

Widespread patterns of gene loss in the evolution of the Animal Kingdom

Cristina Guijarro¹, Peter W.H. Holland², Jordi Paps^{*1,2,3}

1. School of Biological Sciences, University of Essex, Colchester CO4 3SQ

2. Department of Zoology, University of Oxford, Oxford OX1 3SZ

3. School of Biological Sciences, University of Bristol, Bristol BS8 1TQ

*Corresponding author

The Animal Kingdom shows an astonishing diversity, the product of over 550 million years of animal evolution. The current wealth of genome sequence data offers an opportunity to better understand the genomic basis of this disparity. Here we analyse a sampling of 102 whole genomes including >2.6 million protein sequences. We infer major genomic patterns associated with the variety of animal forms from superphylum to phylum level. We show a remarkable amount of gene loss that occurred during the evolution of two major groups of bilaterian animals, Ecdysozoa and Deuterostomia, and further loss in several deuterostome lineages. Deuterostomes and Protostomes also show large genome novelties. At the phylum level flatworms, nematodes and tardigrades show the largest reduction of gene complement, alongside gene novelty. These findings paint a picture of the evolution within the Animal Kingdom in which reductive evolution at protein-coding level played a major role in shaping genome composition.

The Metazoa encompass an astonishing diversity of body forms. More than 30 animal phyla have been defined, the evolutionary relationships between which are well understood in broad outline, although some are still a matter of debate¹⁻⁴. Understanding how their genomes have evolved can help us to better comprehend the origin of this disparity and reconstruct their evolutionary history. The Metazoa comprise sponges, ctenophores, cnidarians, placozoans, and bilaterians, with most of animal diversity found in the last of these. The Bilateria can be split into three major groups or superphyla —Deuterostomia, Lophotrochozoa and Ecdysozoa⁵, the latter two forming the Protostomia.

Gene losses and gains play major roles in evolution. The gain of new functions via assembly of modules from older genes or emergence of de novo coding regions has been proposed to be important during major evolutionary transitions such as the origin of animals⁶⁻⁸. Gene loss has been associated with loss of anatomical structures in evolution, consistent with views that evolution can lead to both increases and decreases in complexity⁹⁻¹¹. Previous studies have shown the prevalence of gains and losses of genes and protein domains in the dawning of different groups of animals¹⁰⁻¹⁴. The increasing availability of genome data is giving new opportunities to investigate animal genome evolution. For example, recent analyses have shown the importance of using large taxon sampling and a range of outgroups to reconstruct the minimum protein coding genome present in the ancestor of a clade^{6,15,16}.

To analyse origins of genes during the early evolution of mammals and Metazoa, a bioinformatics pipeline was introduced that used state-of-the-art methods of homology assignment^{15,16}. We have previously outlined the limitations to the approach and differences to other methods such as phylostratigraphy^{15,16}. The approach focuses on protein-coding genes, but other genomic elements (non-coding RNA genes, regulatory regions, transposable elements, etc) most likely also contributed to the diversification of metazoans. Here we apply a new version of this pipeline to a large collection of metazoan genomes (59 genomes from 16 animal phyla) and a greatly expanded representation of outgroups (43 genomes) specifically to investigate patterns of gain and loss of genes across major lineages of bilaterian animals. The

use of complete genome sequences is particularly key to determining gene loss, since its inference from incomplete sources is problematic.

The pipeline developed is given in the supplementary materials: Material & Methods and Supplementary Figure 1. Briefly, we assembled a dataset of 102 previously sequenced eukaryotic genomes (Supplementary Table 1 and Supplementary Figure 2), chosen for their phylogenetic position and quality as assessed using BUSCO¹⁷ (Supplementary Figure 4). Over 2.6 million proteins were compared using BLAST¹⁸ all-vs-all, and clustered with MCL¹⁹ into homology groups (HG). An HG is a group of protein-coding genes that differ from others consistently, independently of their mechanism of origin (divergence, *de novo* origin etc). The extent of gene misassignment in HG clustering was assessed using metazoan and eukaryotic BUSCO gene sets for benchmarking, as well as performing receiver operating characteristic analyses (Supplementary Information).

The HG were analysed in a MySQL database, which tabulates all species in the study classified following phylogenetic relationships. For each node of the phylogenetic tree, MySQL will find HG that are gained or lost by combining taxon presence/absence in each member of that clade. For example, an HG present only in species from the clade Vertebrata is considered a vertebrate novelty.¹⁵ (Supplementary Figure 5, Supplementary Data 1 and 2). ‘Novel HG’ (denoted +) are sets of related genes that emerged in the stem lineage or last common ancestor (LCA) of an ingroup, ‘Core novel HG’ (++) are Novel HGs highly retained in the ingroup (refractory to gene loss), ‘Lost HG’ (denoted -) specify HGs lost on the stem lineage of the clade (prior to the LCA), and ‘Core lost HG’ (--) are Lost HG that are highly retained in outgroup taxa. We propose that both categories of ‘core’ HG perform essential biological functions in the groups they are found, underpinning their preservation. However, the values of Core Novel HG are also affected by the number of genomes included in a clade (e.g. a clade with two genomes will display higher values of Core Novel HG than a clade with 10 representatives), and their evolutionary relatedness (clades composed by closely related genomes will show higher proportions of Core Novel HG. Core Novel HG for all nodes were further validated by BLAST against the RefSeq database²⁰ (Supplementary Data 4 and 6). In the case of phyla with a single genome sampled (e.g. rotiferans, orthonectids, brachiopods), HG values may not be representative of

the group; these values are not shown although these genomes are still important to infer HG categories in sister groups.

Figure 1 shows the numbers of Novel HG (+), Core novel HG (++), Lost HG (-) and Core lost HG (--) inferred for major evolutionary branches across the Bilateria. Values in the LCA of Metazoa are consistent with previous work¹⁵⁰, with minor discrepancies explained by expanded taxon sampling; for example, the number of metazoan Novel HG (+25) remains the same. Looking at the patterns of gene gain, Bilateria show the largest number of Novel HG (+1699) among the major animal clades indicating extensive origin of novel gene types. Bilaterians are characterised by a centralised nervous system with anterior elaboration (a ‘brain’). Consistent with this morphological change, our results reveal nervous system functions amongst novel bilaterian genes including genes encoding a diversity of neuropeptide receptors (e.g., *orexin*, *neuropeptide FF* and *neuropeptide Y*) and transcription factors (*oligodendrocyte transcription factor 3*, protein turtle and homeobox protein prospero; Supplementary Data 4). Despite the high levels of gene novelty, no Core novelties were detected suggesting genomic flexibility after the origin of Bilateria.

Further gene novelty is inferred in the evolutionary lineages leading to protostomes (+734) and deuterostomes (+280); within protostomes, lower novelty is detected on the stem lineages of lophotrochozoans (+60) and ecdysozoans (+97) nodes (Figure 1). Our sampling does not include representatives of Scalidophora (Priapulida, Loricifera, and Kinorhyncha) which are the sister group to the ecdysozoans sampled in this study. Low numbers of Core Novel HG are seen in the other major bilaterian clades (Figure 1 and Supplementary Table 3, Supplementary Data 6). At the phylum level (Figure 2), particularly high levels of phylum-specific novelty are found in flatworms (+856), nematodes (+1187) and tardigrades (+945); the latter HG are all shared between two tardigrade genomes, including a version of *Hypsibius dujardini* annotated excluding potential contaminations^{22,23}. Many vertebrate novelties are related to immunity and signalling pathways (Supplementary Data 3). Figure 3 shows the most abundantly gained and lost molecular functions assigned by gene ontologies²¹⁰ (GOs) across all clades (Figure 3,

Supplementary Data 3, Supplementray Figures 8-15); however, we caution against over-interpretation of these since there is a bias in the quantity and quality of GO annotations between organisms. Core novel HG show GPCRs, receptors, and nucleic acid binding as some of the functions gained more often across clades (Figure 3A). Most clades show a broad spread of GO functions gained, while others concentrate gains in a few (e.g., echinoderms gained GPCRs and transporters, panarthropods gained transfer/carrier proteins and receptors).

Gene loss shows a particularly interesting pattern. We deduce that very extensive gene loss, in excess of 1000 HG, occurred on the stem lineages of each of the three major bilaterian superphyla: Ecdysozoa (-4677), Lophotrochozoa (-1760) and Deuterostomia (-4231) (Figure 1). These values are in excess of the amount of gene novelty, suggesting that loss of genes or gene functions was important in shaping the distinctive biological characters of these clades. Similar patterns are not seen in the bilaterian node, where novelty is deduced to be more dominant in genome evolution (+1699 vs -745). The loss of bilaterian genes in the ecdysozoan lineage has been pointed in previous studies^{24,25}. The HG lost in ecdysozoans include several membrane proteins and signal transduction components; deuterostomes lost HGs include functions related to transmembrane proteins such as those that form gap junctions in invertebrates (Supplementary Data 3). Many phyla within these groups also show high degrees of HG loss, with many losses being of genes otherwise highly retained in outgroups (Core Lost HG): echinoderms (--680), urochordates (--845), nematodes (--401), tardigrades (--967), flatworms (--858), and annelids (--3179) (Figures 1 and 2). Occasional examples of convergent gene loss are detected, such as protein LEG1 homolog involved in multicellular development and small ubiquitin-related modifiers (SUMO), both lost in echinoderms and urochordates. Among molecular functions more often lost (Figure 3B) are transfer/carrier proteins, ribosomal proteins, and nucleic acid binding proteins; there are also differences between clades, with urochordates, ambulacrarians and tardigrades losing genes with very diverse GO classifications.

Here we used a comprehensive taxon sampling together with comparative methods to infer the patterns of gene gains and losses of ancestral animal genomes. Our analyses support a major role of gene novelty in the origins of animals and bilaterians, consistent with origin of new biological characters, but in contrast we also deduced there was an exceptional amount of gene loss on the stem lineages of the major bilaterian supergroups: Ecdysozoa, Lophotrochozoa and Deuterostomia. Further gene loss occurred in the evolution of phyla within these groups, although in some cases loss seems largely balanced by novelty. The three animal phyla with the largest levels of gene loss - flatworms, nematodes, and tardigrades - also show remarkable levels of genomic novelty. This pattern could be explained by high gene turnover in the genome of their respective ancestors. Alternatively, it may be influenced by interaction between their biology and our methodology: these are 'fast-evolving' lineages, thus some of their genes may be highly divergent and have formed their own clusters. Gene loss has been suggested as an important force in the evolution of different groups of organisms, including in Metazoa and Fungi¹⁴. This study highlights the importance of rich taxon sampling to understand the evolution of animals and sheds new light on the part that reductive evolution of gene complements has played in evolution of animal diversity.

Material & Methods

Genome collection

Canonical proteins from whole genomes were downloaded (Supplementary Table 1) including 59 animal species, from ~16 phyla (Supplementary Figure 2), and 43 non-animal eukaryotes. Genome annotation completeness was determined by BUSCO analysis¹⁷ using the eukaryote dataset of 303 orthologs (Supplementary Figure 3). The cut-off criteria for genome quality was absence of more than 15% BUSCO orthologs in animals, unless the genome in question shared a phylum/subphylum with another high quality (>85% complete) genome.

Comparative genomics

The selected genomes were compared using a reciprocal BLASTp¹⁸ of all sequences against all sequences (Supplementary Figure 1), with an e-value threshold of $\times 10^{-6}$. Markov Cluster Algorithm¹⁹ (MCL) was used to infer HGs from the BLAST output, with default inflation parameter ($I=2$). GOs²¹ were assigned to

the different HGs using the Uniprot API for the sequences downloaded by Uniprot²⁶ (Supplementary Data 2 and 3). Missing GOs were annotated using Interproscan²⁷.

Definition of homology groups

Following a consensus phylogeny from well-known studies (Supplementary Figure 2)^{4,28,29}, the different types of HG (novel, core novel, etc) were inferred for the different clades through an in-house custom MySQL database (Supplementary Data 1-3; Supplementary Figure 4). For the phyla with only two taxa, the definition of core novel HG and novel HG meet the same criteria, and so the HGs values are equal. The HG values for phyla represented by a single species (rotifers, orthonectids and brachiopods) are not comparable with other groups due to an excess of orphan genes, but they are useful to establish values for the other clades. For each type of HG, GOs were mined from Uniprot or obtained using InterProScan. Not all HGs were assigned GOs due to limited annotations in lesser studied phyla (e.g., Lophotrochozoa). A reliability check was performed on the core novel HGs to assess their absence in the outgroups, using larger sampling. The RefSeq protein database²⁰ was used, which has a comprehensive taxon and sequence sampling derived from studies ranging from single-gene analyses to transcriptomes and complete genomes. All the sequences from Novel Core HG were searched in RefSeq using BLASTp, (Supplementary Data 3); expected cut-off value was 1e-6 and identity >50% in the BLAST parameters, and the option “-negative_glist” was used to exclude hits against the ingroup in the output files. Only in one case, Novel Core HG in cephalochordates, did two genes recover hits in other animals.

References:

1. Egger, B. et al. *Curr. Biol.* **25**, 1347–1353 (2015).
2. Jékely, G., Paps, J. & Nielsen, C. *Evodevo* **6**, 1 (2015).
3. Marlétaz, F., Peijnenburg, K.T.C.A.C.A., Goto, T., Satoh, N. & Rokhsar, D.S. *Curr. Biol.* **29**, 312–318.e3 (2019).
4. Giribet, G. *Org. Divers. Evol.* **16**, 419–426 (2016).
5. Halanych, K.M. *Annu. Rev. Ecol. Evol. Syst.* **35**, 229–256 (2004).
6. Richter, D.J., Fozouni, P., Eisen, M.B. & King, N. *Elife* **7**, 1–3 (2018).

- 176 7. Paps, J. *Integr. Comp. Biol.* **58**, 654–665 (2018).
- 177 8. Grau-Bové, X. et al. *Elife* **6**, (2017).
- 178 9. Lankester, E.R.**12**, (Macmillan and Company: 1880).
- 179 10. Denoeud, F. et al. *Science* (80-.). **330**, 1381–1385 (2010).
- 180 11. Tsai, I.J. et al. *Nature* **496**, 57–63 (2013).
- 181 12. Zmasek, C.M. & Godzik, A. *PLoS Comput. Biol.* **8**, (2012).
- 182 13. Moore, A.D. & Bornberg-Bauer, E. *Mol. Biol. Evol.* **29**, 787–796 (2012).
- 183 14. Albalat, R. & Cañestro, C. *Nat. Rev. Genet.* **17**, 379–391 (2016).
- 184 15. Paps, J. & Holland, P.W.H. *Nat. Commun.* **9**, 1–8 (2018).
- 185 16. Dunwell, T.L., Paps, J. & Holland, P.W.H. *Proc. R. Soc. B Biol. Sci.* **284**, (2017).
- 186 17. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E.M. *Bioinformatics*
187 **31**, 3210–3212 (2015).
- 188 18. Camacho, C. et al. *BMC Bioinformatics* **10**, 421 (2009).
- 189 19. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
- 190 20. Pruitt, K.D., Tatusova, T. & Maglott, D.R. *Nucleic Acids Res.* **35**, 61–65 (2007).
- 191 21. Carbon, S. et al. *Nucleic Acids Res.* **45**, D331–D338 (2017).
- 192 22. Arakawa, K. *Proc. Natl. Acad. Sci.* **113**, E3057–E3057 (2016).
- 193 23. Yoshida, Y. et al. *PLoS Biol.* **15**, (2017).
- 194 24. Simakov, O. et al. *Nature* **493**, 526–531 (2013).
- 195 25. Luo, Y.-J.J. et al. *Nat. Ecol. Evol.* **2**, 141–151 (2018).
- 196 26. Bateman, A. et al. *Nucleic Acids Res.* **45**, D158–D169 (2017).
- 197 27. Jones, P. et al. *Bioinformatics* **30**, 1236–1240 (2014).
- 198 28. Laumer, C.E. et al. *Curr. Biol.* **25**, 2000–2006 (2015).
- 199 29. Kocot, K.M. *Org. Divers. Evol.* **16**, 329–343 (2016).
- 200 30. Kocot, K.M. et al. *Syst. Biol.* **66**, 256–282 (2017).

201

202 End Notes

203 **Supplementary information** is available in the online version of the paper.

204

205 **Data availability:** Publicly available genomes are listed in the supplementary materials
206 (Supplementary Table 1).

207

208 **Code availability:** All the scripts used in this study can be found in:
209 <https://github.com/CristiGuijarro/ComparativeGenomics>

210

211 **Acknowledgements:** The authors would like to thank Nacho Maeso for comments on the
212 manuscript. CG and JP received funding from the School of Biological Sciences (University of Essex).

213

214 **Author contributions:** CG, JP, and PWHH designed the study and analyses. CG performed the
215 analyses. All the authors wrote the manuscript. CG drew additional animal outlines in Figure 1.

216 **Competing Interests Statement:** The authors declare no competing interests.

217 **Author information:** Reprints and permissions information is available at
218 www.nature.com/reprints. Readers are welcome to comment on the online version of the paper.
219 Correspondence and requests for materials should be addressed to JP (jordi.paps@bristol.ac.uk).

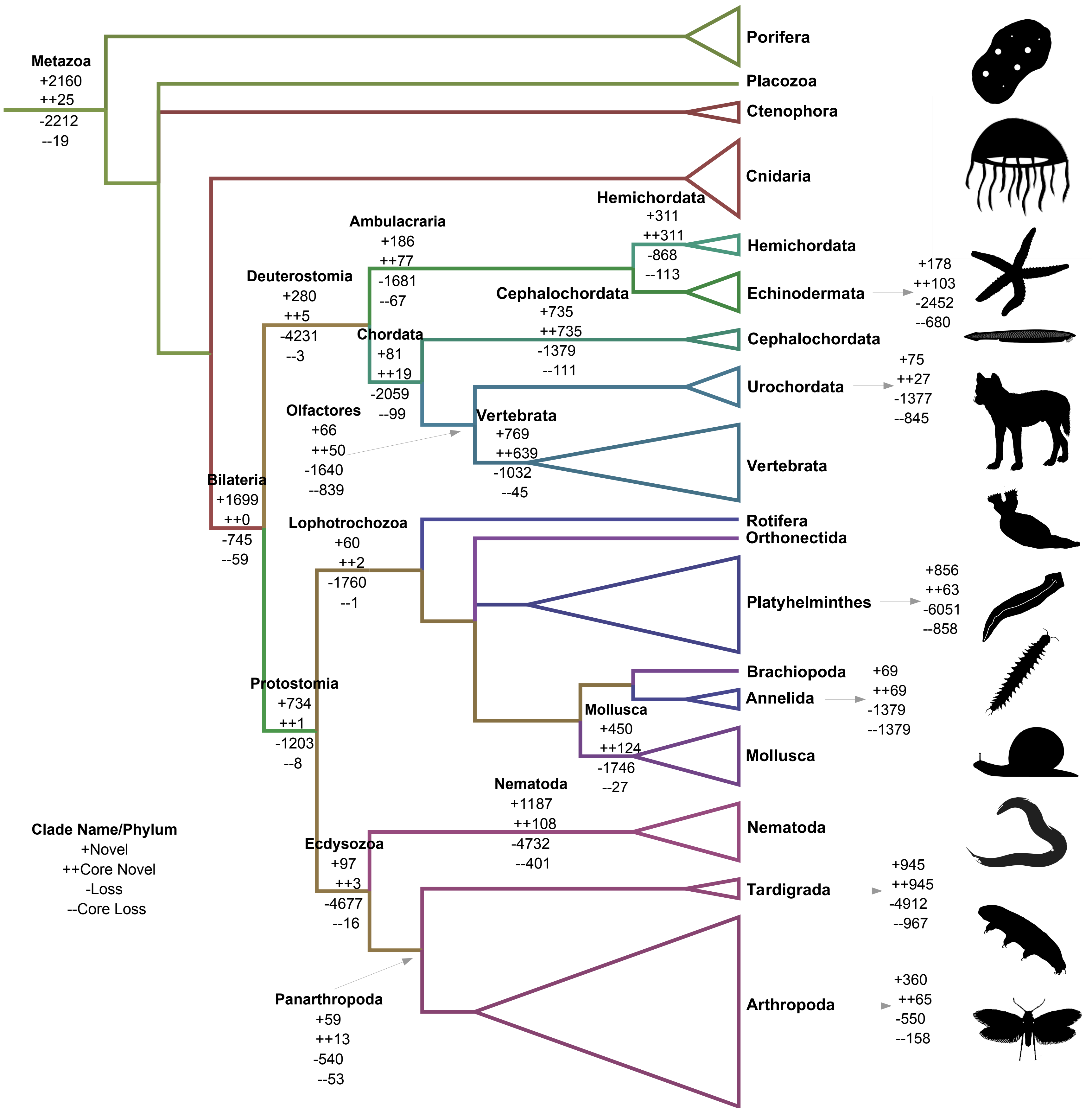
Figure Legends

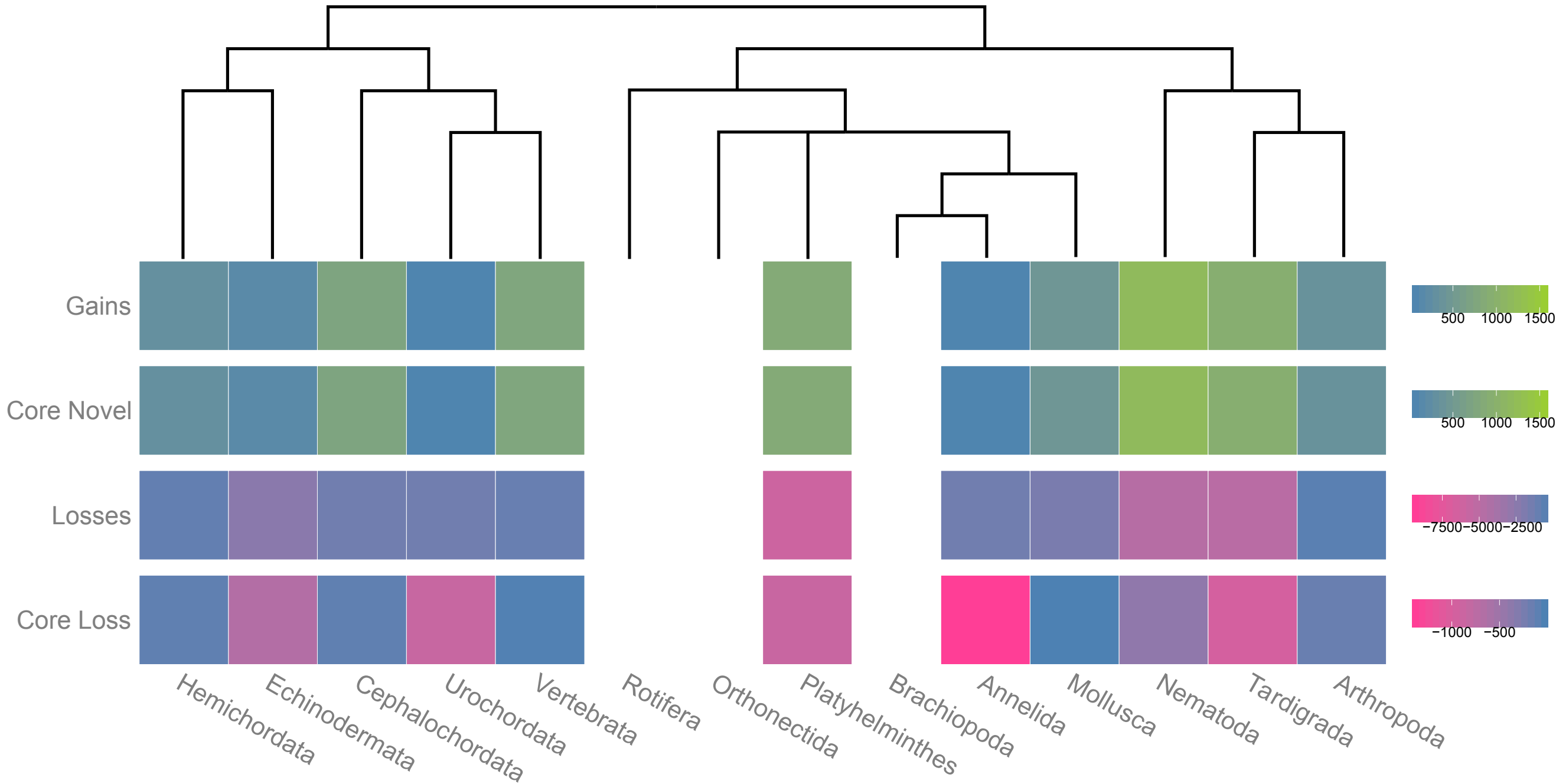
Figure 1. Reconstruction of ancestral genomic gains and losses in the Animal Kingdom.

Evolutionary relationships of the major groups included in this study^{5,28,30}. Different categories of homology groups (HG) are indicated in each node, from top to bottom, Novel HG (cluster of genes that emerged in that node), Core Novel HG (novel clusters that are present in 95% of the ingroup taxa), Lost HG (clusters of genes lost in that node), and Core Lost HG (lost clusters present in 95% of the outgroup species). Organism outlines from phylopic.org and from the author (submitted to phylopic.org).

Figure 2. Levels of gene gains and losses at phylum level. Heatmap normalised by row displaying the amount of gene gains (green at highest numbers, blue at lower numbers) and loss (pink at highest loss, blue at fewer losses) for the animal phyla in this study.

Figure 3. Most abundantly lost and gained molecular functions (GOs). (A) Heatmap for core novel (++) GO molecular functions. Scale corresponds to percentage (%) of each molecular function in each core novel (++) HG per clade, calculated over the total spread of GO molecular functions (Extended Data Figure 5). (B) Loss within a molecular function is indicated by filled blue circle (not necessarily loss of entire GO category). While different clades (columns) may have gained or lost the same functions, the actual HG gained or lost may be different. GO gained or lost in a clade refer to a subset of HG that perform that function, not all the HG associated with it. Full list of GO functions can be found in Supplementary Figures 8-15.





A



B

