nature genetics

Regulatory changes underlying expression differences within and between *Drosophila* species

Patricia J Wittkopp[1,2], Belinda K Haerum[1,2] & Andrew G Clark[1]

Differences in gene expression are an important source of phenotypic variation, and can be caused by changes in *cis* and/or *trans* regulation. *cis*-regulatory variants alter allele-specific expression, whereas *trans*-regulatory variants influence expression of both alleles in a diploid cell. Because of this difference, we hypothesize that natural selection may favor one type of change over the other. Here, we investigate contributions of *cis*- and *trans*-regulatory changes to variable intra- and interspecific gene expression using four strains of *Drosophila melanogaster*, three strains of *D. simulans* and a total of 78 genes. We show that *cis*-regulatory changes account for a greater proportion of the expression differences observed between rather than within species. These data are inconsistent with a neutral model assuming equal probabilities of fixation for *cis*- and *trans*-regulatory polymorphisms, suggesting that natural selection influences the molecular mechanisms underlying divergent gene expression. Specifically, *cis*-regulatory changes seem to accumulate preferentially over time.

The contributions of *cis*- and *trans*-regulatory changes to variable gene expression within and between species have previously been examined[1–7]; however, the use of different techniques and different species to measure intraspecific polymorphism and interspecific divergence has precluded a comparison. This study uses the same species, the same technique, and an overlapping gene set to directly compare the molecular mechanisms underlying expression differences within and between species.

The relative contribution of *cis*- and *trans*-regulatory changes to variable gene expression is estimated by comparing the relative gene expression between two strains to the relative allelic expression in F_1 hybrids produced from crossing the two strains[2]. This strategy allows the combined effects of *cis*-regulatory (that is, allele-specific) differences to be separated from the combined effects of *trans*-regulatory differences. Asymmetric expression of two alleles under the same cellular conditions indicates a difference in *cis*-regulatory activity and can result from changes in enhancer or promoter sequences that affect transcription, changes in the transcribed region that affect mRNA stability, or allele-specific epigenetic changes that alter chromatin structure. In this study, we infer *trans*-regulatory differences when the relative *cis*-regulatory activity of strain-specific alleles does not fully explain the expression differences between the two strains; these differences can result from genetic and epigenetic changes affecting the activity or availability of proteins and RNAs that mediate gene expression.

Extensive variation in gene expression exists both within and between *Drosophila* species[8–12]. To compare the genetic bases of these intra- and interspecific differences, we examined the relative contributions of *cis*- and *trans*-regulatory differences to variable gene expression between strains and species. Gene expression was examined in seven crosses involving four inbred lines of *D. melanogaster*, three crosses involving three inbred lines of *D. simulans*, and five interspecific crosses using a subset of these inbred lines. These two species diverged approximately 2.5 million years ago[13], yet some strains are still able to mate and produce offspring in a laboratory setting, permitting this analysis.

We examined 45 genes within species and 49 genes between species, with 16 genes analyzed in both intra- and interspecific comparisons (**Supplementary Table 1** online). These genes were divided into two sets (**Supplementary Table 2** online). Gene set 1 contained a subset of randomly selected genes examined in our previous study[2], with the specific genes analyzed in each cross dictated primarily by allele sharing. Gene set 2 was created to increase sample size and test for consistency among gene sets. It contained genes selected from prior studies of intra-[8] and interspecific[14] expression that encompass the variety of inheritance patterns observed (**Supplementary Table 2**). Two genes from our previous study[2] were also included with gene set 2 for analysis because they were only sampled in one interspecific cross. Additional discussion of the gene sets is provided in the **Supplementary Note** online.

We determined relative gene expression between strains and relative allelic expression in F_1 hybrids using Pyrosequencing[2], a technique that yields measurements of gene expression comparable to those from microarrays and quantitative PCR[15]. Measurements of gene expression were reproducible among assays, replicate samples, and reciprocal crosses, but they varied among developmental stages, as expected (**Supplementary Note**). *cis*-regulatory differences (C) were

[1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. [2]Department of Ecology and Evolutionary Biology, Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, Michigan 48109, USA. Correspondence should be addressed to P.J.W. (wittkopp@umich.edu).

Received 15 June 2007; accepted 11 December 2007; published online 17 February 2008; doi:10.1038/ng.77

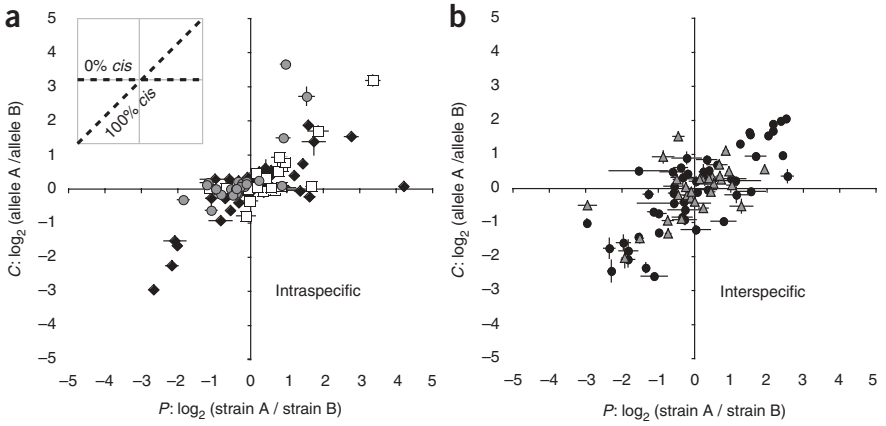

Figure 1 *cis*-regulatory changes explain more of the expression differences between rather than within species. (**a**,**b**) Relative expression between parental strains (*P*) is plotted against the relative allelic expression in F_1 hybrids (*C*) for intraspecific (**a**) and interspecific (**b**) comparisons. As shown in the inset of **a**, genes for which *cis*-regulatory differences can explain all (100% *cis*) of the expression differences between strains fall on the diagonal $y = x$ line. Genes with no difference in *cis*-regulatory activity between strains (0% *cis*) will fall along the horizontal $y = 0$ line. In **a**, black diamonds represent comparisons between *D. melanogaster* strains for gene set 1, white squares represent comparisons between *D. simulans* strains for gene set 1, and gray circles represent comparisons from gene set 2. In **b**, black circles and gray triangles represent interspecific comparisons from gene sets 1 and 2, respectively. Error bars show standard errors for each measurement of *P* and *C*. The distribution of standard errors for *P* was similar for intra- and interspecific comparisons ($P_{MWU} = 0.77$), with three interspecific and two intraspecific comparisons having standard errors >0.5. Standard errors of *C* were significantly larger between than within species, decreasing power to detect small interspecific *cis*-regulatory difference. This may reflect greater instability of gene expression in interspecific hybrids and/or increased environmental variance from more replicate vials. This difference in power introduces a conservative bias, suggesting the true excess of *cis*-regulatory changes between relative to within species may be even greater than reported in **Supplementary Table 4**.

calculated as the ratio of allelic expression in F_1 heterozygotes (C = allele 1 / allele 2), and the expression difference between parental strains (P) was calculated as the ratio of total transcript abundance between strains (P = strain 1 / strain 2). We estimated *trans*-regulatory differences (T) as the difference between these two measures ($T = P - C$).

Ratios of expression for each gene were normalized using genomic DNA[2,16], $\log_2$ transformed, and fitted to a mixed linear model (**Supplementary Methods** online). We used least-squares means and their confidence intervals from the mixed model to test for significant differences between parental strains in total expression ($P \neq 0$), *cis* regulation ($C \neq 0$) and *trans* regulation ($T \neq 0$). An experiment-wise ($n = 510$) false discovery rate (FDR) of 3% (Q value ≤ 0.03, P value ≤ 0.078) was used to determine significance[17]. All least-squares means and significance tests are summarized in **Supplementary Table 3** online. We detected significant changes in both *cis* and *trans* regulation more often between rather than within species, although this difference was not significant for all subsets of genes (**Supplementary Table 4** online).

We used estimates of *cis*-, *trans*- and total regulatory differences to determine whether the genetic basis of regulatory variation was similar within and between species. **Figure 1** shows the relative contributions of *cis*- and *trans*-acting changes as a plot of the relative allelic expression (C) against the relative expression difference between parental strains (P). If *cis*-regulatory changes are solely responsible for the expression differences between parental strains, a linear regression of C on P will result in a regression coefficient of 1, whereas if *trans*-regulatory changes are solely responsible for the expression differences between parental strains, the regression coefficient will be 0 (**Fig. 1a**, inset). Given two datasets with similar distributions of expression differences (**Supplementary Fig. 1** online) and similar measurement errors, *cis*-regulatory changes account for more of the total expression divergence in the dataset with the larger regression coefficient[18]. Regression analyses comparing the contribution of allele-specific, *cis*-regulatory differences to variable gene expression within and between species suggest that *cis*-regulatory differences account for more of the total expression divergence between rather than within species. We observed larger *cis*-regulatory effects between as compared to within species using multiple regression models as well as subsets of the data (**Supplementary Table 5** online).

To further investigate whether *cis*-regulatory effects were larger between rather than within species, we restricted our analysis to genes with a significant expression difference between parental strains ($P \neq 0$) and used nonparametric tests to compare the magnitude of *cis*-regulatory differences within and between species. Again, we observed that the median *cis*-regulatory effects were larger between than within species (**Fig. 2**, left), although scaling by the magnitude of expression differences between parental strains reduced the magnitude of this difference (**Fig. 2**, middle). Scaling *cis*-regulatory differences by total regulatory divergence ($|C| + |T|$) further reduced the difference within and between species by lowering the interspecific measures (**Fig. 2**, right), indicating that compensatory *cis*- and *trans*-regulatory changes are more common between than within species. Although all of the distributions are consistent with larger *cis*-regulatory differences between species, the nonparametric Mann-Whitney U tests do not show statistically significant differences in all cases (**Fig. 2** and **Supplementary Fig. 2** online). This may reflect differences in the gene sets examined within and between species.

To avoid the comparison of different gene sets, we carried out an independent analysis that allowed us to compare intra- and interspecific regulatory variation using the same genes, the same alleles, and even the same samples. This analysis used data only from interspecific crosses; intraspecific regulatory differences were extracted from interspecific comparisons using one species as a common 'reference' strain. That is, we calculated the difference between strain A and strain B from differences between strains A and C and strains B and C, where C was a different species than strains A and B. This analysis directly compares intraspecific (A/B) and interspecific (A/C and B/C) differences for the same genes. We calculated the magnitude of each intraspecific *cis*-regulatory difference as the difference between the $\log_2$ expression ratios for interspecific comparisons: $C_{intra} = |C_{inter\ 1} - C_{inter\ 2}|$. Similarly, we calculated the difference in *trans* regulation between strains as the difference between the $\log_2$ estimates of *trans*-regulatory effects: $T_{intra} = |T_{inter\ 1} - T_{inter\ 2}|$.

Our data contained 59 cases in which the same gene was analyzed in two interspecific crosses using a common allele of *D. melanogaster* or

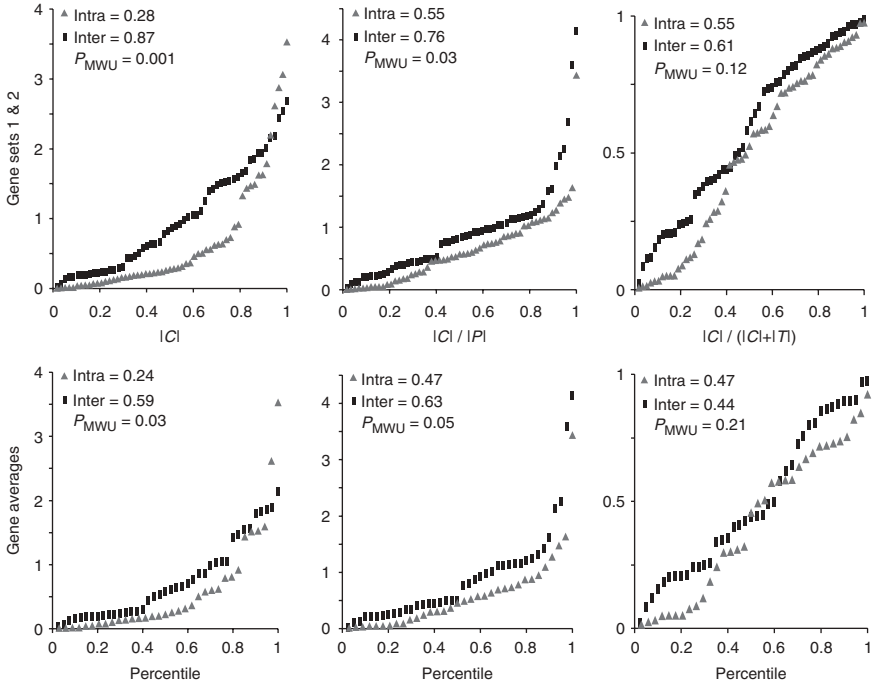

Figure 2 Distributions of *cis*-regulatory effects within and between species. For genes with a significant expression difference between parental strains, ranked values of *cis*-regulatory differences (|*C*|) are shown for intraspecific (gray triangle) and interspecific (black square) comparisons (left). Ranked distributions of *cis*-regulatory differences scaled by total expression divergence (middle, |*C*| / |*P*|) and total regulatory divergence (right, |*C*| / (|*C*| + |*T*|)) are also shown. The top row displays the full dataset, whereas the bottom row displays data in which each gene analyzed is represented exactly once within and between species. These 'gene averages' were calculated to eliminate the pseudo-replication introduced by analyzing some genes in more than one intraspecific (or interspecific) cross (see **Supplementary Table 1**). Separate analysis of gene sets 1 and 2 show the same patterns (**Supplementary Fig. 2**). Median values are shown with the significance of a Mann-Whitney U test comparing the distributions. Note that for data scaled by total regulatory divergence (right), the medians underestimate differences present in most of the rest of the distribution. These data are based on a 3% FDR used to identify genes with significant differences between strains. Reanalysis of these data using a more stringent 1% FDR excluded only 16 of the 115 cases and had very little effect on the overall distributions (data not shown). The comparison of |*C*| / |*P*|, however, did become nonsignificant ($P = 0.18$) despite similar medians (intraspecific: 1% FDR = 0.57, 3% FDR = 0.55; interspecific: 1% FDR = 0.76, 3% FDR = 0.76).

D. simulans. For each case, the proportion of total regulatory divergence resulting from *cis*-regulatory differences, $|C| / (|C| + |T|)$, was calculated for the one intraspecific and two interspecific comparisons. We again found that *cis*-regulatory differences explained more of the total regulatory variation between than within species (**Fig. 3a**, Friedman's test of related samples, $P \leq 0.001$ (χ^2 test, 2 degrees of freedom)). The median percentage of regulatory divergence explained by *cis*-regulatory differences was 35% within species and 64% between species. The mean rank of the intraspecific comparisons was 1.31, whereas the mean ranks of the two interspecific comparisons were 2.36 and 2.32.

To assess the significance of these ranks, we created 100 permuted datasets by randomly pairing interspecific measurements from the 59 cases and calculating hypothetical intraspecific data. These 'intraspecific' measurements had the smallest mean rank in only 18 of the 100 permuted datasets (minimum rank = 1.69), with $P > 0.05$ for Friedman's test of related samples in all 18 cases. The distribution of *cis*-regulatory effects observed in one of these permuted datasets is shown in **Figure 3b** for comparison.

The observation that *cis*-regulatory changes account for more of the expression divergence between than within species is inconsistent with a simple additive, neutral model of regulatory evolution in which all new regulatory mutations (both *cis* and *trans* acting) have an equal probability of fixation. Even if *cis*- and *trans*-acting regulatory mutations arise at different frequencies and have different distributions of effect sizes (which we suspect is true), in the absence of selection, the proportion of regulatory divergence attributable to *cis*-regulatory changes should remain similar over time (this logic is analogous to the McDonald-Kreitman test used to compare the rate of synonymous and nonsynonymous changes within and between species[19]). Therefore, we infer that natural selection influences the molecular mechanisms underlying regulatory variation

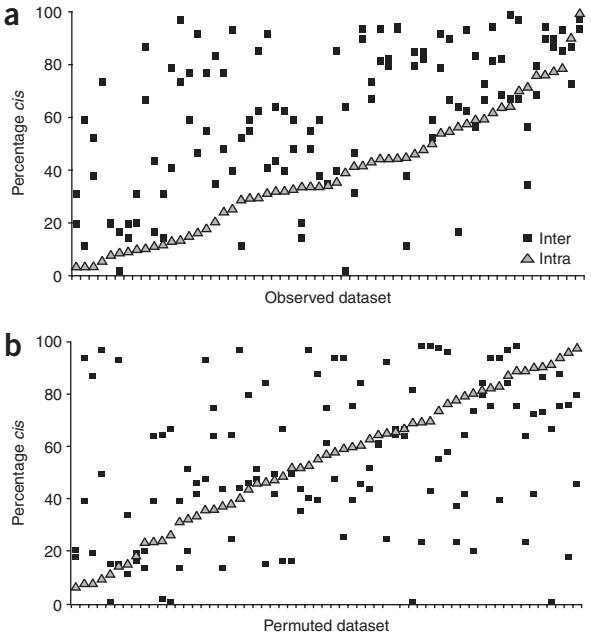

Figure 3 The percentage of regulatory divergence attributable to *cis*-regulatory changes is larger between than within species. (**a**) Ranked values of percent *cis*-regulatory divergence ($|C| / (|C| + |T|) \times 100$) are shown for intra- and interspecific comparisons estimated using data from interspecific crosses. Each column contains one intraspecific and two interspecific comparisons. (**b**) One of the 100 datasets created by randomly pairing interspecific data from the 59 cases analyzed in **a** is shown. In both plots, gray triangles represent intraspecific comparisons, and black squares represent interspecific comparisons.

between species. Specifically, our data indicate that *cis*-regulatory changes contribute more to interspecific expression divergence than predicted by intraspecific variation.

To explain the observed patterns of *cis*- and *trans*-regulatory effects within and between species, we propose that *trans*-acting mutations arise more frequently than *cis*-acting mutations (because of a larger mutational target size[20,21]), but that *cis*-regulatory mutations are more likely to become fixed within a population. Results from a recent study in *Caenorhabditis elegans* are consistent with this proposal: a comparison of regulatory divergence between mutation accumulation lines and natural isolates revealed that *trans*-acting changes arise frequently, but are selected against in natural populations[22]. We anticipate that new *trans*-acting mutations will include both highly pleiotropic changes (presumably associated with large selection coefficients) as well as mutations with more limited effects. The buffering and stability inherent to regulatory systems may allow this latter class of *trans*-regulatory mutations to be maintained within species by mutation-selection balance. *cis*-regulatory divergence may accumulate preferentially between species because of negative selection against *trans*-acting variants (as observed in *C. elegans*) and/or positive selection for *cis*-regulatory changes[23]. The relative impact of these two forces on mechanisms of regulatory evolution is yet to be determined.

Although we believe that this is the best explanation for our data based on published work, it is currently unknown whether population genetic factors, such as demography, could also alter the relative contribution of *cis*- and *trans*-regulatory variation over time. A second factor to consider is that, because whole flies were used for this study, the effects of *cis*- and *trans*-acting variants were combined across tissues. (However, we expect the effects of pooling tissues to be similar within and between species and find it unlikely that this pooling is responsible for the difference in intra- and interspecific measurements.) Regional differences in levels of polymorphism for genes sampled within and between species are also unlikely to explain our results (**Supplementary Note** and **Supplementary Fig. 3** online). Finally, it remains to be seen whether a similar pattern will be observed in other species groups.

This study improves our understanding of regulatory evolution by showing that selection may influence the relative contribution of *cis*- and *trans*-regulatory changes to expression differences over time. It emphasizes that selection acting on phenotypes (for example, gene expression levels) alters the genetic mechanisms used to generate these phenotypes and strongly supports the hypothesis that *cis*-regulatory mutations are an important source of evolutionary change[23].

METHODS

Fly strains, crosses, RNA collections and cDNA synthesis. The strains, crosses and genes surveyed in each cross are summarized in **Supplementary Table 1**. For each cross, we set up 5–20 replicate mating vials, each containing three male and three female virgin flies (interspecific crosses required more replicates because of reduced mating success). We collected 7- to 10-d-old mated female flies, combined them among replicates, and divided them into four replicate pools of 14 flies each for F_1 hybrids and seven flies of each parental type for comparisons of total expression between parental strains. Total RNA and genomic DNA were extracted sequentially from each pool[2]. We synthesized cDNA in duplicate for hybrid pools and in triplicate for parental pools using a polyT primer. Note that expression levels were measured in whole flies and reflect the combined expression in all tissues; this may mask offsetting regulatory changes in different tissues.

Measuring gene expression. We carried out pyrosequencing reactions as previously described[2]. For each gene analyzed in each cross, we measured relative expression in the parental strains once in each of the 12 cDNA samples and in duplicate for the four genomic DNA samples. Relative allelic expression in hybrids was measured for the 8 replicate cDNA samples and once for each hybrid genomic DNA sample. We selected these levels of replication on the basis of variance components observed in our prior work[2]. Consistent with our prior study[2], biological replicates (pools of flies) were the largest source of variance for most analyses (**Supplementary Table 3**). This variance does not reflect a single outlier sample: the pool that deviated most from the mean differed among comparisons (data not shown). Reciprocal crosses were also analyzed and showed no evidence of genomic imprinting (**Supplementary Fig. 4** online), consistent with previously published results[15].

Genes analyzed. Gene set 1 consisted of 23 of the 29 genes with a significant expression difference in our previous study[2] that were found to be suitable for Pyrosequencing analysis within and between species (**Supplementary Methods**). Gene set 2 consisted of 2 additional genes from our previous study[2] and 24 (ref. 8) and 29 genes[14] from other prior studies (**Supplementary Table 2**). Genes were selected from these studies[8,14] to sample the variety of inheritance patterns observed for gene expression in these studies. Potential implications of this gene selection are discussed in the **Supplementary Note**.

Statistical analysis. Ratios of allelic expression were normalized using measurements of genomic DNA and fitted to a mixed linear model using SAS v8.2 as described previously[16] and in **Supplementary Methods**. We carried out additional parametric and nonparametric tests using SPSS 11 and orthogonal regressions[24], and we calculated false discovery rates used to assess statistical significance using QVALUE[17].

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank R. Haygood, K. Montooth, H.A. Orr, B. Payseur, D. Presgraves, G. Wray and anonymous reviewers for helpful discussions of this work. We thank D. Barbash, H. Hollocher, C. Landry, S. Nuzhdin and H.A. Orr for sharing fly strains, and K. Montooth, J. Zhang, B. Payseur, and G. Kalay for comments on the manuscript. Funding for this project was provided by US National Institutes of Health grants to A.G.C. P.J.W. was a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation, and B.K.H. was funded by a Howard Hughes Undergraduate Research award.

AUTHOR CONTRIBUTIONS

P.J.W. conceived and designed the experiments. B.K.H. and P.J.W. performed fly crosses and collected the allelic expression data. P.J.W. performed statistical analyses and prepared the manuscript in consultation with A.G.C., who also provided reagents and other resources for the project.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Yvert, G. *et al.* Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* **35**, 57–64 (2003).
2. Wittkopp, P.J., Haerum, B.K. & Clark, A.G. Evolutionary changes in *cis* and *trans* gene regulation. *Nature* **430**, 85–88 (2004).
3. West, M. *et al.* Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* **175**, 1441–1450 (2007).
4. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
5. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
6. Li, Y. *et al.* Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.* **2**, e222 (2006).
7. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
8. Gibson, G. *et al.* Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. *Genetics* **167**, 1791–1799 (2004).
9. Meiklejohn, C.D., Parsch, J., Ranz, J.M. & Hartl, D.L. Rapid evolution of male-biased gene expression in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **100**, 9894–9899 (2003).
10. Michalak, P. & Noor, M.A. Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. *Mol. Biol. Evol.* **20**, 1070–1076 (2003).
11. Ranz, J.M., Castillo-Davis, C.I., Meiklejohn, C.D. & Hartl, D.L. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* **300**, 1742–1745 (2003).
12. Rifkin, S.A., Kim, J. & White, K.P. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* **33**, 138–144 (2003).
13. Powell, J.R. *Progress and Prospects in Evolutionary Biology: the Drosophila Model* (Oxford Univ. Press, New York, 1997).

14. Ranz, J.M., Namgyal, K., Gibson, G. & Hartl, D.L. Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. *Genome Res.* **14**, 373–379 (2004).
15. Wittkopp, P.J., Haerum, B.K. & Clark, A.G. Parent-of-origin effects on mRNA expression in *Drosophila melanogaster* not caused by genomic imprinting. *Genetics* **173**, 1817–1821 (2006).
16. Landry, C.R. *et al.* Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. *Genetics* **171**, 1813–1822 (2005).
17. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
18. Osada, N., Kohn, M.H. & Wu, C.I. Genomic inferences of the *cis*-regulatory nucleotide polymorphisms underlying gene expression differences between *Drosophila melanogaster* mating races. *Mol. Biol. Evol.* **23**, 1585–1591 (2006).
19. McDonald, J.H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
20. Wittkopp, P.J. Genomic sources of regulatory variation in *cis* and in *trans*. *Cell. Mol. Life Sci.* **62**, 1779–1783 (2005).
21. Landry, C.R., Lemos, B., Rifkin, S.A., Dickinson, W.J. & Hartl, D.L. Genetic properties influencing the evolvability of gene expression. *Science* **317**, 118–121 (2007).
22. Denver, D.R. *et al.* The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat. Genet.* **37**, 544–548 (2005).
23. Wray, G. The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216 (2007).
24. Akritas, M. & Bershady, M. Linear regression for astronomical data with measurement errors and intrinsic scatter. *Astrophys. J.* **470**, 706–714 (1996).