LETTER

doi:10.1038/nature09634

Gene expression divergence recapitulates the developmental hourglass model

Alex T. Kalinka[1]*, Karolina M. Varga[1]*†, Dave T. Gerrard[2], Stephan Preibisch[1], David L. Corcoran[3], Julia Jarrells[1], Uwe Ohler[3], Casey M. Bergman[2] & Pavel Tomancak[1]

The observation that animal morphology tends to be conserved during the embryonic phylotypic period (a period of maximal similarity between the species within each animal phylum) led to the proposition that embryogenesis diverges more extensively early and late than in the middle, known as the hourglass model[1,2]. This pattern of conservation is thought to reflect a major constraint on the evolution of animal body plans[3]. Despite a wealth of morphological data confirming that there is often remarkable divergence in the early and late embryos of species from the same phylum[4–7], it is not yet known to what extent gene expression evolution, which has a central role in the elaboration of different animal forms[8,9], underpins the morphological hourglass pattern. Here we address this question using species-specific microarrays designed from six sequenced *Drosophila* species separated by up to 40 million years. We quantify divergence at different times during embryogenesis, and show that expression is maximally conserved during the arthropod phylotypic period. By fitting different evolutionary models to each gene, we show that at each time point more than 80% of genes fit best to models incorporating stabilizing selection, and that for genes whose evolutionarily optimal expression level is the same across all species, selective constraint is maximized during the phylotypic period. The genes that conform most to the hourglass pattern are involved in key developmental processes. These results indicate that natural selection acts to conserve patterns of gene expression during mid-embryogenesis, and provide a genome-wide insight into the molecular basis of the hourglass pattern of developmental evolution.

The notion that early development is similar among related animal species has been a guiding principle in comparative embryology since von Baer (1828) formalized the observation as his third law[10]. Darwin (1859) believed this to be the most compelling evidence in favour of common descent, reasoning that adult life-stages will afford the greatest opportunity for natural selection to operate, and thus adult structures should show signs of species-specific adaptations more than earlier stages[11]. These earlier stages, where adaptive opportunities are limited, will ultimately represent the 'pruned' but necessary features of ancestral differentiation[12].

Despite its intuitive appeal, the principle of early embryonic conservation has not been supported by morphological studies[2]. Counter to the expectations of early embryonic conservation, many studies have shown that there is often remarkable divergence between related species both early and late in development, often with little apparent influence on adult morphology[4–7]. The extensive variation that is seen in early and late development is contrasted by a period of conserved morphology occurring in mid-embryogenesis. This is known as the phylotypic period because it coincides with a period of maximal similarity between the species within each animal phylum[13].

The morphological conservation evident in the phylotypic period motivated a proposal of the hourglass model[1,2] as a revised formulation of von Baer's third law. The hourglass model predicts that early and late divergence is separated by a 'waist' corresponding to the phylotypic period. One of these studies argues that an increase in the number of global interactions between genes and developmental processes during the phylotypic period renders any evolutionary modification highly deleterious due to their damaging side-effects[2], whereas the other study views conservation during this period as a consequence of the need for precise coordination between growth and patterning, which is seen to be reflected in the genomic organization of the vertebrate Hox genes[1].

Support for the hourglass model has been found at the morphological[7,14] and sequence levels[15–17]. However, both the model and the concept of the phylotypic period remain controversial subjects in the literature[3,18], with some studies of heterochrony in vertebrates indicating that divergence peaks at the phylotypic period[19] or that there is no temporal pattern of phenotypic conservation[20].

Although it is generally appreciated that gene expression divergence has a key role in the evolution of morphological diversity[8,9], no studies so far have addressed the extent to which expression divergence underpins the morphological hourglass pattern at the genome-wide level. Here, we test the molecular basis of the hourglass model of developmental evolution using gene expression data from six *Drosophila* species with sequenced genomes (*D. melanogaster*, *D. simulans*, *D. ananassae*, *D. persimilis*, *D. pseudoobscura* and *D. virilis*), thereby enabling unambiguous quantitative comparisons across orthologous genes for a set of species separated by up to 40 million years. Gene expression levels were measured for 3,019 genes, known to be expressed during embryonic development from RNA *in situ* data[21], at 2-h intervals for the majority of embryogenesis using a microarray time course with three biological replicates per species and four species-specific probes per gene (Supplementary Figs 1 and 2).

For each gene in each species we generated a gene expression time course, corrected for differences in developmental time (Supplementary Information, Section 2.2), and measured the correlation of the resulting temporal profiles for each pair of species (Fig. 1a, b). The distribution of the correlation coefficients shows that whereas most genes are positively correlated in their temporal expression, the divergence in embryonic gene expression follows the known phylogenetic relationships[22]. These results clearly demonstrate that there is evolutionary signal across the data set as a whole.

To quantify gene expression divergence rigorously we fitted a linear model to the expression data. This approach enables us to quantify the divergence between species by measuring the influence that different species have on the expression of individual genes at specific times during development. We extract two different measures of divergence from the model: quantitative divergence, which reflects differences in expression across the whole time course; and temporal divergence, which reflects divergence of temporal profiles at specific time points (Supplementary Information, Section 2.6). We show that both of these measures of

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany. [2]Faculty of Life Sciences, The University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK. [3]Institute for Genome Sciences and Policy, Duke University, 101 Science Drive, Durham NC 27708, USA. †Present address: New Biochemistry Building, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK.
*These authors contributed equally to this work.

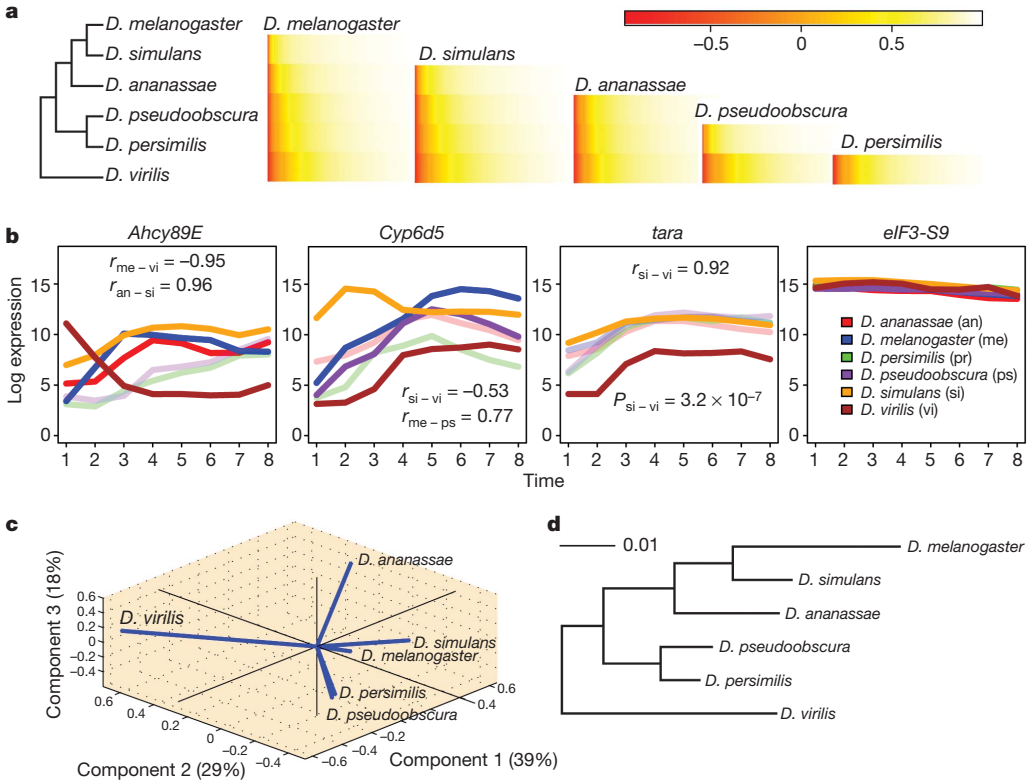

Figure 1 | Gene expression during *Drosophila* embryogenesis recapitulates the known phylogeny. **a**, Between-species pairwise correlation coefficients for temporal profiles are depicted using a colour gradient. **b**, Species profiles are shown for two genes with both positive and negative correlations between different species pairs (*Ahcy89E* and *Cyp6d5*) and two genes that are temporally conserved (*tara* and *eIF3-S9*). Log_2 expression profiles are averaged over probes and replicates. Selected correlation coefficients are shown on the plots, and the *P*-value refers to quantitative divergence. Time points along the *x* axis are 2-h intervals starting from 0–2 h (1) and ending at 14–16 h (8). **c**, The first three principal components for quantitative divergence. **d**, A maximum likelihood phylogeny based on temporal expression divergence across all genes.

divergence recapitulate the known evolutionary relationships between the species when the phylogeny is constructed using all of the genes simultaneously (Fig. 1c, d). However, despite producing an identical topology to the known phylogeny, we see relatively long terminal branches in the phylogram, indicating that gene expression divergence does not scale with the amount of time separating pairs of species (Supplementary Fig. 3)[23].

If temporal expression divergence has saturated through time then we would expect to find a reduced capacity for reconstructing the known phylogeny at the level of individual genes. To explore this possibility we estimated the phylogenetic signal for each gene using a statistic that compares the observed phylogenetic signal to what would be expected under a process of random evolutionary change[24]. A random evolutionary process produces a phylogeny where closely related species resemble each other more than distantly related species as lineages inherit the random changes of their ancestors. The results show that at each time point the majority of genes exhibit a weaker phylogenetic signal than expected under random evolution (Supplementary Fig. 5a).

Phylogenetic signal may be eroded by stabilizing selection[23], and to test for this possibility we compared different evolutionary scenarios by fitting four alternative models to the expression data. The models were purely random evolutionary change and three stabilizing selection models where the optimal expression level may vary between groups of species, allowing us to model adaptive changes in expression (Supplementary Fig. 4)[25]. The stabilizing selection models describe the change in expression as a combination of random changes and stabilizing selection curtailing the accumulating variance. The results show that at each time point at least 80% of genes fit best to models that incorporate stabilizing selection (Supplementary Fig. 5b). We also see that a substantial fraction of the genes fit best to models where there are adaptive changes in expression, indicating that a combination of both stabilizing and directional selection may be acting on a large fraction of the genes[26].

The variance between species in the behaviour of a particular gene at a particular time point provides a measure of the divergence of a gene's temporal dynamics (see Methods). Plotting these values across all genes as a function of time shows that temporal expression divergence follows an hourglass pattern with maximal conservation occurring at time point 5 (8–10 h), a period that corresponds to the extended germband stage, generally regarded as the arthropod phylotypic period (Fig. 2a). We confirmed that the hourglass pattern is not an aggregate behaviour of the data set, but is present on a gene-by-gene basis for the majority of genes (Supplementary Figs 6 and 7), and also that this pattern is evident in the absolute, untransformed gene expression levels (Supplementary Fig. 8b and Supplementary Information, Sections 2.6 and 2.7). For genes that fit best to models where the optimal expression level is the same across all species we calculate a measure of selective constraint[23] (Fig. 2b). This shows that for genes whose evolutionary optimum is the same across species, selective constraint is maximized during the phylotypic period when gene expression divergence is minimized. Therefore, natural selection conserves gene expression patterns during the phylotypic period.

To discover the functional classes of genes responsible for driving the hourglass pattern in the data, we correlated each gene's divergence profile with the average across all genes, thereby allowing us to rank genes by their tendency to follow the global hourglass divergence profile. We find that these genes are enriched for biological processes involved in cellular and organismal development and gene expression (Supplementary Tables 1–3). Moreover, functional characterization of genes that follow an absolute expression hourglass (Supplementary Fig. 8b) shows that they are also enriched for developmental and gene expression processes (Fig. 3a and Supplementary Tables 4–6). Taken together, these results show that genes involved in core developmental processes conform

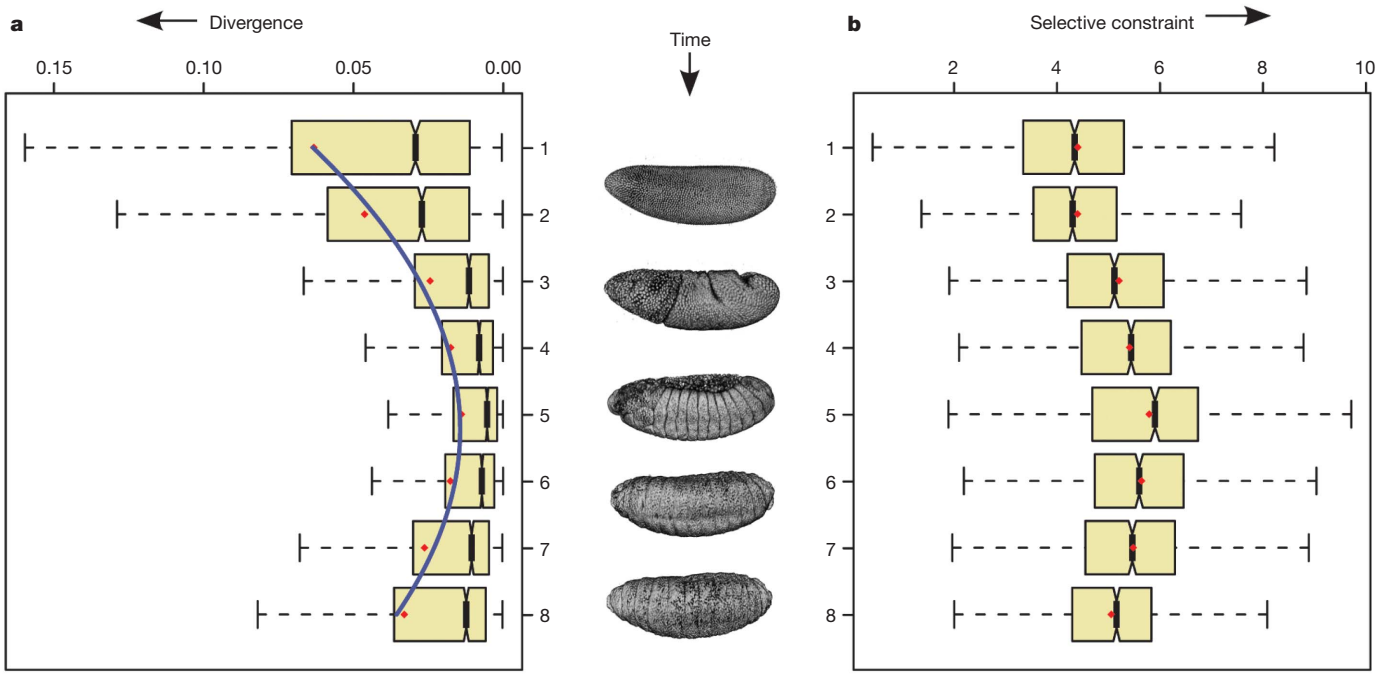

Figure 2 | Temporal expression divergence is minimized during the phylotypic period. **a**, Temporal divergence of gene expression at individual time points during embryogenesis. The curve is a second-order polynomial that fits best to the divergence data. Embryo images are three-dimensional renderings of time-lapse embryonic development of *D. melanogaster* using Selective Plane Illumination Microscopy (SPIM). **b**, Selective constraint for genes that fit best to single optimum stabilizing selection models, calculated as the negative log of the equilibrium variance (see Methods and Supplementary Fig. 5b). Time points are 2-h intervals starting from 0–2 h (1) and ending at 14–16 h (8). Red diamonds indicate the mean; error bars encompass data within 1.5 times the inter-quartile range, and the boxes show the lower and upper quartiles together with the median.

strongly to the global hourglass divergence pattern both in terms of temporal dynamics and absolute expression differences.

We also asked whether there are sets of genes that don't follow the global hourglass pattern and found genes enriched for processes involved in secondary metabolism, the immune system, and responses to oxidative and wounding stresses (Supplementary Fig. 9a and Supplementary Table 7). These are processes that are upregulated late in development, such as pigment or chitin metabolism, or processes that will be upregulated in response to changes that are independent of the developmental program, such as a change in the external environment or the presence of a parasite. The transcript levels of genes in this latter category will reflect the particular challenges faced by individual embryos and so we would not expect these genes to follow a clear temporal pattern of conservation and divergence. These genes tend to be zygotically expressed and are largely present in the yolk (Supplementary Table 7).

Independent of the hourglass patterns, our measures of quantitative and temporal expression divergence exhibit similar functional associations; housekeeping processes tend to be conserved and metabolic processes tend to be divergent (Supplementary Tables 8–11 and Supplementary Information, Sections 2.8 and 2.9). Given these broad functional similarities, it is of interest to ask whether genes in these categories of divergence also share similar genomic and gene-level features. We observe that genes that diverge quantitatively tend to have short introns and 5′ intergenic regions (Fig. 3b) whereas genes that diverge temporally have long introns and 5′ intergenic regions consistent with the notion that increased regulatory complexity in long noncoding regions[27] may provide opportunities for temporal expression divergence (Fig. 3c). This increased regulatory complexity is also supported by a strong positive correlation between temporal divergence and tissue specificity. Additionally, temporal divergence is negatively correlated with mRNA length, raising the possibility that the

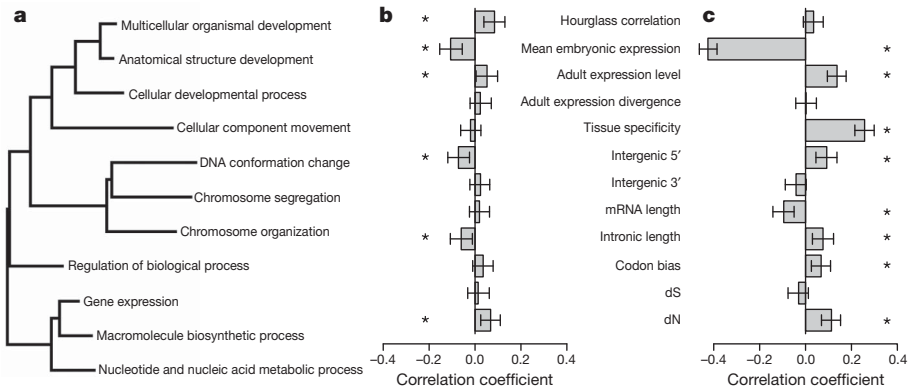

Figure 3 | Properties of genes with different divergence patterns. **a**, A neighbour-joining dendrogram of enriched functional processes for genes that follow an hourglass pattern of divergence. **b**, **c**, Correlation of gene-level variables with quantitative divergence (**b**) and temporal divergence (**c**). Error bars are 95% confidence intervals based on 1,000 bootstraps. Asterisks indicate significant correlations. dN, non-synonymous substitution rate; dS, synonymous substitution rate.

proteins of these shorter genes are engaged in fewer protein–protein interactions. We also observe a positive correlation between rates of amino acid evolution (dN) and both quantitative and temporal divergence, supporting similar findings based on adult expression levels[28], and providing further evidence that embryonic expression divergence is measuring biologically relevant signals.

Our results show that gene expression is more resistant to evolutionary change during mid-embryogenesis than either early or late periods of *Drosophila* development. Evolutionary analyses support the notion that this conservation is the result of natural selection acting to maintain expression levels and their temporal relationships during mid-embryogenesis for genes involved in building up the body plan of the larva. These results complement a recent finding suggesting that the pupal stage in *Drosophila* is under strong stabilizing selection due to the complexity of the processes that occur during metamorphosis, a process that parallels many aspects of embryonic development[29]. These findings seem to support the hypothesis of ref. 2 that an increase in global interactions constrains evolutionary change of the phylotypic period; however, neither study directly addresses the coordination of growth and patterning proposed by ref. 1. Such a relationship may be best examined in the context of gene regulatory networks. Future studies will also need to address the mode and strength of selection acting on gene expression with greater resolution by coupling interspecific expression divergence with intraspecific variation during embryogenesis[30].

METHODS SUMMARY

RNA was extracted from embryos from six *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. ananassae*, *D. persimilis*, *D. pseudoobscura* and *D. virilis*) reared at 25 °C. The embryos were aged at 2-h intervals to form a time course. Sixty-base-pair-long, species-specific microarray probes (four per species) were selected by choosing regions of the orthologous genes of each species that were maximally conserved according to an information entropy measure. Candidate probes with a G+C content higher than 50% were penalized and hence were less likely to be chosen. After scaling the time courses and normalizing replicates, the following linear model was fitted to log expression levels:

$$\log(y_{ijklmn}) = \mu + G_i + S_j + T_k + r_{l(j)} + p_{m(ij)} + GS_{ij} + GT_{ik} + ST_{jk} + rG_{l(j)i} + pT_{mk(ij)} + rp_{l(j)m(ij)} + rT_{kl(j)} + GST_{ijk} + e_{n(ijklm)}$$

where μ is the global average, G_i is the gene effect, S_j is the species effect, T_k is the time effect, $r_{l(j)}$ is the replicate effect nested in species, $p_{m(ij)}$ is the probe effect nested in genes and species, and $e_{n(ijklm)}$ is the residual error. Values are averaged over missing subscripts. Divergence per time point was measured as the between-species variance in *GST* values for each gene separately. We fitted four different evolutionary models to the *GST* values for each gene using the R package 'ouch' and ranked them by their Akaike Information Criterion (AIC). The models were Brownian motion plus three stabilizing selection models with between one and three selective optima (Supplementary Fig. 4). Genomic features of genes were retrieved from FlyBase release 5.14, adult expression level and tissue specificity were retrieved from FlyAtlas, and tissue expression data were retrieved from APOGEE (http://fruitfly.org/cgi-bin/ex/insitu.pl). Partial correlations were calculated and 95% confidence intervals for each partial correlation were generated from 1,000 bootstraps.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 5 September; accepted 2 November 2010.

1. Duboule, D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev. Suppl.* 135–142 (1994).
2. Raff, R. A. *The Shape of Life: Genes, Development and the Evolution of Animal Form* (Univ Chicago Press, 1996).
3. Hall, B. K. Phylotypic stage or phantom, is there a highly conserved embryonic stage in vertebrates? *Trends Ecol. Evol.* **12,** 461–463 (1997).
4. Sander, K. Specification of the basic body plan in insect embryogenesis. *Adv. Insect Physiol.* **12,** 125–238 (1976).
5. Wray, G. A. & Raff, R. A. Rapid evolution of gastrulation mechanisms in a sea urchin with lecithotrophic larvae. *Evolution* **45,** 1741–1750 (1991).
6. Goldstein, B., Frisse, L. M. & Thomas, W. K. Embryonic axis specification in nematodes: evolution of the first step in development. *Curr. Biol.* **8,** 157–160 (1998).
7. Schmidt, K. & Starck, J. M. Developmental variability during early embryonic development of zebra fish, *Danio rerio*. *J. Exp. Zoolog. B* **302,** 446–457 (2004).
8. Carroll, S. B. Evolution at two levels: on genes and form. *PLoS Biol.* **3,** e245 (2005).
9. Wray, G. A. The evolutionary significance of *cis*-regulatory mutations. *Nature Rev. Genet.* **8,** 206–216 (2007).
10. von Baer, K. E. *Uber Entwickelungsgeschichte der Thiere: Beobachtung und Reflektion* (Königsberg, 1828).
11. Darwin, C. *On the Origin of Species* (Murray, 1859).
12. Garstang, W. The theory of recapitulation: a critical restatement of the biogenetic law. *Linn. J. Zool* **35,** 81–101 (1922).
13. Sander, K. The evolution of patterning mechanisms: gleanings from insect embryogenesis and spermatogenesis. In *Development and Evolution* 137–159C (Cambridge Univ. Press, 1983).
14. Galis, F. & Metz, J. A. Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservation. *J. Exp. Zool.* **291,** 195–204 (2001).
15. Hazkani-Covo, E., Wool, D. & Graur, D. In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. *J. Exp. Zoolog. B* **304,** 150–158 (2005).
16. Davis, J. C., Brandman, O. & Petrov, D. A. Protein evolution in the context of *Drosophila* development. *J. Mol. Evol.* **60,** 774–785 (2005).
17. Cruickshank, T. & Wade, M. J. Microevolutionary support for a developmental hourglass: gene expression patterns shape sequence variation and divergence in *Drosophila*. *Evol. Dev.* **10,** 583–590 (2008).
18. Richardson, M. K. *et al.* There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat. Embryol. (Berl.)* **196,** 91–106 (1997).
19. Bininda-Emonds, O. R. P., Jeffery, J. E. & Richardson, M. K. Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proc. R. Soc. Lond. B* **270,** 341–346 (2003).
20. Poe, S. & Wake, M. H. Quantitative tests of general models for the evolution of development. *Am. Nat.* **164,** 415–422 (2004).
21. Tomancak, P. *et al.* Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* **8,** R145 (2007).
22. Zhang, Y., Sturgill, D., Parisi, M., Kumar, S. & Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* **450,** 233–237 (2007).
23. Bedford, T. & Hartl, D. L. Optimization of gene expression by natural selection. *Proc. Natl Acad. Sci. USA* **106,** 1133–1138 (2009).
24. Blomberg, S. P., Garland, T. & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57,** 717–745 (2003).
25. Butler, M. A. & King, A. A. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am. Nat.* **164,** 683–695 (2004).
26. Lemos, B., Meiklejohn, C. D., Cceres, M. & Hartl, D. L. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. *Evolution* **59,** 126–137 (2005).
27. Nelson, C. E., Hersh, B. M. & Carroll, S. B. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5,** R25 (2004).
28. Nuzhdin, S. V., Wayne, M. L., Harmon, K. L. & McIntyre, L. M. Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol. Biol. Evol.* **21,** 1308–1317 (2004).
29. Artieri, C. G. & Singh, R. S. Molecular evidence for increased regulatory conservation during metamorphosis, and against deleterious cascading effects of hybrid breakdown in *Drosophila*. *BMC Biol.* **8,** 26 (2010).
30. Rifkin, S. A., Kim, J. & White, K. P. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nature Genet.* **33,** 138–144 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Alexa for providing modified code for his topGO R package, A. Larracuente and T. Sackton for sharing data with us, M. Weber for generating the embryo images for Fig. 2, and Carl Zeiss MicroImaging for providing the SPIM microscope. We also thank N. Barton, T. Bedford, D. Hartl, J. Howard, A. Oates and D. Robertson for providing useful comments and discussion on the manuscript. This work was funded by The Human Frontier Science Program (HFSP) Young Investigator's Grant RGY0084.

Author Contributions K.M.V. and P.T. conceived the experiment, and K.M.V. and J.J. carried it out. K.M.V., P.T. and S.P. designed the microarray. P.T. conducted the interspecies correlation analysis, and S.P. formulated the linear interpolation algorithm. A.T.K. conceived and conducted the statistical analyses. D.T.G. and C.M.B. conducted the genomic correlates analysis. D.L.C. and U.O. carried out the probe orthology assignments. C.M.B. brought the hourglass concept to the attention of the HFSP team. A.T.K. wrote the paper with support from co-authors.

Author Information The expression data are available for download from ArrayExpress under experiment name 'hourglass', accession number E-MTAB-404, and together with the probe sequences at http://publications.mpi-cbg.de/4240-data. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to P.T. (tomancak@mpi-cbg.de).

METHODS

Embryo collections and RNA isolation and labelling. Embryos were collected from a population of well-fed adults reared at 25 °C. To synchronize the age of the embryos in each sample we pre-laid the flies twice for 1 h with a fresh apple juice plate with yeast paste before every collection. Another fresh plate with yeast was used to collect embryos. The plate was removed from the cage after a 2-h interval and aged in the same incubator for the remaining time required by each time point. After ageing, embryos were collected and rinsed with water to remove yeast paste, and then dechorionated in 100% bleach for 2 min and then washed in desalinated water. The embryos were then transferred into a 1.5-ml tube and snap-frozen in liquid nitrogen and stored at −80 °C.

When isolating RNA, embryos were thawed on ice and homogenized with a pellet pestle and a pellet pestle cordless motor (Kontes). RNA was isolated with the RNeasy Mini kit (Qiagen) and eluted with 30 µl of water. The RNA concentration was measured with the NanoDrop spectrophotometer and RNA quality was assessed with Bioanalyser using the Agilent RNA 6000 Nano kit.

To prepare samples for hybridization to the chip, we followed the Agilent One-Colour Microarray-Based Gene Expression Analysis protocol version 5.5. The starting amount of RNA was normalized to 600 ng for all samples. Samples of a given time-course were processed on the same day.

Probe selection. Probe selection was limited to 60-mers that started within 1 kb from the 3′ end of the transcript. The two main factors that influenced subsequent probe selection were the similarity of orthologous probes in six species determined by information entropy (Supplementary Fig. 1) and the specificity of a probe estimated by the G+C content-weighted BLAST score (Supplementary Information, Section 1.1).

Additionally, we incorporated the distance from the 3′ end of the transcript into the information entropy measure by means of weighting the information entropy (Supplementary Fig. 2). Hence, the further away the candidate probe was from the end of the transcript, the higher the final information entropy measure became.

Probe specificity was verified in two steps. We first rejected candidate probes that did not have a 60-nucleotide-long match to the respective genome assembly. By doing so we eliminated probes that fell on the border of two exons in the transcript sequence. For the remaining candidate probes the G+C content-weighted BLAST score was calculated. This score was the sum of nucleotides that were identical to the query 60-mer and were found in short, unspecific hits. The sum was weighted by the G+C content of hits shorter than 60 nucleotides. If the G+C content exceeded 50% the sum was multiplied by a factor greater than 1 and as a consequence the probe was penalized. Four probes were selected for each gene in each species and we calculated the base-pair overlap between the probes and, where possible, tried to minimize this value (Supplementary Fig. 13).

Time-course registration and correlation analysis. To register the time courses for different species with different developmental time periods[31] onto a common time axis we scaled the non-*melanogaster* time courses to *D. melanogaster* by maximizing the similarity among the profiles across all genes. The selection of genes on the array resulted in a progressive shift of signal intensity distributions from bimodal (mixture of non-expressed and expressed genes early) to unimodal (most genes expressed late) across the time course (Supplementary Fig. 18). Therefore we normalized the replicates for each time point in each time course separately using quantile normalization. Next we averaged the probe signal intensities using the Tukey biweight algorithm to obtain a single expression value per gene and time point in each species while removing outliers. We then re-sampled each time course to 100 time points using cosine transform interpolation (DCT)[32]. Subsequently, all 3,019 expression profiles of the non-*melanogaster* species were scaled by factors ranging from 0.4 to 1.6 in 0.01 increments to find the optimal scaling factor. We calculated squared sums of average differences between all the scaled profiles and *D. melanogaster* profiles and plotted these sums as a function of the scaling factor applied (Supplementary Fig. 16). The global minimum in the graph corresponded to the scaling factor at which all the profiles of the two species were most similar to each other. We applied the optimal scaling factors (Supplementary Table 12) to the averaged non-*melanogaster* profiles with the DCT interpolation resulting in registered time courses (four example genes, before and after registration, are shown in Supplementary Fig. 17).

To compare the overall shape of the profiles among species we row normalized the gene expression values for each gene in each time course and calculated pairwise correlation coefficients for all pairs of orthologous genes. Genes ordered by this simple measure of similarity give an intuitive impression of the amount of conservation of temporal profiles among each pair of species (Fig. 1a).

For the statistical analysis described below, we applied the optimal scaling factors to the raw $\log_2$ Agilent array signal intensities and subsequently quantile-normalized each time point in each time course separately, as described above.

Linear models. A global ANOVA model was fitted to the data to partition the main effect variables and interactions of biological interest from random factors and residual error. This normalizes the gene expression values and provides a single coherent statistical framework in which to explore the variance and covariance structure of the data[33]. The model for gene expression, y_{ijklmn}, is a five-factor, partially nested, mixed-model ANOVA

$$\log(y_{ijklmn}) = \mu + G_i + S_j + T_k + r_{l(j)} + p_{m(ij)} + GS_{ij} + GT_{ik} + ST_{jk} + rG_{l(j)i} + pT_{mk(ij)} + rp_{l(j)m(ij)} + rT_{kl(j)} + GST_{ijk} + e_{n(ijklm)}$$

where μ is the global average, G_i is the gene effect, S_j is the species effect, T_k is the time effect, $r_{l(j)}$ is the replicate effect nested in species, $p_{m(ij)}$ is the probe effect nested in genes and species, and $e_{n(ijklm)}$ is the residual error. Values are averaged over missing subscripts. The probe and replicate effects are random factors in the model and account for error variance arising from different probes and from different samples of within-strain genotypes respectively.

The remaining terms are two- and three-way interactions between the main factors. The gene-by-species, $GS_{ij} = \log(y_{ij}) - G_i - S_j - \mu$, and gene-by-species-by-time, $GST_{ijk} = \log(y_{ijk}) - G_i - S_j - T_k - GS_{ij} - GT_{ik} - ST_{jk} - \mu$, effects contain information about divergence between species. Here we treat time as a categorical variable so that we can extract variances at discrete time points in different species. Divergence at each time point is then measured as the between-species variance in *GST* values per gene and per time point (Fig. 2a). Mean sums of squares were estimated for each variable in the model after subtracting the mean from the data, and the resulting ANOVA table is shown in Supplementary Information, Section 2.4. A Principal Component Analysis (PCA) of the gene × time (*GT*) effect from the above model was computed and the results are shown in Supplementary Information, Section 2.3.

A reduced version of the above model was fitted as a linear regression to each gene separately (the gene effect was dropped) using the R package 'limma' version 3.2.2 (ref. 34). Limma uses an empirical Bayesian approach to infer differential expression in individual genes, producing moderated *t*-statistics with Bayesian-adjusted denominators that incorporate information across the entire ensemble of genes[35]. By fitting a linear model to each gene separately, limma allows for gene-specific error distributions. The probe effect was also dropped from the ANOVA model as the probes were normalized using Tukey's median polish method to fit a linear model for gene expression to each gene, $y_{ij} = \exp_i + a_j + e_{ij}$, where $\exp_i$ is the normalized gene expression value for gene i, a_j is the probe effect for the jth probe, and e_{ij} is the residual error[36]. The species effect from limma is equivalent to the *GS* effect from the global ANOVA, and this value was used for assessing quantitative divergence between species (Fig. 3b).

The temporal profiles of genes were compared across species using a PCA-based approach[37]. This method quantifies pairwise species differences in temporal profiles for individual genes using the Mahalanobis distance, which is calculated using *GST* values for all time points estimated from the global ANOVA model. The Mahalanobis distance is calculated as

$$D_i^2 = (\Delta Z_i - Z_C)\mathrm{cov}(\Delta Z)^{-1}(\Delta Z_i - Z_C)^T,$$

where ΔZ_i is the species *GST* score contrast for gene i, Z_C is the centroid for all of the *GST* score contrasts, and $\mathrm{cov}(\Delta Z)$ is the covariance matrix for the difference matrix ΔZ. This metric is distributed according to a chi-squared distribution with k degrees of freedom where k is the number of principal components included in the contrast. We used the Mahalanobis distances as a measure of temporal divergence between species across all time points (Fig. 3c). The distances were calculated using the first three principal components, which together account for 89% of the total variance.

Phylogenetic analyses and evolutionary models. A maximum likelihood phylogeny was constructed with *GST* values from every gene and every time point using the 'contml' continuous character restricted maximum likelihood approach implemented in PHYLIP version 3.69 (ref. 38) with *D. virilis* identified as the outgroup, and the resulting phylogram was plotted using Dendroscope version 2.0 (ref. 39) (Fig. 1d).

We estimated the phylogenetic signal for the *GST* values at each gene by calculating the K statistic described in ref. 24 using the R package 'picante' version 1.1-1. The tree used for this purpose was a phylogram based on median dS values for ~10, 000 orthologous genes[40] which was then converted to a chronogram in the R package 'ape' version 2.5-1[41].

Ornstein–Uhlenbeck (OU) and Brownian motion models were fitted to the six species-specific *GST* values for each gene at each time point using the R package 'ouch' version 2.6-1 (ref. 25). The OU models fitted to each gene describe evolutionary change in a trait X over an infinitesimally small increment of time as $\mathrm{d}X(t) = \alpha(\theta - X(t))\mathrm{d}t + \sigma \mathrm{d}B(t)$ where $\mathrm{d}B(t)$ describes Brownian motion (independent and identically distributed normal random variables with mean 0 and variance $\mathrm{d}t$), σ is the strength of Brownian motion, α is the strength of stabilizing selection, and θ is the trait optimum[42]. Under a purely Brownian process of

evolutionary change the first term on the right-hand side is absent. This model was extended by ref. 25 to include branch-specific values for θ, thereby allowing for adaptive evolution along specific branches. We fitted four models to each gene: Brownian motion plus three OU models with between one and three stabilizing selection optima (Supplementary Fig. 4), based on the chronogram mentioned above. Here we did not engage in an exhaustive model fitting endeavour as our intention was to demonstrate two things: (1) models incorporating stabilizing selection fit best to the majority of the genes, and (2) models incorporating adaptive changes in trait optima often out-perform non-adaptive models.

To avoid treating time points as if they are independent we fitted models to subsets of time points. These subsets were chosen by bootstrapping (1,000 bootstraps) hierarchical clusters of time points based on *GST* values for each gene separately using the R package 'pvclust' version 1.2-1 (ref. 43) and selecting clusters of time points with *P*-values below 0.05. This approach allows different modes of selection to operate across different periods of each gene's time course.

For each gene the model that showed the best fit to the data was defined as the model with the lowest Akaike Information Criterion (AIC), calculated as $2k - 2\log(\text{likelihood})$ where k is the number of degrees of freedom in the model in question. AIC scores balance the likelihood of a model against its complexity (the number of parameters in the model).

After ranking models by their AIC scores, genes for which Brownian motion was not ranked first were tested to see if the top-ranked model showed a significantly better fit to the data than Brownian motion using a log-likelihood ratio test. The resulting *P*-values were adjusted using the Benjamini–Hochberg false discovery rate correction in the R package 'multtest' version 2.4.0 (ref. 44) and models with adjusted *P*-values above 0.05 were dropped down into the Brownian category. We then repeated this process, but treating single optimum models as the null model and testing models that ranked best with two or three optima against this null model to ensure that the resemblance to the phylogeny for these genes was not the result of chance under a single optimum across all species. If single optimum models showed a better fit then they were, in turn, tested against the Brownian model.

We extracted a measure of selective constraint from the genes that fitted best to single optimum models (Fig. 2b), calculated as the negative log of the equilibrium variance, $\frac{\sigma^2}{2\alpha}$ (ref. 23).

Gene Ontology and tissue expression enrichment. Gene Ontology (GO) analyses were conducted using the R package 'topGO' version 1.14.0 (ref. 45). Three enrichment methods were used. For genes that were ranked by a real number score (such as a correlation coefficient) a Kolmogorov–Smirnov ranking test was applied and GO terms with distributions among the genes that showed significant departure from a uniform distribution in a particular direction were deemed to be enriched. Unranked sets of genes were tested for enriched GO terms using the 'elim' and 'parent-child' algorithms in topGO. The 'elim' algorithm decorrelates the local GO graph structure to take into account local dependencies between terms so that more biologically relevant terms are enriched[45] and the 'parent-child' algorithm controls for the inheritance bias between parent and child terms in the GO hierarchy[46]. Fisher's exact test was then used to determine enrichment *P*-values for both of these algorithms. The same approach was used to identify enriched tissue expression terms from a controlled vocabulary based on *in situ* expression data[21] by using modified code from the topGO package.

P-values from Kolmogorov–Smirnov tests were adjusted using the Benjamini–Hochberg false discovery rate correction, but no correction was applied to the 'elim' and 'parent-child' *P*-values because they are not calculated independently for each GO term in these algorithms and are effectively already adjusted. For defined sets of genes, the reference set was all of the genes on the chip.

We plotted a neighbour-joining tree of enriched, non-redundant GO terms[47] for Fig. 3a using the R package 'ape' and Dendroscope. Terms were enriched for 1,188 genes that show an hourglass profile in both temporal dynamics and absolute expression levels using Fisher's exact test and selecting terms with adjusted *P*-values below 0.05.

Correlation of divergence with gene-level variables. Quantitative and temporal divergence measures were generated for each of the 15 pairwise species comparisons. Following ref. 48, we converted these to nine branch lengths on the known phylogeny using the Fitch–Margoliash least squares method (implemented in the PHYLIP program 'fitch'[38]). Negative branch lengths were set to zero. Total expression divergence for each gene is the sum of branch lengths and constitutes our 'quantitative' and 'temporal' measures using the limma or Mahalanobis distances, respectively.

We collated structural, functional and expression data for all of the genes on the chip from public databases and previous genome-level studies. These data were generated from gene coordinates retrieved from FlyBase Release 5.14 (January 2009). Only protein-coding genes were retained (as all genes on our chip are protein coding) and genes from the heterochromatic portions of the otherwise 'euchromatic' chromosome arms were discarded (168 genes from the genome, including 25 from our chip data set)[49].

In addition, data on further variables for 8,500 *D. melanogaster* genes compiled by ref. 48 were obtained from the authors. These data could be assigned to 2,526 of the genes on the chip. Gene expression was described by adult expression level and tissue specificity (both from FlyAtlas[50]), expression divergence between adults (measured in a very similar set of species by ref. 22), and we added the mean embryonic expression level from our own data. Gene sequence evolution was described by codon bias (the frequency of optimal codons) in *D. melanogaster* and by dN and dS, the rates of non-synonymous and synonymous nucleotide substitutions, respectively.

As many of our variables of interest were correlated with one another, we calculated partial correlations between each variable and expression divergence while controlling for the other variables. The set of variables included are described in Supplementary Information, Section 1.3. We only used filtered genes for which we had information on all the variables ($n = 1{,}832$). Partial correlations were calculated from Spearman's rank correlation matrices using the R package 'corpcor'. Ninety-five per cent confidence intervals for each partial correlation were generated by boot-strapping (random sample with replacement) the set of genes contributing to the correlation. One thousand bootstraps were performed using the R package 'boot'.

31. Markow, T. A. & O'Grady, P. M. *Drosophila* biology in the genomic age. *Genetics* **177,** 1269–1276 (2007).
32. Ahmed, N., Natarajan, T. & Rao, K. R. Discrete cosine transform. *IEEE Trans. Comput.* **C-23,** 90–93 (1974).
33. Kerr, M. K., Martin, M. & Churchill, G. A. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **7,** 819–837 (2000).
34. Smyth, G. K. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor* 397–420 (Springer, 2005).
35. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3,** Article3 (2004).
36. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15 (2003).
37. Jonnalagadda, S. & Srinivasan, R. Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data. *BMC Bioinformatics* **9,** 267 (2008).
38. Felsenstein, J. PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics* **5,** 164–166 (1989).
39. Huson, D. H. *et al.* Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8,** 460 (2007).
40. Heger, A. & Ponting, C. P. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. *Genome Res.* **17,** 1837–1849 (2007).
41. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–290 (2004).
42. Hansen, T. F. & Martins, E. P. Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. *Evolution* **50,** 1404–1417 (1996).
43. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22,** 1540–1542 (2006).
44. Ge, Y., Dudoit, S. & Speed, T. P. *Resampling-Based Multiple Testing for Microarray Data Analysis.* Technical Report (Univ. California, 2003).
45. Alexa, A., Rahnenfhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22,** 1600–1607 (2006).
46. Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* **23,** 3024–3031 (2007).
47. Agudelo-Romero, P. *et al.* Changes in the gene expression profile of *Arabidopsis thaliana* after infection with Tobacco etch virus. *Virol. J.* **5,** 92 (2008).
48. Larracuente, A. M. *et al.* Evolution of protein-coding genes in *Drosophila. Trends Genet.* **24,** 114–123 (2008).
49. Smith, C. D., Shu, S., Mungall, C. J. & Karpen, G. H. The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. *Science* **316,** 1586–1591 (2007).
50. Chintapalli, V. R., Wang, J. & Dow, J. A. T. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nature Genet.* **39,** 715–720 (2007).