# Pervasive evolution of tissue-specificity of ancestral genes differentially shaped vertebrates and insects

Federica Mantica [1], Luis P. Iñiguez [1], Yamile Marquez [1], Jon Permanyer [1], Antonio Torres-Mendez [1], Josefa Cruz [2], Xavi Franch-Marro [2], Frank Tulenko [3], Demian Burguera [1], Stephanie Bertrand [4], Toby Doyle [5], Marcela Nouzova [6], Peter Currie [3,7], Fernando G. Noriega [8,9], Hector Escriva [4], Maria Ina Arnone [10], Caroline B Albertin [11], Karl R Wotton [5], Isabel Almudi [12], David Martin [2], Manuel Irimia [1,13,14,15]

1 - Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain.
2 - Institute of Evolutionary Biology (IBE, CSIC-Universitat Pompeu Fabra), Passeig de la Barceloneta 37, 08003 Barcelona, Catalonia, Spain.
3 - Australian Regenerative Medicine Institute, Level 2, 15 Innovation Walk, Monash University, Wellington Road, Clayton, Victoria 3800, Australia.
4 - Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins, BIOM, F-66650, Banyuls-sur-Mer, France.
5 - Centre for Ecology and Conservation, University of Exeter, Cornwall Campus, Penryn, UK.
6 - Institute of Parasitology, CAS, České Budějovice, Czech Republic.
7 - EMBL Australia, Victorian Node, Level 1, 15 Innovation Walk, Monash University, Wellington Road, Clayton, Victoria 3800, Australia.
8 - Biology and BSI, Florida International University, Miami, USA.
9 - Department of Parasitology, University of South Bohemia, České Budějovice, Czech Republic.
10 - Stazione Zoologica Anton Dohrn, villa comunale 80121, Napoli, Italy.
11 - Eugene Bell Center for Regenerative Biology and Tissue Engineering, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA.
12 - Department of Genetics, Microbiology and Statistics and IRBio, Universitat de Barcelona, Av. Diagonal 643, 08028 Barcelona, Spain.
13 - Universitat Pompeu Fabra, Barcelona, Spain.
14 - ICREA, Barcelona, Spain.
15 - Corresponding author.

**Manuel Irimia**
Centre for Genomic Regulation
Dr. Aiguader, 88, 08003 Barcelona, Spain
e-mail: mirimia@gmail.com
Phone: +34933160212 Fax: +34933160099

**Abstract**

Regulation of gene expression is arguably the main mechanism contributing to tissue phenotypic diversity within and between species. Here, we assembled a vast RNA-seq dataset covering twenty bilaterian species and eight tissues, selecting a specular phylogeny that allowed both the combined and parallel investigation of gene expression evolution between vertebrates and insects. We specifically focused on widely conserved ancestral genes, identifying strong cores of pan-bilaterian tissue-differential genes but even larger groups that diverged to define vertebrate and insect tissues. Consistently, systematic inferences of tissue-specificity gains and losses show that nearly half of all ancestral genes have been recruited into tissue-specific transcriptomes. This occurred during both ancient and, especially, recent bilaterian evolution, with numerous gains being associated with the emergence of unique phenotypes. Such pervasive evolution of tissue-specificity was boosted by gene duplication coupled with specialization, including an unappreciated prolonged effect of whole genome duplications during recent vertebrate evolution.

## Introduction

Fossil records reconstruct the image of the last common ancestor (LCA) of all bilaterian animals as a small, marine creature crawling on the seafloor approximately 539-571 million years ago (MYA) [1]. Despite its apparent simplicity, this ancestral organism introduced remarkable biological novelties, including a revolutionary body plan, defined by two perpendicular symmetry axes, and a third embryonic germ layer [2]. In addition, it already possessed at least a draft of the main tissue types homologous across extant bilaterian species, including a nervous system, skeletal muscle, female and male gonads, through guts and an excretory system [3]. How did this ancient organism specify such a great variety of biological structures? Since all its cells shared the same genome, gene expression regulation was likely key for the generation of unique transcriptomes across these ancestral tissue types, and consequently for the emergence of their distinctive biological functions.

The bilaterian ancestor gave rise to the vast majority of extant animals, in which the original body plan and tissues have been greatly diversified and modified. Determinants of animal evolution include changes in gene complements (i.e., gene gains/losses and gene duplications) and divergence of protein-coding sequences [4–6]. In addition, given the central role of tissue-specific transcriptomes in defining distinct intra-species phenotypes, modifications of gene expression patterns were likely to be employed during bilaterian evolution to generate inter-species phenotypic diversity. In fact, it is now appreciated that expression divergence of conserved genes has a major impact on phenotypic diversity, and it is arguably the main evolutionary mechanism among multicellular organisms [7,8]. Examples of such occurrences underlie important phenotypic novelties in key bilaterian lineages: the vertebrate endocrine pancreas emerged following the co-option of ancestrally neural-specific genes [9], and it has been suggested that insect wings evolved thanks to ectopic expression of ancient genes originally involved in gill specification and proximal leg segments [10–12]. Still, even if some remarkable cases linked to biological novelties have been identified, the specific role that the evolution of expression of ancestral genes played in shaping homologous, yet often highly divergent, tissues between distant bilaterian lineages has never been thoroughly assessed.

In the same way as fossil records provide glimpses of ancestral morphologies, comparative transcriptomic data can offer invaluable insights into ancestral molecular states and their subsequent history. Here, we studied the evolution of tissue-specific transcriptomes based on a vast RNA-seq dataset covering eight tissue types from twenty different bilaterian species,

including novel data for fifteen of them. We focused on vertebrates and insects, selecting a symmetric phylogeny for both groups that allowed us to perform sound ancestral inferences as well as to uncover parallel, convergent and divergent evolutionary trajectories in these early branching bilaterian lineages. We first characterized global gene expression patterns and reconstructed ancestral bilaterian tissue-differential modules that are still widely conserved across extant species. Second, we investigated conserved genes that specifically evolved divergent expression profiles between individual vertebrate and insect tissues. Lastly, we systematically inferred and analyzed gains and losses of tissue-specific expression throughout our bilaterian phylogeny. Overall, our work sheds light into the highly plastic nature of deeply conserved genes in terms of tissue-specific expression patterns, which we find is tightly linked to gene duplication, specialization and the emergence of unique tissue-related phenotypes.

## Results

### Characterization of global patterns of gene expression across bilaterian tissues

In order to reliably investigate the evolution of tissue transcriptomes in two key bilaterian lineages, we selected twenty representative species (eight gnathostome vertebrates, eight insects and two pairs of relative outgroups) evenly divided into two monophyletic branches with specular phylogenetic structures **(Fig. 1a)**. After correcting for broken/chimeric genes and enriching the annotations of the majority of the species (see Supplementary Methods, **Extended Data Fig. 1a-c** and **Supplementary Table 1**), we derived gene orthologous relationships among all of them and isolated 7,178 bilaterian-conserved gene orthogroups, which were unambiguously present in the bilaterian LCA and overall widely conserved across vertebrates and insects (see Supplementary Methods; **Extended Data Fig. 1d,e** and **Supplementary Table 2)**. We then assembled a comprehensive bulk RNA-seq dataset covering up to eight tissues in all species (**Fig. 1a** and **Supplementary Table 3**). Compared to previous studies of transcriptome evolution, mostly focused on mammals, this dataset extends the phylogenetic coverage beyond vertebrates and provides the first comparative framework for insects. Moreover, while some of the included tissue types (neural, testis, ovaries, muscle and excretory system) had been considered in previous studies of gene expression evolution among bony vertebrate species [13–16], others (digestive tract, epidermis and adipose) have never been analyzed in such a context. In total, we generated 95 RNA-seq samples across 15 species, which we combined with publicly available data into a final dataset of 346 RNA-seq meta-samples (see Methods for meta-samples definition), including up to three meta-samples for each tissue and species **(Fig. 1a** and **Extended Data Fig. 2)**.

Previous milestone studies had highlighted the greater transcriptomic similarity between homologous tissues in different vertebrate species (inter-species) rather than between non-homologous tissues in the same species (intra-species) [13,16,17]. In our dataset, the first two components of a principal component analysis (PCA) showed a clear distinction between meta-samples from vertebrates and non-vertebrates (including all outgroups), independently of their tissue identity **(Fig. 1b)**. Nevertheless, subsequent components significantly separated groups of meta-samples based on their tissue of origin, starting from neural and testis **(Extended Data Fig. 3a-d)**. Moreover, after z-scoring gene expression within species, in order to exclusively evaluate inter-tissue variability, we obtained clusters largely corresponding to tissue groups **(Extended data Fig. 3e)**. All together, these results suggest at least partial conservation of ancestral bilaterian tissue-differential expression modules.

**Reconstruction of ancestral bilaterian tissue-differential expression modules**

In order to characterize the ancestral tissue-differential expression modules that are still widely conserved across extant bilaterians, we implemented a strategy based on a sparse least square discriminant analysis (sPLS-DA) [18], which allowed us to isolate the genes with the most distinctive expression profiles in each tissue compared to the rest across all species (see Methods, **Fig. 2a** and **Supplementary Fig. 1a-f**). This produced eight sets of ancestral genes with conserved tissue-differential expression, which overall comprised 506 (~7%) of all bilaterian-conserved orthogroups **(Fig. 2b** and **Supplementary Table 4)**.

We next investigated in detail the gene groups corresponding to the neural and testis modules, whose expression profiles are shown in **Fig. 2c,d** (see **Supplementary Fig. 1g-l** for other modules). The neural module presented strong over-representation of gene ontology (GO) categories related to synaptic transmission, neuronal morphology and other associated terms **(Fig. 2e)**, reflecting the high conservation of the specialized neuronal gene complement across eumetazoan nervous systems [19,20]. The testis module showed significant enrichments for cilium and cytoskeleton-related functions **(Fig. 2f),** likely determined by the axoneme, a highly conserved microtubule-based structure located at the core of most bilaterian spermatozoa flagella and indispensable for their mobility [21]. The putatively conserved role that these ancestral genes play in the respective tissue is supported by their significantly greater association with validated neural- or testis-related phenotypes either in mammals or fly,

compared to all bilaterian-conserved genes (**Fig. 2g,h** and **Supplementary Table 5**; p-value < 1e-05 for both tissues, Fisher's exact tests).

In addition to the neural and testis modules, all other sets exhibited GO enrichments coherent with the deep-rooted functions of each tissue (**Fig. 2i** and **Supplementary Table 6**). For instance, genes in the ancestral ovary module comprised several key meiotic genes and were enriched in cell cycle and DNA-replication/repair functions (**Supplementary Tables 4** and **6**). Some examples included *CCNB2*, a cyclin necessary for timely oocyte maturation and correct metaphase-to-anaphase transition in mice [22,23], *MOS*, a serine-threonine kinase which mediates metaphase II arrest during meiosis and whose deletion causes human female infertility [24], and *CPEB*, a protein involved in regulation of translation prior fertilization [25]. As another example, the most significant GO categories for the excretory system module, mainly ion transport and amino acid metabolism, reflected the basic shared functions of ultrafiltration-based excretory systems [26]. Moreover, even in those tissues in which the homology status of the specific cellular components is more ambiguous/complex (e.g., epidermis, digestive system, adipose), we still obtained a few significant enrichments linked to core molecular programs underlying fundamental functions of each tissue type **(Fig. 2i)**.

Finally, we investigated which transcription factors (TFs) might have regulated these ancestral modules since the bilaterian LCA. Thus, we specifically tested if the TFs included in each module presented a significant over-representation of predicted binding sites in the regulatory regions of all the other genes within the same ancestral set (see Methods). We obtained significant results for multiple TFs, comprising several known master regulators of the respective tissues such as PAX4/6 [27] or FEZF1 [28] in neural, MEF2A-D [29] in muscle and GRHL1/2 [30] in epidermis **(Fig. 2j)**.

In summary, we reconstructed eight confident modules of ancestral genes with highly conserved tissue-differential expression profiles, offering a high-resolution snapshot of the tissue-specific transcriptomes in the bilaterian LCA.

**Ancestral genes divergently shape tissue transcriptomes in vertebrates and insects**

Since only a relatively small proportion of bilaterian orthogroups showed globally conserved tissue-differential expression patterns, we next aimed at identifying ancestral genes subjected to different evolutionary forces and/or co-opted for distinct tissue-related tasks specifically

between vertebrates and insects, which we could readily compare thanks to the symmetric structure of our phylogeny.

First, we evaluated the expression conservation within tissues by comparing the relative rates of transcriptomic evolution within both vertebrates and insects. Specifically, as previously performed [13], we built expression-based trees for each tissue in each clade, where higher total branch length corresponds to greater transcriptomic divergence. Interestingly, testis resulted to be the fastest evolving tissue both in vertebrates and insects, extending previous findings in amniotes [13] **(Fig. 3a,b)**. However, while neural was one of the slowest evolving tissues in vertebrates **(Fig. 3a)**, in line with [13], it was unexpectedly the second fastest evolving tissue in insects **(Fig. 3b)**.

Second, we computed the average expression correlation across tissues (Spearman's rho) for bilaterian-conserved orthogroups within either vertebrates or insects, and checked how this measure relates to other relevant gene features such as coding sequence evolution and duplication status. In line with previous studies, and consistent with their shorter generation time [31], protein sequence similarities of ancestral genes among insect species were significantly lower than those among vertebrates (**Fig. 3c;** Wilcoxon Rank-Sum test < 2e-16). However, their expression correlation across tissues presented similar distributions between the two clades **(Fig. 3d)**, suggesting comparable rates of tissue transcriptomic evolution. Notably, these patterns emerged also when restricting the analysis to the 1,312 single copy orthogroups **(Extended Data Fig. 4a,b)**.

Next, we aimed to isolate ancestral genes that differentially evolved in vertebrates and insects either in terms of expression profiles or sequence. For this purpose, we compared the expression correlation and the protein sequence similarity within the same orthogroups between the two clades**.** Remarkably, differences in each of the features were very lowly correlated with one another (linear model beta: 0.039; **Fig. 3e**), implying no association between sequence and expression conservation, in line with a more restricted comparison between zebrafish and frog [32] but in contrast to observations within mammals [33]. Thus, to investigate the nature of genes whose evolution was largely shaped by changes in either one or the other feature, we defined four groups with the top 500 ancestral genes with more conserved sequence (SeqCons; **Fig. 3f**) or correlated expression profiles (ExprCons; **Fig. 3g**) in one lineage compared to the other **(Supplementary Table 7)**. These represent bilaterian-conserved

genes which, through distinct evolutionary modalities (sequence or expression evolution), might have been adapted to differential uses in vertebrates and insects. Consistent with a putatively higher relevance of vertebrate SeqCons and ExprCons genes only in vertebrates, these genes exhibited a significantly greater proportion of experimentally validated phenotypes exclusively in mammals compared to flies (**Fig. 3h;** p-values = 5e-04 and 2e-10, respectively, two-sided Fisher's Exact tests), while the opposite was true for the sets of insect SeqCons and ExprCons genes (**Fig. 3h;** p-value = 0.011 and 2e-07, respectively, two-sided Fisher's Exact tests) (see **Supplementary Table 8** for complete phenotypic classification). Moreover, vertebrate and insect SeqCons genes presented opposite duplication patterns among vertebrate species, which underscore the known impact that gene duplication has on sequence divergence: whereas genes with higher sequence conservation in vertebrates compared to insects showed significantly less duplications among vertebrates relative to the background, the converse was true for genes with lower sequence conservation **(Fig. 3h)**. Finally, we asked if these orthogroups with divergent features (i.e., sequence similarity or expression correlation) between vertebrates and insects comprised functionally related genes. GO enrichment analyses revealed different functional categories among each set **(Fig. 3i)**. Strikingly, vertebrate ExprCons genes were strongly enriched for developmental terms, whereas insect ExprCons genes top significant categories were related to rRNA preprocessing/nucleolus and ATPase-coupled ion transmembrane transport. These results were not affected by the chosen GO annotation (human) or set size, as comparable results were obtained with the fly GO annotation and with sets composed of the top 250 or 750 genes (see Methods and **Supplementary Table 9**).

Next, we separately applied the sPLS-DA methodology for each tissue to identify those ancestral genes with tissue-differential expression exclusively in vertebrates or in insects **(Extended Data Fig. 4c, Supplementary Fig. 2** and **Supplementary table 10)**. For many tissues, the identified vertebrate-specific and insect-specific genes were involved in distinct biological processes, possibly highlighting co-option of these genes for novel functions in at least one of the clades and/or loss of the ancestral tissue-specific function in the other **(Supplementary Table 11)**. As an example, vertebrate-specific ovary-differential genes are significantly enriched in several categories related to development of gonads, reproductive systems and primary sexual characteristics, while the insect-specific ones are mainly enriched for functions associated with cell cycle regulation and cytoskeleton organization.

**Pervasive evolution of tissue-specificity of ancestral genes impacts both ancient and recent bilaterian history**

We next unbiasedly defined genes with tissue-specific expression profiles throughout our phylogeny using the Tau metric [34]. As expected, the overall proportion of tissue-specific genes (Tau ≥ 0.75) in each species was lower for bilaterian-conserved genes compared to all genes **(Fig. 4a** and **Supplementary Fig. 3)**. We assigned each bilaterian-conserved, tissue-specific gene to the tissue(s) with the highest relative expression (see Methods, **Fig. 4b** and **Extended Data Figs. 5a and 6**), providing a comprehensive characterization of their tissue-specificity across all species and tissues **(Fig. 4c).** Neural- and testis-specific genes were the most abundant throughout our phylogeny, followed by genes with restricted expression in two different tissues **(Fig. 4d)**. Overall, the number of tissue-specific genes was significantly higher among vertebrate species **(Fig. 4e;** p-value = 2e-04, Wilcoxon rank sum test), likely as a consequence of the two whole genome duplications (WGDs) at the base of vertebrates (see next section).

We then set out to investigate the conservation of the identified tissue-specific profiles. Remarkably, we found that these profiles were overall poorly conserved. For instance, for the orthogroups that are tissue-specific in mouse, only a median of 6 out of the other 19 species had orthologs with Tau ≥ 0.75, and this number merely increased to 9 for Tau ≥ 0.5 **(Fig. 4f)**. This pattern was observed for tissue-specific genes from all species **(Fig. 4g** and **Extended Data Fig. 7)**, suggesting that tissue-specificity is highly dynamic and that a high proportion of these profiles may have a recent evolutionary origin. In fact, while only between 4% and 15% of bilaterian-conserved orthogroups are tissue-specific in each species (barplot in **Fig. 4h**), more than 47% of them contain at least one tissue-specific gene in at least one species (line plot in **Fig. 4h**). As expected, these orthogroups with tissue-specificity potential (i.e., those that are tissue-specific in at least one studied species) presented a significantly higher proportion of gene duplications compared to orthogroups that are never tissue-specific (p-value = 2e-04, Fisher's exact test), which in turn were strongly enriched for housekeeping functions such as RNA processing/binding and translation (extra boxes in **Fig. 4h** and **Supplementary Table 12**).

**Evolution of tissue-specificity is tightly linked to gene duplication and specialization**

We next performed a systematic phylogenetic inference of tissue-specificity gains and losses for each tissue since the last common bilaterian ancestor (see Methods, **Extended Data Figs. 5b-e** and **8a** and **Supplementary Table 13**). Remarkably, the Vertebrata node presents the highest average proportion of inferred gains across tissues **(Fig. 5a)**, even if some of the most important gain waves were considerably more recent (e.g., in fruit fly testis and frog ovary, with 198 and 109 gains, respectively; **Extended Data Fig. 8a**). Comparison of the proportions of tissue-specificity gains and losses within each node and species shows that testis had the highest turnover, as it presented the greatest proportion of gains and/or losses in 30/39 (77%) of nodes/species **(Fig. 5b)**. On the contrary, neural gains are mainly prevalent in the most ancestral nodes on both branches (i.e., Euteleostomi/Neoptera or older), but while they seem to have little impact in later vertebrate evolution, they still dominate the gain landscape in several more recent insect nodes (e.g., Holometabola, Oligoneoptera and Diptera). Notably, this result is in line with the differential conservation of the neural transcriptomes in vertebrates and insects **(Fig. 3a)**. Finally, as opposed to the most ancestral nodes, the tissue-specific transcriptome of few recent nodes and of the majority of single species is predominantly shaped by losses of tissue-specificity rather than by gains (i.e., we observed an average of 11% of losses on the total number of inferences for Tetrapoda/Holometabola and more ancient nodes, compared to 43% for more recent nodes and single species).

We then compared the total numbers of tissue-specificity gains between phylogenetically equivalent nodes on the two branches **(Fig. 5c)**. The gain signal reached the overall maximum in the Vertebrata ancestor, consistent with a strong impact of the two rounds of WGD at the origin of this group. Strikingly, although this effect progressively decreased, the fraction of gains remained high throughout all subsequent vertebrate nodes, in clear contrast with the relative flat signal observed on the insect side. Moreover, gains in both the Vertebrata and subsequent nodes, as well as in the species-specific branches, showed a much higher proportion of orthogroups involving paralogs derived from the vertebrate WGDs (2R-ohnologs) compared to phylogenetically equivalent Insecta nodes and species (**Fig. 5c** and **Extended Data Figure 8b**). Altogether, this suggests the existence of a previously unappreciated prolonged evolutionary impact of vertebrate WGDs on the rewiring of tissue-specific transcriptomes.

Importantly, the association between the gain of tissue-specificity and gene duplication extended beyond the vertebrate WGDs. For all nodes and extant species, we found that

orthogroups with inferred tissue-specificity had a higher proportion of duplicates compared to the corresponding background **(Fig. 5d** and **Extended Data Fig. 8c)**. Moreover, we found evidence that the acquisition of tissue-specific expression occurred to a large extent through the process of specialization [35], where the specialized paralog reduces its expression in most tissues, while the other paralog(s) conserve the ancestral broader expression pattern. For example, under this model, the Eutheria testis-specific genes are expected to have lower expression across the other tissues (i.e., all tissues but testis) in eutherians compared to non-eutherian species, as readily seen in our dataset **(Fig. 5e)**. This pattern was consistently observed across all nodes and species, as quantified by the number of other tissues (from 0 to 7) in which the expression of a given ortholog is lower in the set of species with tissue-specificity compared to those without, and far more extensively than expected by chance **(Fig. 5f** and **Extended Data Fig. 8d,e)**. Moreover, although stronger for duplicated genes, this pattern held true when separately considering duplicated or not duplicated tissue-specific genes **(Fig. 5g)**, supporting a widespread specialization landscape associated with tissue-specificity gains throughout our bilaterian phylogeny.

**Tissue-specificity gains are associated with emergence of unique tissue-related phenotypes**

Exploiting the symmetric structure of our phylogeny, we also identified 156 bilaterian-conserved orthogroups that acquired a single but distinct tissue-specificity on the vertebrate and insect sides (e.g., in neural and in ovary, respectively), thus fulfilling their functional potential in divergent contexts **(Extended data Fig. 9a)**. The most frequent pairs of parallel tissue-specificity gains were neural and testis together with testis and ovary **(Extended data Fig. 9b)**, in agreement with the compartments among which expression shifts are more likely to occur also within vertebrates [16]. In addition to these parallel tissue-specificity gains, we also characterized independent convergent acquisitions of the same tissue-specific expression profiles in both the vertebrate and the insect sides **(Fig. 6a)**. Such convergent gains were most abundant in testis, probably as a consequence of the faster rates of transcriptomic evolution that this tissue experiences both in vertebrates and in insects (see **Fig. 3a**). One exemplary case is represented by *TESMIN* and *tomb* **(Fig. 6b)**. These are paralogs of the ancestral *LIN54/mip120* gene that independently originated within the vertebrate and insect lineages and convergently acquired testis-specific expression in amniotes and the fruit fly, respectively, and whose importance for testis development and function is proven by spermatogenesis disruption upon gene perturbation both in mouse [36] and fruit fly [37].

Next, we aimed to functionally characterize the tissue-specificity gains in each node and species (**Supplementary Table 14**). We found that a few gene functions were significantly and repeatedly enriched in multiple nodes/species on both phylogenetic branches **(Fig. 6c**; see Methods**)**. GO categories such as double-strand DNA binding, cation transmembrane transport and animal organ morphogenesis were over-represented throughout all tissue types, but we also identified a few functions specifically enriched only across somatic organ tissues (e.g., plasma membrane region) or reproductive ones (mainly related to meiotic division). On the contrary, each tissue presented categories exclusively enriched across gains in a single node/species **(Fig. 6d** and **Supplementary Table 15)**, several of which could be linked to the concurrent emergence of novel phenotypic traits. For instance, only vertebrate neural-specific gains were significantly enriched in categories related to oligodendrocyte differentiation and ensheathment of neurons (i.e., the myelination of neuronal axons operated by the oligodendrocytes), consistent with the origin of these glial cells in the gnathostome vertebrate ancestor [38]. In another example, mammalian ovary-specific gains exhibited a unique enrichment for response to BMP (bone morphogenetic protein), a molecule shown to be involved in the regulation of mammalian oogenesis and folliculogenesis [39].

Finally, we focused on the functions of species-specific, tissue-specific gains, which represent 59% of all our inferred gains. In order to identify functional categories potentially over-represented among these species-specific inferences, we plotted a distribution of GOs based on the proportion of their bilaterian-conserved orthogroups experiencing at least one of such species-specific gains **(Fig. 6e)**. Strikingly, the top 5% of this distribution includes cell-cell signaling, tissue development and several other morphogenesis-related or developmental categories, which are significantly over-represented in the upper tail compared to the lower percentiles **(Fig. 6e** and **Extended Data Fig. 9c,d)**. One remarkable example of a developmental gene included in these species-specific gains is *FGF17*, which is neural-specific only in human **(Fig. 6f)**. *FGF17* is a fibroblast growth factor broadly expressed during the embryonic and postnatal brain development of multiple species, but which was co-opted in the adult brain only in human (**Extended Data Fig. 10**; data from [14]). Remarkably, a recent study [40] showed that the Fgf17 contained in the cerebrospinal fluid of young mice activates a transcriptional program leading to proliferation of oligodendrocyte progenitors and, when injected into aged mice, slows down brain aging and improves memory functions **(Fig. 6g).** Thus, even if Fgf17's potential to induce oligodendrocyte proliferation seems to be ancestral,

this gene became part of the adult neural-specific transcriptome only during recent human evolution, where it might contribute to the preservation of cognitive abilities in old age.

## Discussion

In this study, we have assembled an unprecedented dataset of RNA-seq samples spanning twenty bilaterian species and eight tissues, with the goal of tracing the evolution of gene expression in homologous tissues within and between vertebrates and insects. Therefore, in terms of phylogenetic range, our study represents a considerable step forward compared to previous works where a similar framework of tissue transcriptional evolution has been applied [13–16], as it extends the investigation range to a large panel of vertebrate and non-vertebrate species, including organisms which diverged ~600 MYA. Moreover, we explicitly designed our study around a symmetric phylogeny for the vertebrate and insect branches. This allowed us to identify not only ancestral features, but also parallel, convergent and divergent evolutionary trajectories of ancestral genes between and within these two major bilaterian lineages. Using this phylogenetic framework, we performed a pioneering analysis of the evolutionary dynamics of tissue-specific expression among ancestral bilaterian genes. Strikingly, we found that nearly half of the ancestral bilaterian gene complement has acquired tissue-specific expression in at least one of the studied species, an unexpectedly high fraction that reveals a surprising plasticity for this transcriptomic trait. We thus investigated the timings and mechanisms behind the pervasive evolution of these tissue-specificity patterns, as well as their functional impact.

Before discussing these aspects, however, we acknowledge that a major limitation of our study is the use of bulk RNA-seq data, which merges the signals originating from the different cell types present in each tissue. This issue is particularly relevant, given that we are analyzing distantly related species with highly divergent tissue histologies. Thus, differences in cell type composition might be a confounding factor in our comparisons of gene expression dynamics, especially those estimating quantitative differences among species across the entire tissue panel (e.g., correlations, PCAs, etc.). Notwithstanding, we aimed at minimizing these considerations by explicitly studying tissue-specific patterns, which are largely qualitative in nature (i.e., presence/absence), and thus should be more robust to quantitative variations in cell type composition. Indeed, changes in tissue-specific transcriptomes can provide information about evolutionary events such as the origin of novel cell types. For instance, we detected enrichment

for oligodendrocyte differentiation exclusively in the neural-specific genes acquired in the vertebrate node, concomitantly with the emergence of this cell type [38].

With regards to evolutionary timings, our phylogenetic inference revealed that most ancestral genes acquired tissue-specific expression during late bilaterian evolution. Despite this, we found that at least ~7% of all ancestral orthogroups have been expressed in a tissue-specific manner since the bilaterian LCA. Importantly, all the ancestral tissue-differential modules we identified are linked to core and shared functions within each tissue type, even in those tissues that have greatly diverged at the histological and cell type level (e.g., digestive system [41]) or mainly originated by convergent morphological trajectories (e.g., fat-rich tissues [42]). Moreover, we found that validated phenotypic effects both in mammals and in fly supported their extant, wide physiological relevance. Remarkably, neural, muscle and reproductive organ transcriptomes present the largest ancestral tissue-differential modules; this suggests they have more distinctive and conserved transcriptomic signatures compared to other bilaterian tissues, likely related to the high complexity and specialization of the main cell types that form them (neurons, myocytes and meiotic cells, respectively).

At the mechanistic level, we showed that tissue-specific gains have a strong association with gene duplication across the entire bilaterian phylogeny, as previously reported for more restricted lineages [35,43]. Furthermore, we investigated how often evolution of tissue-specificity generally occurred through specialization, by which the specialized paralog reduced its expression in the tissues without tissue-specificity compared to the broadly expressed ancestral patterns. This mechanism had previously been identified for more restricted groups, including paralogs originated from vertebrate [35] and salmon [44] WGDs, together with gene duplicates specific in the pea aphid [45], primates and rodents [46]; here, we expanded the search space and provided evidence that specialization is strongly associated with tissue-specificity gains throughout the entire bilaterian phylogeny. Another particularly remarkable finding in the context of gene duplication is the seemingly prolonged effect of the vertebrate WGDs on the amount of tissue-specificity gains throughout recent vertebrate evolution. Specifically, the Vertebrata node showed the highest level of tissue-specificity gains in the phylogeny, especially among paralogs retained from the WGDs (ohnologs), as expected from a direct causal effect of these events. However, subsequent ancestral nodes and extant species within the vertebrate lineage also exhibited substantially higher number of gains, as well as a higher fraction of affected ohnologs, compared to other phylogenetically equivalent non-vertebrate

nodes and species. While this could be partly due to loss of tissue-specificity in early branching vertebrate species, we suggest that this pattern reflects the increased likelihood of orthogroups with retained ohnologs in vertebrates to evolve tissue-specificity even millions of years after the WGDs that generated the genetic redundancy. If so, this unexpected finding implies that the evolutionary impact of WGDs on phenotypic diversification may go beyond immediately subsequent effects, providing an additional potential explanation for the lag observed between the timing of WGDs and their purported consequences in multiple lineages [47–49].

Finally, we assessed the functional impact of the rewiring of tissue-specific transcriptomes. At one extreme of the phylogenetic range, we identified sets of genes with unique tissue-differential expression patterns either in all vertebrates or in all insects; in other words, gene sets that differentially define the equivalent tissues in each lineage. For instance, the vertebrate-specific and insect-specific ovary-differential sets are significantly enriched for gonad development and cell cycle related functions, respectively, exemplifying how changes in ancient genes have contributed to differentially shape the female reproductive systems in the two lineages. At the other extreme, almost 60% of our inferred tissue-specificity gains occurred in specific species and were often associated with the emergence of unique phenotypes, highlighting the great potential of novel tissue-specific expression patterns to underlie organismal novelties [50]. For instance, we detected a distinctive enrichment in sensory perception of light stimulus in the octopus' skin, consistent with the unique presence of light-activated chromatophores organs all over the cephalopod's body surface [51]. Strikingly, we also uncovered a significant tendency for developmental genes to retain adult tissue-specific expression in a species-specific manner. Cases like *FGF17* in humans that we reported here point to a potential widespread, functional co-option of ancestral developmental genes within distinct tissue-specific transcriptomes throughout the most recent bilaterian evolution. Future research should elucidate the functional significance of the adult tissue-specific expression of these developmental genes as well as of the myriad of other tissue-specific genes we identified, and how they ultimately contribute to animal evolution.

## Methods

### Genome annotation and sequence files

For eight species, we downloaded the GTF and genome FASTA files from Ensembl [https://www.ensembl.org/], selecting the following assemblies and versions: human (*Homo sapiens, Hsa*: hg38, v88), mouse (*Mus musculus, Mmu:* mm10, v88), cow (*Bos Taurus, Bta*:

bosTau9, v99), opossum (*Monodelphis domestica, Mdo*: monDom5, v86), chicken (*Gallus gallus, Gga*: galGal6, v99), tropical clawed frog (*Xenopus tropical, Xtr*: XenTro9, v101), zebrafish (*Danio rerio, Dre*: danRer10, v80), elephant shark (*Callorhinchus milii, Cmi:* 6.1.3, v99). For other eight species, we downloaded GTF and genome FASTA files from Ensembl Metazoa [https://metazoa.ensembl.org/]: sea urchin (*Strongylocentrotus purpuratus:* Spu: Spur_5.0 v51), fruit fly (*Drosophila melanogaster, Dme*: dm6, v26), yellow fever mosquito (*Aedes aegypti, Aae*: AaegL5, v46), domestic silk moth (*Bombyx mori, Bmo*: ASM15162v1, v45), red flour beetle (*Tribolium castaneum, Tca*: Tcas5.2, v45), honey bee (*Apis mellifera, Ame*: Amel_4.5, v35), centipede (*Strigamia maritima*, Sma: Smar1, v26), California two-spot octopus (*Octopus bimaculoides, Obi*: PRJNA270931). For other three species, we used the GTF and genome FASTA files from the relative publications: amphioxus (*Branchiostoma lanceolatum*, Bla: [52]), marmalade hoverfly (Episyrphus balteatus, Eba: [53]), mayfly (*Cloeon dipterum*, Cdi: [10]). For the cockroach (*Blattella germanica*, Bge) we downloaded the GFF3 (blager_OGSv1.2.1.gff3) and the genome FASTA (GCA_000762945.2_Bger_2.0_genomic.fna.gz) from https://i5k.nal.usda.gov/content/data-downloads. We then converted the GFF3 to GTF with gffread and we modified it to match the Ensembl format using custom scripts. Finally, for Bmo and Bge, we enriched the annotations using the RNA-seq data we generated as described in **Supplementary Methods**.

**Gene orthology calls**

We used *Broccoli (v1.2)* [54] to infer gene orthogroups among all protein coding genes from the 20 species. In order to avoid redundant gene homology calls, we selected one representative protein isoform for each gene in each species (i.e., the isoform with the longest coding sequence). See **Extended Data Fig. 1** and **Supplementary Table 2** for gene orthogroups statistics and **Supplementary Dataset** for gene orthogroup files.

**Genome annotation refinements**

We corrected broken genes (a single gene annotated as two or more separated entities) and chimeric genes (independent genes annotated as a fused element) from almost all genome annotations (excluding human, mouse and fruit fly), since they can respectively result in erroneous gene duplication inferences and incomplete ortholog detection. For this purpose, we used the information provided by the gene orthology call as well as pairwise alignments, as described in detail in the **Supplementary Methods**.

**Gene Ontology (GO) annotation and transfers**

Comparative analyses relying on functional GO annotations risk being biased by the different GO annotation qualities existing between species. In order to avoid such biases, we generated a unified annotation for each of the species starting from the assumption that orthologous genes likely share functional properties. First, we built comprehensive GO annotations both for human and fruit fly. We downloaded the GO annotations (GeneID-GO correspondence) from *Ensembl (v106)* [55] for the two species, and we combined the human annotations with those from *clueGO v2.5.5 level 5* [56]. Then, to build a human-based GO annotation file, we assigned the GO annotations of each human gene to all the genes from the other species belonging to the same orthogroup whenever the number of human genes within the orthogroup with that GO annotation was ≥ 1/4 of the total human genes in that orthogroup. A similar procedure was followed to build a fruit fly-based GO annotation file based on the fruit fly GO annotation. Then, we selected only the GO categories with a number of genes included between 3 and 1,500 for the human-based annotation and 3 and 500 for the fruit fly one. The annotation files resulting from these "ontology transfers" were used for all GO analyses **(**see **Supplementary Dataset).**

**RNA-seq sample dataset**

We downloaded in total 1,130 individual RNA-seq samples across 18 species. All downloaded samples and relative information can be found in **Supplementary Table 3**. Moreover, we generated 95 RNA-seq samples for 15 species covering different tissues that were missing in public resources. See **Supplementary Table 3** for more details (all the samples generated for this project report “in_house” in the SRA field). All these samples were dissected from adult animals and the RNA extracted using the most suited protocol for the organism and tissue, namely TRIzol™ Reagent (Thermo Fisher) or Qiagen RNeasy kit (QIAGEN) (see GEO series GSE205498 for more details on sample extraction and processing protocols). These RNA samples were used to construct standard Illumina RNA-seq libraries at the CRG Genomics Unit, and an average of ~78 million 125-nucleotide paired-end reads were generated for each of them in a HiSeq2500. In the case of octopus, the sequencing was performed at the University of Chicago, with NovaSeq. In total, we generated ~7.6 billion individual reads. All read and mapping statistics for all RNA-seq samples are also provided in **Supplementary Table 3**.

**RNA-seq quantification**

We quantified expression using *Kallisto quant* [57], setting parameter *-b 100 --single -l 190 -s 20* for single end RNA-seq samples and *-b 100* for paired-end RNA-seq samples (with -b=number of bootstrap samples; -l=estimated average fragment length; -s=estimated standard deviation of fragment length). For each species, we quantified gene expression for each sample by summing the raw counts of all its corresponding annotated transcripts. We next normalized the expression with *DESeq2* [58] and used the effective lengths returned by Kallisto to compute the Transcript Per Million (TPMs). For all analyses, log2(TPM+1) was used as the final expression measure for each sample.

**Meta-samples and tissue expression measures**

When multiple datasets were available for a given tissue and species, we grouped the RNA-seq samples into a maximum of three meta-samples. This was done to: (i) increase read depth per meta-sample, (ii) dilute potential batch effects from publicly available samples, (iii) facilitate downstream analyses and comparisons by having a comparable number of replicates across tissues and species. Meta-sample groups are detailed in the column "Metasample" in **Supplementary Table 3**. In particular, we followed the approach that we previously described for human, mouse, cow, zebrafish and drosophila in *VastDB* [https://vastdb.crg.eu/] [59], where samples from comparable experiments based on clustering approaches are pooled. For this study, we computed the median expression across all the samples included in each meta-sample, which we used as its representative measure.

**Best-hits orthogroups definition**

Most comparative analyses need matrices with a single gene per species as input, which often restricts the studies to single 1-to-1 orthologs, which are known to have different evolutionary and expression biases [35]. To expand our gene sets beyond 1-to-1 orthologies, we defined a "best representative ortholog" (best-hit) per species in each bilaterian-conserved orthogroup (see scheme in **Extended Data Fig. 3a**). For this purpose, we first calculated the pairwise protein sequence similarity for all pairs of genes from different species within an orthogroup based on BLOSUM62, using *mafft* [60] with default parameters and comparing the best representative isoform per gene. Next, for any species with multiple paralogs in a given orthogroup, we selected as best-hit the gene with the highest average sequence similarity (with respect to the genes from all other species in the orthogroup) if this similarity value was at least 0.2 higher than the sequence similarity of all the other paralogs from the same species. If this requirement was not fulfilled, we discarded the genes with ≥0.2 average sequence similarity difference from

the gene with the top average sequence similarity, and we selected as best-hit among the remaining genes the one with the highest expression similarity across tissues (or a random one in case of equal expression similarity). This expression similarity was defined as the average of all pairwise Pearson's expression correlations across tissues between the target gene and each gene from all other species in the orthogroup, where expression in each tissue was represented by the averaged log2(TPM+1) expression values among all meta-samples of that tissue.

**Principal component analysis (PCA) and clustering analysis**

To investigate the interrelation among our meta-samples, we first quantile normalized their log2(TPM+1) expression values across the best-hit orthologs of all species per orthogroup using *limma* in R [61]. Next, we subsetted this normalized expression matrix to only the 2,436 orthogroups with at least one ortholog in each species (in order to avoid imputation) and applied the *prcomp* function in R (center=TRUE, scale=TRUE) to perform a PCA. To assess the biological nature of each principal component, we performed one-sided ANOVA tests between species and tissue groups employing the *aov* function in R (**Extended Data Fig. 3c,d**; shown p-values were Bonferroni corrected). The heatmap in **Extended Data Fig. 3e** was generated by the *pheatmap* R package with *ward.D2* clustering method on the same input matrix used for the PCA, but where the expression values were z-scored across tissues of the same species in order to minimize the inter-species variability.

**Definition of ancestral bilaterian tissue-differential modules**

As summarized in **Fig. 2a**, we first performed a sPLS-DA with the *splsda* function from the *mixOmics* package [62] in R, using as input the quantile-normalized expression table for all best-hit orthologs and meta-samples described above. We specifically compared all tissue groups versus each other, selecting the optimal number of components and loadings per component by running the *tune.splsda* function on the same expression table with the following parameters: *ncomp = 10, validation = 'Mfold', folds = 4, dist = 'max.dist', measure = "BER", test.keepX = c(1:10, seq(20, 300, 10)), nrepeat=10*. Since each of the resulting components specifically separated the meta-samples of each tissue group **(Supplementary Fig. 1)**, we used the corresponding loadings (which represent orthogroups with the most distinctive expression profiles in the isolated tissue compared to the others) to define the respective ancestral bilaterian tissue-differential modules. Importantly, contrary to a PCA, the proportion of variance explained by consecutive components does not necessarily decrease, as the aim is not

to maximize the variance. As an extra filter, we further selected only those orthogroups that had the highest median expression in the isolated tissue both among vertebrates and insects. Expression values per tissue in each species were derived by averaging the quantile-normalized log2(TPM+1) expression values of all meta-samples per tissue (i.e. one expression value per tissue and species). These values were then z-scored within species to be able to pool their relative tissue expression across species. Finally, the values plotted in **Fig. 2c,d** and **Supplementary Fig.1** correspond to the median of these measures among all vertebrates, all insects or all outgroups (i.e., only three values instead of 20 are plotted per orthogroup and tissue).

**Characterization of ancestral bilaterian tissue-differential modules**

GO enrichment analyses were performed with the *gprofiler2* [63] R package, using the human orthology transfers as GO annotation and all bilaterian-conserved orthogroups as background. All p-values were FDR corrected. For the representation of GO networks of significantly enriched categories (adjusted p-value ≤ 0.05) in **Fig. 2i**, only significant categories containing at least 5 genes in the tested set were considered. The networks were obtained from *Revigo* (http://revigo.irb.hr/) [64], selecting large output lists (90% of the input list; option 0.9) for all modules except the neural-differential (for which 0.4 [40%] was selected). To characterize the phenotypic impact of these genes, we downloaded all validated gene-phenotype associations from Ensembl ([55]; v105) for human and mouse and from FlyBase [65] updated in January 2020. Neural phenotypes were defined as anything matching "neuro", "behavior", "brain", "glia", "CNS" (case insensitive), while testis phenotypes were defined as anything matching "sperm", "infert", "sterile", "testis" (case insensitive). Orthogroups with positive matches in either species were considered for the plots shown in **Fig. 2e,f**. Neural and testis phenotypes associated with the relative ancestral tissue-differential module are reported in **Supplementary Table 5**, while all phenotypic associations mapped to the relative bilaterian-conserved orthogroup are available in the **Supplementary Dataset**.

**Enrichment of TF binding motifs in ancestral bilaterian tissue-differential modules**

To investigate which TFs might be behind ancestral bilaterian tissue-differential modules, we first built a database of Positional Weight Matrices (PWMs) combining motifs from 56 vertebrate and non-vertebrate species, for a total of 8016 original motifs with direct evidence in CIS-BP [66]. We clustered similar motifs into consensus motifs by running *gimme_cluster* [67] with parameter *-t 0.9999* as described in [35], obtaining a final set of 1406 PWMs of length ≥ 5.

We then mapped each of these consensus motifs to the respective bilaterian-conserved orthogroup(s) based on the genes (in all species) from which the original motifs in the group were derived. The database containing the orthogroup - motif cluster associations is provided in the **Supplementary Dataset**.

We then checked if the TFs included in the ancestral bilaterian tissue-differential modules presented binding sites enriched in the regulatory regions of the other genes in the same modules. For this, we considered as regulatory region the 3kb upstream of each gene's annotated transcription start site, or the intergenic distance from the closest upstream gene (last poly-adenilation site) if this was lower than 3kb. For all bilaterian-conserved genes in each species, we then calculated the di-nucleotide frequency with *rsat oligo analysis -2str -noov* [68]. We scanned all the sequences in each species, using the di-nucleotide frequencies as background and running *rsat matrix-scan* with the following parameter: *-quick -pseudo 1 -decimals 1 -2str -log_pseudo 0.01 -uth pval 0.01 -n score*. From the returned hits, we only selected the highly confident matches in each sequence (p-value ≤ 1e-05). Then, for all ancestral tissue-differential TFs, we performed two tests on the corresponding module, using as background all other bilaterian-conserved genes: (i) a Fisher's exact test, to check if the proportion of sequences with at least one hit in the tested module was significantly higher compared to the background proportion; and (ii) a negative binomial regression test, to check if the number of hits per sequence was increased among the genes in the tested module compared to the background. We performed each test three times, selecting in turns all, only vertebrate or only insect species. In **Fig. 2j** we represent all TFs which have: (i) either p-value ≤ 0.05 and positive beta in the regression test or p-value ≤ 0.05 in the Fisher's exact test performed on all species; and (ii) p-value ≤ 0.05 for both vertebrates and insect species separately either in the regression or Fisher's exact test.

**Characterization of relative rates of tissue transcriptome evolution in vertebrates and insects**

We built separate gene expression trees for each tissue in vertebrates and insects as described in [13]. We adopted a neighbor–joining approach based on pairwise distance matrices between the species of each clade in the considered tissue. The distance between species was defined as 1-ρ, where ρ is the Spearman's rho between the expression of the two species in the considered tissue. We computed ρ with the *cor* function of the *stats* R package, and we used as input the quantile-normalized best-hits log2(TPM+1) expression matrix where we averaged all meta-

samples per tissue/species (i.e. one expression value per tissue and species). The total tree length was calculated by summing the lengths of all branches. The reliability of branching patterns was assessed with bootstrap analyses (1000 random sampling with replacement). Adipose tissue was excluded from this analysis since it was missing for elephant shark.

**Definition of SeqCons and ExprCons gene sets**

The SeqCons and ExprCons genes (**Fig. 3f,g**) were defined based on the relative difference (i.e., delta) in average protein sequence similarity and expression correlation of each bilaterian gene orthogroups within vertebrate and insect species. More specifically, we defined four groups with the top 500 bilaterian-conserved orthogroups with the highest relative average sequence similarity (SeqCons) or average relative expression correlation (ExprCons) in one clade (vertebrates or insects) compared to the other. Importantly, this definition implies that such genes are either highly conserved in the target clade and/or highly divergent in the other. For instance, a vertebrate SeqCons gene can have either high values of sequence similarity among vertebrates, low similarity among insects, or a combination of the two. Sequence similarity within each clade was defined as the average pairwise protein sequence similarity between all pairs of best-hits orthologs in that clade (computed as described in "Best-hits orthogroups definition"). Expression correlation was represented by Spearman's rho computed on the quantile-normalized best-hits average log2(TPM+1) expression matrix. Lists of SeqCons and ExprCons orthogroups are available in **Supplementary Table 7**.

**Characterization of SeqCons and ExprCons genes**

We used the phenotypic annotation we built for human, mouse and fly (see "Characterization of ancestral bilaterian tissue-differential modules") to evaluate the phenotypic impact of SeqCons and ExprCons genes in different clades (vertebrates or insects). We considered as a mammalian or fly phenotype any phenotype annotated only in human/mouse or fly, respectively. The relative numbers and proportions for all groups compared to the background (i.e. all ancestral bilaterian orthogroups) were reported in **Fig. 3h** (left side), and all phenotypic labels are provided in **Supplementary Table 8**. In addition, for the genes in each SeqCons and ExprCons set, we counted the number of vertebrate species in which they have at least one duplicate. These numbers range between 0 (implying single copy orthologs in all vertebrates) and 8 (where the orthogroup presents at least two paralogs in all vertebrate species). The relative proportions of duplication levels in vertebrates for each group compared to the background were represented in **Fig. 3h** (right side). GO enrichment analyses were performed

with the *gprofiler2* [63] R package, using the human orthology transfers as GO annotation and all bilaterian-conserved orthogroups as background. All p-values were FDR corrected. GO networks including up to the top 20 significant categories for vertebrate and insect SeqCons and ExprCons sets. Only significant GOs containing at least 5 genes in the tested set were considered. Networks were obtained from *Revigo* (http://revigo.irb.hr/) [64], selecting large output lists (90% of the input list; option 0.9). Complete GO enrichment results are provided in **Supplementary Table 9**.

**Definition of vertebrate and insect tissue-differential modules**

We performed an sPLS-DA run per tissue as described in "Definition of ancestral bilaterian tissue-differential modules". For each tissue, we compared three different groups: (i) vertebrate meta-samples of the query tissue, (ii) insect meta-samples of the query tissue, and (iii) all other meta-samples. We then selected the components specifically separating either the vertebrate or the insect meta-samples of the query tissue from all others **(Supplementary Fig. 2)**, using the corresponding loadings to define vertebrate- and insect-specific tissue-differential modules. We selected as vertebrate-specific, tissue-differential orthogroups only those loadings that showed the highest median expression in the query tissue among vertebrates but not among insects (or the other way round for insect-specific, tissue-differential orthogroups). Moreover, we required the vertebrate/insect-specific orthogroups to have higher median expression in the query tissue of vertebrates/insects compared to the other clade. Expression values per tissue in each species were derived by averaging the quantile-normalized log2(TPM+1) expression values of all meta-samples per tissue (i.e. one expression value per tissue and species). These values were then z-scored within species to be able to pool their relative tissue expression across species.

**Tissue specificity calls**

To perform the tissue-specificity calls, we first computed the Tau [34] for all genes separately in each species. Tau is a measure of tissue specificity ranging from 0 (ubiquitous genes) to 1 (highly tissue-specific genes). For each species, we employed as input a quantile-normalized expression matrix of log2(TPMs+1) values averaged by tissue (i.e., one value per tissue). We defined as tissue-specific in each species all genes with Tau $\geq$ 0.75 and maximum expression $\geq$ log2(5). To associate these tissue-specific genes with one or two tissues ("Associated tissue(s)" in **Fig. 4b,c** and **Extended Data Fig. 5a**), we evaluated the expression proportion per tissue (tissue_expr / all_tissue_expr), where "tissue_expr" is the average normalized

log2(TPMs+1) expression of the gene in the target tissue and "all_tissue_expr" the sum of the average normalized log2(TPMs+1) expression values across all tissues. Specifically, we applied the following steps for each gene in each species (**Extended Data Fig. 5a**): (i) if the difference in expression proportion between two most-highly expressed tissues was ≥ 0.10 and their ratio ≥ 1.7, we associated the gene only with the top tissue. (ii) If the above conditions are not fulfilled, but the difference in expression proportion between the second and third most highly-expressed tissues was ≥ 0.15, we associated the gene with the two top tissues (multi-tissue specificity). (iii) Else, the gene was not considered as tissue-specific and not associated with any tissue. In addition, for the gain/loss inferences (see next section), we more loosely defined the "Top tissue(s)", corresponding to the "Associated tissue(s)", when available, or simply to the two tissues with the highest expression (**Extended Data Fig. 5a**, last example).

**Phylogenetic inference of tissue specificity gains and losses**

We performed the phylogenetic inferences of tissue specificity gains and losses for each tissue separately, starting from all bilaterian orthogroups presenting at least one tissue-specific call in that tissue (see previous section). For each of these orthogroups, we considered one gene per species per tissue inference, independently selecting the representative ortholog in each tissue in case of multiple paralogs. In particular, we selected the paralog with the strongest association with the query tissue, according to the following prioritization (**Extended Data Fig. 5b**): (i) a paralog called as tissue-specific in that tissue as defined above; if there were multiple tissue-specific paralogs, we selected the one with the highest Tau. (ii) Else, a paralog in which the target tissue is in the Top tissue(s), as defined above; if there were multiple such paralogs, we selected the one with the highest Tau, giving priority to those passing the expression cutoff (max expression ≥ log2(5). (iii) Else, the paralog with the highest Tau, giving priority to those passing the expression cutoff.

Then, once the set of gene representatives was selected for the query orthogroup and tissue, we implemented two subsequent inference approaches independently for each major branch (deuterostome and protostome). First, we performed a "strict approach", inferring a maximum of one tissue specificity gain for each major branch. Here, we inferred a gain in a node if (**Extended Data Fig. 5c**, left): (i) the first-branching species in the node was tissue-specific in the query tissue (as defined in the previous section); (ii) at least 50% of the node's descendant species with an ortholog had Tau ≥ 0.60 and were associated with the query tissue; and (iii) none of the outgroup species to that node on the same branch that passed the expression cutoff

had Tau ≥ 0.60 and were associated with the query tissue. Exceptionally, in the case of the most internal nodes (i.e., Euarchontoglires: human and mouse, Cyclorrhapha: fruit fly and hoverfly) we required Tau ≥ 0.6 and association with the query tissue in both species and a tissue-specific call in that tissue for at least one of them. Second, for all the orthogroups that could not be classified with the first strict approach for a given branch, we inferred gains with less stringent requirements ("relaxed approach"; **Extended Data Fig. 5c**, right). Here, we inferred gains in the last common ancestor of all species with Tau ≥ 0.60 that are associated with the query tissue as long as at least one tissue-specific gene is present. However, the relaxed approach inferred multiple gains on each branch if the minimum distance between two species or nodes respecting those tissue-specificity cutoffs was higher than 3 nodes (e.g., in human and in chicken, or in Eutheria and in zebrafish). Also, if no inference of gain in an ancestral node could be done by either approach, tissue-specific genes (as defined in the previous section) were considered species-specific gains. Finally, from the combined output of both approaches, we inferred a bilaterian gain whenever a gain was identified in both Deuterostoma/Chordata/Vertebrata and Protostoma/Arthropoda/Insecta with either strict or relaxed criteria (**Extended Data Fig. 5d**). As an exception, since shark testis samples showed poor correlation with other testis samples, we also inferred bilaterian gains for testis in case of gain inferences in Euteleostomi and Protostoma.

We then inferred tissue-specificity losses starting from the nodes in which gains were inferred for each tissue (**Extended Data Fig. 5e**). In case of bilaterian gains, the inferences were conducted separately on the deuterostome and protostome branches. We considered as potential losses all species (internal to the node with the inferred gain) where either: (i) Tau ≤ 0.45; or (ii) the query tissue is not among the top tissue(s), as defined above (**Extended Data Fig. 5a**); or (iii) the difference in expression proportions between the query tissue and the third highest tissue is ≤ 0.1. Then, starting from the innermost species with a potential loss, if there were two or more consecutive such species, we inferred a loss in the node corresponding to their LCA and a novel gain in the node of the LCA of their consecutive inner species if: (i) all these species are tissue-specific as described above; (ii) the ancestral loss is separated by at least one node from the most ancestral gain, and (iii) the total number of these new inferences (including single losses in all the species excluded from the ancestral loss inference) is lower than the number of original inferences (i.e. independent losses for each potential loss species). Otherwise, separated losses for each single species are inferred.

**Characterization of non-tissue-specific orthogroups**

The GO enrichments for the non-tissue-specific orthogroups shown in **Fig. 4h** and **Supplementary Table 12** were performed as described in "Characterization of ancestral bilaterian tissue-differential modules". The duplication status of these orthogroups compared to the tissue-specific and all orthogroups was evaluated by counting the number of species in each of them that presented at least two paralogs (also shown in **Fig. 4h**). The difference between the TS and non-TS orthogroups in terms of duplications was assessed with Fisher's exact test (*fisher.test* function in R, with *simulate.p.value = TRUE*).

**Duplication and specialization patterns of tissue-specificity gains**

Each orthogroup's duplicated proportion was defined as the number of species with at least two paralogs over the total number of considered species (which depends on the tested node). The mean duplicated proportion for the orthogroups with gains in each node compared to the relative background (i.e. all orthogroups in that node) is shown in **Fig. 5d**. The same logic was also used to divide the species with tissue-specificity between duplicated species (with at least two paralogs) and non-duplicated species (with only one gene in the orthogroup) for the plot shown in **Fig. 5g**. The proportion of orthogroups with gains including 2R-onhologs (**Fig. 5c**) was based on the list of 2R-ohnologs provided by [69]. We then checked how each tissue-specific gain fitted the specialization hypothesis. We started from the same expression matrices used for the tissue specificity call (see above), comparing the median expression in each tissue between species with tissue-specificity and species without tissue-specificity (including species with inferred tissue-specificity losses). For each gain, we counted for how many tissues (excluding the tissue with tissue-specificity) this median expression was higher in the species without tissue-specificity (ranging 0-7; relative proportions across nodes and species in **Fig. 5f** and **Extended Data Fig. d,e**). For the gains in each node and species, we performed 100 randomizations of the tissue-specificity labels among all species in the relative orthogroup. For each of these randomization rounds, we counted the proportion of gains in which the number of tissues (excluding the tissue with tissue-specificity) with higher median expression in the species without tissue-specificity was $\geq 5$. We plotted in red the distributions of these proportions for all randomizations overlaying the relative observed distributions in **Extended Data Fig. d,e** or their collapsed distributions in **Fig. 5e**.

**Functional characterization of tissue-specificity gains**

Parallel and convergent gains of tissue-specificity (**Extended Data Fig. 9a** and **Fig. 6a**) were evaluated exclusively among those orthogroups that present tissue-specificity gains in only one tissue on each of the main branches (deuterostome or protostome). The GO enrichment analysis on the orthogroups with gains in each node/species reported in **Fig. 6c,d** and **Supplementary Table 14,15** were performed as described in "Characterization of ancestral bilaterian tissue-differential modules". For the heatmap in **Fig. 6c**, we exclusively considered GO categories that were either (i) significantly enriched in the gains of at least 15 nodes/species across all tissues or (ii) significantly enriched in the gains of at least 8 nodes/species in one tissue exclusively; in this last analysis, ovary and testis were grouped in order to catch a combined signature from the reproductive organs. The plotted values (log2(observed/expected+1)) were computed starting from the proportion of gains in each node/species belonging to the tested category (observed) and the proportion of all bilaterian-conserved orthogroups with a functional annotation belonging to the same category (expected). Highly redundant categories were manually removed. For **Fig. 6d** and **Supplementary Table 15**, we only consider the GO categories that are exclusively enriched in one node or species. Then, we moved to the characterization of species-specific gains, where we evaluated if developmental GOs were more represented in these recent gains compared to ancestral ones. Developmental GO categories were defined starting from the human transferred GO annotation (see above) as any term matching "develop", "differentiation", "determination", "morphogen", "commitment", "specification", "regionalization", "formation", "genesis". For the plot shown in **Fig. 6e**, only the GO categories including at least 10 bilaterian-conserved orthogroups were considered. The GSEA analysis in **Extended Data Fig. 9b** was performed with the *fgsea* package in R [70], and distribution shown in **Extended Data Fig. 9d** resulted from 1000 randomizations of the GO categories labels across the proportions of orthogroups in each category that included at least one species-specific gain.

## Code availability

All the code used for the analysis is available on GitHub at https://github.com/fedemantica/bilaterian_GE.

## Data availability

The FASTQ and processed files of the RNA-seq samples generated for this project are available at GEO under series GSE205498. The Supplementary Dataset is available at https://data.mendeley.com/drafts/22m3dwhzk6.

## Authors' contribution

FM performed most analyses and generated most figures and tables. LPI built the motif dataset, designed and performed all motif-related analysis, and contributed with intellectual discussion. YM and ATM performed additional analyses and contributed with intellectual discussion. JP, ATM, JC, XFM, FT, DB, SB, TD, MN, PC, FN, HE, MIA, CA, KW, IA, DMC contributed RNA and/or tissue samples. FM and MI wrote the manuscript.

## Acknowledgements

We thank Queralt Tolosa Ramon and Niccolo' Arecco for their original drawing of tissue and cell-type icons. We also thank Niccolo' Arecco, Nuno Barbosa-Morais, Ignacio Maeso, and Arnau Sebé-Pedrós for their critical feedback on the manuscript. We thank the CRG Genomics Unit for the RNA sequencing. This research has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-StG-LS2-637591 and ERCCoG-LS2-101002275 to MI), by the Spanish Ministry of Economy and Competitiveness (BFU-2017-89201-P and PID2020-115040GB-I00 to MI) and by the 'Centro de Excelencia Severo Ochoa 2013-2017'(SEV-2012-0208). FM holds a FPI fellowship associated with the grant BFU-2017-89201-P. Additional support for this research was provided by the Spanish MINECO (PGC2018-098427- B-I00 to DM and XF-M), the Czech Science Foundation (22-21244S to MN), and the National Institutes of Health-NIAID (grant R21AI167849 to FGN).

## References

1. Evans, S. D., Hughes, I. V., Gehling, J. G. & Droser, M. L. Discovery of the oldest bilaterian from the Ediacaran of South Australia. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 7845–7850 (2020).
2. Technau, U. Brachyury, the blastopore and the evolution of the mesoderm. *Bioessays* **23**, 788–794 (2001).
3. Brusca, Moore & Shuster. Introduction to the Bilateria and the Phylum Xenacoelomorpha Triploblasty and Bilateral. in *Invertebrates* (ed. Sinauer Associates, Inc) (2016).
4. Paps, J. & Holland, P. W. H. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat. Commun.* **9**, 1730 (2018).
5. Fernández, R. & Gabaldón, T. Gene gain and loss across the metazoan tree of life. *Nat Ecol Evol* **4**, 524–533 (2020).

6. Lopez-Bigas, N., De, S. & Teichmann, S. A. Functional protein divergence in the evolution of Homo sapiens. *Genome Biol.* **9**, R33 (2008).
7. King, M.-C. & Wilson, A. C. Evolution at Two Levels in Humans and Chimpanzees. *Science* **188**, 107–116 (1975).
8. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
9. Arntfield, M. E. & van der Kooy, D. β-Cell evolution: How the pancreas borrowed from the brain: The shared toolbox of genes expressed by neural and pancreatic endocrine cells may reflect their evolutionary relationship. *Bioessays* **33**, 582–587 (2011).
10. Almudi, I. *et al.* Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings. *Nat. Commun.* **11**, 2631 (2020).
11. Clark-Hachtel, C. M. & Tomoyasu, Y. Two sets of candidate crustacean wing homologues and their implication for the origin of insect wings. *Nat Ecol Evol* **4**, 1694–1702 (2020).
12. Bruce, H. S. & Patel, N. H. Knockout of crustacean leg patterning genes suggests that insect wings and body walls evolved from ancient leg segments. *Nat Ecol Evol* **4**, 1703–1712 (2020).
13. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
14. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).
15. Chen, J. *et al.* A quantitative framework for characterizing the evolutionary history of mammalian gene expression. *Genome Res.* **29**, 53–63 (2019).
16. Fukushima, K. & Pollock, D. D. Amalgamated cross-species transcriptomes reveal organ-specific propensity in gene expression evolution. *Nat. Commun.* **11**, 4459 (2020).
17. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012).
18. Lê Cao, K.-A., Boitard, S. & Besse, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253 (2011).
19. Burkhardt, P. & Sprecher, S. G. Evolutionary origin of synapses and neurons - Bridging the gap. *Bioessays* **39**, (2017).
20. Sebé-Pedrós, A. *et al.* Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* **173**, 1520–1534.e20 (2018).
21. Inaba, K. Sperm flagella: comparative and phylogenetic perspectives of protein

components. *Mol. Hum. Reprod.* **17**, 524–538 (2011).

22. Daldello, E. M., Luong, X. G., Yang, C.-R., Kuhn, J. & Conti, M. Cyclin B2 is required for progression through meiosis in mouse oocytes. *Development* **146**, (2019).
23. Li, J., Ouyang, Y.-C., Zhang, C.-H., Qian, W.-P. & Sun, Q.-Y. The cyclin B2/CDK1 complex inhibits separase activity in mouse oocyte meiosis I. *Development* **146**, (2019).
24. Zeng, Y. *et al.* Bi-allelic mutations in MOS cause female infertility characterized by preimplantation embryonic arrest. *Hum. Reprod.* **37**, 612–620 (2022).
25. Tay, J., Hodgman, R., Sarkissian, M. & Richter, J. D. Regulated CPEB phosphorylation during meiotic progression suggests a mechanism for temporal control of maternal mRNA translation. *Genes Dev.* **17**, 1457–1462 (2003).
26. Gąsiorowski, L. *et al.* Molecular evidence for a single origin of ultrafiltration-based excretory organs. *Curr. Biol.* **31**, 3629–3638.e2 (2021).
27. Thakurela, S. *et al.* Mapping gene regulatory circuitry of Pax6 during neurogenesis. *Cell Discov* **2**, 15045 (2016).
28. Eckler, M. J. & Chen, B. Fez family transcription factors: controlling neurogenesis and cell fate in the developing mammalian nervous system. *Bioessays* **36**, 788–797 (2014).
29. Taylor, M. V. & Hughes, S. M. Mef2 and the skeletal muscle differentiation program. *Semin. Cell Dev. Biol.* **72**, 33–44 (2017).
30. Mathiyalagan, N. *et al.* Meta-Analysis of Grainyhead-Like Dependent Transcriptional Networks: A Roadmap for Identifying Novel Conserved Genetic Pathways. *Genes* **10**, (2019).
31. Thomas, J. A., Welch, J. J., Lanfear, R. & Bromham, L. A Generation Time Effect on the Rate of Molecular Evolution in Invertebrates. *Mol. Biol. Evol.* **27**, 1173–1180 (2010).
32. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).
33. Warnefors, M. & Kaessmann, H. Evolution of the correlation between expression divergence and protein divergence in mammals. *Genome Biol. Evol.* **5**, 1324–1335 (2013).
34. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
35. Marlétaz, F. *et al.* Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
36. Oji, A. *et al.* Tesmin, Metallothionein-Like 5, is Required for Spermatogenesis in Mice†. *Biol. Reprod.* **102**, 975–983 (2020).
37. Jiang, J., Benson, E., Bausek, N., Doggett, K. & White-Cooper, H. Tombola, a

tesmin/TSO1-family protein, regulates transcriptional activation in the Drosophila male germline and physically interacts with always early. *Development* **134**, 1549–1559 (2007).

38. Hines, J. H. Evolutionary Origins of the Oligodendrocyte Cell Type and Adaptive Myelination. *Front. Neurosci.* **15**, 757360 (2021).
39. Rossi, R. O. D. S. *et al.* The bone morphogenetic protein system and the regulation of ovarian follicle development in mammals. *Zygote* **24**, 1–17 (2016).
40. Iram, T. *et al.* Young CSF restores oligodendrogenesis and memory in aged mice via Fgf17. *Nature* **605**, 509–515 (2022).
41. Hartenstein, V. & Martinez, P. Structure, development and evolution of the digestive system. *Cell Tissue Res.* **377**, 289–292 (2019).
42. Ottaviani, E., Malagoli, D. & Franceschi, C. The evolution of the adipose tissue: a neglected enigma. *Gen. Comp. Endocrinol.* **174**, 1–4 (2011).
43. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. Tissue-Specificity of Gene Expression Diverges Slowly between Orthologs, and Rapidly between Paralogs. *PLoS Comput. Biol.* **12**, e1005274 (2016).
44. Lien, S. *et al.* The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
45. Fernández, R. *et al.* Selection following Gene Duplication Shapes Recent Genome Evolution in the Pea Aphid Acyrthosiphon pisum. *Mol. Biol. Evol.* **37**, 2601–2615 (2020).
46. Farré, D. & Albà, M. M. Heterogeneous patterns of gene-expression diversification in mammalian gene duplicates. *Mol. Biol. Evol.* **27**, 325–335 (2010).
47. Clark, J. W. & Donoghue, P. C. J. Constraining the timing of whole genome duplication in plant evolutionary history. *Proc. Biol. Sci.* **284**, (2017).
48. Macqueen, D. J. & Johnston, I. A. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc. Biol. Sci.* **281**, 20132881 (2014).
49. Donoghue, P. C. J. & Purnell, M. A. Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.* **20**, 312–319 (2005).
50. Almudí, I. & Pascual-Anaya, J. How Do Morphological Novelties Evolve? Novel Approaches to Define Novel Morphologies. in *Old Questions and Young Approaches to Animal Evolution* (eds. Martín-Durán, J. M. & Vellutini, B. C.) 107–132 (Springer International Publishing, 2019).
51. Ramirez, M. D. & Oakley, T. H. Eye-independent, light-activated chromatophore expansion (LACE) and expression of phototransduction genes in the skin of Octopus

bimaculoides. *J. Exp. Biol.* **218**, 1513–1520 (2015).

52. Brasó-Vives, M. *et al.* Parallel evolution of amphioxus and vertebrate small-scale gene duplications. *bioRxiv* 2022.01.18.476203 (2022) doi:10.1101/2022.01.18.476203.
53. Doyle, T. *et al.* Genome-wide transcriptomic changes reveal the genetic pathways involved in insect migration. *Mol. Ecol.* **31**, 4332–4350 (2022).
54. Derelle, R., Philippe, H. & Colbourne, J. K. Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment. *Mol. Biol. Evol.* **37**, 3389–3396 (2020).
55. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
56. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).
57. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
58. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
59. Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res.* **27**, 1759–1768 (2017).
60. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
61. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
62. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
63. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res.* **9**, (2020).
64. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
65. Gramates, L. S. *et al.* FlyBase: a guided tour of highlighted features. *Genetics* **220**, (2022).
66. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
67. Bruse, N. & van Heeringen, S. J. GimmeMotifs: an analysis framework for transcription factor motif analysis. *bioRxiv* 474403 (2018) doi:10.1101/474403.
68. Nguyen, N. T. T. *et al.* RSAT 2018: regulatory sequence analysis tools 20th anniversary.

*Nucleic Acids Res.* **46**, W209–W214 (2018).

69. Touceda-Suárez, M. *et al.* Ancient Genomic Regulatory Blocks Are a Source for Regulatory Gene Deserts in Vertebrates after Whole-Genome Duplications. *Mol. Biol. Evol.* **37**, 2857–2864 (2020).
70. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.

## Main Figures

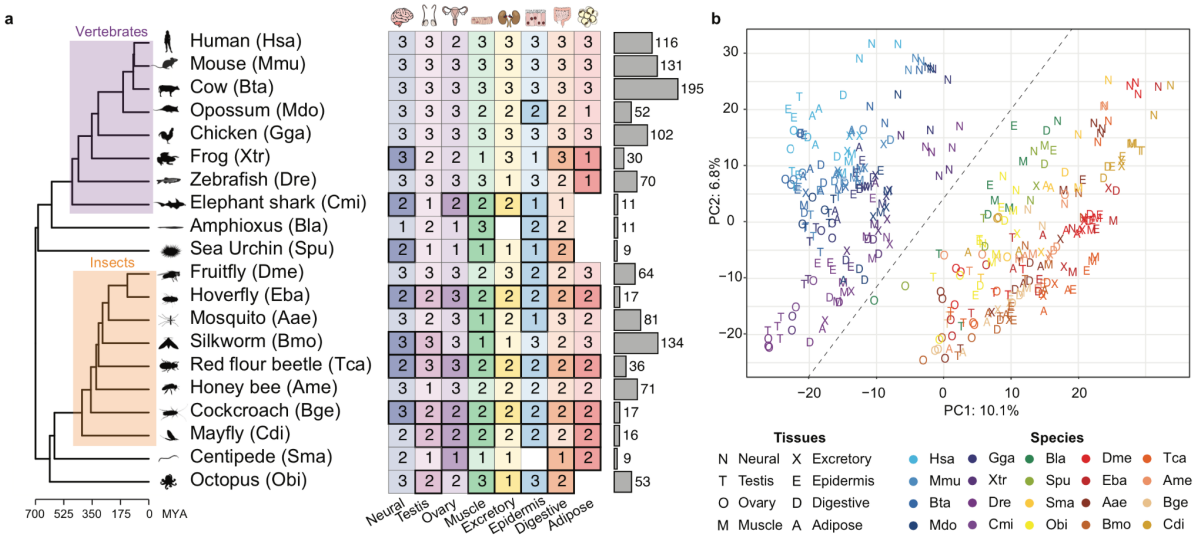



**Fig. 1: Dataset overview and global patterns of gene expression across bilaterian tissues.** **a.** RNA-seq dataset overview**.** Left: phylogenetic tree including the common names and scientific acronyms of the 20 bilaterian species considered in this study. Evolutionary distances were derived from http://www.timetree.org/ (MYA: million years ago) and animal silhouettes downloaded from http://phylopic.org/. Center: scheme of RNA-seq meta-samples. The number of meta-samples for each species (rows) and tissue (columns) is reported. The cell color corresponds to the tissue, while its intensity distinguishes between cases where at least one RNA-seq sample has been generated for this project (full color) from cases where all the included samples are publicly available (transparent color). Right: barplot with the total number of processed RNA-seq samples per species. **b.** Coordinates of the first (PC1; x axis) and second (PC2; y axis) principal components from a PCA performed on normalized gene expression values of best-hit bilaterian-conserved orthogroups across all species' meta-samples (see Methods). Only the 2436 orthogroups conserved in all species were considered. Tissue identity is represented by letters and species by colors. The percentage of variance explained by each PC is reported on the relative axis.

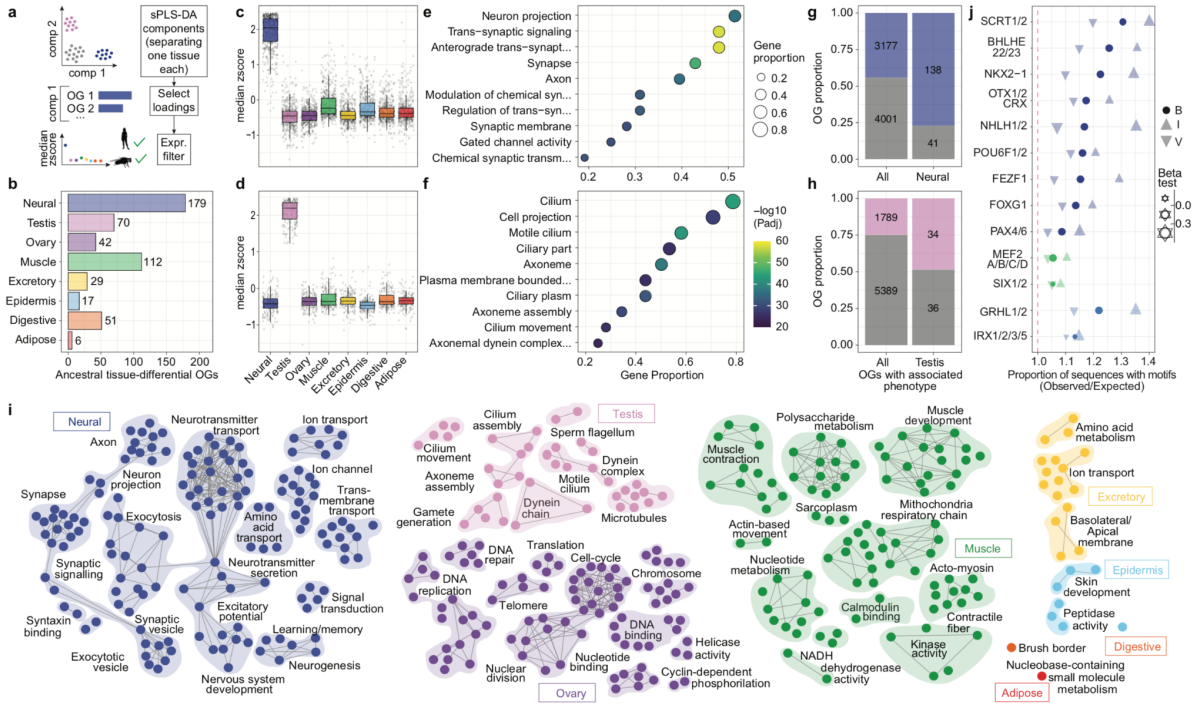



**Fig. 2: Reconstruction of ancestral bilaterian tissue-differential expression modules. a.** Scheme depicting the procedure for the definition of the ancestral tissue-differential modules. We performed an sPLS-DA on the expression of the ancestral bilaterian best-hits orthogroups, where each component specifically separated the meta-samples of one tissue group from all the others. We selected loadings of each component, which corresponded to orthogroups with the most distinctive expression profile in the isolated tissue compared to the rest. Finally, we required these orthogroups to show the highest median expression (after z-scoring expression within each species) in that particular tissue among both vertebrates and insects. **b:** Number of orthogroups included in each ancestral tissue-differential module. **c,d.** Expression profiles across tissues of best-hits orthogroups in the ancestral neural- (c) and testis- (d) differential modules. Expression values were first z-scored by species, and each dot represents the median expression among vertebrates, insects or outgroups. **e,f.** Top 20 most significantly enriched GO categories in the ancestral neural- (e) and testis- (f) differential modules. Only GOs containing at least 5 genes in the tested set were considered. **g,h**: Proportion of all bilaterian-conserved orthogroups (left) and ancestral neural (g) or testis (h) modules (right) associated with experimentally validated phenotypes (in the respective tissue) in mammals and/or fruit fly. **i:** Representation of GO networks of significantly enriched categories for all ancestral tissue-differential modules, where only categories containing at least 5 genes in the tested set were considered (see Methods). **j.** TFs included within each tissue-differential module whose known

binding motifs are significantly over-represented in the regulatory regions of the genes in the corresponding module (see Methods). Each TF was tested (Fisher's exact and regression tests) on all sequences (B: bilaterian), only vertebrate (V) or only insect (I) sequences within the module. TFs in each tissue are ordered by the ratio of the proportion of sequences with at least one predicted binding site in the tested module (observed) compared to the proportion in all other bilaterian-conserved genes (expected). The size of each dot reflects the beta from the regression test in the corresponding group, and tissue colors refer to panel b.

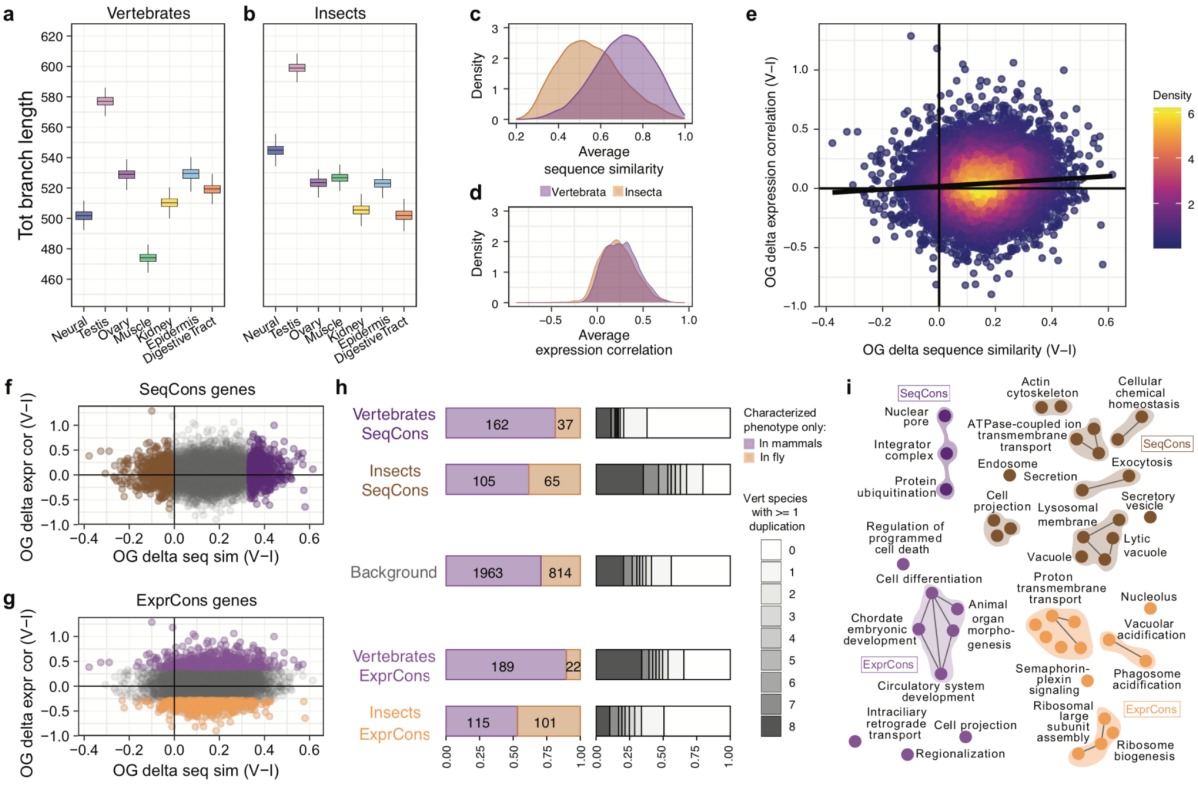



**Fig. 3: Sets of ancestral genes differentially shape tissue transcriptomes in vertebrates and insects. a,b.** Distributions of total branch lengths of expression-based trees built from best-hits orthogroups (see Methods). The values for each tissue and lineage result from 1000 bootstraps. **c,d**. Distribution of average sequence similarity (c) and average expression correlation (Spearman's rho) (d) among best-hits orthogroups across vertebrates and insects. **e:** Scatter plot representing best-hits orthogroups in function of the difference in sequence similarity (x, Delta seq sim) and in expression correlation (y, Delta expr cor) between their vertebrate (V) and insect (I) genes. **f,g.** Scatterplot as in (e) where the vertebrate (purple shades) and insect (brown shades) SeqCons (f) and ExprCons (g) orthogroups are highlighted. **h.** Left: relative proportions of SeqCons (up) or ExprCons (down) orthogroups with validated phenotypes exclusively in mammals (human, mouse) or fly compared to the background (middle). Right: stratification of the same orthogroup sets in function of the number of vertebrate species presenting at least one duplication. **i.** GO networks including up to the top 20 significant categories for vertebrate and insect SeqCons and ExprCons sets. Only GOs containing at least 5 genes in the tested set were considered. Networks were obtained from *Revigo* (http://revigo.irb.hr/) [64], selecting large output lists (0.9 of the input list). Colors refer to panel (f,g)

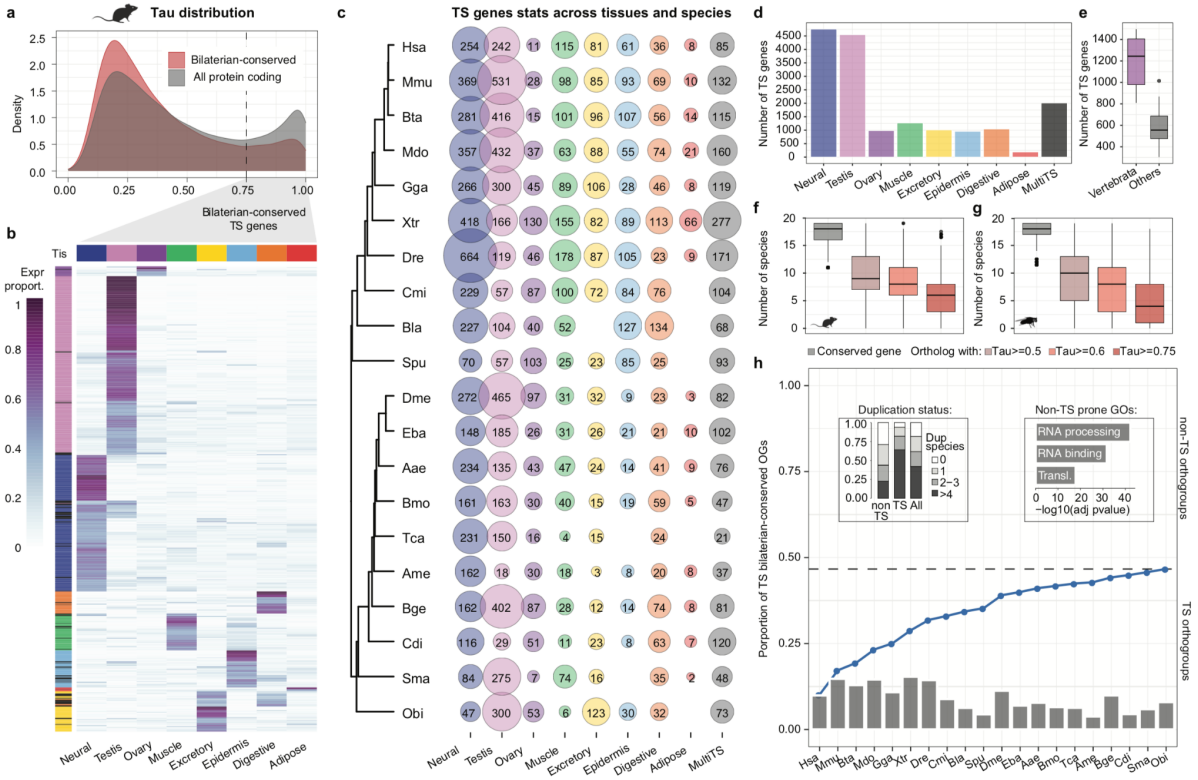



**Fig. 4: Tissue-specificity patterns across species reveal low conservation of tissue-specific expression profiles. a.** Tau distributions of all (gray) or bilaterian-conserved (red) mouse protein coding genes. **b.** Heatmap showing the clustering of mouse bilaterian-conserved, tissue-specific genes (rows) based on their expression proportion (tissue_expr / all_tissue_expr) across tissues (columns). The heatmap was generated by *pheatmap* in R with default parameters, and the complete dendrogram is shown in **Extended Data Fig. 6**. Black indicates multi-tissue specificity. **c,d.** Number of bilaterian-conserved, tissue-specific genes across all species (rows) and tissues (columns) (c) and collapsed by tissue (d). **e.** Distribution of the number of bilaterian-conserved, tissue-specific genes in vertebrates versus all other species (p-value = 2e-04; Wilcoxon rank sum test). **f,g.** Distribution of the number of species in which mouse (f) or fruit fly (g) bilaterian-conserved, tissue-specific genes have at least one ortholog (gray) and this ortholog(s) has a Tau value higher than a specific cutoff (0.5, 0.6 and 0.75; other shades). **h.** Barplot: proportion of bilaterian-conserved orthogroups including at least one tissue-specific gene in each given species. Line plot: cumulative distribution of the proportion of unique bilaterian-conserved orthogroups containing at least one tissue-specific gene across species. The dashed line marks the total proportion. The two boxes include information on the duplication status of the non-tissue-specific (non-TS) and tissue-specific (TS) orthogroups (left) and the top GO enrichments for the non-TS orthogroups (right).

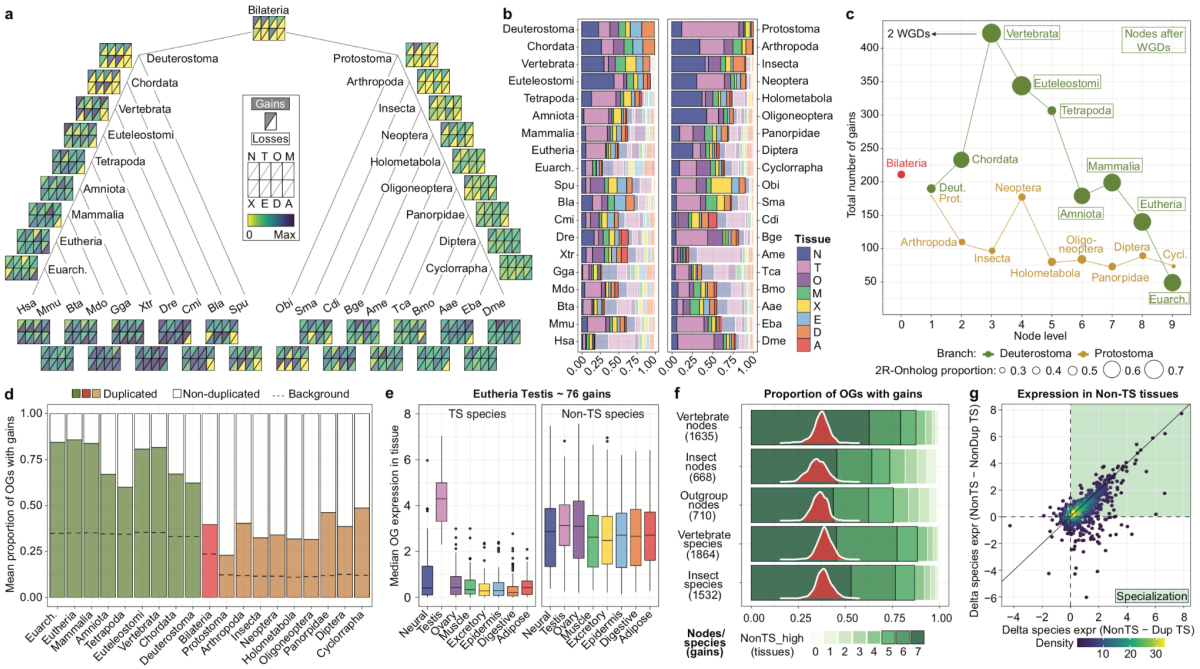



**Fig. 5: Tissue-specificity gains are associated with gene duplication and specialization. a.** Relative proportions of tissue-specificity gains and losses within each tissue across all nodes and species. Proportions were increased by one, log2-transformed and linearly scaled for each tissue with respect to the maximum value in that tissue. **b.** Relative proportions of tissue-specificity gains and losses across tissues within each node and species. Full/transparent shades of tissue colors represent gains/losses, respectively. **c.** Total number of orthogroups showing tissue-specificity gains across nodes on both phylogenetic branches. The size of the dots represents the proportion of 2R-onholog orthogroups in each gain group. **d.** Average proportions of duplicated and non-duplicated species among the species with tissue-specific expression in the orthogroups that gain tissue-specificity in each node. The background line represents the expected proportion based on all bilaterian-conserved orthogroups for the same sets of species. **e.** Median gene expression across tissues for bilaterian-conserved orthogroups with testis-specific gains in Eutheria (76 orthogroups) for the species with (left) or without (right) inferred tissue-specificity. **f.** For each set of tissue-specificity gains, distribution of the number of tissues in which the gene is not tissue-specific where the median expression of the species without tissue-specificity is higher than in the set of species with tissue-specificity (from 0 to 7, "NonTS_high"). The red distribution represents the proportion of gains with NonTS_high ≥ 5 coming from 100 randomizations of the tissue-specificity labels within the respective orthogroups (see Extended Data Fig. 8d,e for full data). **g.** Scatter plot representing each tissue-specific gain in function of the difference in gene expression across the tissues in

which the gene is not tissue-specific between (i) species without tissue-specificity ("non-TS") and (ii) species with tissue-specificity. The latter were divided into species with at least one duplication (x axis, "Dup TS") and species without duplications (y axis, "Non-dup TS"). By definition, only gains where tissue-specific species include species both with and without duplicates were plotted. Abbreviations: N: neural, T: testis, O: ovary, M: muscle, X: excretory system, E: epidermis, D: digestive tract, A: adipose, Euarch: Euarchontoglires. Cycl: Cyclorrapha. Deut: Deuterostoma. Prot: Protostoma.

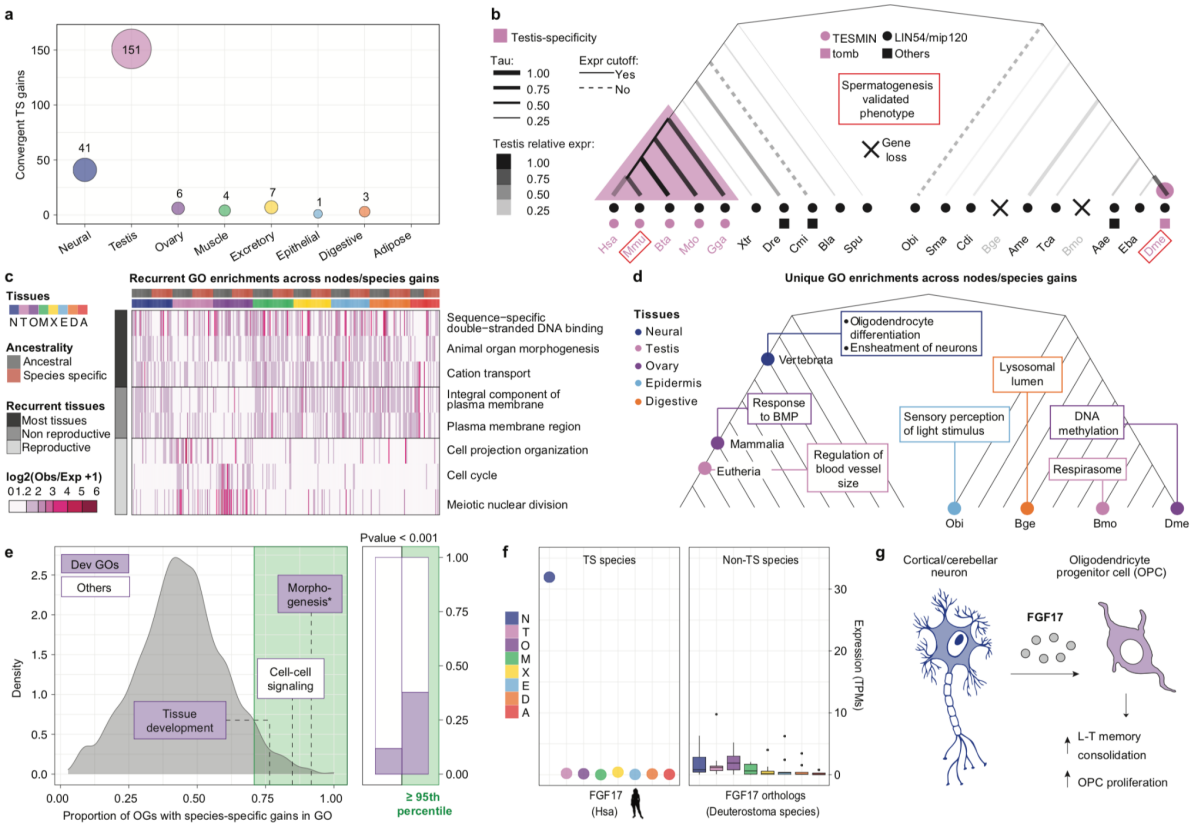
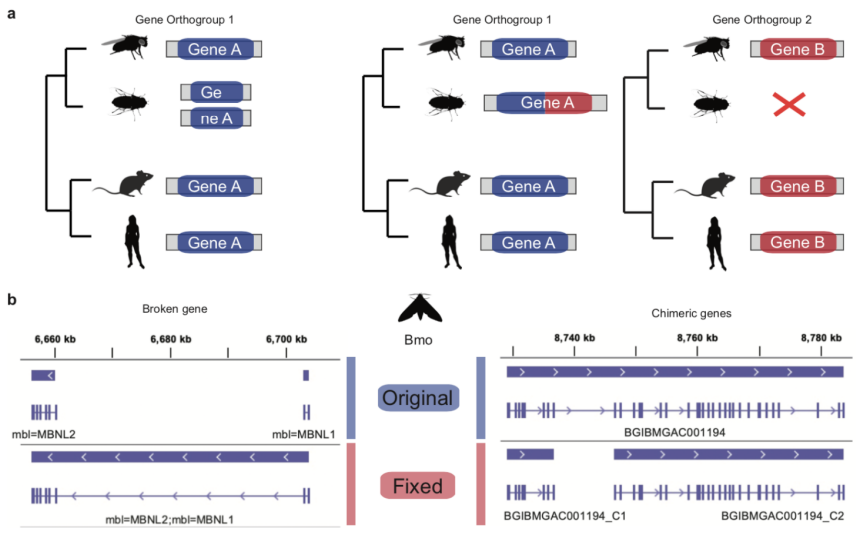



**Fig. 6: Tissue-specific gains are associated with the emergence of unique phenotypes.**

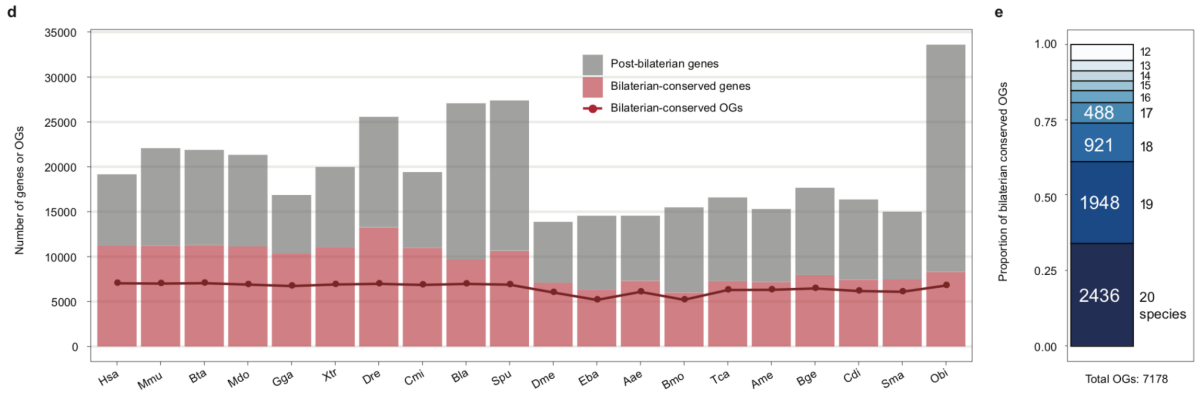
**a.** Number of convergent tissue-specificity gains (on the deuterostome and protostome branches) in each tissue. **b.** Example of a convergent testis-specific gain: *TESMIN*/*tomb*. **c.** Heatmap representing GO categories either (i) significantly enriched in the gains of at least 15 nodes/species across all tissues (most/non reproductive labels) or (ii) significantly enriched in the gains of at least 8 nodes/species in one tissue exclusively (reproductive label, which indicates ovary and testis combined). The plotted values (log2(observed/expected+1)) was computed starting from the proportion of gains in each node/species belonging to the tested category (observed) and the proportion of all bilaterian-conserved orthogroups with a functional annotation belonging to the same category (expected). **d.** Examples of GO categories significantly enriched exclusively among the gains of one node/species. **e.** Left: Distribution of the proportion of orthogroups in each GO category with at least one tissue-specific, species-specific gain. The green area represents categories in the 95th percentile or above. Only GO categories including at least ten bilaterian-conserved orthogroups are shown. Right: Proportions of GO terms below or above the 95th percentile representing developmental functions. The reported p-value is computed out of the proportions of developmental functions in the 95th percentile coming from 1000 randomizations of the GO labels **(Extended Data Fig.**

**9c)**. Morphogenesis* stands for "anatomical structure formation involved in morphogenesis". See Methods for definition of developmental categories. **f.** Expression across tissues for human *FGF17* (left) and its deuterostome orthologs (right). **g.** Schematic summary of FGF17's function in the brain (based on [40]). Abbreviations: N: neural, T: testis, O: ovary, M: muscle, X: excretory system, E: epidermis, D: digestive tract, A: adipose.

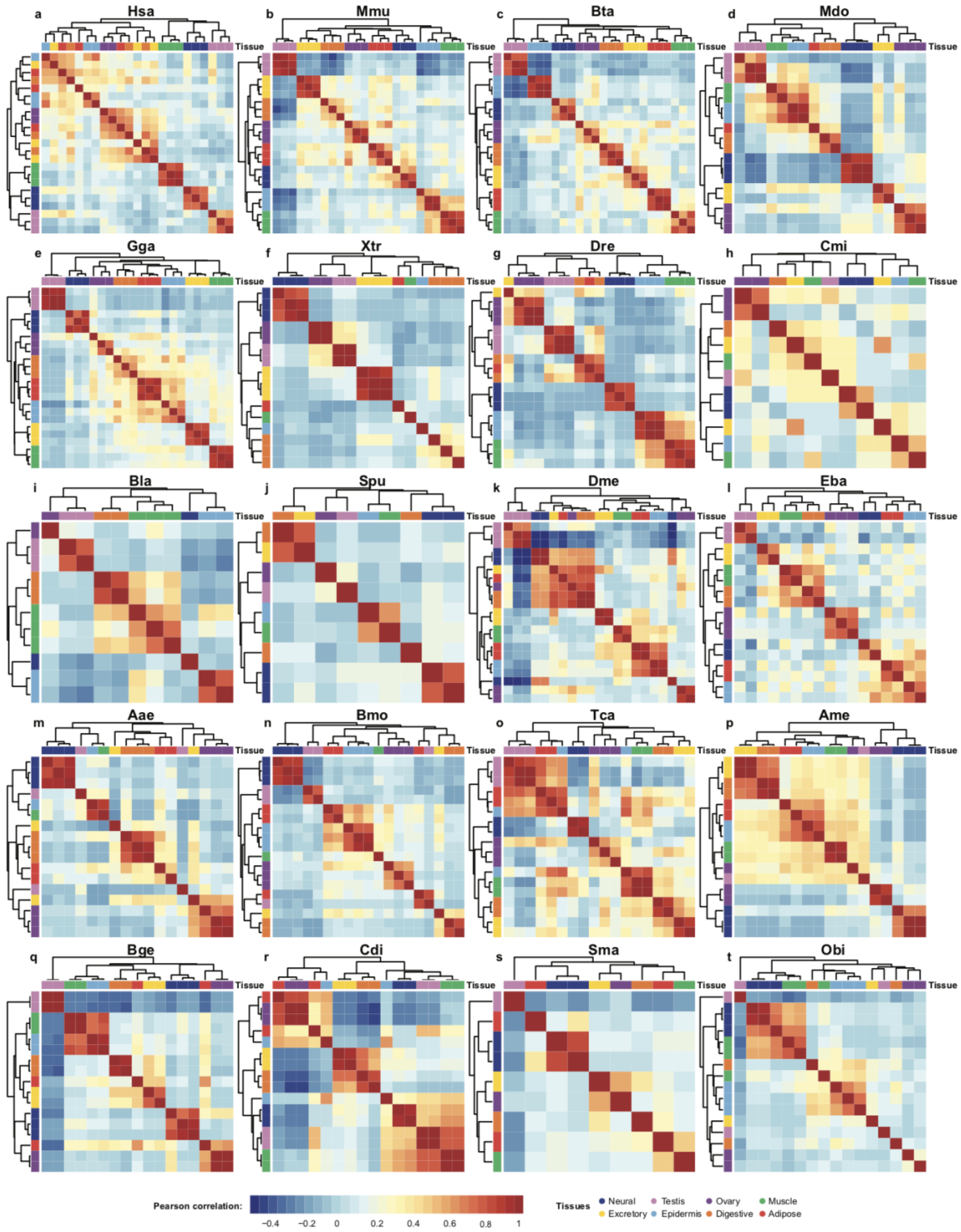
## Extended Data Figures



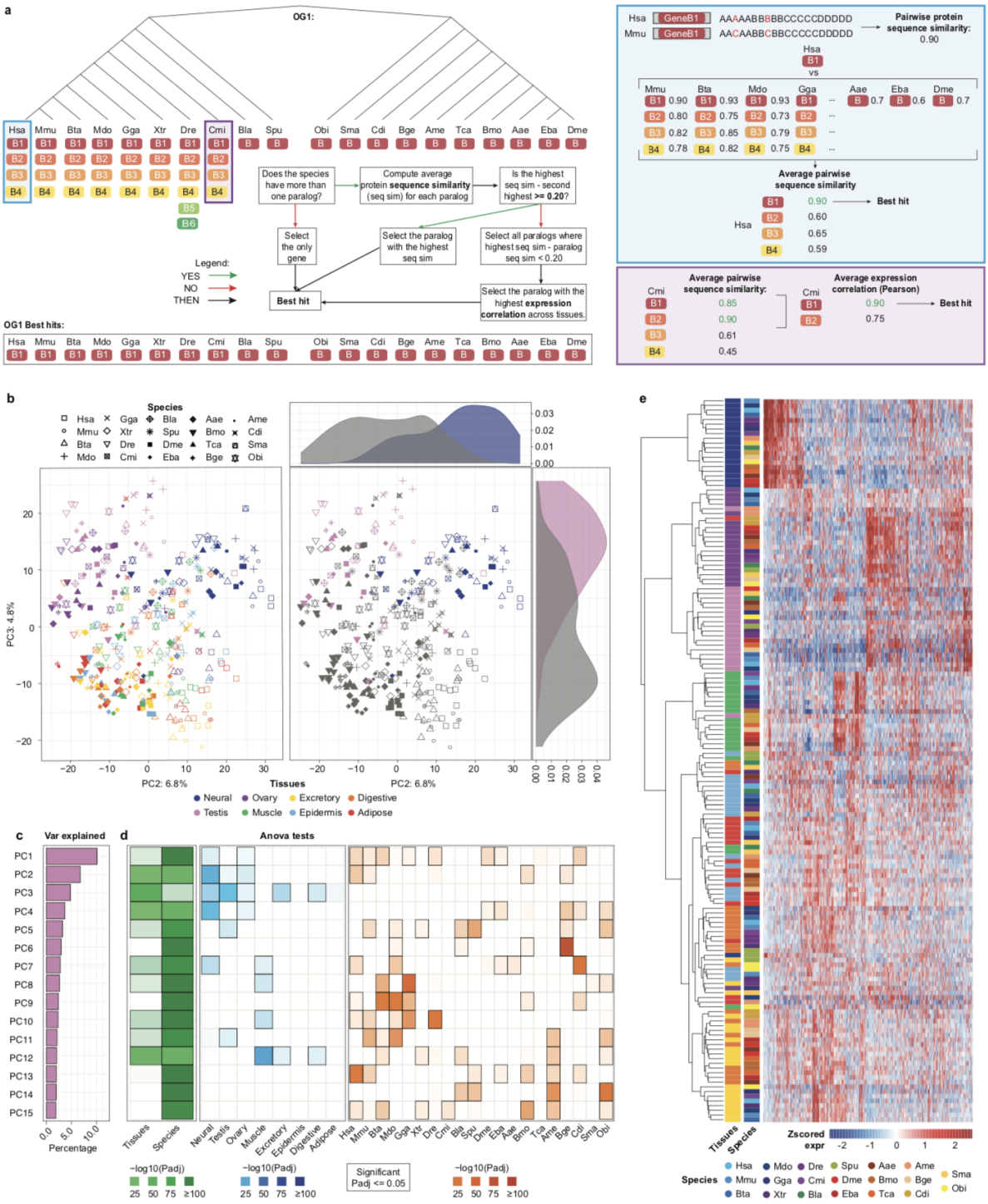

| Species | Resolved broken groups | Resolved chimeric genes | Unresolved chimeric genes |
|---|---|---|---|
| Bta | 132 | 4 | 2 |
| Mdo | 147 | 21 | 27 |
| Gga | 246 | 5 | 11 |
| Xtr | 88 | 5 | 24 |
| Dre | 103 | 0 | 5 |
| Cmi | 335 | 5 | 9 |
| Bla | 77 | 8 | 24 |
| Spu | 365 | 1 | 8 |
| Eba | 95 | 157 | 220 |
| Aae | 38 | 2 | 5 |
| Bmo | 329 | 16 | 40 |
| Tca | 72 | 57 | 56 |
| Ame | 184 | 11 | 26 |
| Bge | 120 | 11 | 18 |
| Cdi | 54 | 133 | 96 |
| Sma | 204 | 160 | 139 |
| Obi | 478 | 0 | 1 |




**Extended Data Fig. 1: a.** Schematic representation of broken (left) and chimeric (right) genes and how they potentially influence gene orthology inferences. **b.** Examples of a broken (left) and chimeric (right) genes corrected in the silkworm gene annotation. **c.** Statistics of corrected and unresolved broken and chimeric genes across all species. **d.** Barplot representing the number of bilaterian-conserved (red) or more recent (gray) protein coding genes across all species. The line plot represents the number of orthogroups (OGs) in which genes from each species are represented. **e.** Proportions of bilaterian-conserved orthogroups based on the number of species in which they are conserved.

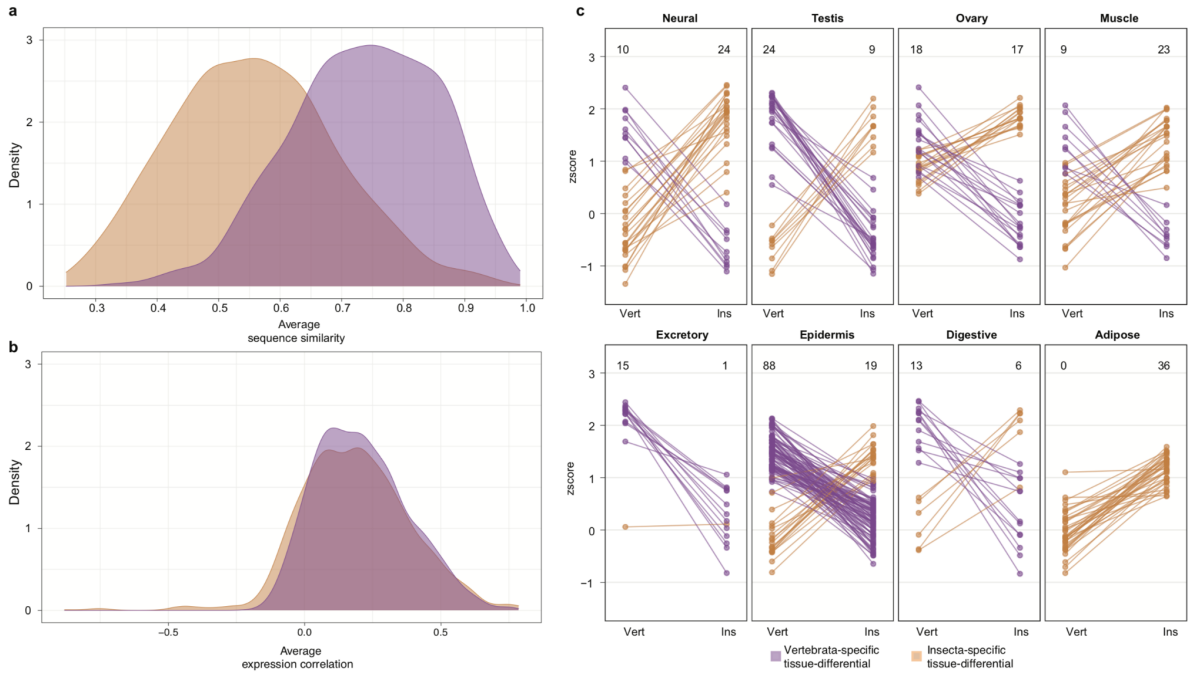



**Extended data Fig.2: a-t.** Clustering of each species' meta-samples based on their expression correlation. Expression correlation is represented by Pearson coefficient computed on log2(TPM+1) meta-sample expression values (see Methods), where only the 2500 genes with highest coefficient of variation were considered in each species. The heatmaps were generated by the *pheatmap* function in R with default clustering parameters.

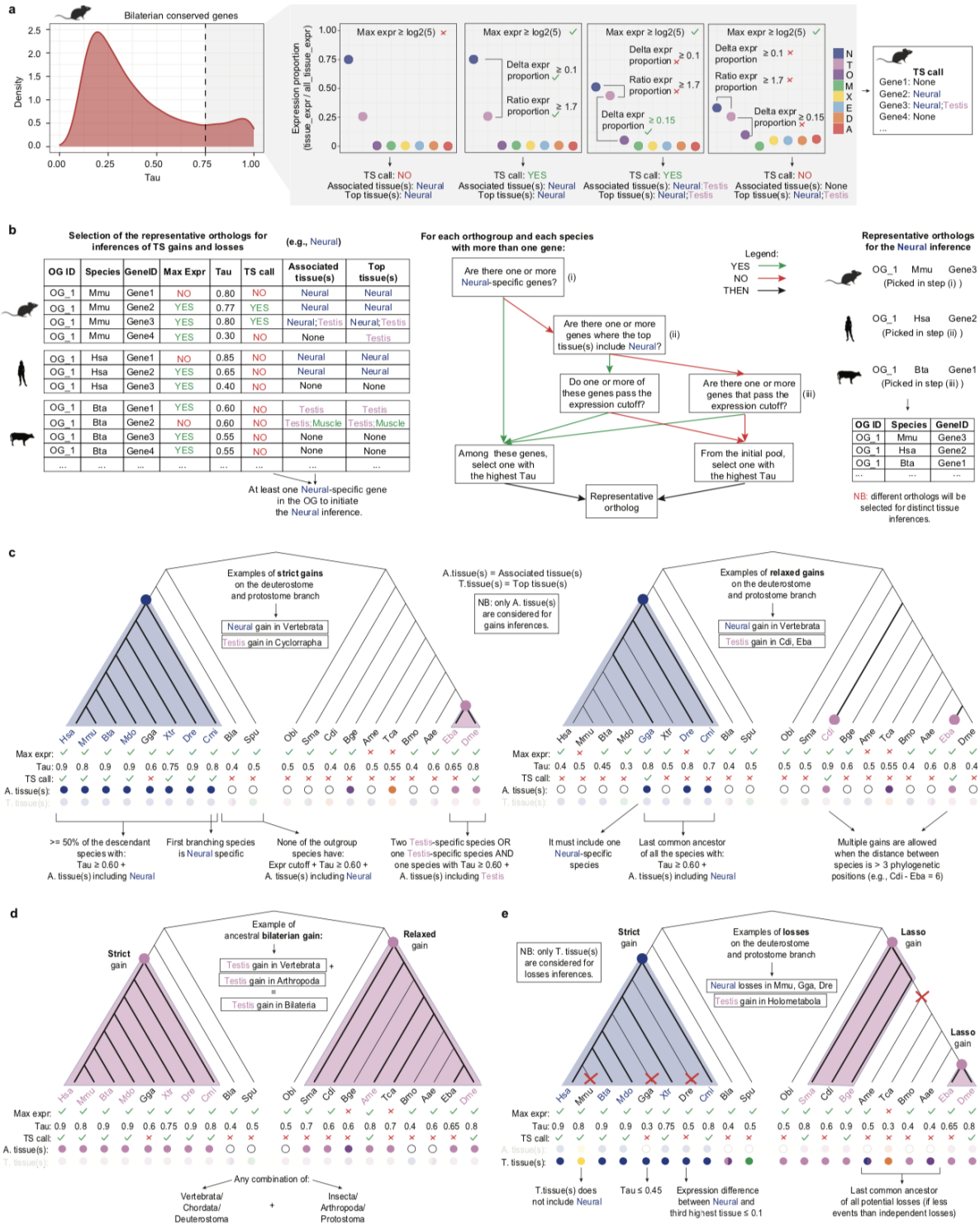



**Extended Data Fig. 3: a.** Scheme and relative example for the selection of bilaterian-conserved best-hit orthogroups (see Methods). **b.** Coordinates of the second (PC2; x axis) and third (PC3; y axis) components of a PCA performed on normalized gene expression values across meta-samples of best-hits orthogroups. Only the 2,436 orthogroups conserved in all species were considered. Tissue identity is represented by colors and species by shape. The left panel shows all tissues, while the right panel highlights neural and testis samples compared to

all others. Coordinate distributions of these three groups of meta-samples are shown on the side of the relative component. The percentage of variance explained by each PC is reported on the relative axis. **c.** Percentage of variance explained by the first 15 principal components from the PCA described in b. **d.** -log10(p-value) of ANOVA tests performed among the coordinates of the specified groups on each component. For the left panel (green) we tested if there was a significant difference between tissues or species groups. For the center and right panel (blue and orange) we tested if there was a significant difference between any query group (i.e., column) versus all other collapsed groups. All tests were performed with the *aov* function in R, and p-values were Bonferroni corrected. **e.** Heatmap showing the clustering of tissues and species (rows) based on the averaged expression across tissues of best-hits bilaterian conserved orthogroups (columns). Expression values were z-scored across tissues of the same species in order to minimize the inter-species variability. Only the 2,436 orthogroups conserved in all species were considered. The heatmap was generated by the *pheatmap* function in R with *ward.D2* clustering method. Tissue colors refer to panel b.

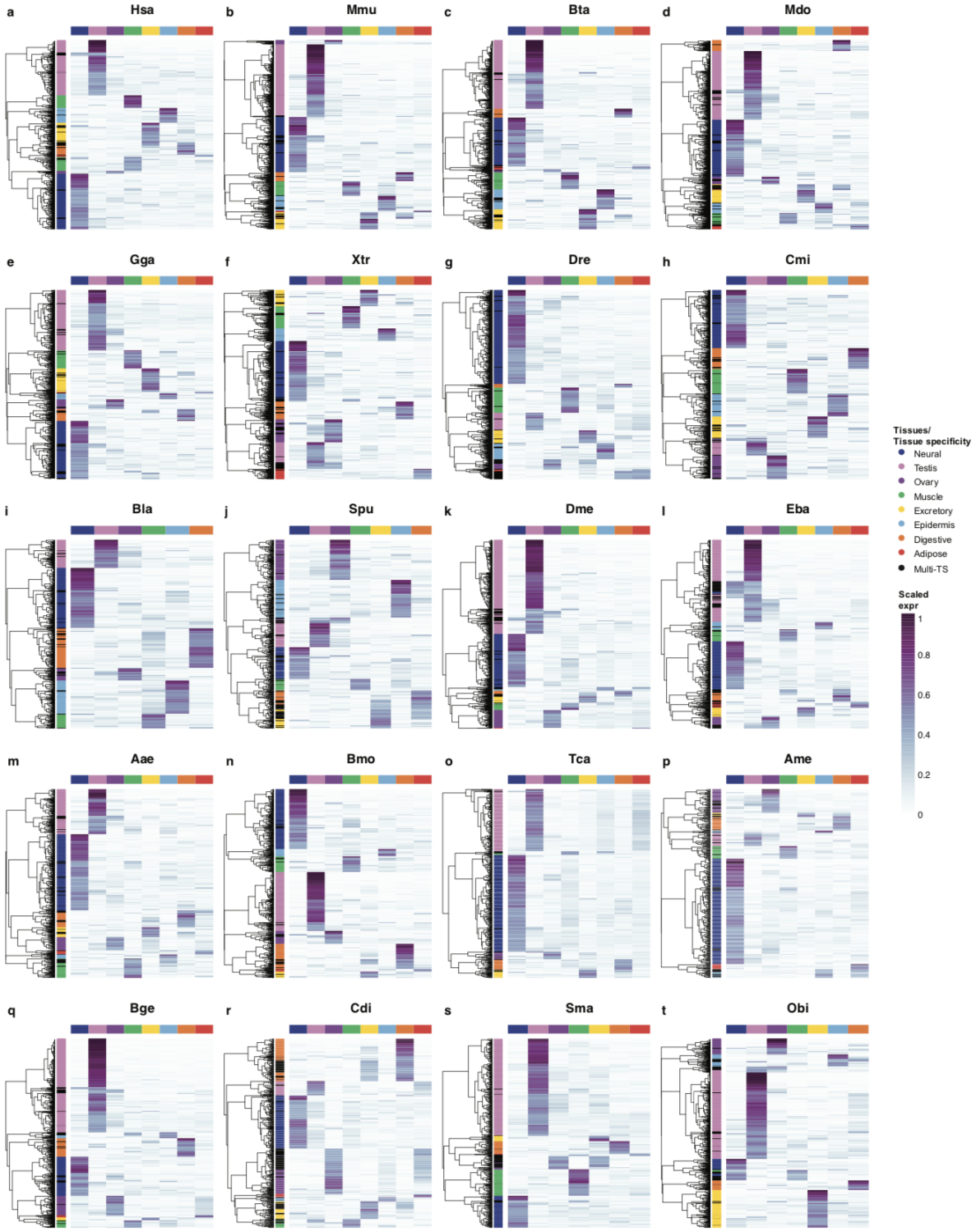



**Extended Data Fig. 4: a,b.** Distribution of average sequence similarity (c) and average expression correlation (Spearman's rho) (d) among best-hits orthogroups across vertebrates and insects (as in Fig. 3c,d) exclusively considering the 1,312 single copy orthogroups (i.e., with one representative gene in each species). **c:** Median expression (z-scored by species) of best-hits orthologs among vertebrate-specific (purple) and insect-specific (brown) tissue-differential orthogroups returned by the relative sPLS-DA run (see Methods).

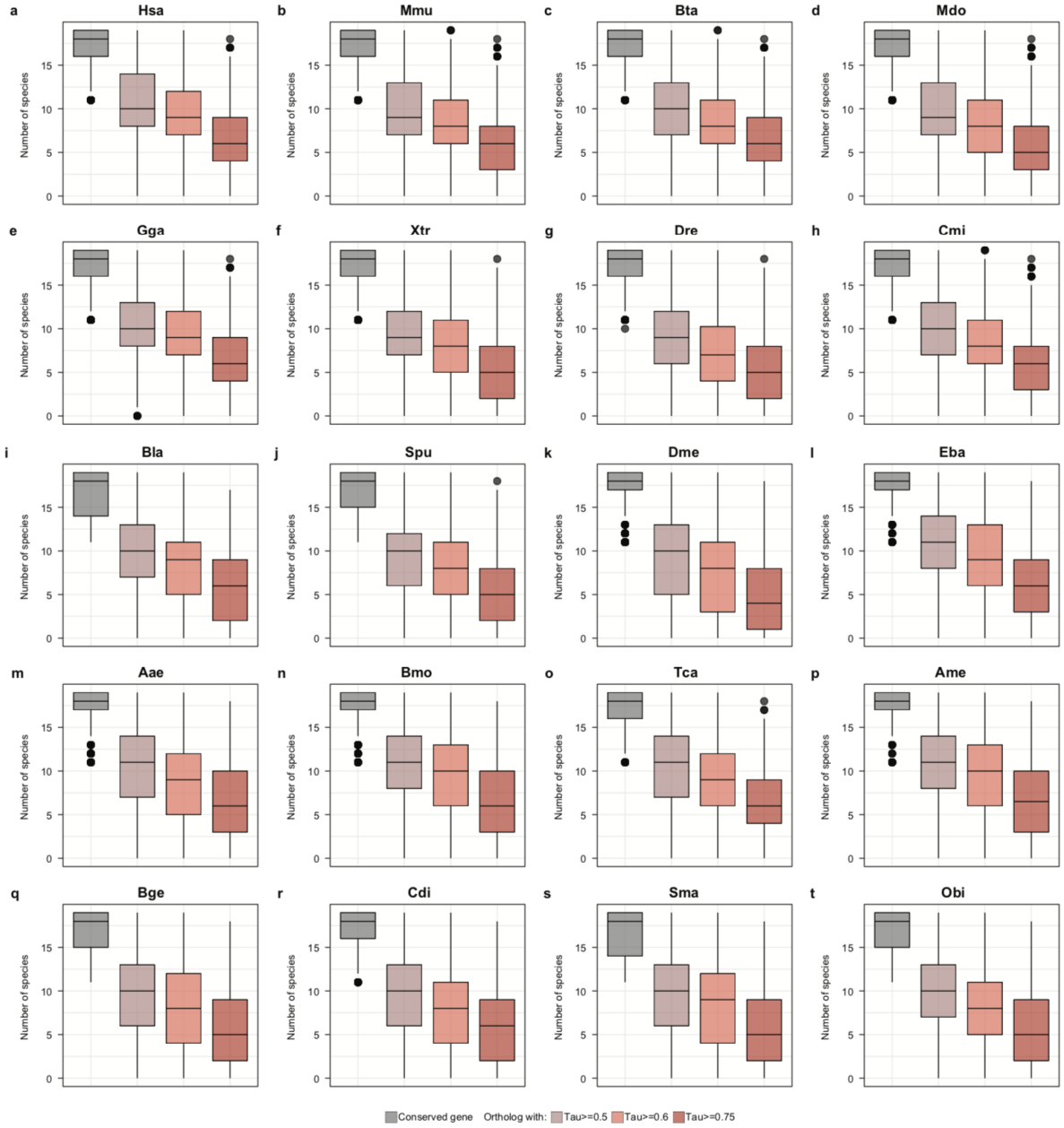



**Extended Data Fig. 5: a.** Schematic of the procedure adopted to associate all tissue-specific genes in each species (Tau ≥ 0.75) with the tissue(s) with tissue-specificity. This association (which we also evaluated for non-tissue-specific genes) will be considered for the inference of tissue-specificity gains (panels c,d). Additionally, we identified the top tissue(s) (i.e., the tissue(s) with the highest expression) for all bilaterian-conserved genes, which will be considered for the selection of the best orthologs and the inference of tissue-specificity losses

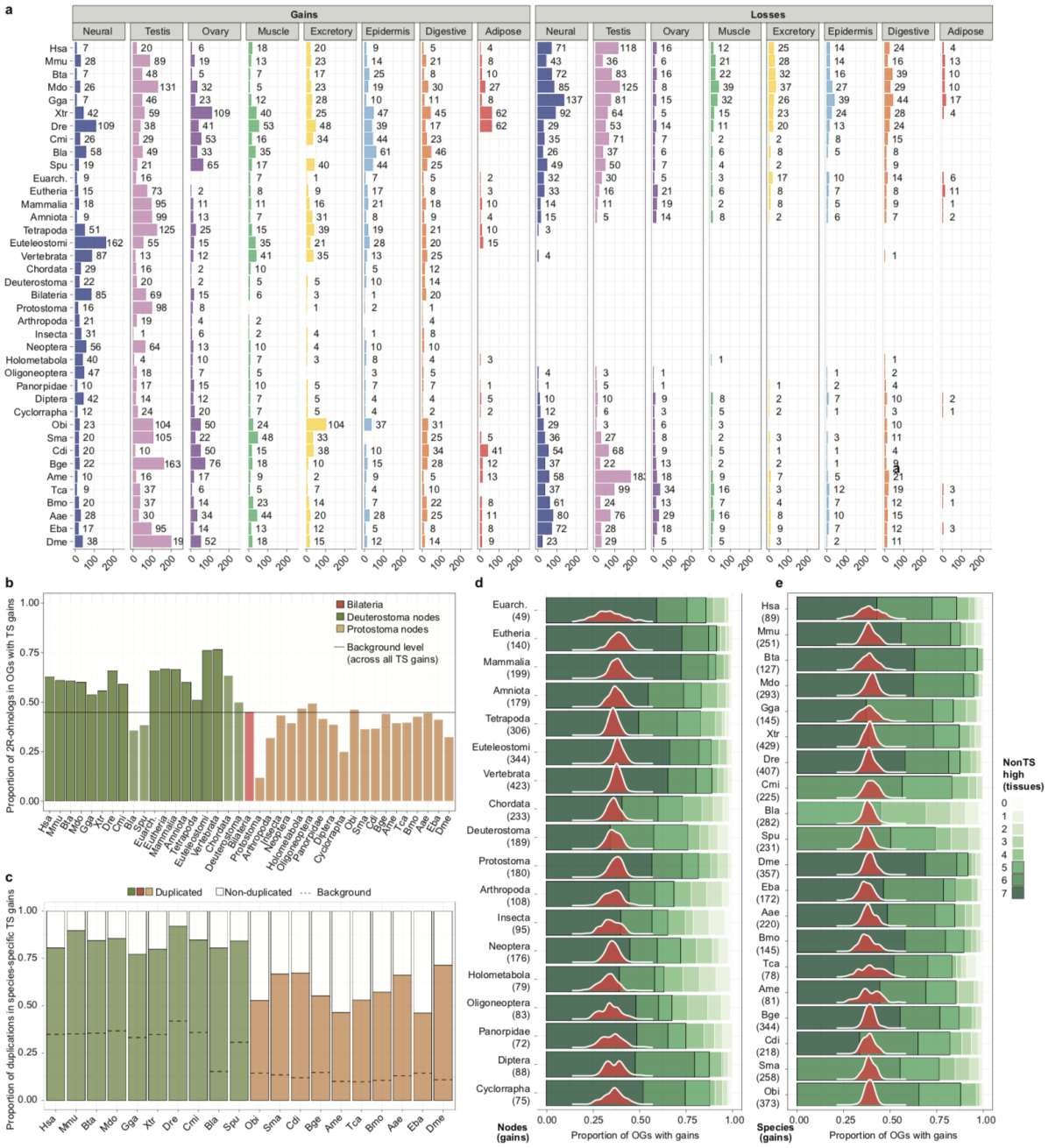
in each tissue (panel b and e, respectively). **b.** Example and schematic of the procedure adopted to select the best representative ortholog in each species for the inference of tissue-specificity gains and losses. **c.** Examples and criteria for the inference of tissue-specificity gains on either the deuterostome or protostome branches with the strict approach (left panel) and the relaxed approach (right panel). **d.** Example and criteria for the inference of bilaterian tissue-specificity gains. **e.** Examples and criteria to infer tissue-specificity losses.




**Extended Data Fig. 6: a-t.** Heatmaps showing the clustering of bilaterian-conserved, tissue-specific genes (rows) based on their expression proportion (tissue_expr / all_tissue_expr) across tissues (columns) in each species. The heatmaps were generated by the *pheatmap* function in R with default clustering parameters.

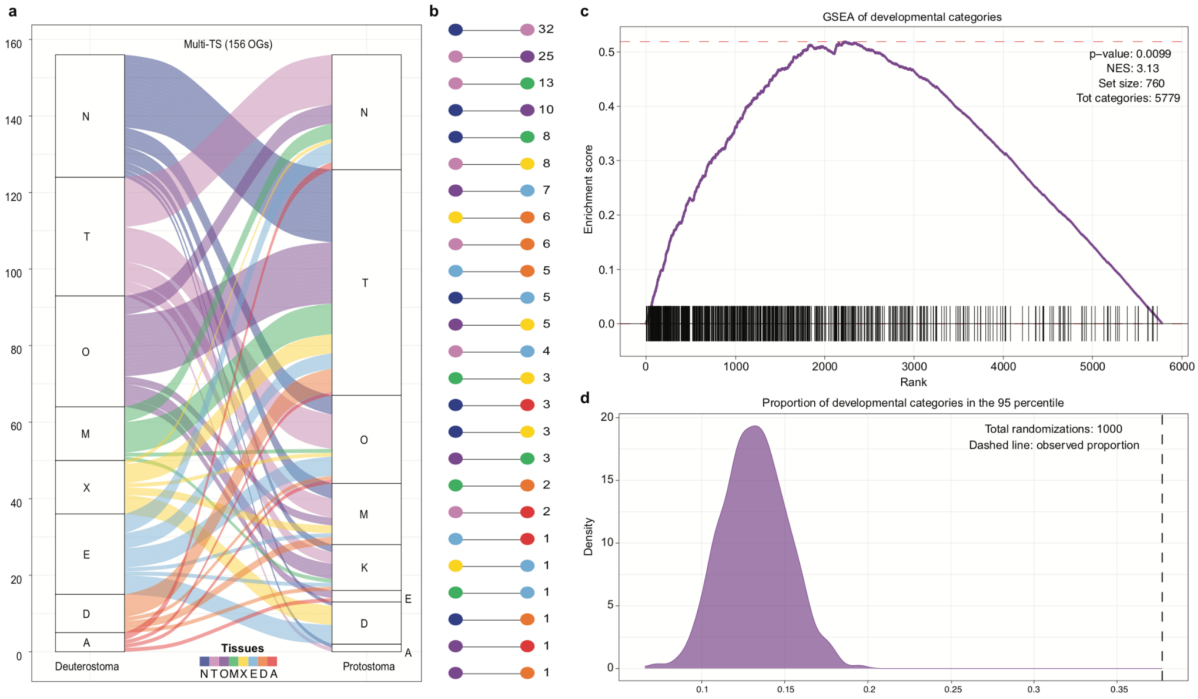



**Extended Data Fig. 7: a-t.** Distribution of the number of species in which each species' bilaterian-conserved, tissue-specific genes have at least one ortholog (gray) and this ortholog(s) has a Tau value higher than the specific cutoff (0.5, 0.6 and 0.75; other shades).

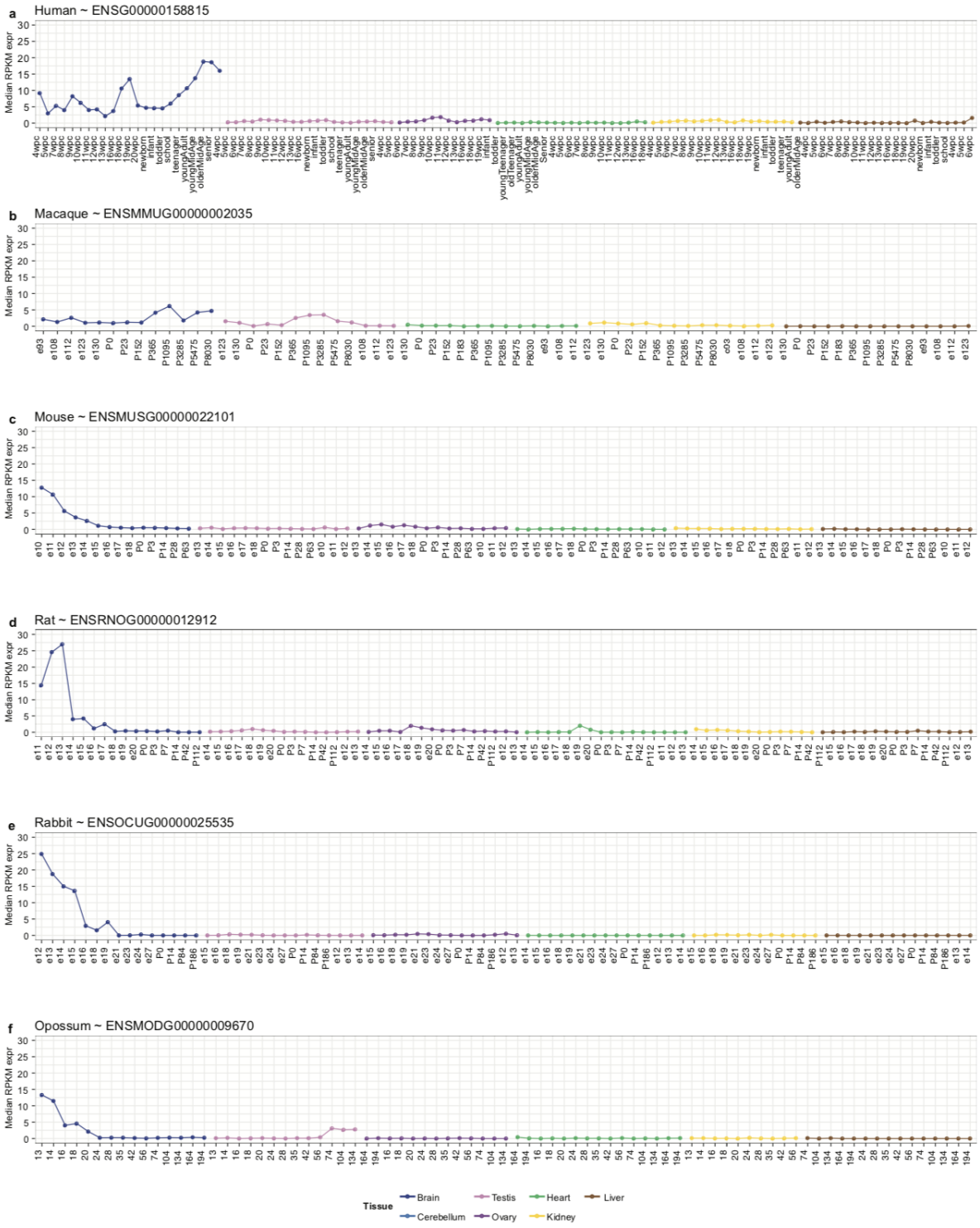



**Extended Data Fig. 8: a**. Barplots representing the number of inferred tissue-specificity gains (left) and losses (right) across all nodes/species (rows) and tissues (columns). **b.** Proportion of tissue-specificity gains in each node/species occurring in orthogroups that include 2R-onhologs. Deuterostome nodes/species are distinguished between those diverging before (transparent color) or after (full color) the two rounds of vertebrate WGDs. The black line represents the proportion of 2R-onhologs across all tissue-specificity gains. **c.** Proportions of duplicated (i.e., with at least one paralog) or non-duplicated (i.e., single-copy) genes with tissue-specific, species-specific gains in all species. The background line represents the overall

proportion of duplicated genes in each species. **d,e.** Same data represented in **Fig. 5f**, but plotted separately across all nodes (d) and species (e). Abbreviations: Euarch: Euarchontoglires.




**Extended Data Fig. 9: a.** Alluvia plot representing the bilaterian-conserved orthogroups with tissue-specificity gains in distinct tissues between deuterostome (left) or protostome (right) nodes and species. Only orthogroups with gains in exclusively one tissue on each branch were considered. **b**. Number of parallel tissue-specificity gains between the deuterostome and protostome branch for all pairs of tissues represented in panel a. **c.** Plot from a Gene Set Enrichment Analysis (GSEA) testing for over-representation of developmental categories (760 out of 5779) among categories with high proportions of orthogroups that undergo species-specific gains of tissue-specificity. **d.** Proportions of developmental GO categories among the top 5% (i.e. 95th percentile) of all GO categories ranked based on the proportions of their annotated orthogroups that undergo species-specific gains. The plotted values derive from 1000 randomization of the developmental labels among all GO categories, with the vertical dashed line corresponding to the observed proportion.




**Extended Data Fig. 10: a-f:** Expression values (RPKMs) for human FGF17 (a) and its orthologs in five mammalian species (b-f) across several developmental and adult timepoints in seven tissues. Data from [14].