Conservation of expression regulation throughout the animal kingdom

Michael Kuhn[1,2]*, Andreas Beyer[3]*

[1] Biotechnology Center, TU Dresden, Dresden, Germany

[2] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

[3] University of Cologne, Cologne, Germany

* To whom correspondence should be addressed: michael.kuhn@biotec.tu-dresden.de (Phone: +49-351-463 40064, Fax: +49-351-463 40061), andreas.beyer@uni-koeln.de (Phone: +49 221-478 84429, Fax: +49 221-478 84045)

Gene expression is often studied in the context of perturbations such as gene knockdown or drug treatment. In most cases, the response of the organism to such perturbation is not under evolutionary selection and the outcomes are therefore largely accidental. In contrast, gene expression changes during development and tissue formation are under strong selection. Here, we present the first cross-species mapping of tissue-specific and developmental gene expression patterns across a wide range of animals, including many non-model species. In our survey of 32 datasets across 23 species, we detected conserved expression programs on all taxonomic levels, both within animals and between the animals and their closest unicellular relatives, the choanoflagellates. We found that the rate of change in tissue expression patterns is a property of gene families. Subsequently, we used the conservation of expression programs as a means to identify neofunctionalization of gene duplication products. We found 1206 duplication events where one of the two genes kept the expression program of the original gene, whereas the other copy adopted a novel expression program. We corroborated such potential neofunctionalizations using independent network information: the duplication product with the more conserved expression pattern shared more interaction partners with the non-duplicated reference gene than the more diverging duplication product. Our findings open new avenues of study for the comparison and transfer of knowledge between different species.

Introduction

Gene functions have traditionally been determined using molecular and cellular approaches involving forward or reverse genetics. Functional annotations that were directly determined through these approaches are, however, not available at all for most species, and incomplete even for model species (Thomas et al. 2012). For non-model species, often only data transferred from other organisms is available. In this case, the degree of conservation of functions is uncertain, especially in the case of gene duplications. Previously, gene expression data has been used to compare genes with similar functional annotations across species, to reveal functional similarity between genes from different species (Chikina & Troyanskaya 2011). However, this approach requires that the two species are well-studied in both gene expression and functional annotation, and will suffer from incomplete and biased annotations (Thomas et al. 2012). Developmental gene expression profiles between closely related species can be compared to find functional links between genes and to detect differences between orthologs (Yanai et al. 2011; Levin et al. 2012; Silver et al. 2012). Existing approaches require that expression datasets have been obtained under comparable conditions for the respective species: For closely related species, homologous tissues can easily be identified (Niknejad et al. 2012), and cross-species correlations between homologous tissues of closely related species have previously been investigated (Piasecka, Kutalik, et al. 2012a; Liao & Zhang 2006). This is however a severe limitation for functional mapping between many species. Even between closely related species, the relative amounts of cell types that make up tissues may change. Across larger evolutionary distances, only few clearly homologous tissues are available. Nonetheless, it is possible to identify deep homologies among tissues (Shubin et al. 2009). For example, homologous structures have been identified in the nervous systems of vertebrates and annelids (Tomer et al. 2010; Strausfeld & Hirth 2013). Other organs show functional convergence, for example mammalian liver and brown fat in flies, which both carry out xenobiotic functions (Chung et al. 2009).

Many gene expression datasets have been generated under experimental conditions that represent non-physiological conditions, such as gene knockouts, that are not under evolutionary selection. Such data is therefore not necessarily suitable for comparing gene expression across species (Seok et al. 2013). In

contrast, the formation of tissues during development and the maintenance of tissue function throughout the life of an animal are crucial for survival and reproduction, and are therefore under direct evolutionary selection (Winter et al. 2004; Gu & Z. Su 2007). Tissue expression data is available for many species, as tissues can be gathered even from non-model species where genetic tools such as transgenesis or RNAi are not available. Previous research has shown that it is possible to predict tissue-specific expression patterns from gene expression experiments within the same species (Chikina et al. 2009). However, it remains challenging to map tissue expression over larger phylogenetic distances. If such mapping was possible, we could substantially improve the annotation of non-model-species genomes, fill annotation gaps in model species and in particular address the problem of gene duplications.

We have developed a method to map a gene's tissue expression pattern from one species to another, creating a virtual tissue expression pattern in the destination species. This predicted, virtual expression pattern can be compared to the observed expression pattern. We show that high correlations in tissue expression across species are predictive for 1:1 orthology, shared structure, and similar function. Subsequently we use our modeling approach for three applications: first, for determining the degree of conservation of tissue-specific gene expression patterns, second, for analyzing the speed of functional divergence after gene duplications and third, for proposing an improved method for the prediction of functionally equivalent orthologs. Many datasets contain both tissues and developmental samples, e.g. different adult organs and embryonic stages. For the sake of brevity, we refer to the all of these samples as "tissues."

Results

Correlation between tissues of distant species

To analyze tissue expression across the entire metazoan kingdom, we gathered genome and tissue expression data from 32 datasets covering 23 different species. Among these were eight chordate species: *Ciona intestinales* (Shoguchi et al. 2011), *Danio rerio* (Domazet-Loso & Tautz 2010), *Gallus gallus* (Chan et al. 2009; Irie & Kuratani 2011), *Homo sapiens* (Lukk et al. 2010), *Mus musculus* (Irie & Kuratani 2011; A. I. Su et al. 2004), *Sus scrofa* (Freeman et al. 2012), *Tetraodon nigroviridis* (Chan et al. 2009), and *Xenopus tropicalis* (Chan et al. 2009; Yanai et al. 2011); two

cnidarians: *Hydra vulgaris* (Hemmrich et al. 2012) and *Nematostella vectensis* (Tulin et al. 2013); two flatworms: *Schistosoma japonicum* (Gobert et al. 2009) and *Schistosoma mansoni* (Nawaratna et al. 2011; Fitzpatrick et al. 2009); three insects: *Anopheles gambiae* (Baker et al. 2011; Dissanayake et al. 2006; Goltsev et al. 2009), *Bombyx mori* (Xia et al. 2007) and *Drosophila melanogaster* (St Pierre et al. 2014; Robinson et al. 2013); seven nematodes: *Ascaris suum* (Wang et al. 2013), *Brugia malayi* and five *Caenorhabditis* species (Levin et al. 2012; Spencer et al. 2011). Furthermore, we added the choanoflagellate *Salpingoeca rosetta* as an outgroup (Fairclough et al. 2013).

To determine orthology relations between genes, we assembled groups of orthologs (OGs) using the eggNOG pipeline (Powell et al. 2014) on the genomes of the choanoflagellate *Salpingoeca rosetta* and 67 animals. We then computed gene trees for all OGs using GIGA (Thomas 2010), which we then analyzed to extract 1:1 orthologs and duplication events. First, we quantified the correlation of gene expression between tissues across species. For each pair of datasets, we built gene expression vectors for all tissues using the expression patterns of 1:1 orthologs. This yielded one vector of expression values for each tissue. We then calculated the correlation of these vectors across species and found that for 89.0% of all dataset pairs, more than half of the tissues in one dataset were significantly correlated with at least one tissue from the other dataset (using a p-value cutoff of 0.05 for each tissue pair). Importantly, this was true even across large phylogenetic distances. For example, between fly and *C. elegans*, the two largest correlations of 0.31 were between ovary and gonad, and between head and L2 glutamate receptor neurons. When we removed the three worst datasets from the analysis (*Nematostella, Hydra* and *Bombyx*), the fraction increased to 98.4% of all dataset pairs. Interestingly, for 70.6% of all dataset pairs in the filtered set, all tissues of one dataset were significantly correlated with at least one tissue from the second dataset. These correlations suggested that it is feasible to map gene expression patterns between tissues of distantly related species, even if a homology relation between the tissues is not apparent.

Mapping gene expression between species

To predict tissue expression patterns across species we chose a simple and transparent method, namely to train linear models for mapping expression values across species (Fig. 1). Given the tissue expression values for a source species,

each linear model predicted the expression value for one tissue from the target species. Thus, for each combination of source and target species, we trained as many linear models as there are tissues in the target species. Importantly, this modeling approach did not require 1:1 relationships of tissues (i.e. the existence of homologous tissues). Rather, the expression in each tissue of the target species was modeled as a combination of the tissues in the source species (see Methods).

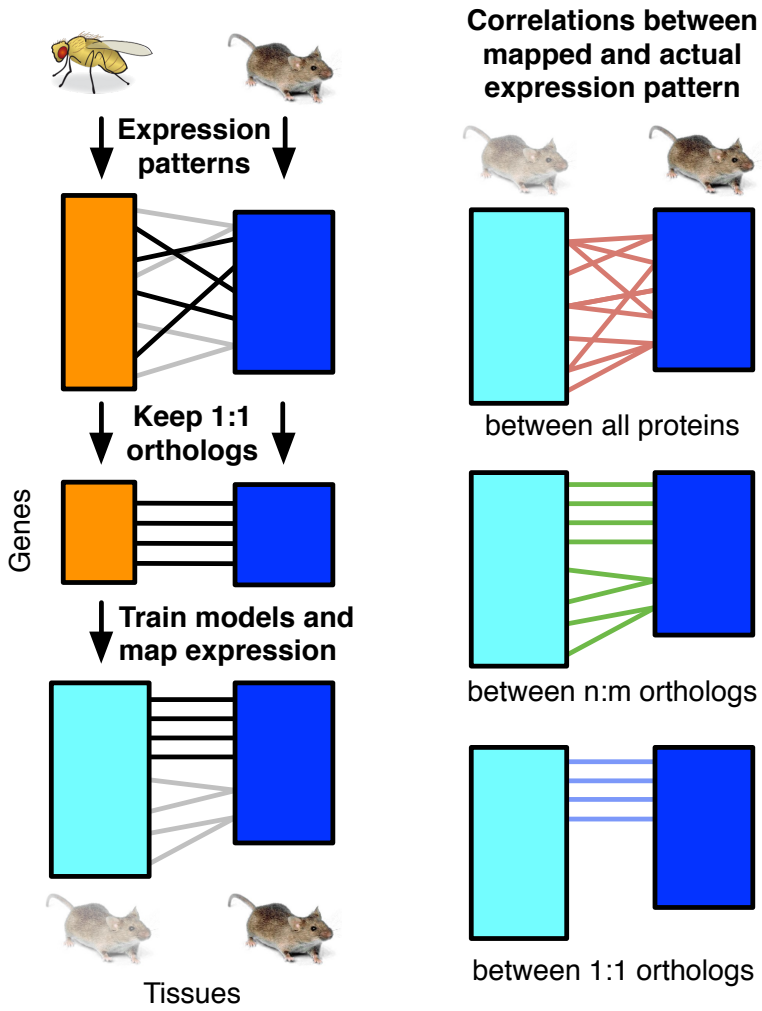


Fig. 1: Mapping expression patterns across species. For each tissue in the target species, models were trained to predict the tissue-specific gene expression pattern from the expression patterns of 1:1 orthologs in a source species. Mapping the expression patterns of all genes created virtual expression patterns, which could then be used to compute correlations between the mapped and actual expression patterns.

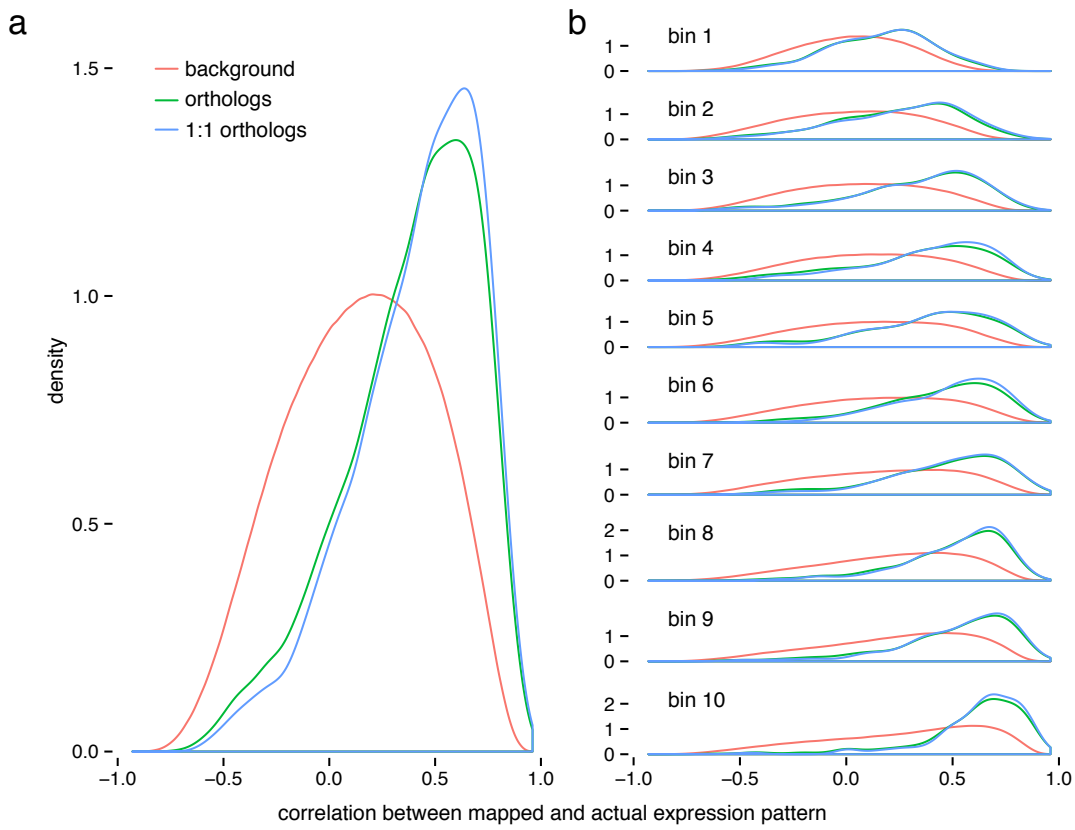


Fig. 2. Distribution of correlations between mapped and actual expression patterns. (a) When mapping expression patterns from fly to *C. elegans*, correlations between orthologs (green) and 1:1 orthologs (blue) were much higher than for background gene pairs (pairs of genes that are not homologous to each other, shown in red). **(b)** Target genes were distributed in bins according to the number of genes with similar expression patterns within the target species. Pairs of background genes had a higher correlation when there were more genes with similar expression patterns, as is evident from the shift towards higher correlations. For this pair of datasets, bins contained between 297 and 316 one-to-one orthologs, with an average of 305.

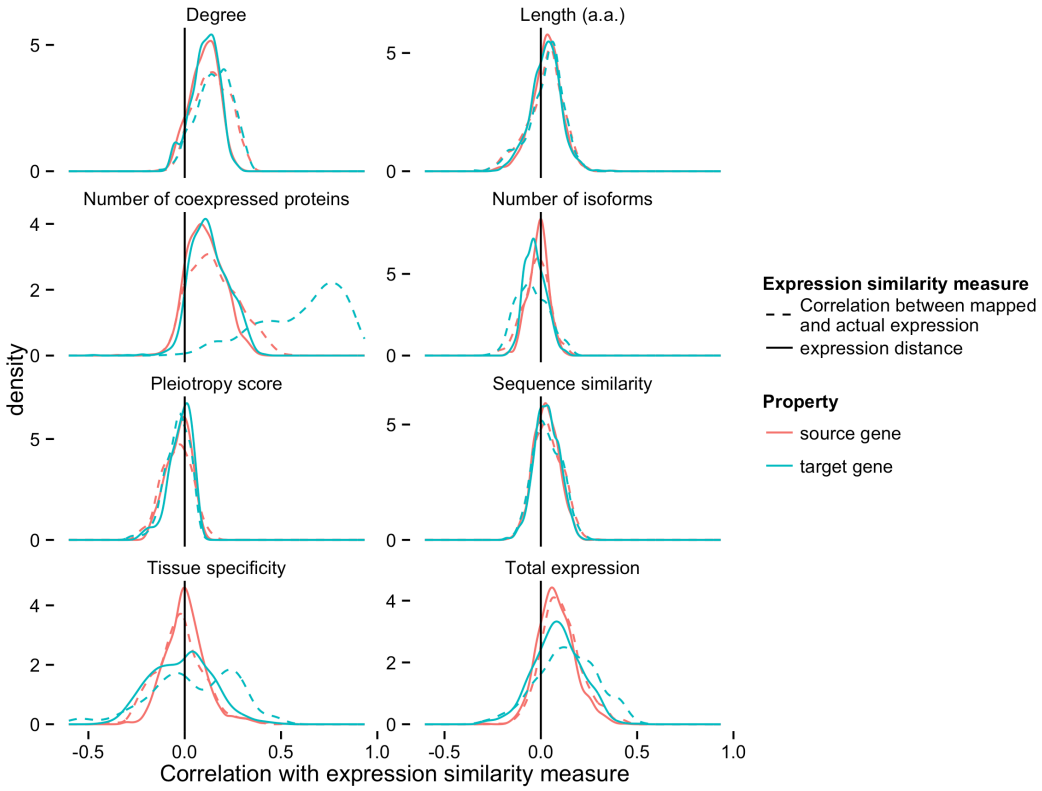


Fig. 3. Gene properties correlated with expression similarity. Different properties of the source (red) or target gene (blue) influenced the distribution of expression distances. To measure this influence, we computed the correlation between the gene properties and the expression similarity. When the correlation between mapped and actual expression patterns was used as the expression similarity (dashed line), there was a very high correlation with the number of coexpressed genes. Correcting for this (Fig. 2b, solid line), this correlation became lower. Genes corresponding to proteins with high degree (i.e. number of interactions) could be mapped better, while target genes with many isoforms resulted in a worse mapping. See Supplementary Information and Fig. S1 for further details on factors that influence the mapping quality.

Using the trained model, we mapped all expression values from the source to the target species. We then calculated the Pearson correlation between the mapped expression values and the actual expression values, for different sets of genes: (1) all genes that have homologs between the two species, (2) orthologous groups and (3) 1:1 orthologs. We restricted the background set to genes with homologs to exclude lineage-specific genes that were found to have much lower correlations than genes with homologs. When analyzing a pair of genes that are 1:1 orthologs, we used expression values predicted by 10-fold cross-validation. From the distribution of correlations, we calculated p-values for all pairs of genes using the null hypothesis that the compared genes belong to the background and thus are not orthologous. During initial tests, we found a strong correlation between these

p-values and the number of genes with similar expression patterns in the target species (Fig. 3, dashed lines). We therefore split target genes into bins according to the number of target genes with similar expression patterns (Fig, 2b). For each bin, we obtain a mapping from correlation to p-values. For a given correlation between the mapped expression pattern of the source gene and the expression pattern of the target gene, we then calculate an expression distance out of the p-values obtained for the adjacent bins (see Methods). Thus, a low expression distance indicates that the expression of this gene in a given target species can be well predicted using the expression of homologous genes in the source species.

The mapping success can be measured in different ways. For each pair of datasets, we first compared the distribution of correlations for background genes and 1:1 orthologs using the Kolmogorov-Smirnov (K-S) test. Controlling for multiple testing with the Benjamini-Hochberg method (Benjamini & Hochberg 1995), 77% of all K-S p-values are significant ($q < 0.05$). As K-S p-values are strongly influenced by the number of gene pairs, we also computed the fraction of 1:1 orthologs that can be mapped at an expression distance threshold of 0.25 (Fig. 4). This fraction was highly correlated with the K-S statistic *D* (Pearson correlation 0.97), but more intuitive. Across all pairs of datsets, the median fraction of 1:1 orthologs with expression distances below 0.25 was 40%, indicating an enrichment of 1:1 orthologs with well-conserved expression patterns.

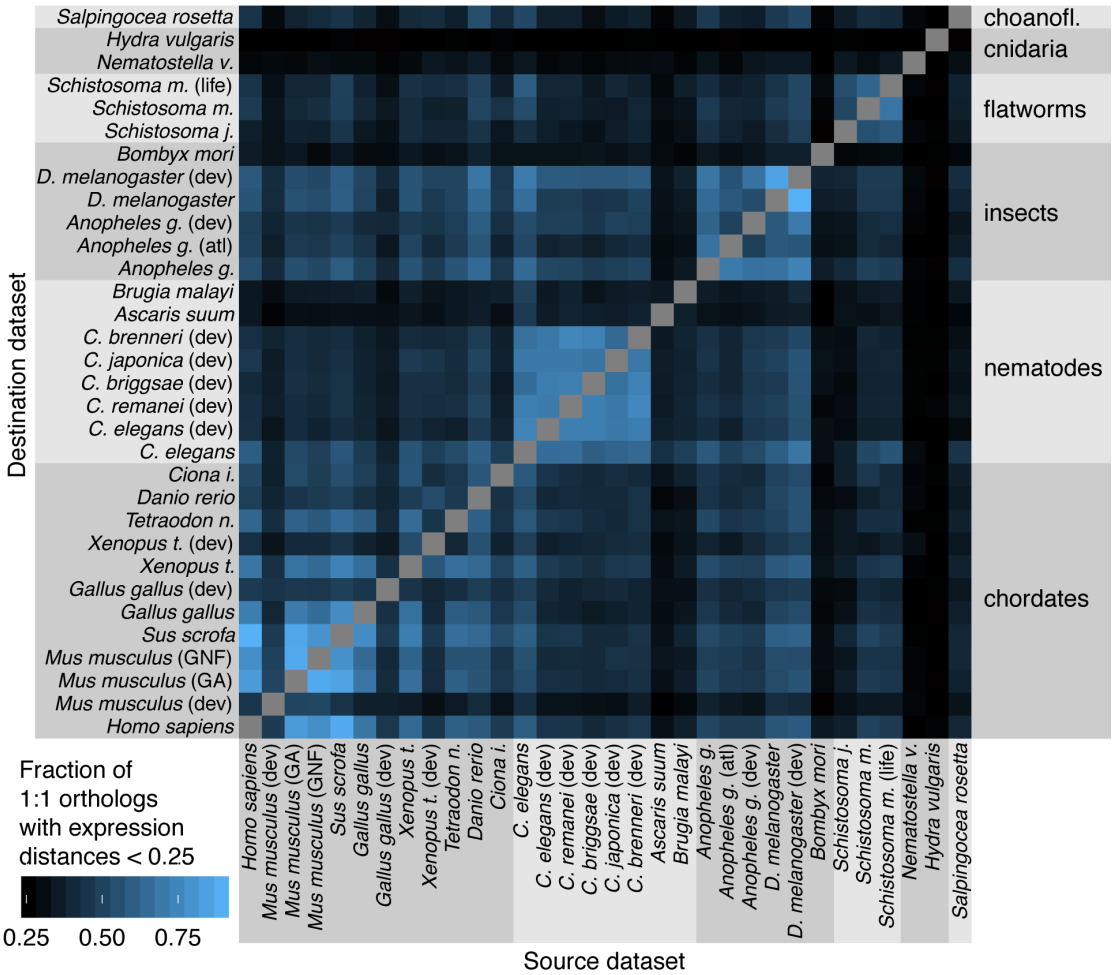


Fig 4. Conservation of expression patterns throughout the metazoans and choanoflagellates. For all dataset pairs, the fraction of 1:1 orthologs with expression distances below 0.25 is shown. Within clades, this fraction becomes very high and approaches 1 in some cases. When there is no enrichment of 1:1 orthologs towards lower expression distances, the distributions of correlations are identical for pairs of background genes and orthologs. In this case, the distribution of expression distances is uniform, and the fraction of orthologs with expression distances below 0.25 is 0.25 (see Fig. 8). Note that there are some datasets with universally low values. Here, the kinds of measured tissues and the quality of the dataset apparently prevented better mapping performance. However, some otherwise distant species had a higher than expected fraction of 1:1 orthologs with well-conserved expression patterns.

Benchmark 1: Identification of 1:1 orthologs

In a first benchmark, we tested whether the expression similarity could be used to identify 1:1 orthologs from top BLAST hits. For each dataset pair, we used BLAST to find the top two hits (bitscore cutoff: 100) for each protein of the source species, discarding proteins with only one hit. After training the expression mapping on an independent set of genes as outlined above, we then computed the expression similarities for the top two hits, and checked whether the gene with the lower expression distance corresponded to the actual 1:1 ortholog. For example, mapping

from fly to *C. elegans*, 67.6% of 1999 one-to-one orthologs could be correctly identified (p-value of Binomial test: 2e-57). Predictions could be ordered in different ways according to the expression distances between the two pairs of genes. For example, they could be ordered by the lowest expression distance, by the difference of the expression distances or their ratio. Of these, the difference between the expression distances performed best in distinguishing confident predictions from less confident predictions (Fig. 5). Between human and mouse (GNF dataset), 76.6% of 7304 1:1 orthologs were correctly mapped. The median fraction across all dataset pairs was 58.9%.

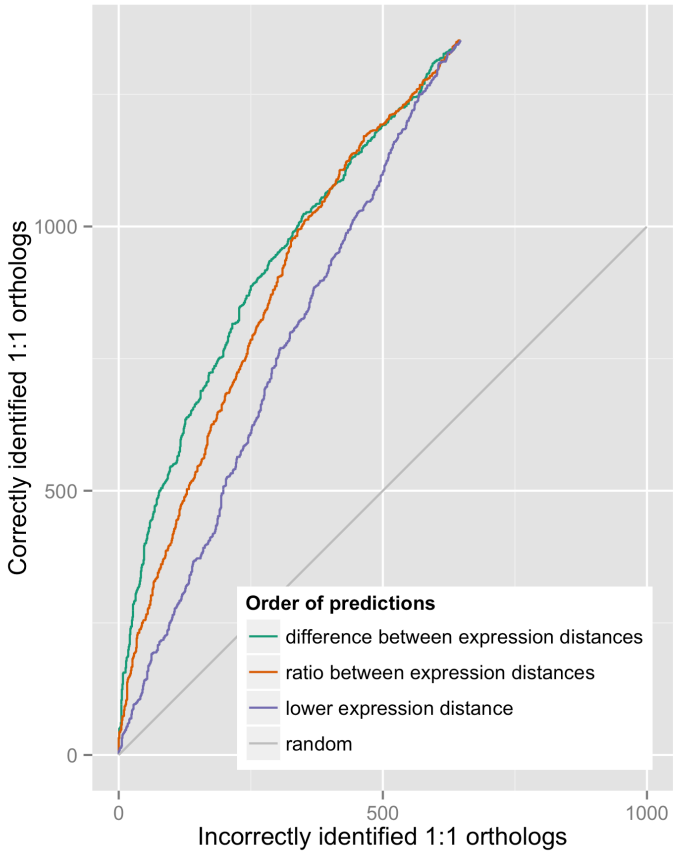


Fig. 5. Prediction of 1:1 orthologs from best hits. For each 1:1 ortholog between fly and *C. elegans*, the source gene's expression pattern is mapped to *C. elegans* and compared to the top two BLAST hits. If the mapped expression pattern is more similar to the actual ortholog, it is counted as correctly identified. Thus, a perfect prediction method would be a vertical line. Ordering the predictions by the difference between the two expression distances is the most successful strategy.

Benchmark 2: Analysis of 3D protein structure

As a further test, we checked if genes corresponding to proteins with the same structure were more likely to have lower expression distances than unrelated proteins. Using the Gene3D database (Lees et al. 2014), we determined CATH folds for all proteins that we could map to the database (resulting in 15 species and 23 datasets). For each dataset pair, we then analyzed each homologous superfamily, computing the median expression distance for all proteins of the superfamily. The superfamilies contain varying numbers of proteins, and we found a correlation between the expression distance and the size of the superfamilies (Fig. 6): Those with many members (and thus more different functions) had more diverse expression patterns. For example, mapping fly to *C. elegans*, the Spearman correlation between the number of proteins per species (using the maximum of the two species) and the median expression distance was 0.40. Between human and mouse (GNF dataset), the Spearman correlation was 0.46. Across all dataset pairs, the median Spearman correlation was 0.28.

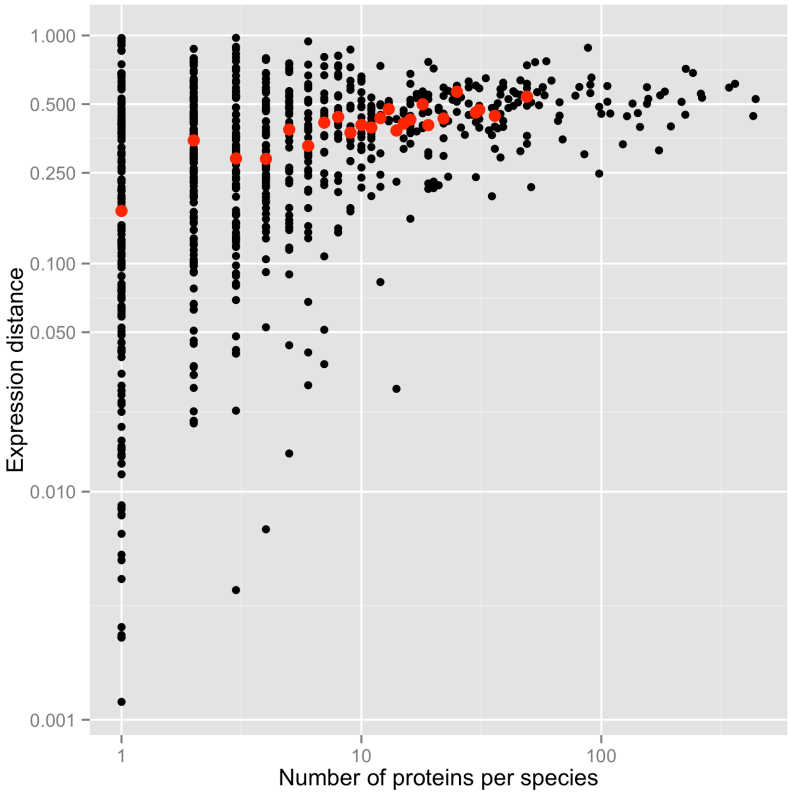


Fig 6. Correlation between expression distance and shared protein folds. Proteins that belong to structural families with few members are more similar in their expression patterns than proteins from large families. Red dots denote the median when at least five superfamilies have the same number of proteins per species. Here, the mapping from fly to *C. elegans* is shown.

Benchmark 3: Phenologs

Finally, we used functional information to evaluate our method. We applied the phenolog concept (McGary et al. 2010) to validate that genes from different species with similar tissue expression are functionally related. Based on orthologous genes, related pairs of functional annotations (Gene Ontology terms, FlyBase and WormBase phenotypes) are predicted by looking for significant overlap between OGs that correspond to the functional annotations. For each pair of well-annotated species (mouse, human, fly, *C. elegans*), we tested all OGs excluding 1:1 orthologs. For each OG, we found the phenolog pair with the lowest p-value. For all gene–gene pairs in this OG, we then determined their expression distance and whether their functional annotation matched the phenolog pair. First, we noted that the distributions of expression distances differed between gene pairs with matched and mismatched annotations: For fly and *C. elegans*, the one-sided K-S p-value was 0.001 and the median K-S p-value across all dataset pairs was 0.009. Second, for each OG, we looked at the gene pair with the lowest expression distance and checked if both genes matched the expected functional annotation based on the phenolog. We ordered OGs by the difference between the lowest and second lowest expression distances. Mapping fly to *C. elegans*, 50% of all top predictions had matching functional annotations, compared to an expected fraction of 43% (Fig. 7). This corresponds to a relative increase of 17% over the expected fraction. Between human and mouse, this increase is 52%. The median increase among all dataset pairs is 11%.

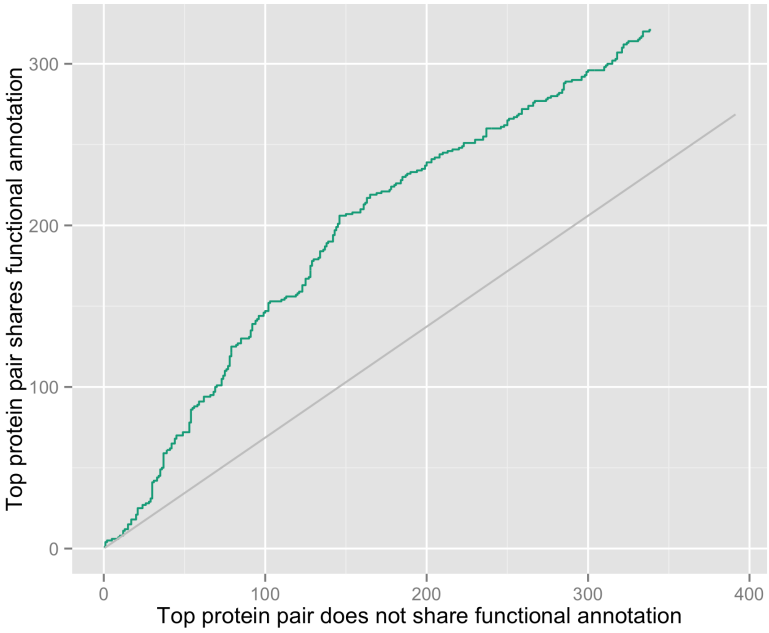


Fig 7. Benchmarking based on phenologs. For each OG, we tested whether the gene pair with the lowest expression distance shared the functional annotation predicted by the phenolog. Predictions are ordered by the difference between the lowest and second lowest expression distance. Randomly choosing gene pairs from the OGs results in the grey line.

Conservation of tissue-specific gene expression

Body plans and tissues change throughout evolution. As a result, gene expression patterns vary from species to species. When the changes are small, then our mapping procedure will yield virtual expression patterns of 1:1 orthologs that are very similar to the actual expression patterns, and the distribution of expression distances is skewed towards lower values. Our mapping procedure becomes less accurate over larger evolutionary distances, and the distribution of expression distances becomes less skewed. It becomes a uniform distribution when 1:1 orthologs cannot be mapped better than background gene pairs. In this section, we tested whether expression distances are indeed shifted towards lower values, even for pairs of species where homologous organs cannot readily be identified.

At all taxonomic levels, we determined the conservation of the expression patterns of 1:1 orthologs. This data then allows us to estimate the degree of conservation of tissue-specific expression patterns, even between groups of species that do not have readily identifiable homologous organs. For all sets of 1:1 orthologs, we computed the median expression distances when mapping across a particular taxonomic split (e.g. for vertebrates, we mapped between fish and tetrapods). First,

we compared the distributions of expression distances to the uniform distribution. With the exception of mappings with cnidaria (*Nematostella* and *Hydra*), all distributions were significantly shifted towards lower p-values (Fig. 8). For some clades, the available data was very uneven on the two sides of the taxonomic split. For example, at the level of eumetazoa, only two species with few tissues were available for cnidarians, whereas most bilaterian species had many tissues measured. Thus, expression distances were higher when mapping from cnidarians to bilaterians. Interestingly, the median divergence between animals and the outgroup choanoflagellates was comparable to the median divergence between major animal clades, e.g. bilateria.

When we chose an expression distance cutoff of 0.25 to designate well-conserved genes, we found that 77% of all 1:1 orthologs could be mapped successfully between mouse and human. For larger clades (like vertebrates), we computed for each OG the median of all pairwise expression distances between the subclades (in this example, tetrapods and fish). Between tetrapods and fish, we found that 55% of all OGs have an expression distance below 0.25 (relative to background gene pairs). Between animals and the outgroup choanoflagellate, 32.7% of all 1:1 orthologs showed conserved expression, a significant increase over the 25% expected when 1:1 orthologs behave like background genes (p-value of one-sided binomial test: 3e-23). Thus, mapping tissue-specific gene expression revealed expression programs conserved for 1 billion years. As the median expression similarities were negatively influenced by datasets of low quality, we also computed the distributions of expression distances and the number of well-conserved OGs for the best dataset pair across each taxonomic split (Fig. 10 and 11).

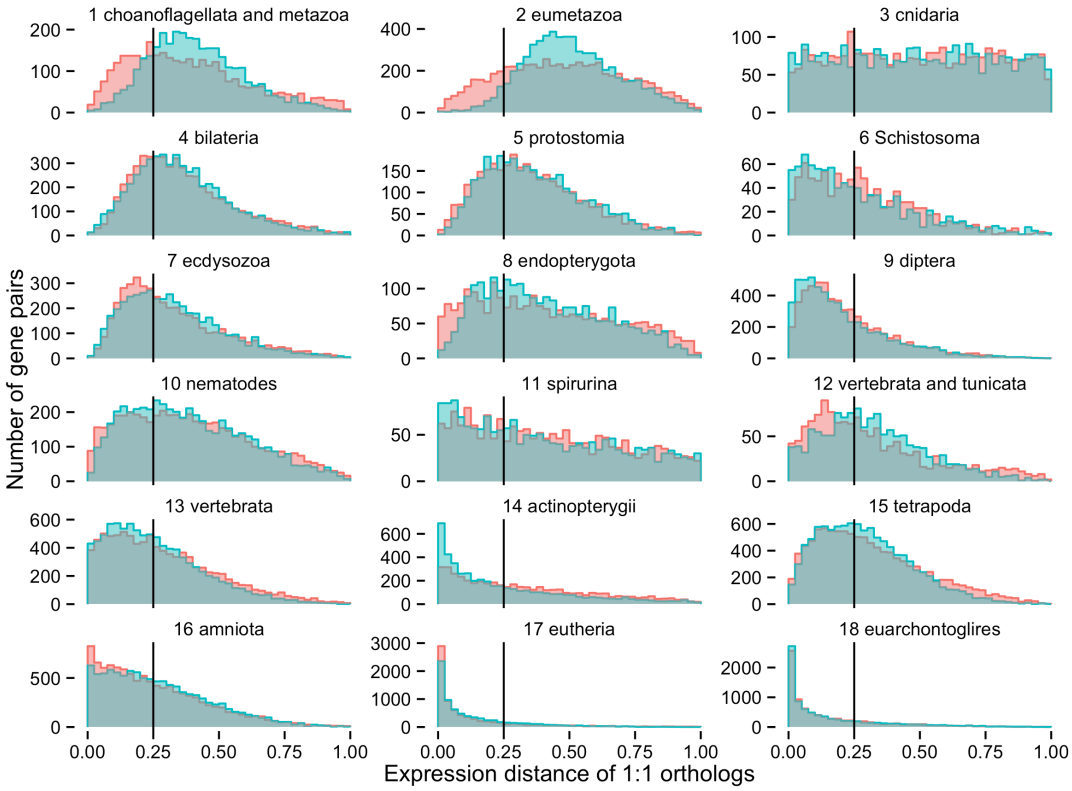


Fig. 8: Distribution of median conserved expression. For each clade, the distribution of expression distances of 1:1 orthologs is shown. Red and blue colors denote the direction of the mapping, either from the first subclade to the second or vice versa. Clades are numbered corresponding to the taxonomic tree in Fig. 9.

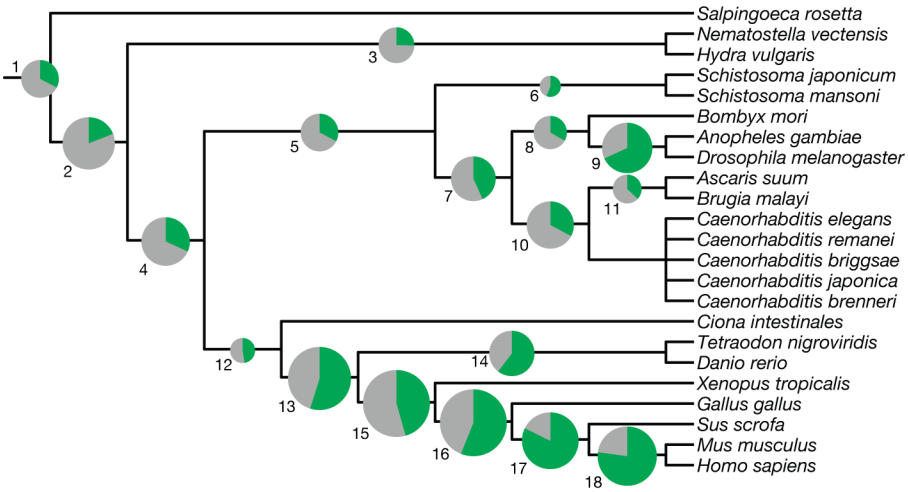


Fig. 9: Median conserved expression across animal clades. At each bifurcation, the pie chart denotes the median fraction of 1:1 orthologs with expression distances below 0.25. The area corresponds to the number of 1:1 orthologs across the taxonomic split. Numbers enumerate the clades for easier identification in the following figure.

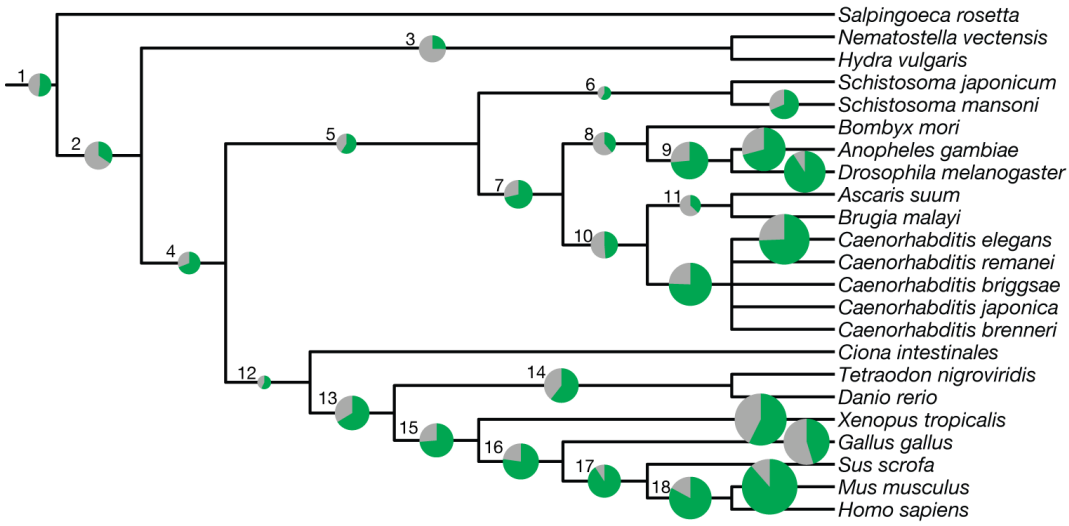


Fig. 10: Most conserved expression across animal clades. As in Fig. 9, the fraction of 1:1 orthologs with expression distances below 0.25 is shown. However, the fraction for the best dataset pair is shown instead of the median fraction for each clade. Thus, charts at species branches show how well expression patterns could be mapped using different datasets for the same species.

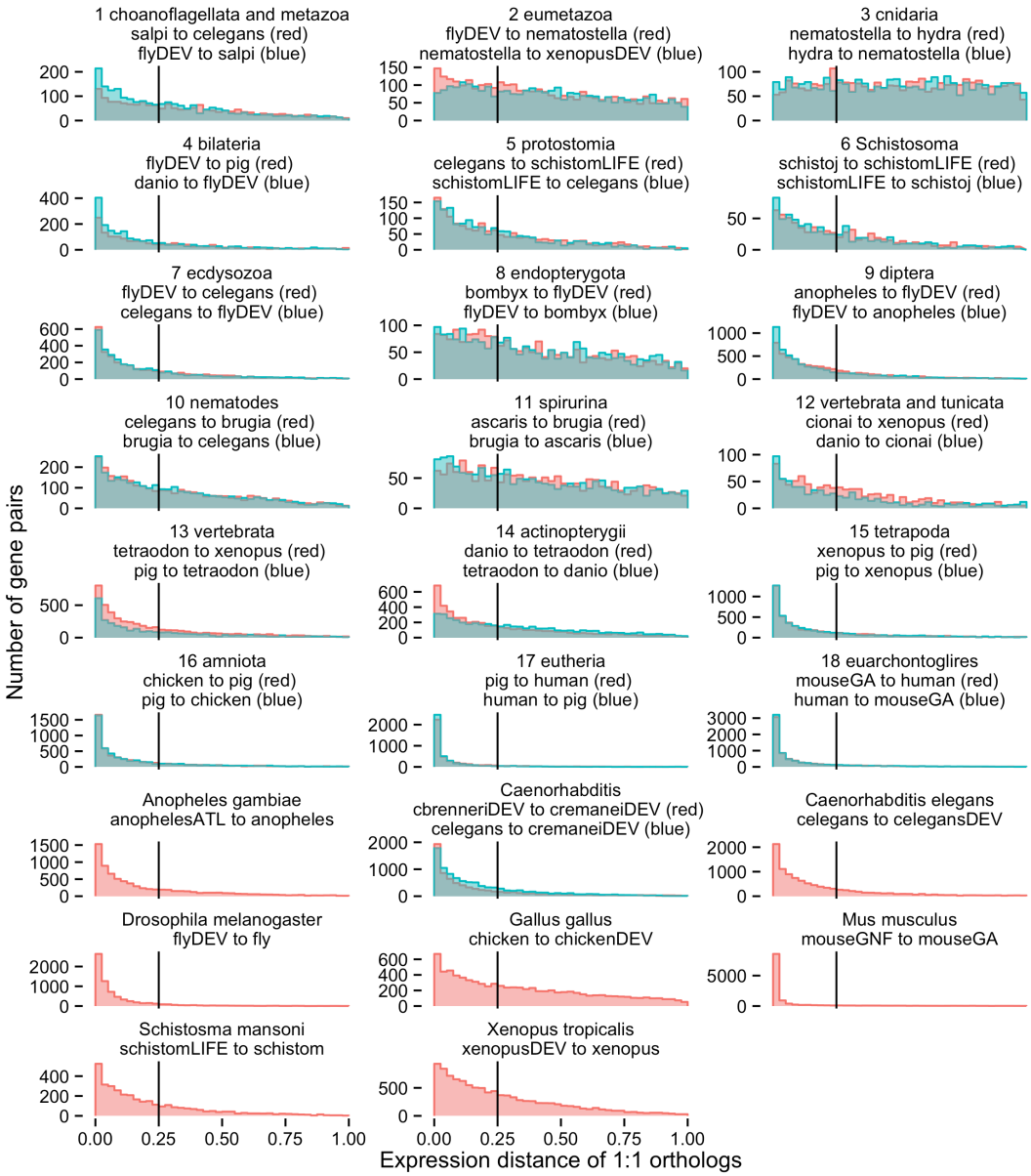


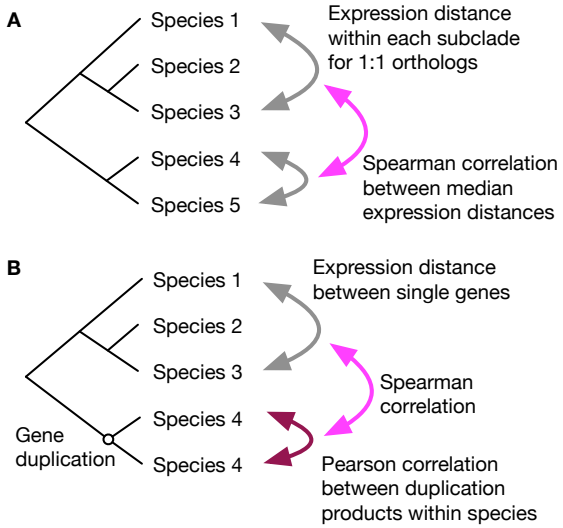
Fig. 11: Distribution of conserved expression for best dataset pairs. For each taxonomic split, the distribution of expression distances is shown for the datasets given. These datasets show the highest degree of conservation for the respective taxonomic split. See Table S1 for a description of the abbreviated dataset names.

The conservation of conservation

In the previous section, we showed that there is an enrichment in conserved expression programs across most taxonomic splits. Here, we analyzed how the rate of expression divergence itself is conserved across taxonomic splits. If the rate of expression divergence is a property of the gene family we expect a correlation between the expression similarities for each family in different clades. In other words, a gene that has a conserved expression pattern in one clade should also have a conserved expression pattern in another clade. For each taxonomic split with

two or more species on either side of the split, we calculated the median expression distance per gene family within each of the two clades. Out of six taxonomic splits with more than one species on both sides, we found significant Spearman correlations (r_s) of median expression similarities for three splits (Fig. 12A): between tetrapods and fishes (r_s=0.18, #13 in Fig. 9), between protostomes and deuterostomes (r_s=0.14, #4), and between nematodes and insects (r_s=0.074, #7). Not significant were the splits involving cnidaria (#2), *Schistosoma* (#5) and spirurina (#10).

The previous analysis was only possible for a subset of the taxonomic splits in our body of data, due to the requirement of having more than one species on either side of the split. We therefore also analyzed the fate of duplicated genes. In this case, we tested whether duplication products are more similar if the non-duplicated members of the gene family have low expression distances across the species outside the duplication event. Indeed, we found significant negative correlations between the median expression distance among the non-duplicated genes and the intra-species correlation of the duplicated genes (Fig. 12B). For example, duplicated genes in fish were more similar (i.e. has a higher correlation) when the corresponding tetrapod genes had more similar expression patterns (i.e. had a low expression distance): r_s=-0.11 for 1999 pairs of duplicated genes, corresponding to a p-value of 1e-7. Taken together, these two observations implied that for a significant fraction of genes, the rates of change in gene expression patterns were correlated between independently evolving clades.



#	clade	r_s	p-value	n
2	eumetazoa	0.04	0.03	2478
4	bilateria	0.14	0	4201
5	protostomia	0.05	0.24	629
7	ecdysozoa	0.07	1E-05	3467
10	rhabditida	0.04	0.13	1458
13	vertebrata	0.18	0	3391

#	clade with duplicated gene	clade with single genes	r_s	p-value	n
1	*Salpingocea*	eumetazoa	-0.14	0.08	164
4	deuterostomia	protostomia	-0.03	0.004	8618
5	*Schistosoma*	ecdysozoa	-0.11	0.02	440
5	ecdysozoa	*Schistosoma*	-0.10	0.02	560
7	endopterygota	rhabditida	-0.07	0.02	1120
12	*Ciona*	vertebrata	-0.39	0.03	29
13	actinopterygii	tetrapoda	-0.12	1E-07	1999

Fig. 12. Correlations between expression conservation rates. A For 1:1 orthologs, the expression distance across taxonomic splits was compared. In the dataset, there were only six splits with at least two species on both sides of the split. For example, when genes were similar within tetrapods, they also tended to be similar within fishes. **B** The rate at which gene duplication products diverge was negatively correlated with the expression distance among single-copy genes in related species. Only correlations with p-values below 0.1 are shown in this table. (# – Number of clade in Fig. 9, n – count of 1:1 orthologs [A] or duplicated genes [B])

Resolving the fate of gene duplication products

As described above, we identified sets of duplicated genes and related, unduplicated genes by constructing gene trees using GIGA (Thomas 2010). We then extracted sub-trees corresponding to a duplication event, and the corresponding related orthologous genes that emerged from speciation events. For each pair of duplication products under consideration, we considered the expression distance among the non-duplicated genes, and the expression distance across the duplication event. We then partitioned gene families based on their expression distance between the non-duplicated genes to distinguish genes that had conserved expression patterns from those that were more variable (Fig. S2). To study the functional divergence after gene duplication, we created an additional expression distance metric that combines measures for expression similarity and dissimilarity, which we term “expression divergence score.” This allowed us to also test if two genes have significantly diverging expression patterns. As above, we used the p-value for the null hypothesis that the genes are not related to each other (p_b) to quantify expression similarity. To measure expression dissimilarity, we used the p-value for the null hypothesis that considered genes are in fact 1:1 orthologs (p_o). We then combined the two p-values into an expression distance score E:

$$E = \begin{cases} -\log_{10} p_o & p_o \leq p_b \\ \log_{10} p_b & p_o > p_b \end{cases}$$

Thus, the expression divergence score *E* was negative for similar, and positive for dissimilar gene pairs. Considering gene duplications, we computed divergence scores for both duplication products (Fig. 13). Using $\log_{10}(0.25)$ as cutoff, we divided pairs of duplication products into three categories: (a) both genes had conserved expression patterns (2226 pairs of duplication products), (b) both genes had diverging expression patterns (911 pairs) and (c) only one of the duplication products had a diverging expression pattern, while the other one was conserved (1206 pairs). When the non-duplicated genes were more similar to each other, the fraction of duplication products that are both conserved was higher (Fig. 14), supporting again that purifying selection acts across large phylogenetic distances.

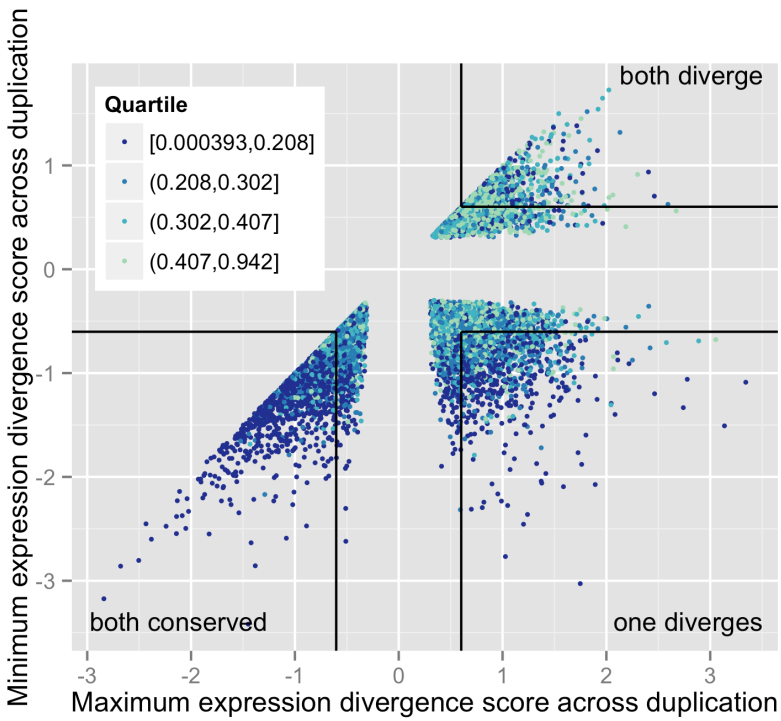


Fig. 13. Expression divergence scores of duplication products. For each duplication event, the expression divergence score of the duplication products to the non-duplicated genes was computed. For each pair of duplication products, the expression divergence scores were sorted. Thus, in the lower left quadrant, both duplication products had conserved expression patterns. In the upper right, both duplication products had diverging expression patterns and in the lower right, the outcome was mixed. Black lines denote an expression divergence score cutoff of 0.25.

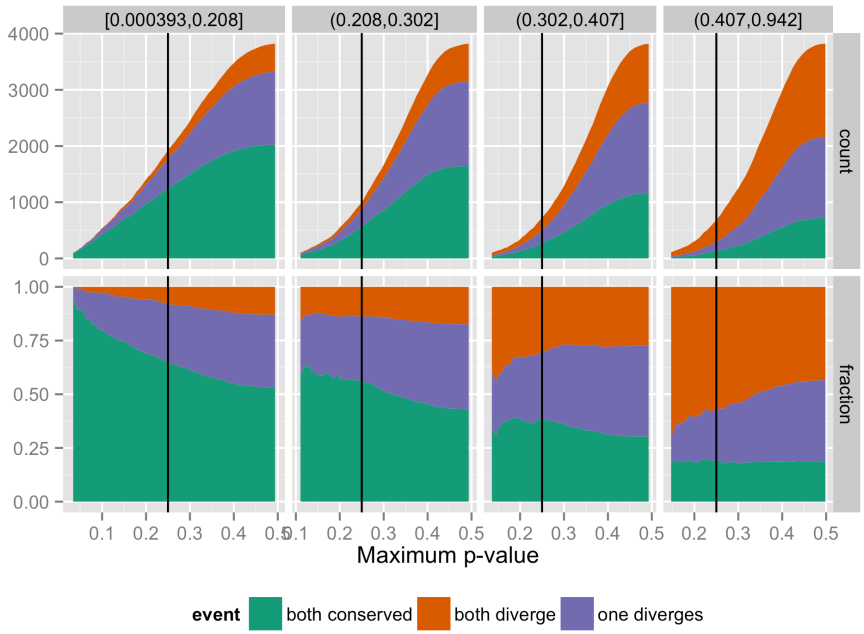


Fig. 14. Counts of duplication outcomes. Duplication events were grouped into quartiles according to the expression distance among the non-duplicated genes (Fig. S2). For each quartile, counts and fractions of the different outcomes are shown as the p-value cutoff is varied. This maximum p-value corresponds to a minimum of the absolute value of the expression divergence score, e.g. $|E| > \log_{10}(0.25)$ for the black lines.

Instead of partitioning duplication events into quartiles according to the expression distances among the non-duplicated genes, it was also possible to group them into smaller bins (Fig. 15). This way, correlations between the expression distance and the behavior of the duplication products became apparent: as the expression distance among the non-duplicated increased, the number of duplications with conserved expression patterns decreased, while more duplication events led to genes that both diverge. For each clade with duplicated genes, we shuffled the expression divergence scores 1000 times to assess the significance of the observed counts of events. We found that for all but one clade, there were significantly less duplications with mixed outcome (one conserved, one diverged) than expected, while there were significantly more duplications with identically behaving duplication products (either both are conserved or both diverge). (P-values: $p < 0.001$ for eight clades; $p < 0.001$ for one clade; $p < 0.05$ for one clade; $p > 0.05$ for one clade, namely duplications occurring in insects compared to non-duplicated genes in nematodes.)

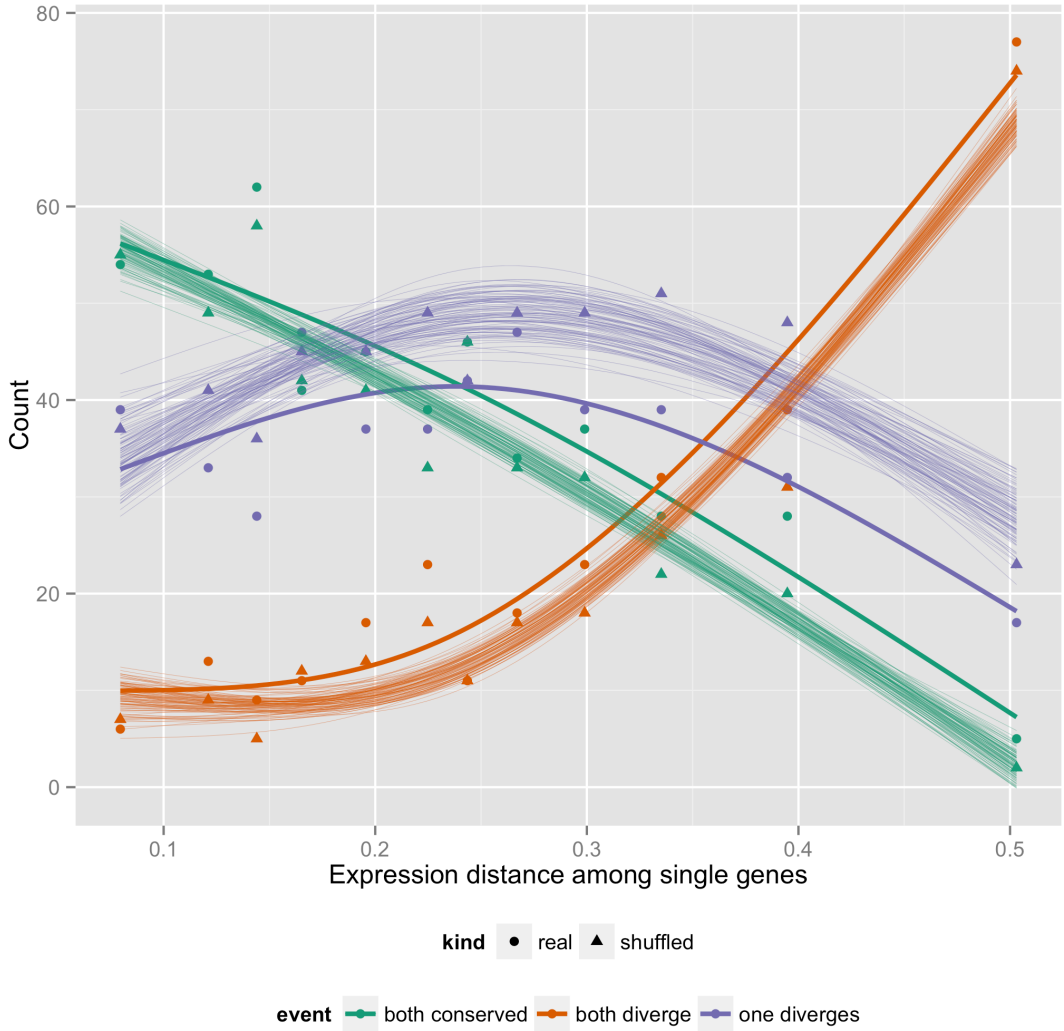


Fig. 15: Duplication products are more similar than expected. For each bin of 100 duplication events that occurred in deuterostomes, the expression pattern was compared to non-duplicated genes in protostomes. Overall, there were less duplication events with mixed outcome (one gene diverged, the other's expression pattern was conserved) than expected. Thin lines correspond to quadratic fits to 100 randomizations, and triangles correspond to the median counts in the randomizations.

Functional implications of diverging expression patterns in duplication products

In order to quantify the functional implications of the observed conservation or divergence of tissue-specific gene expression patterns between duplication products, we looked for differences in the protein–protein interactions (PPI) between the duplication products. To this end, we obtained interaction networks with a variety of selected evidence channels and confidence score thresholds from the STRING database (Franceschini et al. 2013). In this section, we report results for the complete interaction network with a minimum confidence score of 0.5, while the figures show results for other channels and cutoffs.

As discussed above, our method enabled us to distinguish between the duplication products by comparing their tissue-specific gene expression pattern to that of non-duplicated reference genes of the same gene family. Thus, we could designate one of the duplication products as “less diverging”, as it has lower expression distances to the non-duplicated genes than the other duplication product. First, we found that less diverging genes had significantly more PPI than the more diverging genes in 15 of 25 datasets with available STRING data (using a p-value cutoff of 0.05 for one-sided Wilcoxon signed-rank tests, Fig. 16). From this data, it remained unclear if the less diverging protein gained interaction partners, or if the more diverging protein lost interaction partners. However, the latter hypothesis seems more parsimonious: interaction partners are often tissue specific. That is, if the diverged protein got expressed in different tissues it likely lost some of its former interaction partners. To further corroborate this notion, we compared the interaction partners of the duplication products with the interaction partners of the respective non-duplicated genes.

Fig. 16: Less diverging proteins had more interaction partners. For each dataset with sufficient PPI data in the STRING database, we tested which duplication product has more interaction partners using one-sided Wilcoxon signed-rank tests. In most cases, the duplication products with lower expression distances to the non-duplicated reference genes had more PPI (at a p-value cutoff of 0.05), regardless of the chosen evidence channels or score cutoffs.

We mapped PPI across species by counting how many orthologous groups were shared between the interaction partners of the duplication product and the reference protein in a second species. We then calculated the Jaccard indices for the shared interaction partners between the reference protein and either of the duplication products. We found that for 40.5% of 395 dataset–species pairs, less diverging genes had a significantly higher Jaccard index compared to the more diverging genes (Fig. 17), while there were no dataset–species pairs where the more diverging proteins had significantly higher Jaccard indices. Furthermore, when we distinguished the possible outcomes discussed above (both duplication products have conserved expression patterns, one diverges, or both diverge), major differences only occurred for the case of one gene diverging (Fig. 18): for this case, less diverging genes had higher Jaccard indices in 40.0% of 175 dataset–species pairs, and all other outcomes were much less prevalent (<7%). (To be able to compare the p-values between the different outcomes, the same number of duplication products was used for the significance tests by sub-sampling 100 times.) This observation is consistent with earlier findings about the tissue specificity of protein complexes (Börnigen et al. 2013) and strengthens the notion that the duplication product with the more diverging expression pattern lost previous interactions and acquired novel interaction partners.

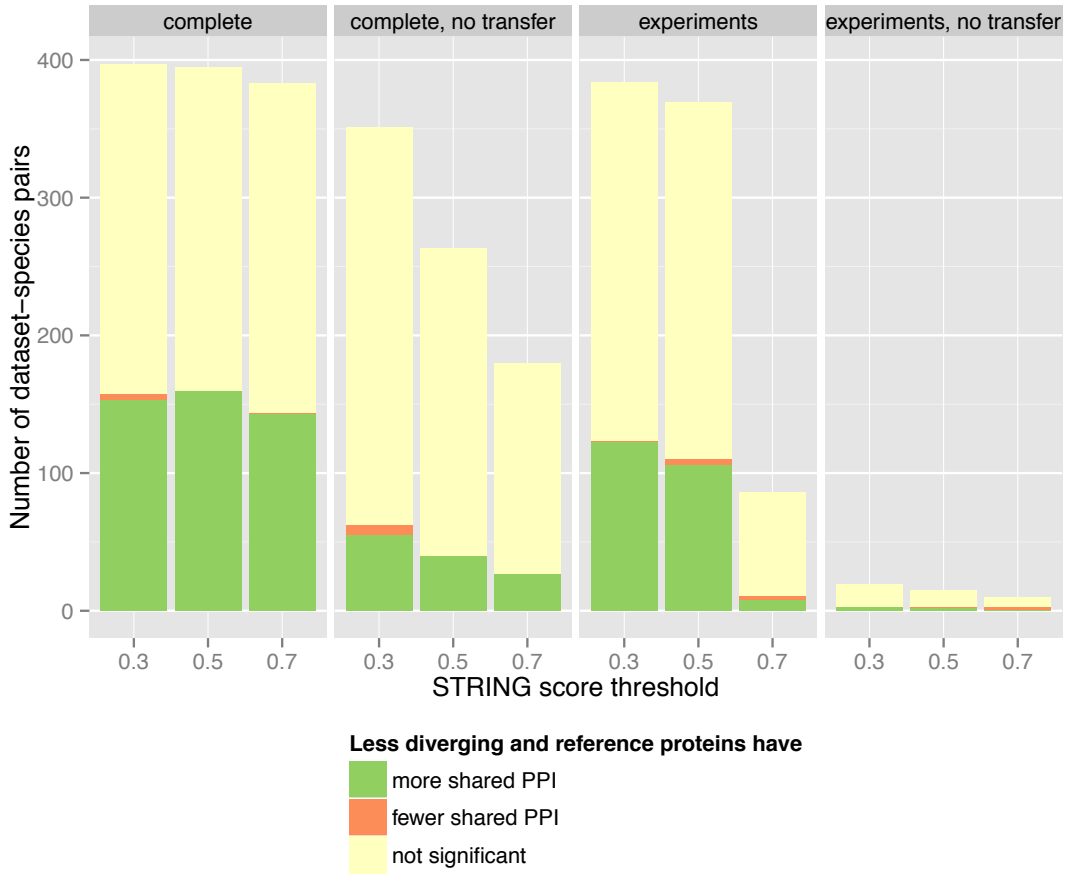


Fig. 17: Less diverging proteins shared more interaction partners with reference proteins. For each dataset-species with sufficient PPI data in the STRING database, we tested which duplication product shares more interaction partners with the non-duplicated reference protein. We compared Jaccard indices using one-sided Wilcoxon signed-rank tests and found a prevalence of duplication products with lower expression distances sharing more interaction partners with the non-duplicated reference gene.

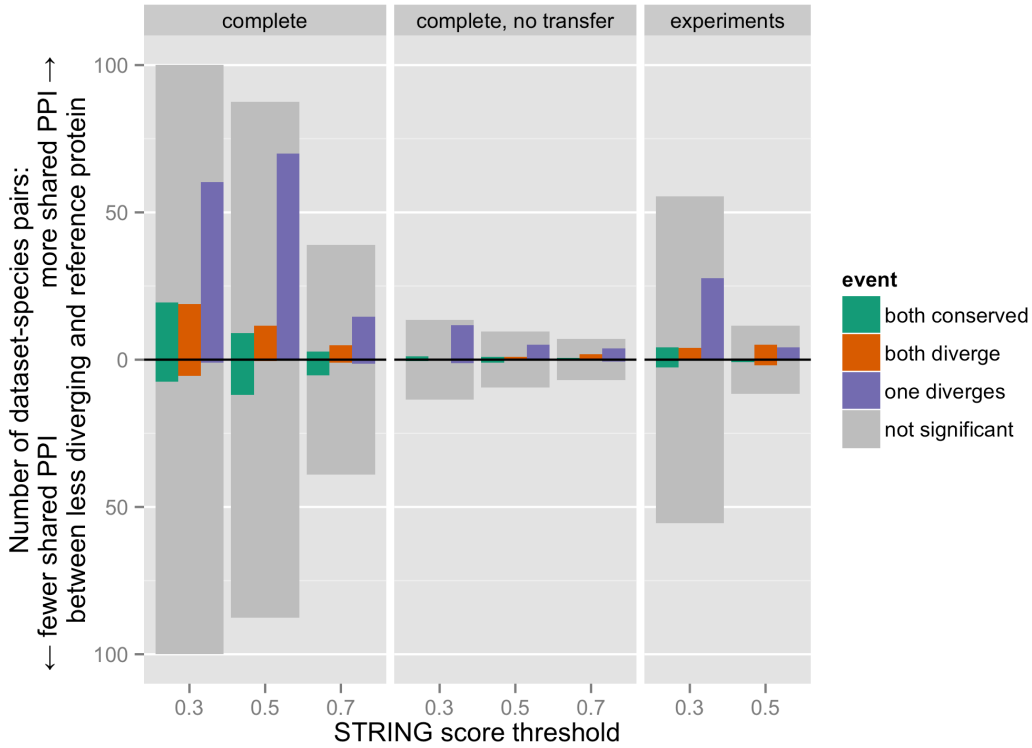


Fig. 18: Divergence of gene duplication products affects their protein–protein interactions. As in Fig. 17, we determined which duplication product shared more interaction partners with the non-duplicated reference protein. We further divided duplication events into three outcome categories (Fig. 14), and found that significant differences in the number of shared interaction partners are mainly observed for the case where one duplication product has a conserved expression pattern, while the other duplication product diverges in its expression pattern. In the other cases, i.e. when both duplication products were conserved of or both diverge, the difference in interaction partners became significant in fewer cases.

Discussion

We have shown that tissue-specific gene expression can be predicted across large evolutionary distances, even in the absence of apparent similarities between the species' tissues. Our approach can be rationalized as follows: we assume that evolution conserves the co-expression of functionally related genes, both on the level of homologous cell types and on the level of functional modules that occur in unrelated tissues. Our analysis demonstrated that the expression patterns of such conserved gene modules can be predicted across species using 1:1 orthologs as "anchors." This approach worked despite the fact that the tissues themselves are only conserved within smaller clades. Control of gene expression by transcription factors, miRNAs and other factors is known to turn over rather quickly (Odom et al. 2007; Bradley et al. 2010; Berezikov 2011). Most probably, over large evolutionary

distances functional dependencies between genes lead to shared expression patterns. Further research will be needed to reveal which expression similarities between tissues are caused by homology and which are caused by convergent evolution.

When we applied the concept of looking for correlations between orthologs across species to an existing dataset (Brawand et al. 2011), we found that many of the reported lineage-specific expression shifts only change the absolute expression levels, while the relative expression patterns stayed conserved (Fig. S3). This suggests that further studies could combine approaches that test absolute and relative expression patterns to identify truly novel expression patterns. In a first step, we investigated products of gene duplication events and found that they seem to have the ability to "opt out" of such gene expression modules to acquire new functions. Such events suggest unidirectional dependences: whereas the duplicated gene does not need (all of) its ancient interaction partners, the partners seem to need the duplicated gene and, thus, one of the two remained in the respective expression module. Between clades, there is a significant amount of conservation in the rates of change of gene expression patterns among the members of gene families.

Methods

Import of expression data

Datasets were obtained either from repositories like ArrayExpress and GEO, from supplementary materials or the respective websites of the resources. Expression profiles were then mapped to our set of genes by one of the following methods (see Table S1): If possible, genes were mapped by given identifiers, such as Affymetrix, Ensembl or WormBase identifiers. If identifiers could not be used for microarrays, we mapped probe sequences to transcripts using exonerate (Slater & Birney 2005), allowing for up to three mismatches and discarding probes that mapped to multiple genes. In the case of RNA-seq data without matching identifiers, we mapped reads to annotated transcripts using tophat2 and cufflinks 2.1.1 (Kim et al. 2013; Trapnell et al. 2010) and used the resulting FPKM counts.

Normalization of expression data

In initial small-scale tests, we tested several normalization methods (Liao & Zhang 2006; Piasecka, Robinson-Rechavi, et al. 2012b), and settled on a *z*-like normalization of expression vectors **x**, which corresponds to the Euclidean normalization of **x** minus its median value. Therefore, we did not look for conserved expression abundance, but rather for conserved relative expression across tissues. Normalizing each gene's expression individually also avoided technical concerns regarding the comparability of absolute expression values between genes. RNA-seq data, e.g. the *Drosophila* modENCODE dataset, contained zeros, which were of course not suitable for logarithmic analysis. For these datasets, we determined the expression value of the 1/1000th quantile of all genes with non-zero expression. All expression values were incremented by this value.

Tissue correlations between species

P-values for tissue correlations were calculated analytically. We performed tests with shuffling of genes to confirm that the analytical p-values correspond to empirical p-values.

Mapping of tissue expression patterns

For each pair of datasets, individual linear models were trained for each tissue of the target species, using the tissues of the source species as input. (Note that due to the normalization, one tissue is redundant and therefore left out. This also implies that the coefficients of the linear model are not directly interpretable.) The set of 1:1 orthologs between the two species was used as a training set. (When there were multiple probes per gene, all combinations of probes were used for training.) When there are many tissues in the source species, but few 1:1 orthologs, there is the danger of over-fitting. We therefore allowed only one predictor (i.e. one tissue from the source species) per 15 samples (1:1 orthologs) (Babyak 2004). For each pair of species, the safe number of predictors was calculated. If there were too many tissues, we combined tissues using *k*-means clustering and used the centers of the clusters as predictors. This situation only occurred for six out of 992 dataset pairs. The trained models are then applied to all genes of the source species, yielding corresponding predicted expression patterns in the target species. Note that since the 1:1 orthologs are used for training, we used predictions from a 10-fold cross-validation for these genes.

Computation of expression distances

For each pair of datasets, we computed a matrix of predicted expression patterns of all genes from the source species. We then calculated the weighted Pearson correlations between the mapped expression patterns and the actual expression patterns of the target species' genes. Weights on the tissues were calculated using the Gerstein-Sonnhammer-Chothia (GSC) weighting scheme to reduce the effect of uneven coverage of different anatomical regions (Gerstein et al. 1994). For example, in the mouse tissue datasets, there are many different brain tissues. Given the matrix of all weighted Pearson correlations, we then calculated expression distances like p-values, i.e. by determining the fraction of unrelated genes that have the same or higher correlation. For technical reasons, we sampled one million pairs of background genes, such that the lowest possible expression distance is 1e-6.

As mentioned in the Results section, there is a strong correlation between the raw expression distances and the number of genes in the target species. To counter this effect, target genes are split into ten bins according to the number of co-expressed genes in the target species. Thus, there exist ten conversion functions from weighted Pearson correlation to an uncorrected expression distance. For a given pair of genes, the final expression distance is interpolated from the two adjacent bins.

We determined the number of co-expressed genes for each target gene as follows: we first computed all pairwise correlations among the target genes of the training set. Then, we determined the correlation cutoff corresponding to the top 10%, and counted for each gene how many other target genes were among the global top 10% correlations. This strong correlation showed that predictions were biased towards the average target gene (i.e. the average expression profile of all genes considered in the target species), which in turn was similar to many target genes. As a consequence, these "close-to-average" target genes had higher correlations with mapped source genes, and thus seemed more conserved. To correct for this, we used the counts of co-expressed genes and split the target genes into ten bins. Distributions of correlations were determined for each bin (Fig. 2b).

Acknowledgements

The authors thank Anthony A. Hyman and Vineeth Surendranath for helpful discussions.

Funding

MK is funded by the Deutsche Forschungsgemeinschaft (DFG KU 2796/2-1).

Author Contributions

AB and MK conceived the study, planned the analyses and wrote the paper. MK conducted all analyses.

Competing Interests

The authors declare that there are no competing interests.

References

Babyak, M.A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, 66(3), pp.411–421.

Baker, D.A. et al., 2011. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, Anopheles gambiae. *BMC Genomics*, 12, p.296. doi:10.1186/1471-2164-12-296.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp.289–300. doi:10.2307/2346101.

Berezikov, E., 2011. Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics*, 12(12), pp.846–860. doi:10.1038/nrg3079.

Börnigen, D. et al., 2013. Concordance of gene expression in human protein complexes reveals tissue specificity and pathology. *Nucleic acids research*, 41(18), p.e171. doi:10.1093/nar/gkt661.

Bradley, R.K. et al., 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS biology*, 8(3), p.e1000343. doi:10.1371/journal.pbio.1000343.

Brawand, D. et al., 2011. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369), pp.343–348. doi:10.1038/nature10532.

Chan, E.T. et al., 2009. Conservation of core gene expression in vertebrate tissues. *Journal of biology*, 8(3), p.33. doi:10.1186/jbiol130.

Chikina, M.D. & Troyanskaya, O.G., 2011. Accurate Quantification of Functional Analogy among Close Homologs. *PLoS computational biology*, 7(2), p.e1001074. doi:10.1371/journal.pcbi.1001074.

Chikina, M.D. et al., 2009. Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. *PLoS computational biology*, 5(6), p.e1000417. doi:10.1371/journal.pcbi.1000417.

Chung, H. et al., 2009. Characterization of Drosophila melanogaster cytochrome P450 genes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14), pp.5731–5736. doi:10.1073/pnas.0812141106.

Dissanayake, S.N. et al., 2006. angaGEDUCI: Anopheles gambiae gene expression database with integrated comparative algorithms for identifying conserved DNA motifs in promoter sequences. *BMC Genomics*, 7, p.116. doi:10.1186/1471-2164-7-116.

Domazet-Loso, T. & Tautz, D., 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, 468(7325), pp.815–818. doi:10.1038/nature09632.

Fairclough, S.R. et al., 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. *Genome biology*, 14(2), p.R15. doi:10.1186/gb-2013-14-2-r15.

Fitzpatrick, J.M. et al., 2009. Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses. *PLoS neglected tropical diseases*, 3(11), p.e543. doi:10.1371/journal.pntd.0000543.

Franceschini, A. et al., 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(Database issue), pp.D808–15. doi:10.1093/nar/gks1094.

Freeman, T.C. et al., 2012. A gene expression atlas of the domestic pig. *BMC Biology*, 10, p.90. doi:10.1186/1741-7007-10-90.

Gerstein, M.B., Sonnhammer, E.L.L. & Chothia, C., 1994. Volume changes in protein evolution. *Journal of Molecular Biology*, 236(4), pp.1067–1078.

Gobert, G.N. et al., 2009. Developmental gene expression profiles of the human pathogen Schistosoma japonicum. *BMC Genomics*, 10, p.128. doi:10.1186/1471-2164-10-128.

Goltsev, Y. et al., 2009. Developmental and evolutionary basis for drought tolerance of the Anopheles gambiae embryo. *Developmental biology*, 330(2), pp.462–470. doi:10.1016/j.ydbio.2009.02.038.

Gu, X. & Su, Z., 2007. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proceedings of the National Academy of Sciences of the United States of America*, 104(8), pp.2779–2784. doi:10.1073/pnas.0610797104.

Hemmrich, G. et al., 2012. Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity.

Molecular Biology and Evolution, 29(11), pp.3267–3280. doi:10.1093/molbev/mss134.

Irie, N. & Kuratani, S., 2011. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nature communications*, 2, p.248. doi:10.1038/ncomms1248.

Kim, D. et al., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4), p.R36. doi:10.1186/gb-2013-14-4-r36.

Lees, J.G. et al., 2014. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic acids research*, 42(1), pp.D240–5. doi:10.1093/nar/gkt1205.

Levin, M. et al., 2012. Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Developmental cell*, 22(5), pp.1101–1108. doi:10.1016/j.devcel.2012.04.004.

Liao, B.-Y. & Zhang, J., 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular Biology and Evolution*, 23(3), pp.530–540. doi:10.1093/molbev/msj054.

Lukk, M. et al., 2010. A global map of human gene expression. *Nature Biotechnology*, 28(4), pp.322–324. doi:10.1038/nbt0410-322.

McGary, K.L. et al., 2010. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(14), pp.6544–6549. doi:10.1073/pnas.0910200107.

Nawaratna, S.S.K. et al., 2011. Gene Atlasing of digestive and reproductive tissues in Schistosoma mansoni. *PLoS neglected tropical diseases*, 5(4), p.e1043. doi:10.1371/journal.pntd.0001043.

Niknejad, A. et al., 2012. vHOG, a multispecies vertebrate ontology of homologous organs groups. *Bioinformatics (Oxford, England)*, 28(7), pp.1017–1020. doi:10.1093/bioinformatics/bts048.

Odom, D.T. et al., 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, 39(6), pp.730–732. doi:10.1038/ng2047.

Piasecka, B., Kutalik, Z., et al., 2012a. Comparative modular analysis of gene expression in vertebrate organs. *BMC Genomics*, 13, p.124. doi:10.1186/1471-2164-13-124.

Piasecka, B., Robinson-Rechavi, M. & Bergmann, S., 2012b. Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human. *Bioinformatics (Oxford, England)*, 28(14), pp.1865–1872. doi:10.1093/bioinformatics/bts266.

Powell, S. et al., 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic acids research*, 42(1), pp.D231–9. doi:10.1093/nar/gkt1253.

Robinson, S.W. et al., 2013. FlyAtlas: database of gene expression in the tissues of Drosophila melanogaster. *Nucleic acids research*, 41(Database issue), pp.D744–50. doi:10.1093/nar/gks1141.

Seok, J. et al., 2013. Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 110(9), pp.3507–3512. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23401516&retmode=ref&cmd=prlinks.

Shoguchi, E. et al., 2011. Direct examination of chromosomal clustering of organ-specific genes in the chordate Ciona intestinalis. *Genesis (New York, N.Y. : 2000)*, 49(8), pp.662–672. doi:10.1002/dvg.20730.

Shubin, N., Tabin, C. & Carroll, S., 2009. Deep homology and the origins of evolutionary novelty. *Nature*, 457(7231), pp.818–823. doi:10.1038/nature07891.

Silver, D.H., Levin, M. & Yanai, I., 2012. Identifying functional links between genes by evolutionary transcriptomics. *Molecular BioSystems*, 8(10), pp.2585–2592. doi:10.1039/c2mb25054c.

Slater, G.S.C. & Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, p.31. doi:10.1186/1471-2105-6-31.

Spencer, W.C. et al., 2011. A spatial and temporal map of C. elegans gene expression. *Genome research*, 21(2), pp.325–341. doi:10.1101/gr.114595.110.

St Pierre, S.E. et al., 2014. FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic acids research*, 42(Database issue), pp.D780–8. doi:10.1093/nar/gkt1092.

Strausfeld, N.J. & Hirth, F., 2013. Deep Homology of Arthropod Central Complex and Vertebrate Basal Ganglia. *Science (New York, NY)*.

Su, A.I. et al., 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16), pp.6062–6067. doi:10.1073/pnas.0400782101.

Thomas, P.D., 2010. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*, 11, p.312. doi:10.1186/1471-2105-11-312.

Thomas, P.D. et al., 2012. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS computational biology*, 8(2), p.e1002386. doi:10.1371/journal.pcbi.1002386.

Tomer, R. et al., 2010. Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium. *Cell*, 142(5), pp.800–809. doi:10.1016/j.cell.2010.07.043.

Trapnell, C. et al., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5), pp.511–515. doi:10.1038/nbt.1621.

Tulin, S. et al., 2013. A quantitative reference transcriptome for Nematostella vectensis early embryonic development: a pipeline for de novo assembly in emerging model systems. *EvoDevo*, 4(1), p.16. doi:10.1186/2041-9139-4-16.

Wang, Z. et al., 2013. Gene expression analysis distinguishes tissue-specific and gender-related functions among adult Ascaris suum tissues. *Molecular genetics and genomics : MGG*, 288(5-6), pp.243–260. doi:10.1007/s00438-013-0743-y.

Winter, E.E., Goodstadt, L. & Ponting, C.P., 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome research*, 14(1), pp.54–61. doi:10.1101/gr.1924004.

Xia, Q. et al., 2007. Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, Bombyx mori. *Genome biology*, 8(8), p.R162. doi:10.1186/gb-2007-8-8-r162.

Yanai, I. et al., 2011. Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Developmental cell*, 20(4), pp.483–496. doi:10.1016/j.devcel.2011.03.015.