Gene gain and loss across the metazoan tree of life

Rosa Fernández13,4 and Toni Gabaldón1,2,4*

Although recent research has revealed high genomic complexity in the earliest-splitting animals and their ancestors, the macroevolutionary trends orchestrating gene repertoire evolution throughout the animal phyla remain poorly understood. We used a phylogenomic approach to interrogate genome evolution across all animal phyla. Our analysis uncovered a bimodal distribution of recruitment of orthologous genes, with most genes gained very ‘early’ (that is, at deep nodes) or very ‘late’, representing lineage-specific acquisitions. The emergence of animals was characterized by high values of gene birth and duplications. Deuterostomes, ecdysozoans and Xenacoelomorpha were characterized by no gene gain but rampant differential gene loss. Genes considered as animal hallmarks, such as Notch/Delta, were convergently duplicated in all phyla and at different evolutionary depths. Genes duplicated in all nodes from Metazoa to phylum-specific levels were enriched in functions related to the neural system, suggesting that this system has been continuously and independently reshaped throughout evolution across animals. Our results indicate that animal genomes evolved by unparalleled gene duplication followed by differential gene loss, and provide an atlas of gene repertoire evolution throughout the animal tree of life to navigate how, when and how often each gene in each genome was gained, duplicated or lost.

The animal tree of life includes 36 main phyla that entail an impressive diversity in terms of morphology, physiology and lifestyles. This diversity is the result of radical morphological or physiological innovations such as the development of gonads, muscles or brains. Although the lineages in which these innovations appear can be identified with confidence, we know very little about the genomic changes underlying their appearance. Recent studies based on sequence comparison and clustering of up to 14 animal phyla1-3 have shown that thousands of homologous gene groups could be detected at the level of Opisthokonta, Holozoa, Metazoa, Eumetazoa or Bilateria, among other clades, supporting a shared ancestry of many of the genes present in extant animals and describing a minimal genome content for the hypothetical last common ancestor of each of these lineages. However, besides missing more than half of the metazoan diversity, none of these studies used a phylogenetic approach, and thus they provide limited resolution of the reconstructed gene evolutionary histories, hampering our understanding of the evolutionary dynamics governing gene repertoire evolution. Herein, we undertook a phylogenomic approach to gene family evolution across the animal tree of life including representatives of all animal phyla. Our methodological approach, centred around the use of gene phylogenies to infer pairwise orthology and paralogy relationships for all genes in all genomes, is robust to the effect of horizontal gene transfer, gene loss, genome completeness and variable genome size, factors that strongly affect analyses based on gene family inference. Since we inferred these orthology/paralogy relationships and their evolutionary dynamics for each gene in one species at a time, our results are independent for each lineage, thus providing a robust backbone of global evolutionary patterns of gene evolution across metazoans. Moreover, since orthologous genes have significantly more similar functions than paralogous genes4-5, our approach allowed us to shed light on the evolution of function across different phyla, even if putatively.

Results and discussion
A phylogenomic approach to gene repertoire evolution across the metazoan tree of life reveals unbalanced distributions of gene gains, losses and duplications. We used a phylogenetic approach to reconstruct evolutionary histories for each gene in each genome and infer pairwise orthology/paralogy relationships in a dataset composed of 231 genomes and transcriptomes. For this, we rooted the gene trees with the topology depicted in Fig. 1a, based on our current knowledge on animal phylogeny10-14. The results were not significantly different when rooting with alternative topologies (for example, considering the ‘Porifera-sister’ hypothesis; \(P > 0.05\)). Note that the orthology inference algorithm we used is robust to a non-resolved phylogeny (see Methods); nevertheless, our results are based on a species tree assumption and may vary following future reshaping of the animal tree of life. Remarkably, a consistent pattern of orthologue gain/loss was independently recovered in all lineages (Fig. 1b,c), which consisted of strikingly unbalanced distributions of gene gain, loss and duplication (Fig. 2a,b, see Glossary). The gene gain patterns displayed a bimodal distribution. Approximately one-quarter of the genes in each genome were gained or already present at the node Opisthokonta or gained at Holozoa (Fig. 1b, see Supplementary Mat. S3, S4). The following nodes from Filozoa to Nephrozoa exhibited much lower values of orthologue gain, up to 10%, a value that decreased further towards shallower nodes (Fig. 1b). Around one-third of each genome comprised lineage-specific genes (that is, genes that had no homologues in the rest of the taxa (Fig. 1b)), reflecting that lineage-specific gene gain plays a crucial role during animal genome evolution. Notably, Xenacoelomorpha, Deuterostomia and Ecdysozoa were defined by high values of orthologue gene loss, as compared to all earlier-splitting lineages, whereas virtually no orthologue gene gain was observed in any of the phyla included in these clades (less than 2%; Fig. 1c).

*Corresponding author.
†Bioinformatics and Genomics Unit, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain. †Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. †Present address: Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Barcelona, Spain. †Present address: Barcelona Supercomputing Centre (BSC-CNS) and Institute for Research in Biomedicine (IRB), Barcelona, Spain.
†e-mail: toni.gabaldon.bcn@gmail.com
High gene duplication at the origin of Metazoa. We next calculated genome-wide average ratios of gene duplication for each genome in each tree node (Figs. 1d and 2c). Our results recovered high duplication ratios at Opisthokonta and Holozoa, followed by a low ratio at Chonoanimalia that increased again at Metazoa. The following nodes were characterized by progressively descending ratios towards the tips of the tree, with those at Parahoxozoa and Bilateria being twice those at shallower nodes (Xenoceolomorpha, Deuterostomia, Protostomia, Lophotrochozoa and Ecdysozoa). Finally, lineage-specific duplication ratios were again high, consistent with widespread lineage-specific expansions for all taxa. These results show that both ancient duplication events and a high rate of retention of duplicate genes have contributed to an abundance of duplicate genes in animal genomes. Altogether our results support a scenario of high gene birth rate in the branch leading to metazoans: such is the case for tyrosine kinase receptors, transmembrane proteases and, remarkably, dynein, a family of cytoskeletal motor proteins that move along microtubules in cells and that drive the beat of eukaryotic cilia and flagella.

Orthology-informed putative functions of duplicated genes. To further understand animal gene repertoire evolution at a functional level, we performed an evolutionarily informed interrogation of the putative function of gene duplications, gains and losses in our dataset. The fraction of annotated genes ranged from 50–60% in some phyla (for example, Tardigrada and Rotifera) to 90–100% in others (for example, most deuterostomes), indicating a ‘hidden biology’ effect of non-annotated genes that is biasing our understanding of gene function evolution. This limitation notwithstanding, we need to focus our discussion on the annotated subset of genes. To attenuate this bias, we were careful to limit functional Gene Ontology (GO) term propagation to orthologous, not just homologous, genes, and to focus on high levels of the GO hierarchy, where function is more likely to be conserved45. In all phyla, orthologues gained or already present at Opisthokonta and Holozoa were significantly enriched in complex functions such as DNA and protein modification, signal transduction, cell division and embryogenesis and developmental structures such as muscle, mesoderm or...
neural system formation, these patterns being highly similar across phyla. In Choanimalia, enriched functions among gained genes were related to cilia assembly and movement and chemotaxis (Fig. 3b and Supplementary Mat.). In orthologues gained at Metazoa to Nephrozoa, functions related to synapsis establishment and nervous system development were enriched and prevalent in virtually

Fig. 2 | Gene gain and duplication ratios are high at deeper nodes, and gene loss at shallower ones. a–c. Box plot graphs of gene gain (a), loss (b) and duplication ratios (c) per node as inferred from dataset 1. The mean and standard deviation values are shown for each node.
all genomes, together with various functions related to morphogenesis and reproduction. Virtually no gene gain (and therefore no enrichment) was detected in shallower nodes. These results suggest that orthologues involved in such complex developmental features were gained very ‘early’ in evolution, and support that co-option of existing genes to novel functions was potentially a critical factor generating the increasingly complex morphological variation observed across animal phyla.

For duplications, GO enrichment analyses in most nodes revealed a strong correlation with enriched functions also enriched among gained genes (Supplementary Mat. 6, 7 and 9). In contrast, while no functional enrichment was detected in the few genes gained at relatively shallower levels due to their low number (for example, Deuterostomia, Protostomia, Xenacoelomorpha and so on), genes duplicated at these nodes were strongly enriched in functions related to the neural system, such as synopsis.
Fig. 4 | The core gene repertoire of metazoans includes genes from a plethora of KEGG pathways that have undergone different degrees of duplication. 

**a.** Heatmap of pathway conservation and duplication level in the core gene repertoire of metazoans. The reference pathways were selected from the KEGG pathways collection. The colour in each cell depicts the relative number of duplicates within each KEGG category found in each phyla as inferred for dataset 2 (for the total number of duplicates and specific KEGG pathway annotation per orthogroup, see Supplementary Mat. 9). The tree topology and colours for each clade are as in Fig. 1. The asterisks indicate specific pathways expanded in **b** and **c, d.** Percentage of each KEGG category annotated in all orthogroups from the metazoan core gene repertoire. **c, d,** Heatmaps of KEGG nervous system (**c**) and signal transduction (**d**) pathway conservation and duplication level in the core repertoire of metazoans. The colour in each cell is calculated as in **a.**
neurotransmission or neural development (Fig. 3b). From the recurrent enrichment of related terms across the metazoan tree, we hypothesize that the neural system has been continuously shaped during metazoan evolution, being refined independently in all animal phyla—even in those without a complex nervous system (as defined by morphology in other phyla) such as Porifera and Placozoa. This result was further validated with the finding of a high level of duplications in orthogroups from the meta-

Fig. 5 | Pairwise gene loss is pervasive across phyla. a, A chord diagram representation of the dynamics of gene loss observed between each pair of taxa measured as the percentage of genes in each genome that were lost, as inferred for dataset 3 (phyloome approach). Each seed phyla is represented by one colour. Thicker chords between two given phyla represent higher percentages of gene loss. The directionality of the chords is defined by each colour (that is, if a chord and a phylum share a colour, the percentage of gene loss is calculated on the basis of the genome size of that phylum). b, The percentage values of gene loss as inferred for dataset 3. The values are polarized with the genome size of each seed genome shown in the left-most column.
zoan core gene repertoire (see below), in which genes involved in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways related to neural system development and signal transduction were more highly duplicated relative to other pathways (Fig. 4a,c,d; see also Supplementary Mat. 9). We conclude that ancient gene duplicates related to neural activity may have been co-opted in a convergent manner to generate a growing neural complexity in animal phyla, while subsequent lineage-specific duplications potentially enabled an expanding structural plasticity, neuronal morphology and connectivity.

Rampant gene loss and differential retention of paralogues across animal lineages. On the other hand, we detected virtually no enrichment among genes lost at any node in any phyla, suggesting that gene loss and differential retention of paralogues reshaped the whole biology of the organisms at all evolutionary depths. When exploring absolute gene loss pairwise between taxa (that is, the total number of genes lost in one taxon compared to the other, without taking into account in which node they were lost), we observed that values in all comparisons were remarkably high (Fig. 5). Particularly high values correlated with some recalcitrant positions in the animal tree of life, such as Xenoturbellida (that is, Xenoturbellida has lost a high percentage of orthologues in each pairwise inference with the remaining genomes independently) and Placozoa and Cnidaria (with Cnidaria having lost 44% of orthologues relative to Placozoa, and Placozoa 26% relative to Cnidaria). Our results thus may indicate a pervasive effect of extensive hidden paralogy hampering the phylogenomic reconstruction of deep animal relationships. The relevance of gene loss to metazoan evolution has already been anticipated\cite{10,11}. For instance, homology searches and clustering approaches\cite{12} defined a core bilaterian gene repertoire shared between Lophotrochozoa and Deuterostomia that had been apparently lost in Ecdysozoa and Platyzoa (rotifers, flatworms and their kin), and which is putatively involved in the control of homoeostasis and multicellularity. This study also observed that the number of gene families shared within Ecdysozoa and within Platyzoa was much lower than within Lophotrochozoa or Deuterostomia, pointing to extensive differential gene loss. Nevertheless, a phylogenomic-centred approach based on orthology/paralogy inference had never been explored to date. In addition to placing these findings in a broader evolutionary context, our results further validate the role of differential gene loss following gene duplication as a prevailing force shaping metazoan gene repertoire evolution.

The role of gene duplication in the evolution of the metazoan core gene repertoire. We explored pathway conservation in the metazoan core gene repertoire (defined as the orthogroups gained at the branch leading to metazoa or more basal branches, present in at least 80% of the phyla, and in a minimum of 50% of phyla within Xenacoelomorpha, Deuterostomia, Lophotrochozoa and Ecdysozoa). This core repertoire consisted of 3,196 orthogroups, from which 2,479 yielded KEGG annotations (Fig. 4 and Supplementary Mat. 9). Overall, half of the orthogroups in the metazoan core gene repertoire with KEGG annotations in the main KEGG categories represented (clustered in the overarching categories of metabolism, environmental information processing, genetic information processing, organismal systems and cellular processes) corresponded to genetic information processing, including transcription, translation, DNA replication and repair and folding, sorting and degradation (Fig. 4a). The duplication levels of the genes in these categories for some animal phyla were similar to the ones in the outgroups. In contrast, the remaining categories showed a higher number of duplications in metazoans than in their unicellular outgroups, reinforcing the role of gene duplication and not merely the presence or absence of genes, as a main driver of genome evolution across metazoans. Remarkably, genes involved in pathways related to the development of organismal systems (that is, excretory, nervous, circulatory and excretory systems, among others) showed higher levels of duplication in metazoans, underpinning again the link between gene duplication and morphological complexity. From all animal phyla, deuterostomes (and, within deuterostomes, Craniiota in particular) exhibited the highest levels of gene duplication, pointing again to co-option of gene duplicates as a driving force increasing morphological complexity. In contrast, Porifera also exhibited high levels of gene duplication in multiple KEGG pathways, which indicates that regulatory complexity or other evolutionary processes may be required for such an increase in morphological disparity. Finally, pathways that included genes with particularly high levels of duplication in metazoans were related to neural system development and signal transduction (Fig. 4c,d), indicating again—as shown above in the enrichment analyses—that the evolution of the neural systems involves ancient genes and is potentially strongly influenced by gene duplication across metazoans.

Conclusions Altogether, our results provide support for a consistent pattern of gene repertoire evolution across the animal tree of life, characterized by a bimodal distribution of gene gain and duplications with peaks at both the deepest and phylum-level nodes, and rampant gene loss in the branches leading to the most diverse clades. We show that waves of acquisition and loss of orthologous genes are coupled to different levels of gene duplication, with high gene gain values associated with high duplication ratios and high levels of orthologous gene loss preferentially occurring at nodes where duplication ratios are low. These results underscore the key role of gene duplications but also gene loss in animal genome evolution. In other words, it is not the presence of key innovations that explains animal origins, but rather their complex evolutionary dynamics, providing an arena that potentially fuelled an increase in cell type and tissue complexity throughout the animal kingdom. Our results therefore challenge the idea that the gene repertoire underlying the vast morphological disparity displayed by animals has become more complex through time (for example, refs. 11–13). Losses are compensated by gene gains and duplications in each lineage independently, and thus the size of the gene repertoire is not necessarily altered. Consequently, an increase in morphological complexity must result from other evolutionary phenomena, such as co-option following gene duplication, integration of genes in functional modules or through the potential implication of non-coding sequences leading to an increase of genomic regulatory complexity (for example, through the contribution of transposable elements in cis-regulatory elements\cite{14,15}). All in all, our results point towards a highly complex evolutionary history of each individual gene and lineage. The correlation between morphological and genetic complexity across the animal tree of life—epitomized by the early gain and continuous duplication of genes involved in the evolutionary development of the nervous system—is therefore strongly challenged. We emphasize that phylogenomic-centred studies are consequently most needed to further understand gene repertoire evolution in non-model organisms.

Methods Taxon sampling. We included in our study 216 genomes and transcriptomes representing all animal phyla, with each phylum being represented by a number of taxa ranging from 1 to 22 (Supplementary Mat. 1). Special attention was paid to maximize lineage representation within each phylum by including representatives of the main clades below the level of phylum whenever possible. Furthermore, we included 15 outgroup species, comprising representatives of the clades Fungi, Nucleotida, Ichthyophooida, Dermocystida, Corallochytrea, Filasterida and Chooanofflagella (Supplementary Mat. 1). Our dataset comprises a total of 231 genomes and transcriptomes considering ingroup and outgroup species.

Glossary and definitions. Gene gain: the branch leading to a clade where a gene was gained corresponds to the branch leading to the last common ancestor of all the orthologues of that gene (that is, homologous genes derived from a speculation...
event) present in our dataset. Thus, by definition that gene lacks orthologues outside the clade, and it has orthologues in at least another lineage within the clade. Note that the presence of subsequent gene duplications and gene losses of that gene family is maintained through the clade.

Gene loss: we consider that a gene was lost in the branch leading to a given clade when that gene has no homologues within that clade, but that homologues of that gene are present in the closest sister lineage of that clade. Note that gene losses must necessarily be evaluated from the perspective of genes present in other clades.

Duplication ratio: the total number of gene duplications mapped to a given branch in the species tree divided by the total number of genes.

Lineage-specific gene: a gene that does not have any detectable homologous orthogroups outside its lineage, regardless of the taxonomic level of that lineage (for example, phylum-specific refers to genes with no detectable homologous genes outside that phylum). Note that it is consequently conditioned on the taxon sampling of each study.

Inference of orthogroups and pairwise orthology relationships. Most comparative genomic studies on metazoan genome evolution so far have focused on gene family inference— that is, the inference of homologous genes clustered in orthogroups (a set of genes orthologous to one another from different species) that descended from a gene in the last common ancestor of a set of species) (for example, ref. 15). While this approach can definitely be most helpful in understanding gross patterns of change in the gene repertoire (for instance, the presence/absence of a gene family in a clade), it does not provide a detailed scenario of how homologous genes are related, particularly through duplication (orthologous) or a duplication event (that is, they are paralogous)16. For instance, this second approach would allow one to understand not only the presence/absence of orthogroups in lineages, but also how (speciation or duplication), when (positioning the event in a node) and how often a certain gene has been gained, duplicated or lost in the context of a certain phylogeny. Furthermore, gene family inference was shown to be the most adequate approach for the metatranscriptomes mapped, as such, their inflation value in the Markov clustering (MCL) step will infer a higher number of families as we increase its value17.

To overcome this methodological limitation, and with the goal of generating a compelling atlas of gene repertoire evolution across the metazoan tree of life that allows us to understand how, when and how often each gene in each taxon was gained, duplicated or lost, we designed a two-step orthology inference protocol. First, we inferred orthogroups for the whole dataset (dataset 1 hereafter, n = 251) with OrthoFinder v. 2.3.19. Since higher inflation values during MCL clustering can result in higher numbers of inferred gene families (that is, an orthogroup that has undergone a high number of duplications will be split into several smaller orthogroups), we chose an inflation value of 1.5 as a conservative approach, as discussed in other studies (for example, ref. 15). The inference of orthogroups allowed us to compare our results with previous studies and, combined with the pairwise orthology inference as described below, to characterize which gene families had undergone higher rates of duplications, expansions or losses. Second, after inferring orthogroups we inferred pairwise orthology relationships for each gene in each genome/transcriptome. Since orthogroups in this whole dataset (dataset 1) contained up to 24,178 sequences, we inferred gene trees with DendroBLAST21—implemented in the OrthoFinder2 pipeline to overcome computational constraints. A species-overlap algorithm, as implemented in ETE v321, was used to infer orthology and paralogy relationships from the multiple alignments and profile hidden Markov model searches have been proved to perform similarly to BLAST searches for phylostratigraphic analysis33, methods based on BLAST continue to be one of the most adequate approaches due to their high robustness, particularly for long regions of protein sequences. In brief, for each protein encoded in the first genome (referred to as the ‘seed’ hereafter), we performed a BLAST search against the complete proteome database built from all of the 18 genomes and transcriptomes, which included a total of 397,671 proteins. The results were filtered using an e-value threshold of 1 × 10⁻³ and a minimum overlapping region of 0.5 (that is, a minimum of 50% overlap between the query and the hit sequence). Multiple sequence alignments (MSAs) were reconstructed in both sequence orientations using three different programs: MUSCLE v3.8.31, MAFFT v76712 and Kalcing18. The resulting six alignment sets were then combined using M-COFFEE19. A trimming step based on non-consistent regions in the consensus alignment was performed using trimAl v1.31 (consistency-score cutoff 0.1667, gap-score cutoff 0.9). The best-fitting model was selected by reconstructing neighbour-joining trees as implemented in BioNJ17 using seven different models (JT, LG, WAG, Blosum62, MIREV, WAG+disp, Dayho). The best-scoring consensus tree event (that is, the one resulting in an Akaike information criterion was selected for tree reconstruction. Trees were reconstructed using PhyML22. Four rate categories were used and invariant positions were inferred from the data. Branch support was computed using an approximate likelihood ratio test based on a chi-square distribution. The same procedure was repeated with the remaining 17 genomes and transcriptomes to generate a total of 18 independent paralogy trees (one for each phylome reconstruction). Resulting trees and alignments—a total of 260,000 considering all phylomes—can be visualized in PhylomeDB 4.023 (http://phylomedb.org).

Similarity searches are not infallible. Any method holds a different compromise between different sources of errors and entails a balance between false positives and negatives (for example, ref. 15). The PhylomeDB approach is not based on an all-by-all BLAST, but it searches for homologues (and then infers orthology/paralogy relationships) for only one proteome at a time, the so-called seed. For instance, in our study we analysed 18 different phylogenomes, meaning that we used each taxon as an seed in an independent phyloeme analysis. This has an important implication: since we are using the same species tree to polarize our results, the values of gene gain and loss for each node are informed by completely different similarity searches (that is, one per taxon). Therefore, this approach supports the high robustness of our results since the values of gene gain, duplication and loss are strongly similar for each node regardless of the seed taxon, as shown in Figs. 1 and 2 (see also Supplementary Mat. 3). This stands even if we compare the percentages of gene gain and loss per node inferred for fast-evolving lineages (such as nematodes, see the phyloeme of Caenorhabditis elegans) with those for slow-evolving ones, such as Drosophila melanogaster. Moreover, in the case of duplications (including expansions), the PhylomeDB pipeline applies a clustering approach, where individual gene trees are merged into orthogroups (that is, a minimum of 50% of the genes). This means that if a BLAST search fails to recover homologues of a given sequence based on length or sequence similarity, which could result in splitting a single gene family into two or more if we were using a program based on gene family inference, it will be corrected through this step. The OrthoFinder2 approach is likewise based on BLAST reciprocal best hits, the main difference from phylomeDB being that the former is based on an all-by-all approach (that is, inference of gene families through a single BLAST search). From the orthology inference programs based on the inference of gene families, OrthoFinder2 has been shown to be among the best ones, as discussed in ref. 18. We favoured OrthoFinder over the commonly used software OrthoMCL23 because despite using a similar approach (BLAST search followed by an orthology inference pipeline), it has been shown that with the OrthoMCL algorithm short sequences suffer from a low recall rate and long sequences suffer from a low precision (that is, many short sequences fail to be assigned to an orthogroup while many long sequences are assigned to the incorrect orthogroup)24.

Benchmarking studies also support the robustness of our chosen methodology. The tree topology obtained with the best methods to correctly retrieve highly curated individual gene trees and their orthology/paralogy relationships in a thorough benchmarking study25, with very high rates of precision (that is, positive predicted value) and recall (that is, true positive rate) (see Fig. 3). Likewise, the score of OrthoFinder2 was among the highest as benchmarked in refs. 15,18 (see Fig. 2 in ref. 18).

While alternative methodologies to BLAST based on Smith–Waterman alignments and profile hidden Markov model searches have been proved to perform similarly to BLAST searches for phylostratigraphic analysis methods based on BLAST continue to be one of the most adequate approaches due to a good trade-off between computational time and error rate. Two parameters...
have been shown to particularly affect BLAST searches: evolutionary rate and length of sequence. For instance, it has been shown that higher error rates in relative age estimation are associated with higher evolutionary rates and shorter sequence lengths. To further test whether our results were robust to these two factors, we selected four of the phylogmes representing three of the main lineages in the metazoan tree of life differing in their genome size (measured as the number of proteins in their proteome) and overall evolutionary rate: Craniana as a representative of Deuterostomia, Mollusca as a representative of Lophotrochozoa, Arthropoda as a representative of a slow-evolving Ecdysozoa, and Nematoda as a representative of a fast-evolving Ecdysozoa. We took all sequence alignment (MSA) alignments of proteins where homologues were detected following the PhylomeDB pipeline and calculated two parameters. First, we calculated the average identity score (average IdSc hereafter) as a proxy of evolutionary rate. The IdSc for each possible pair of sequences in the alignment is the number of identical residues divided by the sum of the lengths of the two sequences. We have calculated the average IdSc for each MSA with trimAL v1.2.1-
flag -ident. The IdSc ranges from 0 (highly dissimilar sequences) to 1 (highly similar sequences). Second, we calculated the average protein length in each MSA. All phylogmes showed a highly similar distribution pattern (Supplementary Mat. 8).

To test whether our estimates of gene gain and loss per node were affected by these two parameters, for each phylogme we divided all MSAs into four quartiles based on the average IdSc and the length of the MSA. Next, for each quartile in each phylogme we realigned orthology/paralogy inference and gene gain, duplication and loss patterns, totalling 32 new analyses for each phylogme. We statistically compared the values of gene gain, duplication and loss between the quartiles and the full set of MSAs for each phylogme (that is, the original analysis) by means of one-way analysis of variance. Specifically, we compared whether the values of gene gain, duplication and loss were different between the five treatments (the full MSA and the four quartiles of each MSA) for both the average IdSc and the length of the MSA. None of the analysis of variance analyses revealed significant differences between any of the treatments (P > 0.05; Supplementary Mat. 8). Therefore, we can conclude that after accounting for error and biases due to evolutionary rate and length of sequence, our results remained largely unchanged, which attests to the high robustness of our approach.

GO term annotation and enrichment analysis. To assign GO terms to all genomes and transcriptomes in datasets 1 and 2, GO terms based on the orthology relationship were propagated with eggNOG-mapper. For that purpose, we selected the eukaryotic eggNOG database (euNOG) and prioritized coverage (that is, the degree of functional annotation in each genome or transcriptome, we calculated the average protein length in each genome or transcriptome) to propagate GO terms, thus prioritizing coverage (that is, providing more accurate information about gene function than other databases. All orthologues were selected to propagate GO terms, thus prioritizing coverage as in the annotation with eggNOG-mapper. With the goal of understanding the degree of functional annotation in each genome or transcriptome, we calculated the percentage of genes annotated. In addition, we clustered together all genomes/transcriptomes per phyla and calculated the mean percentage of annotation to enable their comparison.

We next checked functional enrichment of genes gained at different nodes with FatIGO. For that purpose, we performed all the genes inferred as gained in each node and tested enrichment against the remaining genes in the genome of that taxon. This analysis was performed independently for all taxa. Whenever enrichment analysis was not performed, the values of enriched GO terms were summed up and visualized in REVIGO. The same pipeline was repeated with the gene expansions in each genome in our core dataset to further understand GO term enrichment in each phyla independently compared to the background (that is, the sum of the non-expanded genes in each of the scrutinized genomes), and with genes lost in each node per taxon as well.

Metazoan core gene repertoire inference and annotation. To infer the metazoan core gene repertoire, we selected the orthogroups as inferred in OrthoFinder2 that were present in at least 80% of all phyla, and were represented in at least 50% of the phyla of each main lineage (Xenacoelomorpha, Deuterostomia, Lophotrochozoa and Ecdysozoa). To further understand the pathway conservation of this core, KEGG pathway annotation and mapping was performed with KEGG Mapper and BlastKOALA. KEGG pathway annotation for 30 selected general categories was combined with the total number of duplication values (that is, not referred to as duplications in any specific node) for each annotated orthogroup in each category and visualized through heatmaps with the heatmap R package.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
All data, code and supplementary information are available in the manuscript. The supplementary materials are deposited in the Harvard dataverse repository https://doi.org/10.7910/DVN/ZKDAE2. The accession numbers for all taxa included in each analysis are indicated in Supplementary Mat. 1. All phylomes can be accessed at PhylomeDB 4.0 under the phylome numbers 431, 462, 747, 778, 812, 819, 824, 875, 888, 957, 950 and 953 (that is, example, http://www.phylomedb.org/ phylome_431 for direct access).

Received: 29 April 2019; Accepted: 21 November 2019; Published online: 28 January 2020

References
1. Paps, J. & Holland, P. W. H. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nat. Commun. 9, 1730 (2018).
2. Richter, D. J., Fozouni, P., Eisen, M. B. & King, N. Gene family innovation, conservation and loss on the animal stem lineage. elife 7, e34226 (2018).
3. Domazet-Loso, T., Brąkowiak, J. & Tautz, D. A phylostratigraphic approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23, 533–539 (2007).
4. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. Nat. Rev. Genet. 14, 360–366 (2013).
5. Laurent, J. M. et al. Humanization of yeast genes with multiple human orthologs reveals principles of functional divergence between paralogs. Preprint bioRxiv https://doi.org/10.1101/668335 (2019).
6. Cannon, J. T. et al. Xenacoelomorpha is the sister group to Nephrozoa. Nature 530, 89–93 (2016).
7. Laumer, C. E. et al. Support for a clade of Placozoa and Cnidaria in genes with minimal compositional bias. elife 7, e36278 (2018).
8. Shen, X.-X., Hittinger, C. T. & Rotas, A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1, 0126 (2017).
9. Laumer, C. E. et al. Revisiting metazoan phylogeny with genomic sampling of all phyla. Proc. R. Soc. B 286, 20190831 (2019).
10. Luo, Y.-J. et al. Nemertean and phoronid genomes reveal lophotrochozoan evolution and the origin of bilaterian heads. Nat. Ecol. Evol. 2, 141–151 (2018).
11. Albalat, R. & Cañestro, C. Evolution by gene loss. Nat. Rev. Genet. 17, 379–391 (2016).
12. Lynch, M. & Conery, J. S. The origins of genome complexity. Science 302, 1401–1404 (2003).
13. Shen, X.-X. et al. Tempo and mode of genome evolution in the budding yeast Cryptococcus. Genome Res. 28, 1176–1188 (2018).
14. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. 18, 71–86 (2017).
15. Sebé-Pedrós, A. et al. Early metazoan cell type diversity and the evolution of multicellular gene regulation. Nat. Ecol. Evol. 2, 1176–1188 (2018).
16. Gabaldón, T. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 9, 235 (2008).
17. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, 157 (2015).
18. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238 (2019).
19. Li, L. OrthoMCL: identification of orthologous groups for eukaryotic genomes. Genome Res. 13, 2178–2189 (2003).
20. Kelly, S. & Manni, P. K. DendroBLAST: approximate phylogenetic trees in the absence of multiple sequence alignments. PLoS ONE 8, e6537 (2013).
21. Huerta-Cepas, J., Dopazo, J. & Gabaldón, T. ETE: a python environment for tree exploration. BMC Bioinformatics 11, 24 (2010).
22. Huerta-Cepas, J., Dopazo, H., Dopazo, J. & Gabaldón, T. The human phylogene. Genome Biol. 8, R109 (2007).
23. Marlič, T., Katja, T. C., Goto, T., Satoh, N. & Rokhsar, D. S. A new spiralian phylogeny places the enigmatic arrow worms among gnathostomes. Curr. Biol. 29, 312–318.e3 (2019).
24. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M. & Gabaldón, T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 42, D897–D902 (2014).
25. Altenhoff, A. M. et al. Towards standardized benchmarking in the quest for orthologs. Nat. Methods 13, 425–430 (2016).
26. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113 (2004).
27. Katoh, K. & Standley, D. M. MAFFT: iterative refinement and additional methods. Methods Mol. Biol. 1079, 131–146 (2014).
28. Lassmann, T. & Sonnhammer, E. L. L. Kaligünv and Mumsa: web servers for multiple sequence alignment. *Nucleic Acids Res.* **34**, W596–W599 (2006).

29. Wallace, I. M., O’Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**, 1692–1699 (2006).

30. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

31. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).

32. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

33. Moyers, B. A. & Zhang, J. Toward reducing phylostratigraphic errors and biases. *Genome Biol. Evol.* **10**, 2037–2048 (2018).

34. Moyers, B. & Zhang, J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.* **33**, 3031–3031 (2016).

35. Huerta-Cepas, J. et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

36. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

37. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

38. Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578–580 (2004).

39. Supek, F., Bošnjak, M., Skunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).

40. Kanehisa, M. Enzyme annotation and metabolic reconstruction using KEGG. *Methods Mol. Biol.* **1611**, 135–145 (2017).

41. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**, 726–731 (2016).

**Acknowledgements**

We are grateful to M. Marcet-Houben and I. Julca for multiple discussions that contributed to greatly improve this study. R.F. was funded by a Juan de la Cierva-Incorporación Fellowship (Government of Spain) and a Marie Skłodowska-Curie Fellowship (747607). T.G. group receives funding from the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) grants “Centro de Excelencia Severo Ochoa 2013-2017” SEV-2012-0208 and BFU2015-67107 co-funded by the European Regional Development Fund (ERDF); from the CERCA Programme/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857, and a grant from the European Union’s Horizon 2020 research and innovation programme under the grant agreement ERC-2016-724173 the Marie Skłodowska-Curie grant agreement no. H2020-MSCA-ITN-2014-642095.

**Author contributions**

R.F. and T.G. developed the overall conceptual approach and analysis. R.F. compiled and analysed the data. R.F. and T.G. wrote the manuscript. T.G. supervised the study.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at [https://doi.org/10.1038/s41559-019-1069-x](https://doi.org/10.1038/s41559-019-1069-x).

Correspondence and requests for materials should be addressed to T.G.

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. © The Author(s), under exclusive licence to Springer Nature Limited 2020.
Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics
For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code
Policy information about availability of computer code
Data collection
Accession numbers of all taxa included in the study are indicated in Suppl. Inf. S1.

Data analysis
PhylomeDB v.4, MUSCLE v3.8 13, MAFFT v6.712b 14, Kalign, M-COFFEE, BioNJ, phyML, ETE v.2, fatiGO, REVIGO, OrthoVenn, OrthoFinder

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data
Policy information about availability of data
All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Accession numbers of all taxa included in the study are indicated in Suppl. Inf. S1.

Field-specific reporting
Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences  
- Behavioural & social sciences  
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf
Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Study description | Phylogenomic approach to understand genome evolution across the Metazoa Tree of Life. |
|-------------------|---------------------------------------------------------------------------------------|
| Research sample   | Taxa included in the present study were selected based on availability of genomes/transcriptomes and inclusion of all main animal lineages. |
| Sampling strategy | Samples were downloaded from the public databases. |
| Data collection   | Samples were downloaded from the public databases (Uniprot, NCBI SRA server, etc.). |
| Timing and spatial scale | NA |
| Data exclusions   | No data was excluded. |
| Reproducibility   | All attempts to reproduce the analyses were successful. |
| Randomization     | Randomization is not relevant for this study. All samples were grouped in phyla/superphyla for some analyses based on our current knowledge about animal biodiversity. |
| Blinding          | Blinding is not relevant for this study. All samples were grouped in phyla/superphyla for some analyses based on our current knowledge about animal biodiversity. |
| Did the study involve field work? | Yes | No |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Involved in the study |
|---------------------------------|-----------------------|
| n/a                             | Antibodies           |
| n/a                             | Eukaryotic cell lines |
| n/a                             | Palaeontology        |
| n/a                             | Animals and other organisms |
| n/a                             | Human research participants |
| n/a                             | Clinical data        |

| Methods                        | Involved in the study |
|--------------------------------|-----------------------|
| n/a                            | ChIP-seq              |
| n/a                            | Flow cytometry        |
| n/a                            | MRI-based neuroimaging |