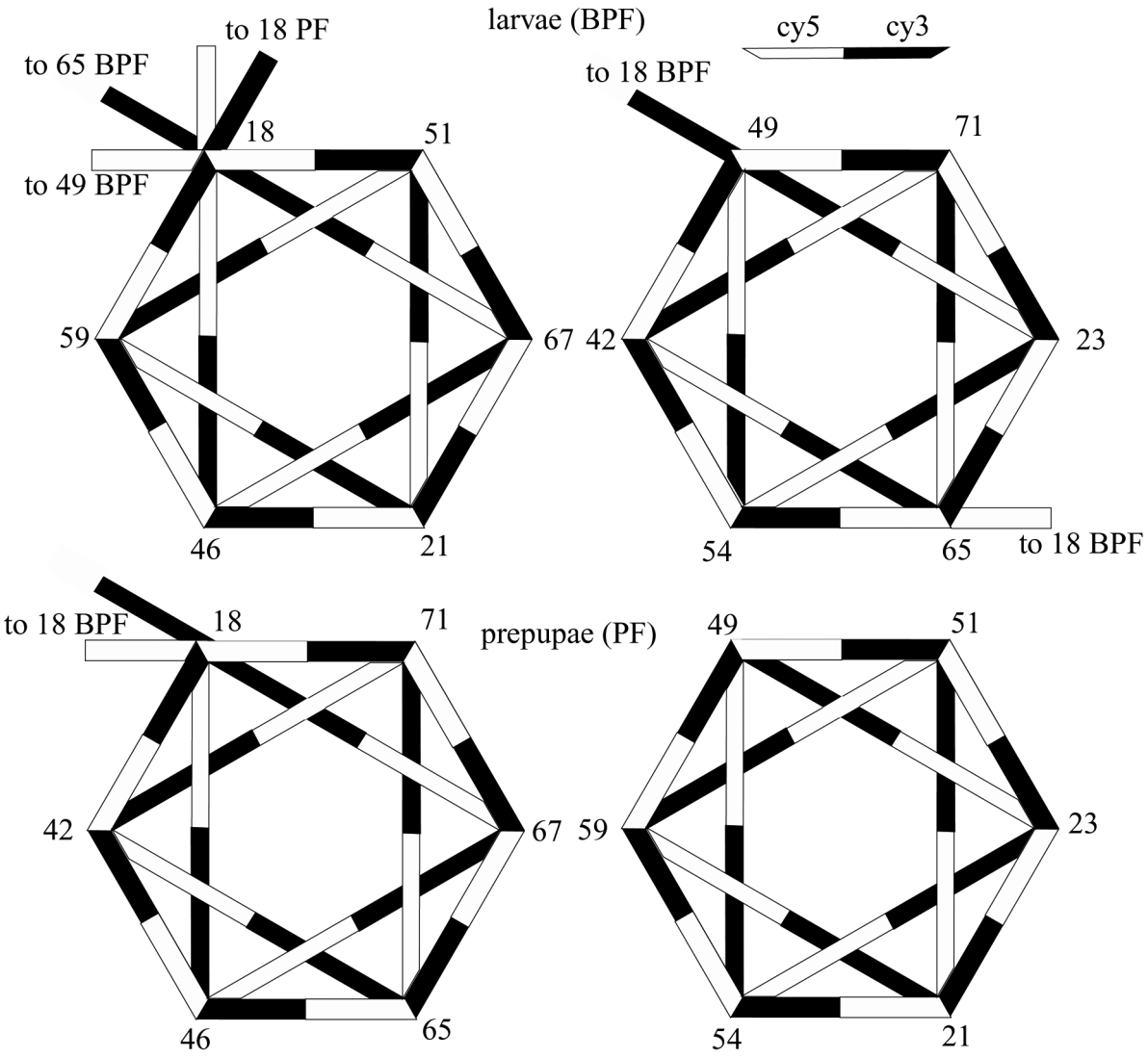
Rifkin, Houle, Kim, and White. A mutation accumulation assay reveals a broad capacity for rapid gene expression evolution.



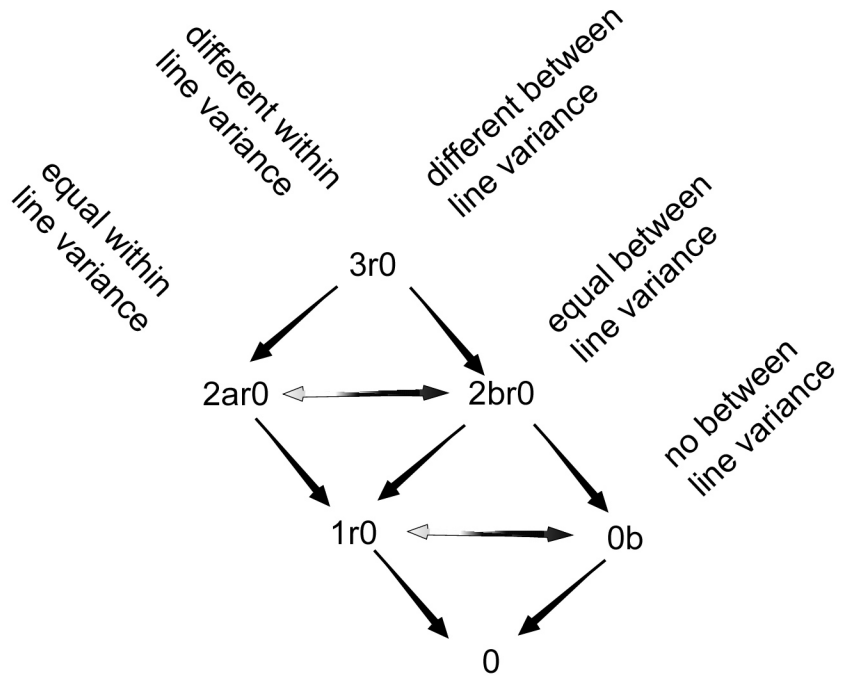
Supplementary figure 1. Schematic of sample collection. Flies were inbred for approximately 40 generations, then separated into 150 different lines. These lines were brother-sister mated and accumulated mutations. After 200 generations, 12 lines picked at random from the surviving lines and rapidly expanded. 8 replicate samples of approximately 30 flies each were collected at stages BPF and PF (see text). mRNA was extracted from these samples, reverse transcribed and fluorescently labelled, and hybridized to microarrays. Microarrays were scanned, normalized and analyzed as described in the text.



Supplementary figure 2. Mutational heritability versus expression level. Each datapoint represents the mutational heritability for a gene at a particular stage and the average expression level across the lines for each gene by stage combination. Only gene by stage combinations with significant mutational heritability are included. There is a slight correlation between mutational heritability and expression level (r_s=0.11).



Supplementary figure 3. Experimental design. For each stage, each line was competitively hybridized with four other lines within its hexagon, two lines in its the other hexagon for its stage, and twice with itself in the other stage. A black and white bar represents one microarray. Connections between hexagons are depicted only for arrays involving line 18 at BPF.



Supplementary figure 4. Model selection procedure. We performed a series of likelihood ratio tests going from more complex to simpler models to find the best estimates of mutational variance for each gene. Solid arrows denote tests of nested models. Graded arrows represent non-nested tests.