Conservation of expression regulation throughout the animal kingdom

Michael Kuhn[1,2,*] and 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)

Abstract

Following the increase in available sequenced genomes, tissue-specific transcriptomes are being determined for a rapidly growing number of highly diverse species. Traditionally, only the transcriptomes of related species with equivalent tissues have been compared. Such an analysis is much more challenging over larger evolutionary distances when complementary tissues cannot readily be defined. Here, we present a method for the cross-species mapping of tissue-specific and developmental gene expression patterns across a wide range of animals, including many non-model species. Our approach maps gene expression patterns between species without requiring the definition of homologous tissues. With the help of this mapping, gene expression patterns can be compared even across distantly related species. In our survey of 36 datasets across 27 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. 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 [1]. 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 when a gene is duplicated in a non-model species, but not in the model species where its function has originally been studied. Previously, gene coexpression data has been used to find conserved coexpressed modules [2, 3] and to uncover functional similarities between genes from different species [4]. However, the latter approach requires that the two species are well-studied in both gene expression and functional annotation, and will suffer from incomplete and biased annotations [1].

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. Developmental gene expression profiles between closely related species can be compared to find functional links between genes and to detect differences between orthologs [5, 6, 7]. For closely related species, homologous tissues can easily be identified [8], and cross-species correlations between equivalent tissues of closely related species have previously been investigated [9, 10, 11]. Existing approaches require that expression datasets have been obtained under comparable conditions for the respective species. Across larger evolutionary distances, only few clearly homologous tissues can be determined. Even between closely related species, the relative amounts of cell types within tissues may change. This reliance on homologous tissue is, therefore, a severe limitation for functional mapping between many species: it is not possible to correlate gene expression patterns across species using the traditional methods. If it was possible to compare expression patterns across large phylogenetic distances, we could substantially improve the annotation of non-model-species genomes, fill annotation gaps in model species, and in particular address the problem of functional conservation after gene duplications.

To ameliorate this situation, we have developed a method to map tissue expression patterns of genes from one species to another, without defining equivalent tissues between the two species. Our hypothesis is that groups of functionally related genes will be coexpressed in very different tissues and species due to the re-use of ancestral functional modules. For example, it is possible to identify deep homologies among tissues [12], like homologous structures in the nervous systems of vertebrates and annelids [13, 14]. Other organs show functional convergence, e.g. mammalian liver and brown fat in flies, which both carry out xenobiotic clearance functions [15]. For each gene of the source species, our approach predicted a virtual tissue expression pattern in the destination species. The correlation between these virtual expression patterns and the actually observed expression could then be used to score how well a gene's expression of a gene in a target species can be predicted from the expression patterns of its orthologs. Importantly, this scheme can be used to determine the extent to which the transcriptional regulation of sets of genes is conserved across large phylogenetic distances.Subsequently we illustrate the potential of our modeling approach with two applications: determining the degree of conservation of tissue-specific gene expression patterns, and for comparing the speed of functional divergence between independently evolving members of protein families.

Results

To analyze tissue expression across the entire metazoan kingdom, we gathered genome and tissue expression data from 36 datasets covering 27 different species (Table 1, Table S1). The datasets contained both developmental time courses (e.g. embryonic stages) and static measurements of different tissues (like adult organs; see Supplementary File 1 for a complete list). For the sake of brevity, we refer to the all of these samples as "tissues." Datasets were imported and normalized per gene (see Methods), i.e. we quantified only relative expression changes of the same gene between tissues, instead of comparing expression differences of genes within the same tissue. (Therefore, housekeeping genes and other genes which are globally expressed did not skew our

Species	1:1 OGs with human	Dataset	Tissues	Timepoints	Kind
Chordates					
Homo sapiens	n/a	[16]	84	0	Microarray
Mus musculus	15008	GNF [17]	59	5	Microarray
		MOE [17]	51	0	Microarray
		dev [18]	0	8	Microarray
Sus scrofa	11156	[19]	57	0	Microarray
Gallus gallus	10900	[20]	20	0	Microarray
		dev [18]	0	15	Microarray
Xenopus tropicalis	9888	[20]	20	0	Microarray
		dev [5]	0	15	Microarray
Tetraodon nigroviridis	7627	[20]	20	0	Microarray
Danio rerio	7851	dev [21]	0	64	Microarray
Ciona intestinalis	2793	[22]	11	0	Microarray
Echinoderms					
Heliocidaris erythrogramma	2780	dev [23]	0	7	RNA-seq
Nematodes					
Ascaris suum	2460	[24]	10	0	Microarray
Brugia malayi	1546	dev [25]	0	7	RNA-seq
Caenorhabditis brenneri	1694	dev [6]	0	10	Microarray
Caenorhabditis briggsae	2509	dev [6]	0	10	Microarray
Caenorhabditis elegans	2589	dev [6]	0	10	Microarray
		[26]	40	0	Microarray
Caenorhabditis japonica	2213	dev [6]	0	10	Microarray
Caenorhabditis remanei	2309	dev [6]	0	10	Microarray
Insects					
Anopheles gambiae	3083	mix [27]	7	7	Microarray
		dev [28]	0	20	Microarray
		[29]	15	0	Microarray
Bombyx mori	2894	[30]	10	0	Microarray
Drosophila melanogaster	3094	mix [31]	26	0	Microarray
		dev [32]	0	30	RNA-seq
Flatworms					
Schistosoma japonicum	2207	dev [33]	0	13	Microarray
Schistosoma mansoni	2211	[34]	13	0	Microarray
		dev [35]	0	15	Microarray
Sponges					
Amphimedon queenslandica	2432	dev [36]	0	7	RNA-seq
Sycon ciliatum	2852	dev [37]	4	16	RNA-seq
Cnidaria					
Nematostella vectensis	3258	dev [38]	0	6	RNA-seq
Hydra vulgaris	2358	[39]	5	0	RNA-seq
Ctenophora					
Pleurobrachia bachei	1435	mix [40]	5	10	RNA-seq
Choanoflagellates					
Salpingoeca rosetta	2259	[41]	8	0	RNA-seq

Table 1: Analyzed species and datasets

analysis.) When we applied the concept of looking for correlations between orthologs across species to an existing dataset [10], we found that many of the reported lineage-specific expression shifts only changed the absolute expression levels, while the relative expression patterns remained conserved (Fig. S1). Normalizing each gene's expression individually also avoided technical concerns regarding the comparability of absolute expression values between genes. However, this gene-wise normalization means that the normalized values are influenced by the complement of tissues that have been measured. For this reason, we only include datasets that survey a whole organism or a wide range of developmental time points. The datasets excluded during quality control (see next section) have between five and ten data points. Therefore, six diverse tissues seemed to be a lower limit for the number of data points.

Quality control

The available datasets differ in their suitability for cross-species mapping. Existing measures for the quality of expression datasets rely on conserved features, e.g. conserved coexpression [42]. Because of the large biological diversity of species in our dataset, relying on conservation of features was not appropriate. We therefore devised a simple measure of dataset quality that only relied on the features of the given dataset. For each normalized dataset, we performed Principal Component Analysis and determined the proportion of variance represented by each eigenvector. We then calculated the fraction v_{50} of components that represent at least half of the total variance. For example, for the *C. elegans* dataset, the first four out of forty principal components explain just above 50% of the variance (hence, $v_{50} = 0.1$). Based on the observed correlation between v_{50} and median mapping quality (Fig. S2), we chose $v_{50} \geq 0.25$ as a filter to remove the five worst datasets from our analysis: *Hydra vulgaris*, *Amphimedon queenslandica*, *Bombyx mori*, *Brugia malayi* and *Ascaris suum*.

Mapping gene expression between species

In order to compare expression patterns across distant species, we first need to map the patterns. Our concept rests on the notion that the expression of a gene in a specific tissue of a target species can be predicted using the expression pattern of that gene across the tissues in the source species. For example, a gene specifically expressed in insect neurons is likely to be expressed also in the mouse brain. Here we show that this concept holds even if these "matching" tissues are not known. Consider the example of mapping gene expression patterns from fly to mouse (Fig. 1). We model $m_{g,t}$, the relative expression of gene g in mouse tissue t, as a linear combination of the relative expression levels in all fly tissues ($f_{\hat{g},s}$):

$$m_{g,t} = \sum_s \beta_{s,t} f_{\hat{g},s} + \beta_{0,t} + \epsilon_{g,t} \quad (1)$$

where $\hat{g}$ is the fly ortholog of gene g and ϵ is the residual error. The regression coefficients $\beta_{s,t}$ and the intercept $\beta_{0,t}$ are fitted using all 1:1 orthologs between mouse and fly. Subsequently, this model can be applied to all fly genes to predict the expression in mouse tissue t. We used linear models in this first description of the method as they are a

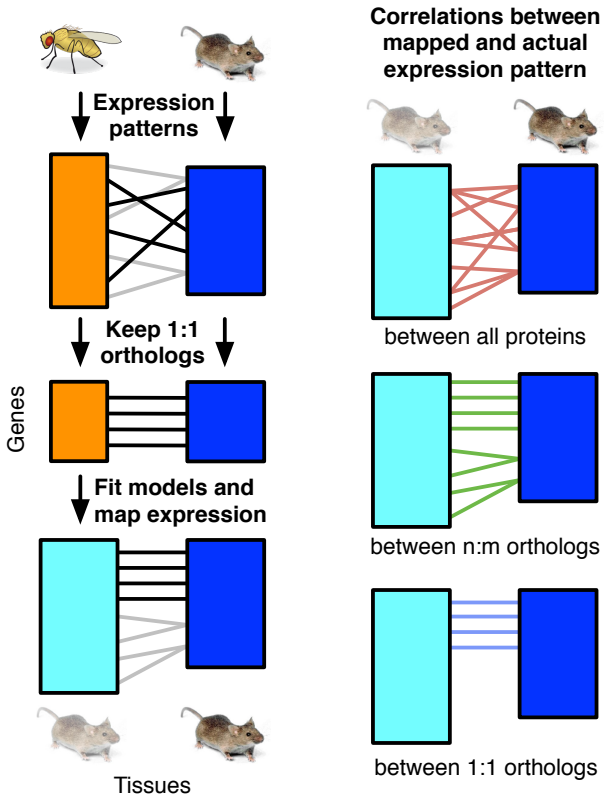


Figure 1: Mapping expression patterns across species. For each tissue in the target species, models were fitted to predict the tissue-specific gene expression pattern from the expression patterns of 1:1 orthologs in a source species. Mapping the expression patterns of all genes created virtual expression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

simple, transparent and efficient method that is relatively robust to over-fitting. Of course, other methods may be used as well. For example, Random Forest regression [43] can deal with non-linearity, while the lasso [44] could be used to deal with redundancy between source tissues.

Expression distances between genes

After mapping expression patterns between species, we quantified how well a gene can be predicted by correlating its predicted expression across all tissues with the respective measurements of the target species using Pearson's correlation coefficient (see Methods). These pairwise correlations between genes could be calculated for different sets of genes: phylogenetically unrelated genes, orthologs and 1:1 orthologs. Of these, 1:1 orthologs had the highest correlations (Fig. 2A). However, the overall distribution of correlations differed between dataset pairs, e.g. due to the varying number of tissues or variable data quality. Therefore, we computed an expression distance based on the quantiles $q_{x,y}$ of the matrix of correlation coefficients $\mathbf{R} = [r_{x,y}]$. We found that lineage-specific genes (i.e. those without homologs between the two species under consideration) tended to have lower correlations than genes with homologs. Therefore, to calculate quantiles, we computed the matrix $\mathbf{R}$ only for genes with homologs between the two species. We first analyzed correlations between 1:1 orthologs and checked if they

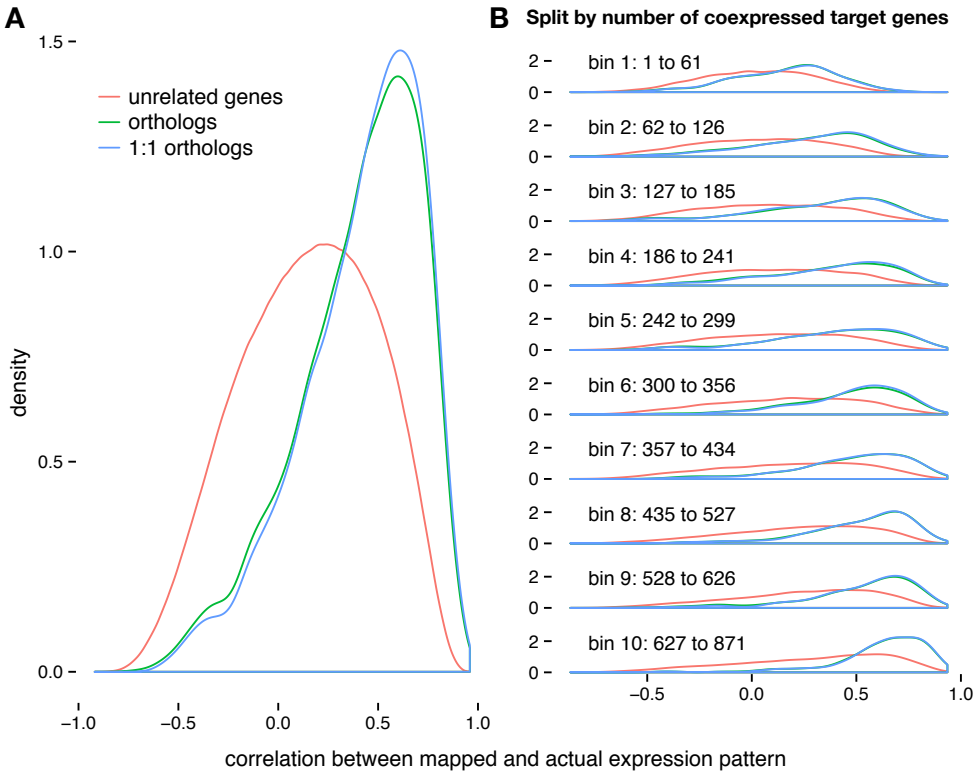


Figure 2: Distribution of correlations between mapped and actual expression patterns. When mapping expression patterns from fly to *C. elegans*, correlations between orthologs (green) and 1:1 orthologs (blue) were much higher than for background gene pairs (pairs of genes that are not homologous to each other, shown in red). **(b)** Target genes were split 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 321 and 326 one-to-one orthologs, with an average of 324.

depended on different properties of the genes (Fig. S3). We found that when the target gene had many coexpressed genes in the same species, the cross-species expression correlation tended to be higher (Fig. S3). To correct for this effect, we considered only target genes with similar numbers of coexpressed genes when computing the expression distance (Fig. 2B and Fig. 9 in Methods section). By design, the expression distance of background gene pairs had an uniform distribution. When making inferences about 1:1 orthologs, we used linear models based on a 10-fold cross-validation.

Benchmarks

In order to establish the biological relevance of our expression distance measure, we applied benchmarks at three levels, namely sequence, structure, and function. On the sequence level, we found that expression distances could be used as a signal to decide which of the top two BLAST hits for a query protein is the true 1:1 ortholog of the query protein in the target species (Fig. 3A and Fig. S4). On the structural level [45], expression distance and the number of proteins belonging to a structural fold were correlated (Fig. 3B and Fig. S5). That is, structural folds with fewer members, and hence lower functional diversity, were more similar in their expression patterns across species.

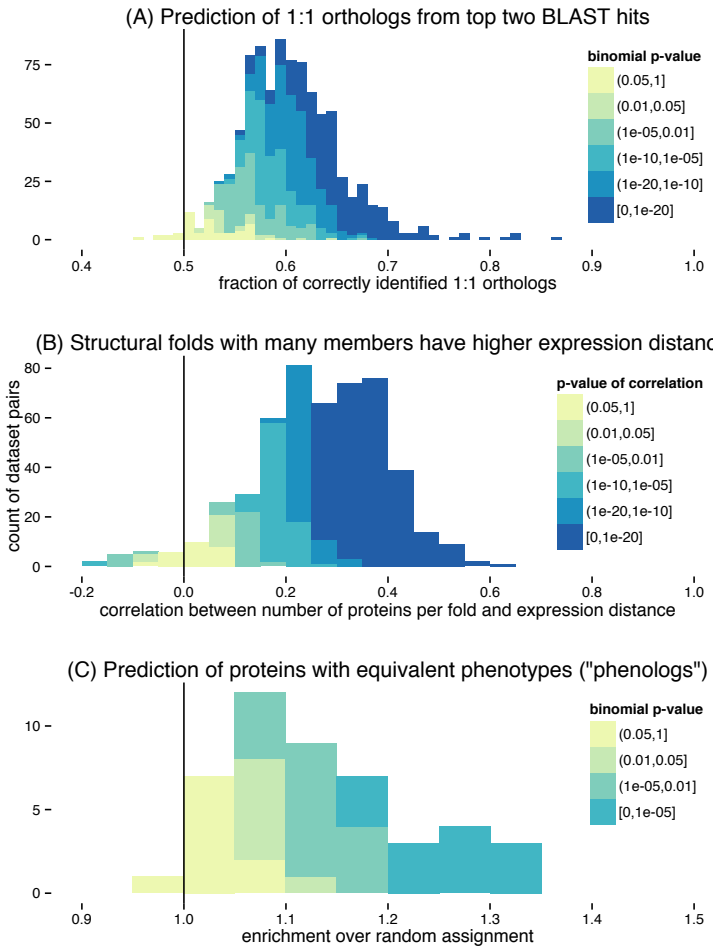


Figure 3: Summary of benchmarking results. For each pair of datasets from different species, the performance in different benchmarks has been computed, along with a p-value. In all three benchmarks, there was a clear shift of the results relative to the random expectation (black line). For details, see Fig. S4, Fig. S5, Fig. S6 and supplementary text. Due to limited structural and functional annotations, there was a lower number of dataset pairs for the two lower panels.

Lastly, on the functional level, we applied the phenolog concept [46] to find equivalent phenotypic annotations across species. We found that expression distances could be used to better predict which member of a protein family has been annotated with a matching phenotype (Fig. 3C and Fig. S6).

Conservation of gene expression programs

At all taxonomic levels, we determined the conservation of the expression patterns of 1:1 orthologs. This data then allowed 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 each pair of datasets, we first computed the median expression distance of 1:1 orthologs (Fig. 4). We then tested for each pair of datasets whether the distribution of expression distances between 1:1 orthologs was shifted towards lower values, i.e. if the median is below 0.5. Using the Wilcoxon signed-rank test and controlling for multiple hypothesis testing with the Benjamini-Hochberg method [47], we found that all dataset pairs had significant shifts to lower expression distances

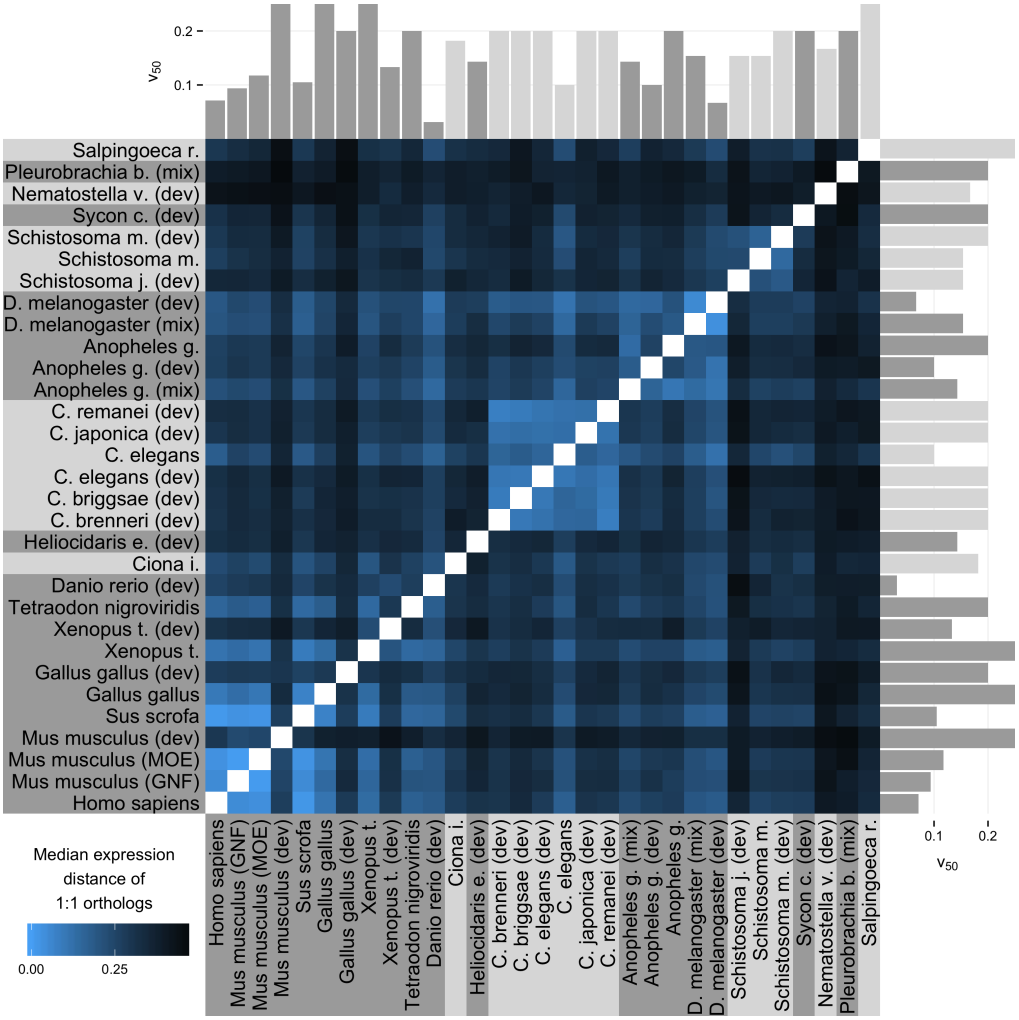


Figure 4: Conservation of expression patterns throughout the metazoans and choanoflagellates. For all dataset pairs, the median expression distance of 1:1 orthologs is shown. Within most clades, this median was very low and approached 0 in some cases. When there was no enrichment of 1:1 orthologs towards lower expression distances, the median was 0.5 (see Fig. 5). See Fig. S7 for a version of this figure without filtering for dataset quality.

($q < 0.05$ and median < 0.5). This analysis revealed both an expected enrichment for closely related species and unexpectedly high enrichments between very distant species, such as between chordates and insects. In general, developmental datasets mapped less well to other species than datasets of adult tissues.

To summarize the data shown in Fig. 4, we computed median expression distances for 1:1 orthologs across all internal nodes of the phylogenetic tree (e.g. for vertebrates, we compared expression patterns between fish and tetrapods). As the median expression distances vary greatly between dataset pairs, we also computed the distribution of expression distances and the number of well-conserved OGs for the best dataset pair across each internal node (Fig. S8 and Fig. S9). Using a Wilcoxon signed rank test, we then tested if the distribution of median expression distances is shifted towards lower values, i.e. if the median of the distribution is lower than 0.5. This was the case for all internal nodes, with the highest p-value (5e-49, median value: 0.39) observed when mapping from the ctenophore *Pleurobrachia* to other animals. This confirmed that our approach could predict expression patterns over large evolutionary distances

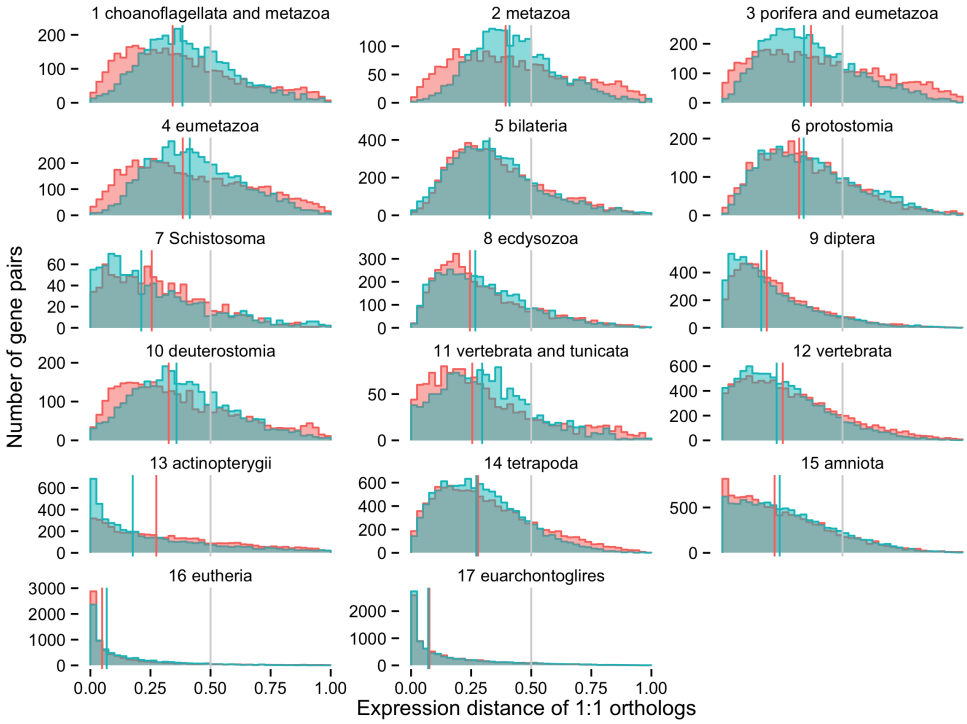


Figure 5: Distribution of median conserved expression. For each clade, the distribution of expression distances of 1:1 orthologs is shown. Red and blue colors denote the direction of the mapping, either from the first subclade to the second or vice versa. For each distribution, the median is shown as a vertical bar. The gray bar corresponds to an expression distance of 0.5, which is the median to be expected by chance. When the mapping is successful, our mapping procedure yields 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. Clades are numbered corresponding to the taxonomic tree in Fig. 6.

(Fig. 5 and 6). For some clades, the available data was very uneven on the two sides of the internal node. For example, at the level of eumetazoa, only one species with few tissues was available for cnidarians, whereas most bilaterian species had many tissues measured. Thus, expression distances were higher when mapping from cnidarians to bilaterians than the other way round. Interestingly, the median divergence between animals and the outgroup choanoflagellates was comparable to the median divergence between major animal clades, e.g. bilateria. Thus, mapping tissue-specific gene expression revealed expression programs conserved for 1 billion years.

Correlations between expression changes of homologs

Next, we addressed the question if conservation of expression programs depends on the functions of genes, i.e. if certain gene functions generally imply a stronger conservation of expression programs than other functions. To this end, we compared the expression distances of gene families in different clades under the assumption that functional constraints would lead to expression conservation in independent clades. If the rate

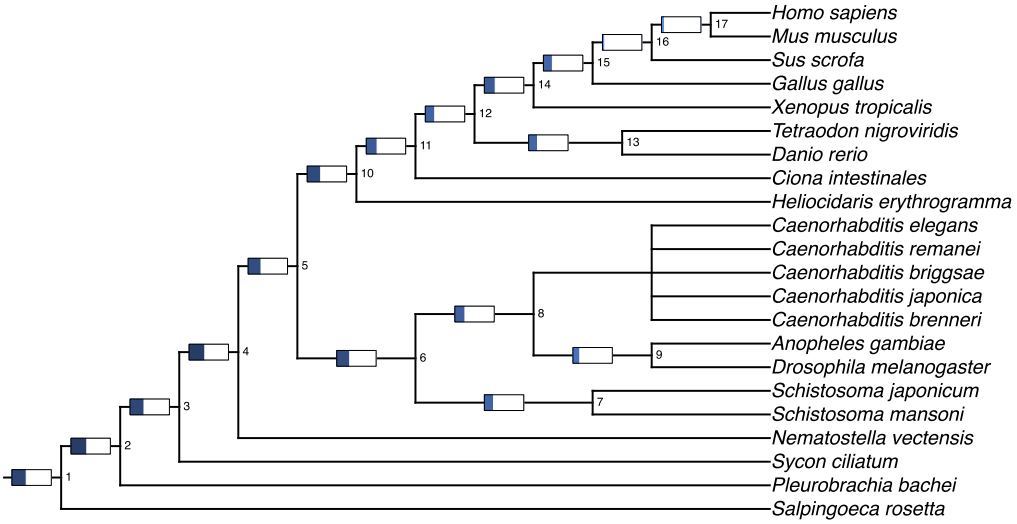
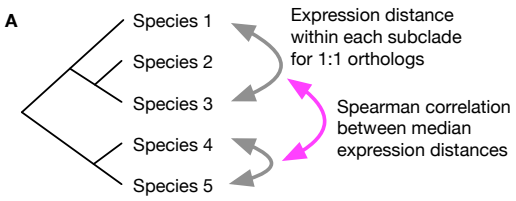


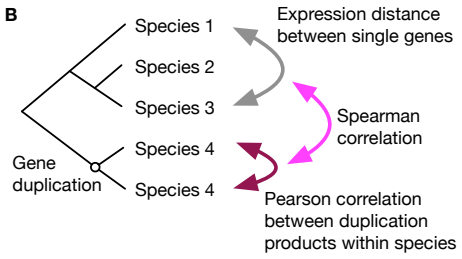
Figure 6: Conservation of expression patterns across clades. The length of the blue bars denotes the median expression distance of 1:1 orthologs across the bifurcation, with values between 0 and 0.5. (0 corresponds to the best possible value, while 0.5 would occur when there is no enrichment of lower expression distances.) The numbers next to the internal nodes refer to the clade numbers in Fig. 5

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 family that has a conserved expression pattern in one clade should also have a conserved expression pattern in another clade. For each internal node 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 four internal nodes with more than one species on both sides, we found significant Spearman correlations (r_s) of median expression similarities for three splits (Fig. 7A): between tetrapods and fishes (r_s=0.18, #12 in Fig. 5), between protostomes and deuterostomes (r_s=0.15, #4), and between nematodes and insects (r_s=0.06, #7).

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. 7B). For example, duplicated genes in fish were more similar (i.e. had 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 2045 pairs of duplicated genes, corresponding to a p-value of 3e-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
5	bilateria	0.15	0	4281
6	protostomia	0.05	0.18	615
8	ecdysozoa	0.06	5E-04	3201
12	vertebrata	0.18	0	3391



#	clade with duplicated gene	clade with single genes	r_s	p-value	n
3	porifera	eumetazoa	-0.11	0.003	750
5	deuterostomia	protostomia	-0.02	0.098	8713
6	ecdysozoa	*Schistosoma*	-0.14	0.003	473
6	*Schistosoma*	ecdysozoa	-0.10	0.033	433
8	diptera	*Caenorhabditis*	-0.07	0.018	1119
8	*Caenorhabditis*	diptera	-0.05	0.051	1800
10	echinodermata	chordates	-0.15	0.024	236
12	actinopterygii	tetrapoda	-0.11	3E-07	2045

Figure 7: Correlations between expression conservation rates. **A** For 1:1 orthologs, the expression distance across internal nodes 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. 5, n – count of 1:1 orthologs [A] or duplicated genes [B])

Evolution of the beta catenin protein family

We seleteced the beta catenin protein family [48] as an example to illustrate the implications of our work. Beta catenin proteins are involved in regulating cell adhesion and gene transcription through the Wnt signaling pathway. Ancestrally, there was a single beta catenin protein, which duplicated independently in the nematode and vertebrate lineages [49]. Hence, *Drosophila*, *Anopheles* and *Schistosoma* only have one beta catenin, armadillo. We found this protein to be similar in its expression patterns with both the vertebrate and nematode beta catenins (Fig. 8), which is indicative of their functional similarities [50]. In vertebrates, two forms exist: beta catenin and plakoglobin. These two proteins have largely overlapping functions [51] and consequently, their observed expression distance was very low. In nematodes, the outcome of the repeated gene duplications [52, 53, 54] is very different: three of the duplication products (*hmp-2*, *wrm-1*, and *sys-1*) are very similar to each other in their expression patterns, which can be explained by their cooperation in in the non-canonical Wnt signaling pathway and the SYS pathway [55]. These three proteins had high expression distances to *bar-1*. In contrast to them, *bar-1* is part of a canonical Wnt signaling pathway [55]. We also observed that *bar-1* had a low expression distance to the vertebrate plakoglobin, while *hmp-2*, *wrm-1*, and *sys-1* had high expression distances. Among the nematode genes, vertebrate beta catenin had the lowest expression distance with *hmp-2*. This example illustrates that our method is able to uncover patterns of functional similarity and divergence both between closely related species and across large evolutionary distances.

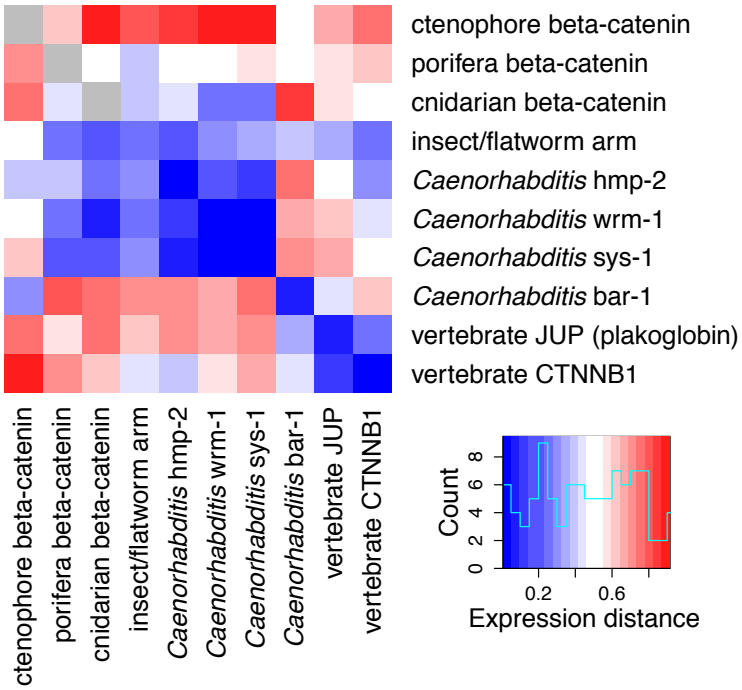


Figure 8: Expression similarity and divergence in the beta catenins. For each group of genes, the median expression distance is shown (see (Fig. S10 for individual expression distances). The unduplicated beta catenins from insects, flatworms and cnidarians a similar to all other protein groups, while functional and expression divergence has occurred independently among nematodes and vertebrates.

Discussion

The presented analysis established and benchmarked a new method, and provided two examples of biological conclusions that can be reached with our method: there is widespread conservation of expression regulation across very large evolutionary distances, and the expression programs of different gene families evolve at distinct rates. Presumably, the latter observation is explained by variable functional constraints between gene families.

In particular, 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 coexpression 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 [56, 57, 58]. Most probably, functional dependencies between genes lead to shared expression patterns over large evolutionary distances. Further research will be needed to reveal which expression similarities between tissues are caused by homology and which are caused by convergent evolution.

Methods

Detection of orthologous proteins

To determine orthology relations between genes, we assembled groups of orthologs (OGs) using the eggNOG pipeline [59] on the genomes of the choanoflagellate *Salpingoeca rosetta* and 67 animals. We then computed gene trees for all OGs using GIGA [60], which we then analyzed to extract 1:1 orthologs and duplication events.

Expression data pre-processing

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 [61], 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 trimmed adapters and mapped reads to annotated transcripts using tophat2 and cufflinks 2.1.1 [62, 63] and used the resulting FPKM counts.

In initial small-scale tests, we tested several normalization methods [11, 64], and settled on a *z*-like normalization of expression vectors **x**, which corresponds to the Euclidean normalization of **x** minus its median value $\tilde{x}$.

$$n_i = \frac{x_i - \tilde{x}}{\sqrt{\sum_i (x_i - \tilde{x})^2}} \tag{2}$$

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 1/1000$^{\text{th}}$ quantile of all genes with non-zero expression. All expression values were incremented by this value.

Mapping of tissue expression patterns

For each pair of datasets, individual linear models were fitted 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 to fit the linear models. When there were multiple probes per gene, all combinations of probes were added to the tissue expression matrix. When there were many tissues in the source species, but few 1:1 orthologs, there was the danger of over-fitting. We therefore allowed only one predictor (i.e. one tissue from the source species) per 15 samples (i.e. 1:1 orthologs) [65]. 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 1260 dataset pairs. The fitted

models were then applied to all genes of the source species, yielding corresponding predicted expression patterns in the target species. Since 1:1 orthologs are used for training, we used predictions from a 10-fold cross-validation for these genes.

Mathematical description of expression mapping

To illustrate our approach, we describe it for a specific pair of datasets, namely mapping expression values from fly to mouse (dataset "MOE"). The same procedure can be applied to all pairs of species. To predict tissue expression patterns of the 51 mouse tissues based on the 26 fly tissues, we fitted 51 separate linear models for each mouse tissue based on 1:1 orthologs.

A given dataset of gene expression values across many tissues of a species can be treated as an expression matrix: Rows correspond to genes and tissues to columns. Hence, it is possible to look at *gene expression vectors* that correspond to a single row, and *tissue expression vectors* that correspond to a single column. Consider the matrices of normalized expression values for fly $\mathbf{F^0}$ and mouse $\mathbf{M^0}$. $\mathbf{F^0}$ contained 13,264 rows corresponding to 12,225 genes and 26 columns. $\mathbf{M^0}$ contained 23,624 rows for 14,307 genes and 51 columns. From the 3120 1:1 orthologs, sub-matrices $\mathbf{F}$ and $\mathbf{M}$ were constructed such that the same row in the two matrices corresponds to a given pair of 1:1 orthologs. When multiple expression measurements per gene were available, the matrices contained all possible combinations of measurements. (E.g. if there were three probes corresponding to one gene, and two for the ortholog, a total of six rows were dedicated to this pair of orthologs.) Due to these combinations, $\mathbf{F}$ and $\mathbf{M}$ each had 4447 rows. A single linear model to fit expression values in mouse tissue t for genes g was thus found by minimizing the errors ϵ:

$$m_{g,t} = \sum_{s=1}^{25} \beta_{s,t} f_{g,s} + \beta_{0,t} + \epsilon_{g,t} \tag{3}$$

Only 25 parameters were needed in the sum, because the normalization produced a matrix with equal row sums. Therefore one variable was redundant. This approach can also be formulated as a matrix multiplication, using $\mathbf{B} = [\beta_{s,t}]$ as parameter matrix, $\mathbf{B_0} = [\beta_{0,t}]$ for the offsets, and $\mathbf{E} = [\epsilon_{g,t}]$ as error matrix:

$$\mathbf{M} = \mathbf{F} \times \mathbf{B} + \mathbf{1} \times \mathbf{B_0} + \mathbf{E} \tag{4}$$

Once $\mathbf{B}$ and $\mathbf{B_0}$ have been determined, they can be applied to the full expression matrix $\mathbf{F^0}$ to create a matrix $\mathbf{V}$ of virtual expression values for fly genes in mouse tissues:

$$\mathbf{V} = \mathbf{F^0} \times \mathbf{B} + \mathbf{1} \times \mathbf{B_0} \tag{5}$$

Cross-species correlations between expression patterns

For each fly gene x with its corresponding gene expression vector $(F^0_{x,1}, F^0_{x,2}, \ldots, F^0_{x,25})$, an expression vector based on mouse tissues had been predicted: $X = (V_{x,1}, V_{x,2}, \ldots, V_{x,51})$.

Thus, for any mouse gene y with expression vector $Y = (M^0_{y,1}, M^0_{y,2}, \ldots, M^0_{y,51})$, the weighted sample Pearson correlation coefficient $r_{x,y}$ could be calculated (Fig. 2A):

$$r_{x,y} = \frac{\sum_{i=1}^{n} w_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} w_i (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} w_i (Y_i - \bar{Y})^2}} \tag{6}$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} w_i X_i \tag{7}$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} w_i Y_i \tag{8}$$

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 [66]. For example, in the mouse tissue dataset, there were many different brain tissues with highly correlated expression patterns. Hence, a gene that was well predicted in one brain tissue was likey to be well-predicted in other brain tissues. When multiple measurements were available for the source or target gene, we reported the maximum of all pairwise correlations.

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 observed a strong correlation between the cross-species expression correlation and the number of coexpressed genes in the target species (Fig. S3). This strong correlation indicated 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 counter this effect, target genes were split into ten bins according to the number of coexpressed genes in the target species (Fig. 9). For each bin, we separately determined the distribution of cross-species expression correlations between all genes. Given this distribution, we determined a conversion function from the cross-species expression correlation to the corresponding quantile.

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 coexpressed 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. For technical reasons, we sampled one million pairs of background genes, such that the lowest possible expression distance is 1e-6.

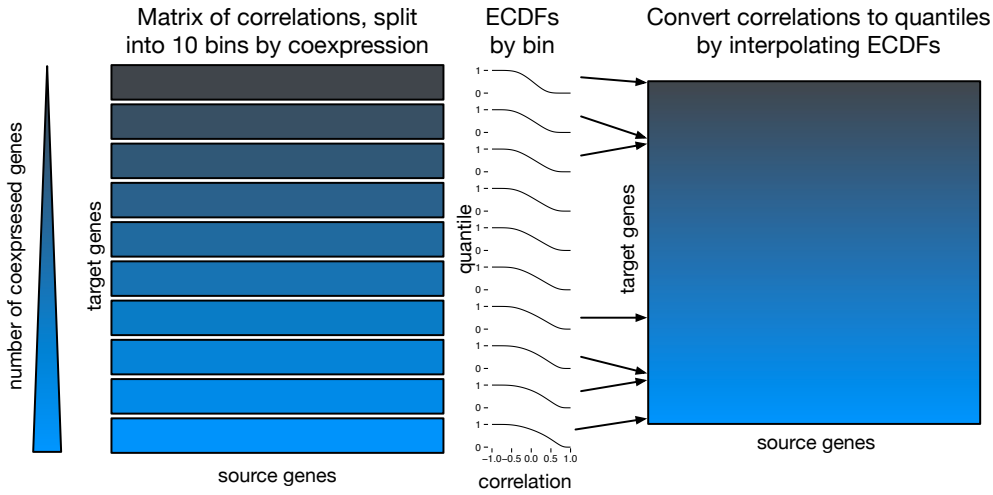


Figure 9: Conversion from correlations to expression distances. For each pair of datasets, all pairwise Pearson correlations between the actual expression patterns and the mapped patterns were computed. This complete matrix of correlations was then split into ten parts, according to the number of coexpressed target genes. For each bin, we separately calculated the ECDF, i.e. the conversion function between correlations and quantile. To retrieve the expression distance of a given pair of source and target genes, the number of coexpressed genes for the target gene was used to select the two closest bins and their ECDFs. To reduce the effect of small differences in the number of coexpressed genes, the expression distance was then computed by interpolating the quantiles returned by the two ECDFs. (Edge cases, where only the first or last bin are appropriate, were treated separately.)

Data access

Protein sequences, normalized datasets, assignments and expression distances of 1:1 orthologs have been deposited at `http://dx.doi.org/10.6084/m9.figshare.1362211`. Separately, all pairwise mappings of expression patterns between datasets are available at `http://dx.doi.org/10.6084/m9.figshare.1362240`.

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). AB receives funding from the Deutsche Forschungsgemeinschaft (DFG CRC 680).

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

[1] Paul D Thomas et al. "On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report." In: *PLoS Comput Biol* 8.2 (2012), e1002386. DOI: 10.1371/journal.pcbi.1002386.

[2] J M Stuart. "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules". In: *Science* 302.5643 (Oct. 2003), pp. 249–255. DOI: 10.1126/science.1087447.

[3] Mark B Gerstein et al. "Comparative analysis of the transcriptome across distant species". In: *Nature* 512.7515 (Aug. 2014), pp. 445–448. DOI: doi:10.1038/nature13424.

[4] Maria D Chikina and Olga G Troyanskaya. "Accurate Quantification of Functional Analogy among Close Homologs". In: *PLoS Comput Biol* 7.2 (Jan. 2011), e1001074. DOI: 10.1371/journal.pcbi.1001074.

[5] Itai Yanai et al. "Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility." In: *Dev Cell* 20.4 (Apr. 2011), pp. 483–496. DOI: 10.1016/j.devcel.2011.03.015.

[6] Michal Levin et al. "Developmental milestones punctuate gene expression in the Caenorhabditis embryo." In: *Dev Cell* 22.5 (May 2012), pp. 1101–1108. DOI: 10.1016/j.devcel.2012.04.004.

[7] David H Silver, Michal Levin, and Itai Yanai. "Identifying functional links between genes by evolutionary transcriptomics." In: *Molecular BioSystems* 8.10 (Oct. 2012), pp. 2585–2592. DOI: 10.1039/c2mb25054c.

[8] Anne Niknejad et al. "vHOG, a multispecies vertebrate ontology of homologous organs groups." In: *Bioinformatics* 28.7 (Apr. 2012), pp. 1017–1020. DOI: 10.1093/bioinformatics/bts048.

[9] Barbara Piasecka et al. "Comparative modular analysis of gene expression in vertebrate organs." In: *BMC Genomics* 13 (2012), p. 124. DOI: 10.1186/1471-2164-13-124.

[10] David Brawand et al. "The evolution of gene expression levels in mammalian organs". In: *Nature* 478.7369 (Oct. 2011), pp. 343–348. DOI: 10.1038/nature10532.

[11] Ben-Yang Liao and Jianzhi Zhang. "Evolutionary conservation of expression profiles between human and mouse orthologous genes." In: *Mol Biol Evol* 23.3 (Mar. 2006), pp. 530–540. DOI: 10.1093/molbev/msj054.

[12] Neil Shubin, Cliff Tabin, and Sean Carroll. "Deep homology and the origins of evolutionary novelty." In: *Nature* 457.7231 (Feb. 2009), pp. 818–823. DOI: 10.1038/nature07891.

[13] Raju Tomer et al. "Profiling by image registration reveals common origin of annelid mushroom bodies and vertebrate pallium". In: *Cell* 142.5 (Sept. 2010), pp. 800–809. DOI: 10.1016/j.cell.2010.07.043.

[14] N J Strausfeld and F Hirth. "Deep Homology of Arthropod Central Complex and Vertebrate Basal Ganglia". In: *Science* (2013).

[15] Henry Chung et al. “Characterization of Drosophila melanogaster cytochrome P450 genes.” In: *Proc Natl Acad Sci USA* 106.14 (Apr. 2009), pp. 5731–5736. DOI: 10.1073/pnas.0812141106.

[16] Margus Lukk et al. “A global map of human gene expression.” In: *Nat Biotechnol* 28.4 (Apr. 2010), pp. 322–324. DOI: 10.1038/nbt0410-322.

[17] Andrew I Su et al. “A gene atlas of the mouse and human protein-encoding transcriptomes.” In: *Proc Natl Acad Sci USA* 101.16 (Apr. 2004), pp. 6062–6067. DOI: 10.1073/pnas.0400782101.

[18] Naoki Irie and Shigeru Kuratani. “Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis.” In: *Nat Commun* 2 (Mar. 2011), p. 248. DOI: 10.1038/ncomms1248.

[19] Tom C Freeman et al. “A gene expression atlas of the domestic pig.” In: *BMC Biology 2010 8:66* 10 (2012), p. 90. DOI: 10.1186/1741-7007-10-90.

[20] Esther T Chan et al. “Conservation of core gene expression in vertebrate tissues.” In: *J Biol* 8.3 (2009), p. 33. DOI: 10.1186/jbiol130.

[21] Tomislav Domazet-Loso and Diethard Tautz. “A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns.” In: *Nature* 468.7325 (Dec. 2010), pp. 815–818. DOI: 10.1038/nature09632.

[22] Eiichi Shoguchi et al. “Direct examination of chromosomal clustering of organ-specific genes in the chordate Ciona intestinalis.” In: *Genesis* 49.8 (Aug. 2011), pp. 662–672. DOI: 10.1002/dvg.20730.

[23] Jennifer A Wygoda et al. “Transcriptomic analysis of the highly derived radial body plan of a sea urchin.” In: *Genome Biol Evol* 6.4 (Apr. 2014), pp. 964–973. DOI: 10.1093/gbe/evu070.

[24] Zhengyuan Wang et al. “Gene expression analysis distinguishes tissue-specific and gender-related functions among adult Ascaris suum tissues.” In: *Mol. Genet. Genomics* 288.5-6 (June 2013), pp. 243–260. DOI: 10.1007/s00438-013-0743-y.

[25] Young-Jun Choi et al. “A deep sequencing approach to comparatively analyze the transcriptome of lifecycle stages of the filarial worm, Brugia malayi.” In: *PLoS Negl Trop Dis* 5.12 (Dec. 2011), e1409. DOI: 10.1371/journal.pntd.0001409.

[26] W Clay Spencer et al. “A spatial and temporal map of C. elegans gene expression.” In: *Genome Res* 21.2 (Feb. 2011), pp. 325–341. DOI: 10.1101/gr.114595.110.

[27] Sumudu N Dissanayake et al. “angaGEDUCI: Anopheles gambiae gene expression database with integrated comparative algorithms for identifying conserved DNA motifs in promoter sequences.” In: *BMC Genomics* 7 (2006), p. 116. DOI: 10.1186/1471-2164-7-116.

[28] Yury Goltsev et al. “Developmental and evolutionary basis for drought tolerance of the Anopheles gambiae embryo.” In: *Dev Biol* 330.2 (June 2009), pp. 462–470. DOI: 10.1016/j.ydbio.2009.02.038.

[29] Dean A Baker et al. “A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, Anopheles gambiae.” In: *BMC Genomics* 12 (2011), p. 296. DOI: 10.1186/1471-2164-12-296.

[30] Qingyou Xia et al. “Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, Bombyx mori.” In: *Genome Biol* 8.8 (2007), R162. DOI: 10.1186/gb-2007-8-8-r162.

[31] Scott W Robinson et al. “FlyAtlas: database of gene expression in the tissues of Drosophila melanogaster.” In: *Nucl. Acids Res.* 41.Database issue (Jan. 2013), pp. D744–50. DOI: 10.1093/nar/gks1141.

[32] Susan E St Pierre et al. “FlyBase 102–advanced approaches to interrogating FlyBase.” In: *Nucl. Acids Res.* 42.Database issue (Jan. 2014), pp. D780–8. DOI: 10.1093/nar/gkt1092.

[33] Geoffrey N Gobert et al. “Developmental gene expression profiles of the human pathogen Schistosoma japonicum.” In: *BMC Genomics* 10 (2009), p. 128. DOI: 10.1186/1471-2164-10-128.

[34] Sujeevi S K Nawaratna et al. “Gene Atlasing of digestive and reproductive tissues in Schistosoma mansoni.” In: *PLoS Negl Trop Dis* 5.4 (2011), e1043. DOI: 10.1371/journal.pntd.0001043.

[35] Jennifer M Fitzpatrick et al. “Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses.” In: *PLoS Negl Trop Dis* 3.11 (2009), e543. DOI: 10.1371/journal.pntd.0000543.

[36] Leon Anavy et al. “BLIND ordering of large-scale transcriptomic developmental time-courses.” In: *Development* 141.5 (Mar. 2014), pp. 1161–1166. DOI: 10.1242/dev.105288.

[37] Sofia A V Fortunato et al. “Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes.” In: *Nature* 514.7524 (Oct. 2014), pp. 620–623. DOI: 10.1038/nature13881.

[38] Rebecca Rae Helm et al. “Characterization of differential transcript abundance through time during Nematostella vectensis development.” In: *BMC Genomics* 14 (2013), p. 266. DOI: 10.1186/1471-2164-14-266.

[39] Georg Hemmrich et al. “Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity.” In: *Mol Biol Evol* 29.11 (Nov. 2012), pp. 3267–3280. DOI: 10.1093/molbev/mss134.

[40] Leonid L Moroz et al. “The ctenophore genome and the evolutionary origins of neural systems.” In: *Nature* 510.7503 (June 2014), pp. 109–114. DOI: 10.1038/nature13400.

[41] Stephen R Fairclough et al. “Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta”. In: *Genome Biol* 14.2 (2013), R15. DOI: 10.1186/gb-2013-14-2-r15.

[42] Yasunobu Okamura et al. “COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems.” In: *Nucl. Acids Res.* 43.Database issue (Jan. 2015), pp. D82–6. DOI: 10.1093/nar/gku1163.

[43] Leo Breiman. “Random Forests”. In: *Machine learning* 45.1 (Oct. 2001), pp. 5–32.

[44] Robert Tibshirani. “Regression shrinkage and selection via the lasso: a retrospective”. In: *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 73.3 (June 2011), pp. 273–282. DOI: 10.1111/j.1467-9868.2011.00771.x.

[45] Jonathan G Lees et al. “Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis.” In: *Nucl. Acids Res.* 42.1 (Jan. 2014), pp. D240–5. DOI: 10.1093/nar/gkt1205.

[46] Kriston L McGary et al. “Systematic discovery of nonobvious human disease models through orthologous phenotypes”. In: *Proc Natl Acad Sci USA* 107.14 (Apr. 2010), pp. 6544–6549. DOI: 10.1073/pnas.0910200107.

[47] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (Jan. 1995), pp. 289–300. DOI: 10.2307/2346101.

[48] M Peifer et al. “The vertebrate adhesive junction proteins beta-catenin and plakoglobin and the Drosophila segment polarity gene armadillo form a multigene family with similar properties.” In: *J Cell Biol* 118.3 (Aug. 1992), pp. 681–691.

[49] Zi-Ming Zhao, Albert B Reynolds, and Eric A Gaucher. “The evolutionary history of the catenin gene family during metazoan evolution.” In: *BMC Evolutionary Biology* 11 (2011), p. 198. DOI: 10.1186/1471-2148-11-198.

[50] P White, H Aberle, and J P Vincent. “Signaling and adhesion activities of mammalian beta-catenin and plakoglobin in Drosophila.” In: *J Cell Biol* 140.1 (Jan. 1998), pp. 183–195.

[51] David Swope, Jifen Li, and Glenn L Radice. “Beyond cell adhesion: the role of armadillo proteins in the heart.” In: *Cell. Signal.* 25.1 (Jan. 2013), pp. 93–100. DOI: 10.1016/j.cellsig.2012.09.025.

[52] Jing Liu et al. “The C. elegans SYS-1 protein is a bona fide beta-catenin.” In: *Dev Cell* 14.5 (May 2008), pp. 751–761. DOI: 10.1016/j.devcel.2008.02.015.

[53] H C Korswagen, M A Herman, and H C Clevers. "Distinct beta-catenins mediate adhesion and signalling functions in C. elegans." In: *Nature* 406.6795 (Aug. 2000), pp. 527–532. DOI: 10.1038/35020099.

[54] L Natarajan, N E Witwer, and D M Eisenmann. "The divergent Caenorhabditis elegans beta-catenin proteins BAR-1, WRM-1 and HMP-2 make distinct protein interactions but retain functional redundancy in vivo." In: *Genetics* 159.1 (Sept. 2001), pp. 159–172.

[55] Ambrose R Kidd et al. "A beta-catenin identified by functional rather than sequence criteria and its role in Wnt/MAPK signaling." In: *Cell* 121.5 (June 2005), pp. 761–772. DOI: 10.1016/j.cell.2005.03.029.

[56] Duncan T Odom et al. "Tissue-specific transcriptional regulation has diverged significantly between human and mouse." In: *Nature Genetics* 39.6 (June 2007), pp. 730–732. DOI: 10.1038/ng2047.

[57] Robert K Bradley et al. "Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species." In: *PLoS Biol* 8.3 (Mar. 2010), e1000343. DOI: 10.1371/journal.pbio.1000343.

[58] Eugene Berezikov. "Evolution of microRNA diversity and regulation in animals." In: *Nature Reviews Genetics* 12.12 (Dec. 2011), pp. 846–860. DOI: 10.1038/nrg3079.

[59] Sean Powell et al. "eggNOG v4.0: nested orthology inference across 3686 organisms." In: *Nucl. Acids Res.* 42.1 (Jan. 2014), pp. D231–9. DOI: 10.1093/nar/gkt1253.

[60] Paul D Thomas. "GIGA: a simple, efficient algorithm for gene tree inference in the genomic age." In: *BMC Bioinformatics* 11 (2010), p. 312. DOI: 10.1186/1471-2105-11-312.

[61] Guy St C Slater and Ewan Birney. "Automated generation of heuristics for biological sequence comparison." In: *BMC Bioinformatics* 6 (2005), p. 31. DOI: 10.1186/1471-2105-6-31.

[62] Daehwan Kim et al. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." In: *Genome Biol* 14.4 (Apr. 2013), R36. DOI: 10.1186/gb-2013-14-4-r36.

[63] Cole Trapnell et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." In: *Nat Biotechnol* 28.5 (May 2010), pp. 511–515. DOI: 10.1038/nbt.1621.

[64] Barbara Piasecka, Marc Robinson-Rechavi, and Sven Bergmann. "Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human." In: *Bioinformatics* 28.14 (July 2012), pp. 1865–1872. DOI: 10.1093/bioinformatics/bts266.

[65] M A Babyak. "What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models". In: *Psychosomatic medicine* 66.3 (2004), pp. 411–421.

[66] Mark B Gerstein, Erik L L Sonnhammer, and C Chothia. "Volume changes in protein evolution". In: *Journal of Molecular Biology* 236.4 (Mar. 1994), pp. 1067–1078.

[67] Andrea Franceschini et al. "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." In: *Nucl. Acids Res.* 41.Database issue (Jan. 2013), pp. D808–15. DOI: 10.1093/nar/gks1094.

[68] Itai Yanai et al. "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification." In: *Bioinformatics* 21.5 (Mar. 2005), pp. 650–659. DOI: 10.1093/bioinformatics/bti042.

[69] Zhi Wang, Ben-Yang Liao, and Jianzhi Zhang. "Genomic patterns of pleiotropy and the evolution of complexity." In: *Proc Natl Acad Sci USA* 107.42 (Oct. 2010), pp. 18034–18039. DOI: 10.1073/pnas.1004666107.

[70] D M Krylov. "Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution". In: *Genome Res* 13.10 (Oct. 2003), pp. 2229–2235. DOI: 10.1101/gr.1589103.

Supplementary Information

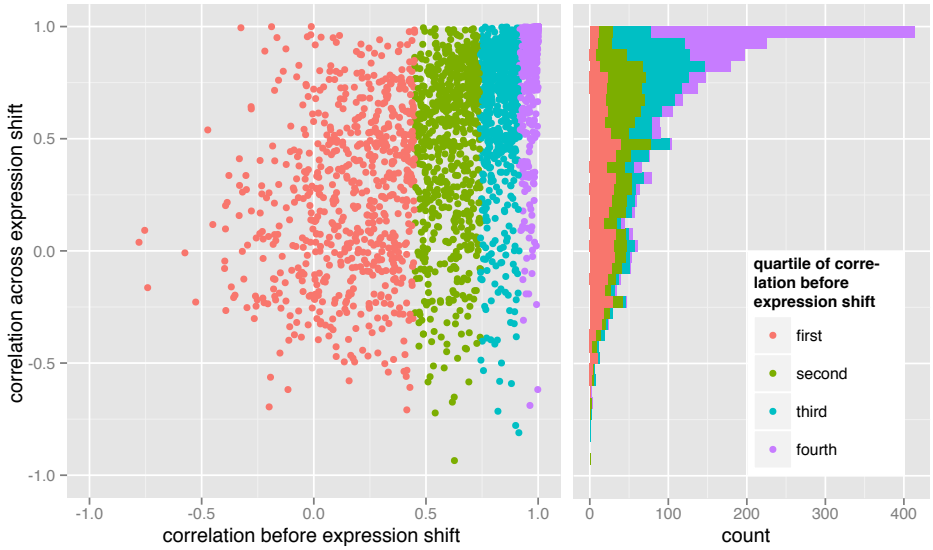


Figure S1: Lineage-specific expression shifts do not change expression patterns. Proteins that have been reported to have lineage-specific changes in expression, e.g. between primates and non-primate species [10] have highly correlated expression patterns even across the expression shift if the expression pattern has been fixed before (fourth quantile, purple).

Lineage-specific expression shifts and relative expression patterns

In the main text, we investigated changing and conserved expression patterns. A previous analysis of expression patterns in six tissues across eight mammals and chicken concluded that while the expression of most genes is under purifying selection, there are also many cases of lineage-specific expression shifts [10]. However, in a re-analysis of this data, we found that these changes occurred mainly on an absolute expression level and that even across the expression shifts, the expression patterns which were reported in the original data set stayed highly correlated (Fig. S1): For the set of genes with significant expression shifts, we found a median correlation of 0.68 between the expression patterns of the species with the expression shift and the species with unchanged expression ("outgroup"). We suspected that for some genes, the expression pattern only becomes fixed after the expression shift. Indeed, when we divided the genes into quartiles according to the median correlation within the set of proteins in outgroup, we found that in the bottom quartile the median correlation across the expression shift is 0.25, while in the top quartile the median correlation is 0.95. In other words, once an expression pattern becomes fixed, it is retained even across lineage-specific expression shifts.

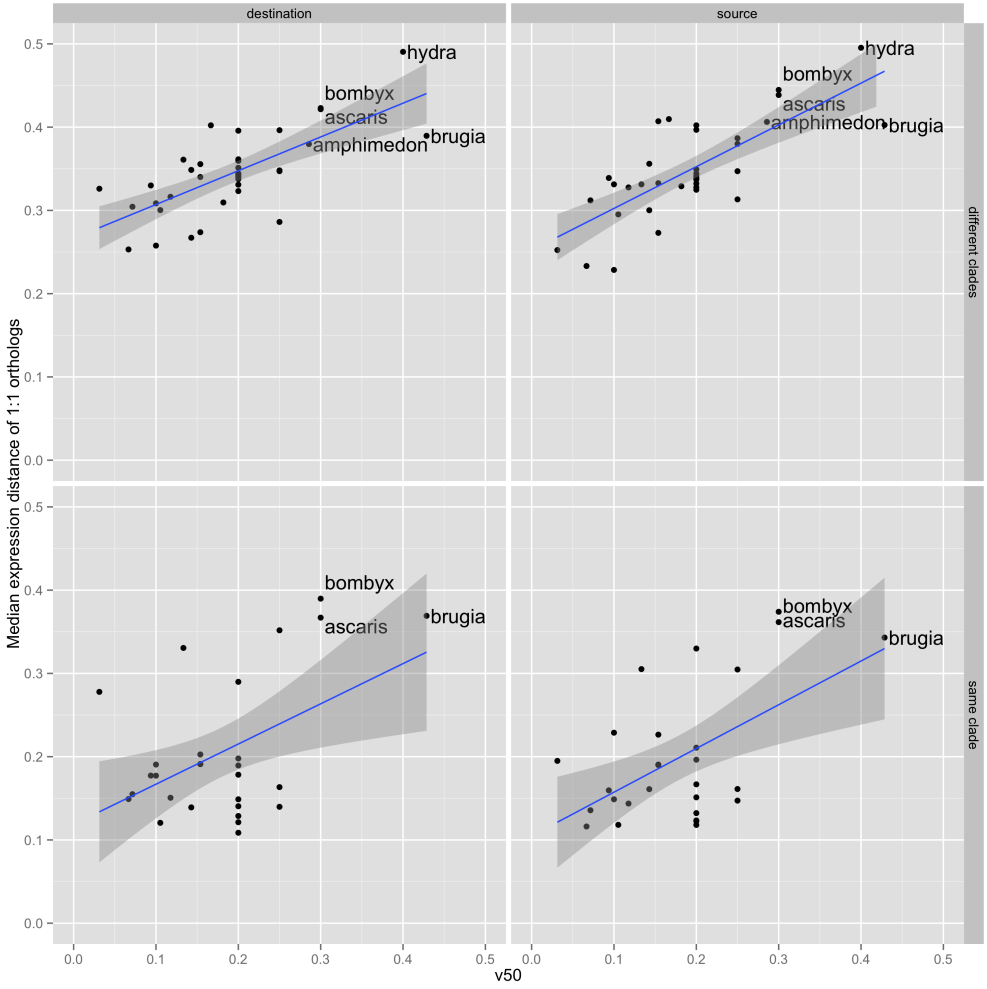


Figure S2: Correlation between dataset quality and mapping quality. For each pair of datasets, the median expression distance of 1:1 orthologs (Fig. 4) can be treated as the mapping quality. For each dataset, we then calculated the median mapping quality over all dataset pairs for which the given dataset is either the source or target dataset. We further distinguish between dataset pairs of the same clade (e.g. two vertebrates) versus pairs from different clades. (Dataset pairs within a clade are only considered for clades of at least three species.) In all combinations, there is a correlation between v_{50} and the mapping quality. We therefore use this measure to exclude five datasets. (Blue line: linear fit; shaded area: 95% confidence interval.)

Properties of proteins that influence the mapping

It is to be expected that properties of the considered genes have an effect on how well the genes' expression patterns can be mapped. For instance, it seems likely that genes that are well-conserved on the protein sequence level should also have conserved expression patterns. Conversely, 1:1 orthologs may appear to have dissimilar expression patterns either due to biological reasons (e.g. functional divergence) or due to technical reasons (e.g. measurement noise, inability to map the expression pattern correctly). We therefore tested eight different properties to which extent they are correlated with expression similarity (using Spearman's rank correlation coefficient). The tested properties were: number of (same-species) proteins with similar expression pattern, degree in the STRING 9.1 protein–protein interaction network (using experimental and text-mining evidence and a confidence score threshold of 0.5) [67], number of isoforms (according to data from Ensembl, WormBase and FlyBase), number of residues, tissue specificity [68], absolute expression level, sequence similarity between the considered proteins, and pleiotropy (for mouse proteins [69]). Almost all properties had a significant influence (Fig. S3). The effects of sequence similarity, total expression level and degree were consistent with previous findings that these factors are inversely correlated with gene loss [70].

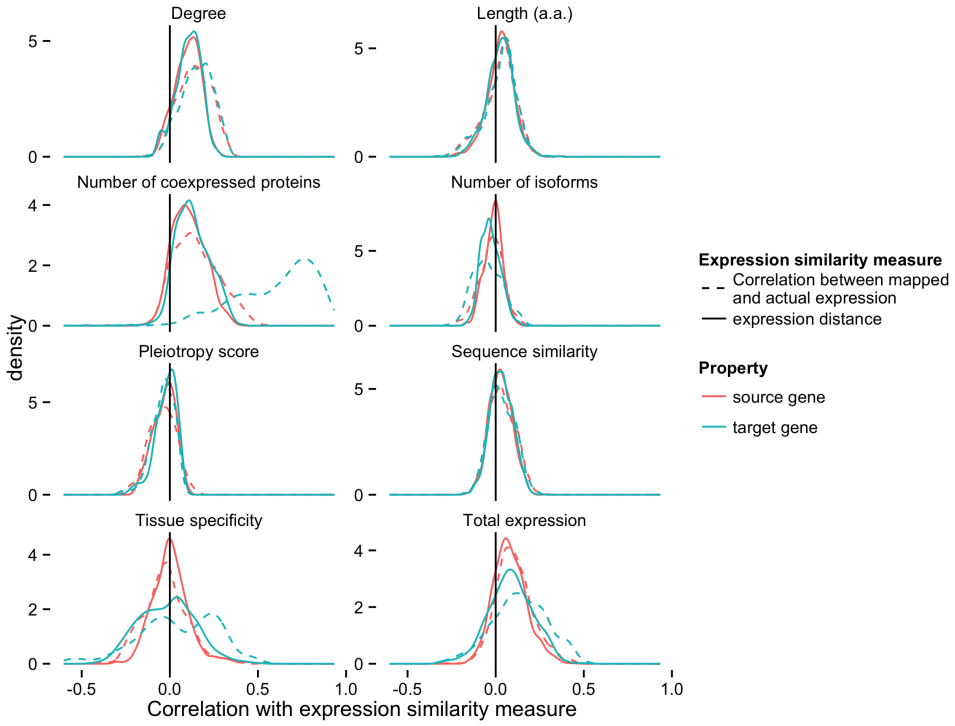


Figure S3: Gene properties correlated with expression similarity. Different properties of the source (red) or target gene (blue) influenced the distribution of expression distances. To measure this influence, we computed the correlation between the gene properties and the expression similarity of 1:1 orthologs. When the correlation between mapped and actual expression patterns was used as the expression similarity (dashed lines), there was a very high correlation with the number of coexpressed target genes. That is, when a target gene had many genes with similar expression patterns, then the expression correlation with its 1:1 ortholog tended to be high. Correcting for this (solid lines), 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.

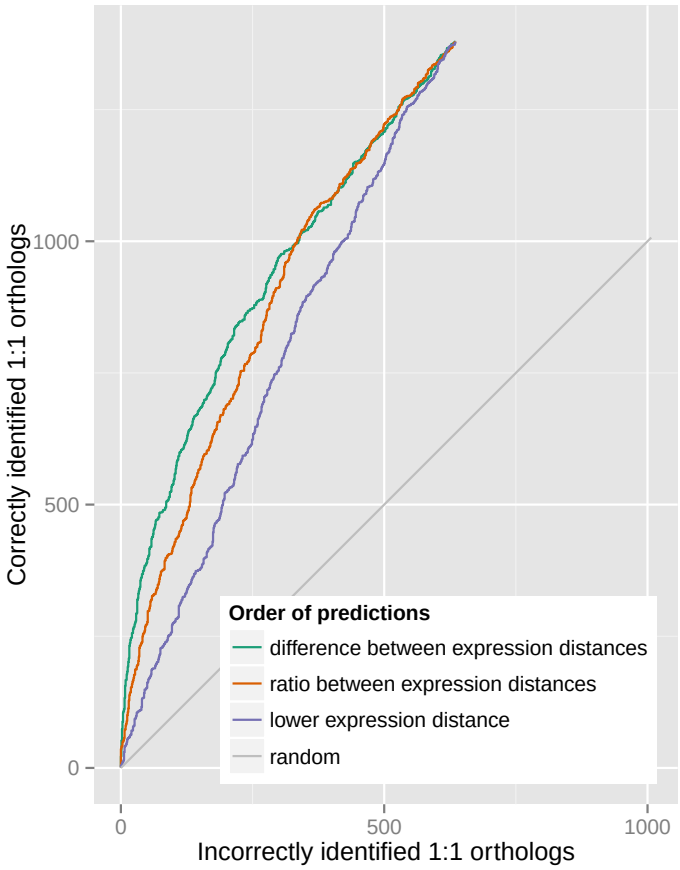


Figure S4: Prediction of 1:1 orthologs from best hits. For each 1:1 ortholog between fly and *C. elegans*, the source genes expression pattern was mapped to *C. elegans* and compared to the top two BLAST hits. If the mapped expression pattern was more similar to the actual ortholog, it was counted as correctly identified. Thus, a perfect prediction method would be a vertical line. Ordering the predictions by the difference between the two expression distances was the most successful strategy.

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 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*, 67.5% of 2014 one-to-one orthologs could be correctly identified (p-value of Binomial test: 5e-57; median across all dataset pairs: 60%). Predictions could be ordered in different ways according to the expression distances between the two pairs of genes: 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. S4).

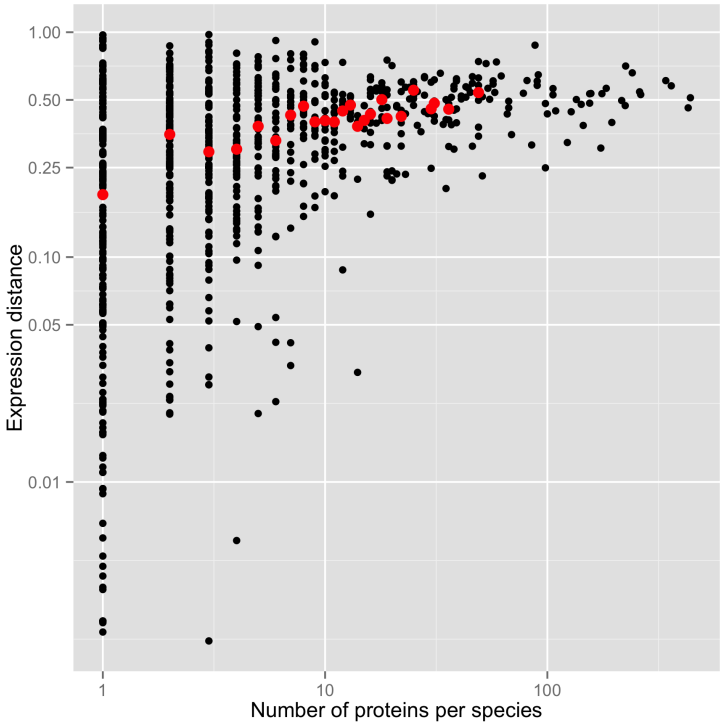


Figure S5: Correlation between expression distance and shared protein folds. Proteins that belong to structural families with few members are more similar in their expression patterns than proteins from large families. Red dots denote the median when at least five superfamilies have the same number of proteins per species. Here, the mapping from fly to *C. elegans* is shown.

Benchmark 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 [45], 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. S5): 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.

Benchmark 3: Phenologs

Finally, we used functional information to evaluate our method. We applied the phenolog concept [46] 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

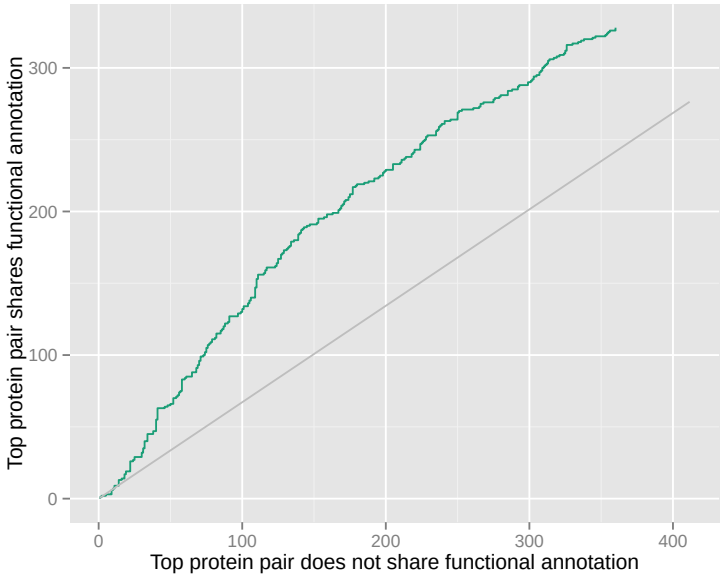


Figure S6: Benchmarking based on phenologs. For each OG between fly and *C. elegans*, we selected the phenolog with the lowest p-value. We then tested whether the gene pair with the lowest expression distance shared the functional annotation predicted by the phenolog. Predictions were ordered by the difference between the lowest and second lowest expression distance. Randomly choosing gene pairs from the OGs results in the grey line.

looking for significant overlap between OGs that correspond to the functional annotations. This leads to phenologs, i.e. pairs of functional annotations with a certain p-value that represents their cross-species similarity. 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 with the lowest p-value. For all cross-species 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 Wilcoxon rank sum test p-value was 6e-19 (median p-value across all dataset pairs: 2e-9). 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*, 47.7% of all top predictions had matching functional annotations, compared to an expected fraction of 40.2% (Fig. S6). This corresponds to a relative increase of 19% over the expected fraction. Between human and mouse (GNF dataset), this increase is 34%. The median increase among all dataset pairs is 10%.

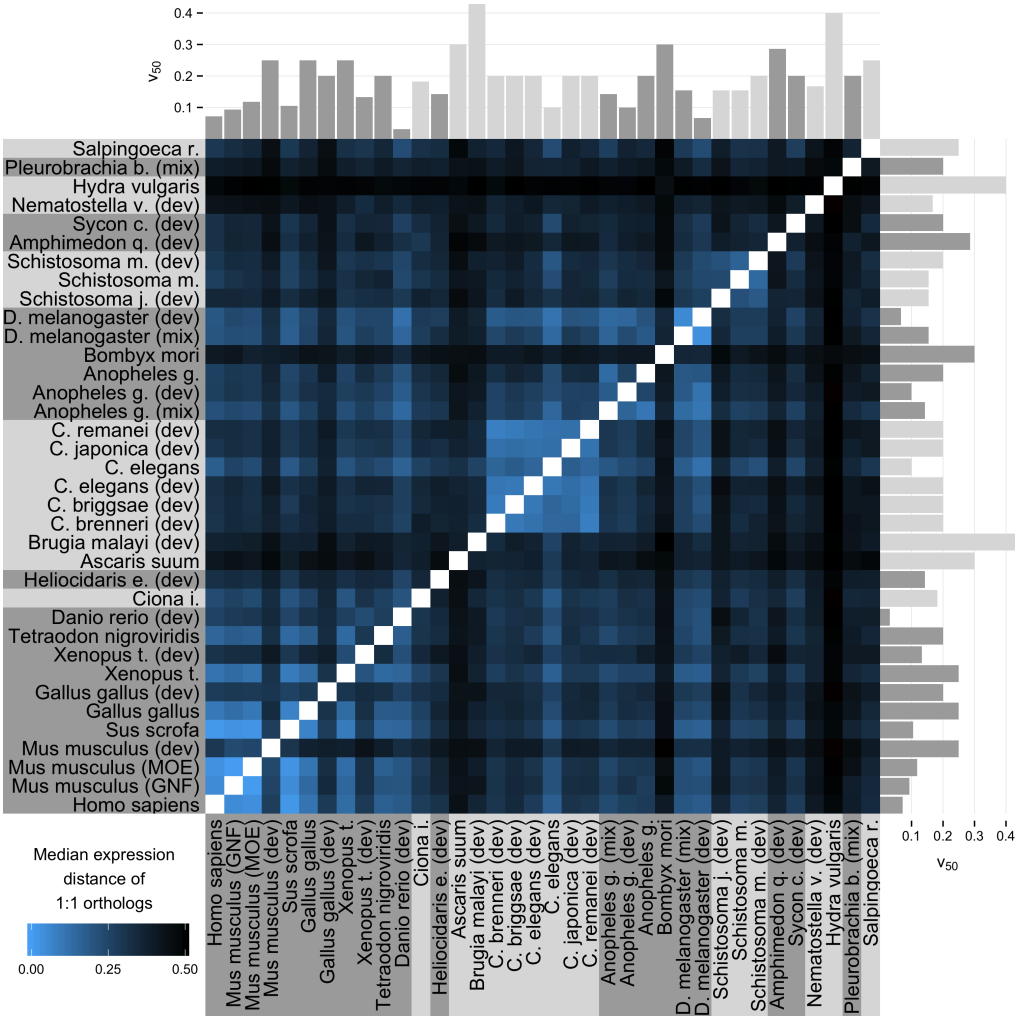


Figure S7: Conservation of expression patterns throughout the metazoans and choanoflagellates. In this version of Fig. 4, all datasets are shown without any filters for dataset quality. Therefore, five additional datasets are shown: *Hydra vulgaris*, *Amphimedon queenslandica*, *Bombyx mori*, *Brugia malayi* and *Ascaris suum*. For these species, the quality of the datasets prevented better mapping performance.

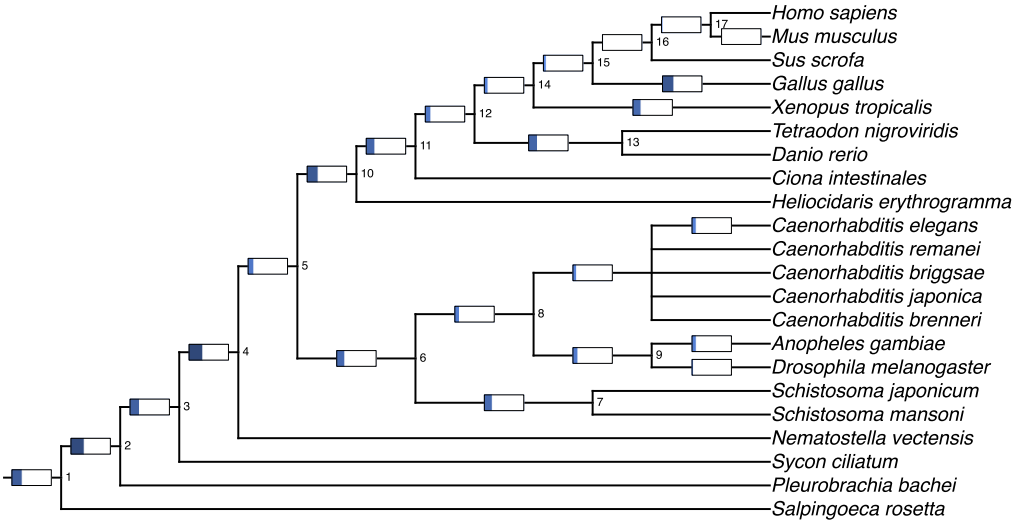


Figure S8: Most conserved expression across animal clades. As in Fig. 5, the median expression distances of 1:1 orthologs are shown. However, instead of the median across all datasets, the median expression distance of the best dataset pair is used. Charts at species branches show how well expression patterns could be mapped between different datasets of the same species.

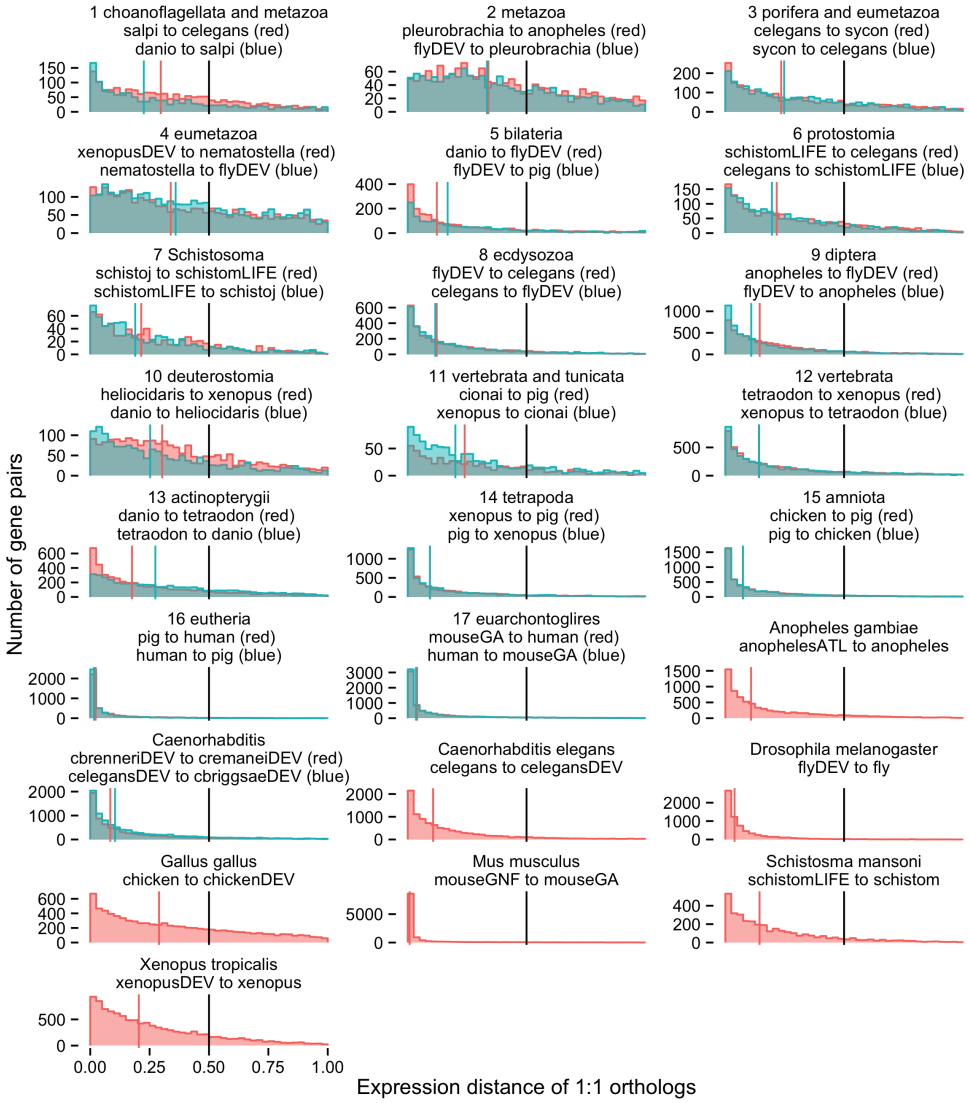


Figure S9: Distribution of conserved expression for best dataset pairs. For each internal node, the distribution of expression distances is shown for the datasets given. These datasets show the highest degree of conservation for the respective internal node. See Table S1 for a description of the abbreviated dataset names.

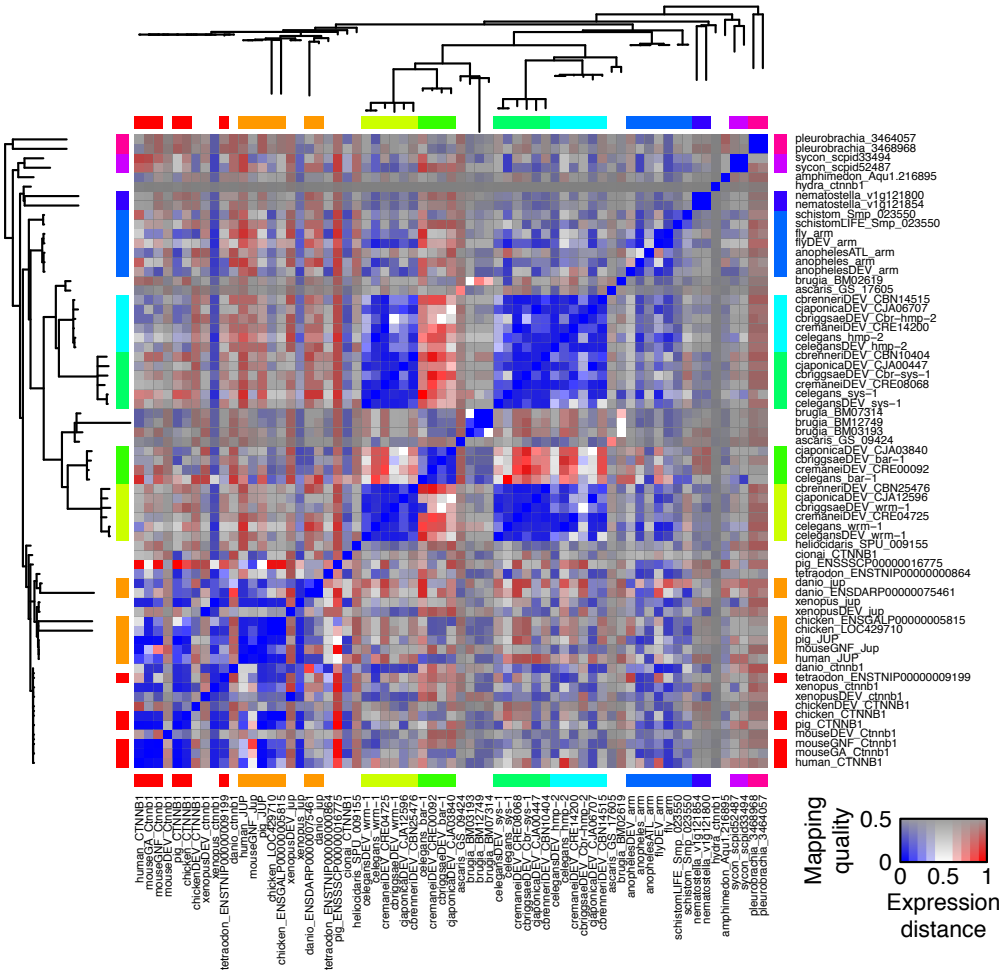


Figure S10: Expression distances of beta catenins. The heatmap shows the expression distances of beta catenins throughout the animal kingdom as color scale. Individual cells are shaded in grey according to a measure of mapping quality of the pair of datasets, namely the median expression distance of 1:1 orthologs (Fig. 4). Genes were clustered according to the gene tree built from a multiple sequence alignment. The colored band along the edge of the heatmap denote the ten groups of proteins which have been analyzed in Fig. 8. Genes that belonged to low-quality datasets or showed divergent expression patterns to their closest neighbors were discarded.