## SI Text

### Ancestral Reconstructions are Not Sensitive to Reconstruction Method

The analyses reported in the main text relied on ancestral gene counts inferred with EvolMap (1), which uses pairwise alignment scores, but not full gene trees. The reasons for choosing EvolMap for the main analyses are discussed below (*SI Methods*). We also inferred consensus gene trees using 100 bootstrap replicates in RAxML (2) and inferred ancestral genome content using gene tree/species tree reconciliation based on parsimony in the package Notung (3). Overall bootstrap support was poor, but reconciliation using the consensus trees recapitulated the EvolMap results (Figs. S2, S3) despite using a different method and a different dataset (*SI Methods*). In particular, the large gene family expansions in the MRCAs of vertebrates, cnidarians, and ctenophores are still found. Other large features, such as the separation of metazoans and non-metazoans on the PCA can also be seen.

The smaller expansion in the MRCA of bilaterians was not as clear, however, nor was the loss event in the MRCA of ecdysozoans. These smaller events are probably not visible because of erroneous overestimates of ancestral genome content that arise from reconciliation using parsimony (4). Thus, even though there was uncertainty in our tree inference, topologies that would result in ancestral reconstructions that differ substantially from the EvolMap analysis were not favored in the bootstraps.

Similar results were found when maximum likelihood (ML) trees were used instead of consensus trees. The ML tree for each gene family is shown in Figure S4, and the independent radiations of $K_v$, LIC, GIC, and ASC channels can be clearly seen there. Our results are therefore robust to the method used to infer ancestral genome content.

### Findings Extend to Other Nervous System Genes

We wondered whether the general patterns found in ion channels extended to other genes, including those not associated with nervous systems. We therefore tested the three main classes of G protein-coupled receptors (GPCRs), which are closely associated with nervous systems, Actin, which is not specific to nervous systems but may correlate with muscular complexity (5), and two protein domains not strongly associated with neuro-muscular function: ubiquitin, and DNA polymerase family A (polA). We found that the patterns of gain and loss in GPCRs were roughly similar to the ion channels, and that the patterns in ubiquitin and polA were not (Figs S5, S6). Remarkably, GPCRs underwent the same loss event in the common ancestor of deuterostomes followed by a gain in the common ancestor of vertebrates. These gain and loss events were not observed in ubiquitin and polA and were only weakly present in Actin, which functions in the musculature and may therefore be expected to correlate with nervous system complexity somewhat. This suggests that the pattern of gain and loss is specific to nervous system-associated genes.

### Choice of Species Tree

The radiation of the major animal lineages was ancient and probably quite rapid. This situation makes the inference of branching order very difficult (6, 7). To see if there was any evidence for a certain species tree in our gene duplication data, we explored which species tree

allowed the most parsimonious reconciliations with our gene trees. We tested all 15 resolutions of the four-way polytomy between *Amphimedon*, *Trichoplax*, ctenophores (*Mnemiopsis*, and *Pleurobrachia*), and cnidarians + bilaterians, as well as the topology found by Philippe *et al.* (8) that places ctenophores and cnidarians together, with sponges branching first, followed by *Trichoplax*. We refer to this tree as the Coelenterata hypothesis. PAUP was used to generate the 15 resolved species trees from the polytomy (9). We then reconciled the 16 topologies with each of our 16 ML gene trees using Notung, and counted up the total gene duplication/loss costs for each species tree using Notung's default rooting method. On the principal of parsimony, the correct species tree would be the one with the lowest incurred cost.

We found that no one tree was clearly favored over the others (Fig. S7). Generally, trees with ctenophores near the base were favored. The best tree had ctenophores as the earliest-branching lineage, but grouped sponges and placozoans as a monophyletic clade, which has never been found in the major phylogenomic studies. The Coelenterata hypothesis, however, was strongly disfavored. Because of these considerations, we used a topology that reflects a growing consensus around early metazoan relationships, with ctenophores branching first, followed by sponges, placozoans, cnidarians, and then bilaterians (10–13). The finding that several species trees are roughly equivalent in terms of duplication/loss costs also has the effect of showing that our results are not heavily dependent on the species tree topology.

## SI Methods

### Gene Collection, Annotation, and Trimming

The workflow for ion channel annotation is shown in Figure S8. Proteomes were downloaded from a variety of sources (*SI Table 1*). Filtered proteomes with one protein per locus were obtained or created using custom python scripts. No proteomes without locus information were used, so all proteins had a one-to-one mapping to genes. We then used a three step process to collect and hand-annotate the data used for all subsequent analyses. We used appropriate hidden Markov models for each protein to search proteomes from each organism (*SI Table 2*) using the *hmmsearch* algorithm in the HMMER package (14). All families had unique HMMs except for the voltage-gated channel superfamily, which includes the families $K_v$, $Na_v$, $Ca_v$, Leak, TPC, TRP, Slo, and CNG/HCN. All hits with a *hmmsearch* e-value below $1 \times 10^{-2}$ were then searched against the Uniprot protein database using Blastp (15) and hand-annotated.

GPCRs, and proteins containing actin, polA, and ubiquitin domains were not reciprocally blasted against Uniprot, but rather were reciprocally searched against PFAM using *hmmscan*. Only proteins hitting the desired domain with an e-value below $1 \times 10^{-4}$ were retained. Proteins from the voltage-gated superfamily were first sorted into families before Uniprot annotation by annotating against the Transporter and Channels Data Base (16) using Blastp. Both of these Blast analyses used $1 \times 10^{-2}$ as an e-value threshold and discarded any sequences with no hit below this threshold. The final result was a hand annotated set of protein sequences for each of the 16 channel families, the GPCRs, and other protein families.

These sequences were then quality filtered by first aligning each family using the *e-ins-i* algorithm in Mafft (17), and then searching for sequences that differed by only one aligned position or less (i.e., not just gaps). If such similar groups were found, only the longest protein sequence was retained. This was the final dataset used for EvolMap analysis, and should represent a conservative estimate of the copy number for each species.

### Ancestral Genome Reconstruction

We used two different methods to reconstruct ancestral genome content. These two methods employ very different techniques, so the results consistent between the two methods should be robust to any biases unique to each method. The two different methods, their potential biases, and the way that these biases were offset by the other analysis will be briefly discussed here.

The first method, implemented in the software EvolMap (1), was used for all the main figures because it has fewer known biases. EvolMap uses Blast to identify putative orthologous groups, and then creates sparse matrices of within-group pairwise alignment scores based on Needleman-Wunsch alignments. This information is then used to identify symmetrical best hits and create estimates for ancestral genome size in a post-order trace of a supplied species tree. Then the tree is traversed in pre-order, and gains and losses are inferred using Dollo parsimony. EvolMap outputs information on ancestral gene copy number, and number of gains and losses for each node. For Figure 1, all channel types were pooled together. To create the data for the other figures, each ion channel family was analyzed by EvolMap separately, and copy number information was collected into genome-by-family matrices using custom scripts. One potential bias in this analysis is that all proteins that passed the hand-annotation and trimming steps were kept, many of which were partial. These partial sequences may have had poor Needleman-Wunsch alignment scores and therefore have been incorrectly characterized as evolutionary novelties in proximal branches. This bias was dealt with in the second analysis by discarding short sequences.

The second method we used was parsimony-based gene tree/species tree reconciliation implemented in Notung (3). Each ion channel family was aligned using the *e-ins-i* algorithm in MAFFT. The original dataset had many partial sequences, as discussed above. This first alignment was used to discard sequences by first trimming columns that were over 50 percent gapped using Trimal (18), and then flagging sequences that had fewer than 150 amino acids in the trimmed alignments. These sequences were then removed from the unaligned data, and all families were realigned and trimmed in the same fashion. These alignments were then used for phylogenetic tree inference using RAxML (2), under the LG + CAT model (19, 20) with the rapid bootstrap and ML tree reconstruction algorithm. The un-rooted gene trees were reconciled using the rooting algorithm in Notung, which finds the rooting point that minimizes gene gains and losses across the species tree, and then outputs information on the number of gains and losses for each branch. We used custom scripts to parse the Notung output and create data matrices of ancestral node counts.

Parsimony-based gene tree/species tree reconciliation is well known to have biases that result from incorrect gene tree inference. Misplaced taxa can artificially inflate the estimates of ancestral genome sizes (4). This bias, however, is not expected to affect the EvolMap analysis. We also note that this bias would tend to lead to the conclusion opposite to ours because the bias artificially inflates ancestral genome size and puts many losses on terminal branches, whereas we find small ancestral genomes and many duplications on terminal branches. Thus, although this bias is present in our Notung analysis (note that ancestral genomes reconstructed by Notung are larger than those reconstructed by EvolMap, despite the fact that some sequences were removed from the Notung analysis), our conclusions are robust with respect to the method used for analysis.

Nevertheless, to try to minimize the effects of this bias within the Notung analysis, we used both ML trees and majority-rule consensus trees as input for Notung. Both strategies supported the idea that the gene content in vertebrates, protostomes, cnidarians, and ctenophores has evolved convergently (Figs S2-4). When consensus trees are used, Notung resolves polytomies

by finding the most parsimonious branching strategy. This technique was recommended by Hahn (2007) as a way decrease the bias discussed above (4). Although our trees mostly had poor bootstrap support, the patterns that support our conclusion were represented in the consensus trees. The general pattern in the gene families with the largest independent expansion events ($K_v$, LIC, GIC, and ASC), was large clusters of genes from within certain lineages, rather than these lineages being interspersed. This pattern places the MRCA of these different lineages (e.g. cnidarians, bilaterians and ctenophores) deeper in the tree and therefore supports independent bouts of gene duplication and a low ancestral copy number in the MRCA (Fig. S8).

Figure S10 shows a reduced tree of $K_v$ channels from a handful of representative organisms inferred with Bayesian sampling in Mr. Bayes under the WAG model with across-site rates modeled with four discretized gamma categories (21, 22). This tree was inferred from an alignment constructed with Maffts *L-ins-i* algorithm using a consensus of five Garli (23) ML replicates as a starting tree. The MCMC settings were two independent runs with four chains each, sampling every 100 generations and printing every 1000 generations for $2x10^6$ total MCMC generations. We discarded the first 50% of these as burn-in. In this tree it can be seen that, while many of the branches have poor support, the bipartitions representing speciation nodes, i.e. those that separate the within-species gene family radiations, are often well-supported, and the conclusion that these gene families have undergone independent bouts of duplication is therefore robust. All trees and alignments used for this analysis can be found the Dryad repository.

**Principal Component Analysis**

We used normalized gene content matrices for the PCA. Each row of the matrix corresponded to one genome, extant or ancestral, and each column to a gene family. The entries were therefore the number of each ion channel type normalized by the total number of ion channels present in each genome. The matrix was then centered and scaled using the *scale* method in the standard R package. The PCA was performed in R using the method *prcomp* and visualized with the package *ggbiplot* (24, 25).

## SI References


1. Sakarya O, Kosik KS, Oakley TH (2008) Reconstructing ancestral genome content based on symmetrical best alignments and Dollo parsimony. *Bioinformatics* 24(5):606–612.

2. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688–2690.

3. Chen K, Durand D, Farach-Colton M (2000) NOTUNG: A Program for Dating Gene Duplications and Optimizing Gene Family Trees. *J Comput Biol* 7(3-4):429–447.

4. Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol* 8(7):R141.

5. Steinmetz PRH, et al. (2012) Independent evolution of striated muscles in cnidarians and bilaterians. *Nature* 487(7406):231–234.

6. Rokas A, Krüger D, Carroll SB (2005) Animal Evolution and the Molecular Signature of Radiations Compressed in Time. *Science* 310(5756):1933 –1938.

7. Philippe H, et al. (2011) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biol* 9(3):e1000602.

8. Philippe H, et al. (2009) Phylogenomics Revives Traditional Views on Deep Animal Relationships. *Curr Biol* 19(8):706–712.

9. Swofford, David L. (2003) *Phylogenetic analysis using parsimony (*and other methods)* (Sinauer Associates, Sunderland, MA).

10. Dunn CW, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452(7188):745–749.

11. Hejnol A, et al. (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc R Soc B Biol Sci* 276(1677):4261 –4270.

12. Ryan JF, et al. (2013) The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. *Science* 342(6164):1242592.

13. Moroz LL, et al. (2014) The ctenophore genome and the evolutionary origins of neural systems. *Nature* 510(7503):109–114.

14. Eddy SR (1998) Profile hidden Markov models. *Bioinforma Oxf Engl* 14(9):755–763.

15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.

16. Saier MH, Tran CV, Barabote RD (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* 34(suppl 1):D181–D186.

17. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33(2):511 –518.

18. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.

19. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21(6):1095–1109.

20. Le SQ, Gascuel O (2008) An Improved General Amino Acid Replacement Matrix. *Mol Biol Evol* 25(7):1307–1320.

21. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinforma Oxf Engl* 17(8):754–755.

22. Whelan S, Goldman N (2001) A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol Biol Evol* 18(5):691–699.

23. Zwickl DJ (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.

24. R Development Core Team (2008) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria) Available at: http://www.R-project.org.

25. Vu VQ (2011) *ggbiplot: A ggplot2 based biplot* Available at: http://github.com/vqv/ggbiplot.

## SI Figures

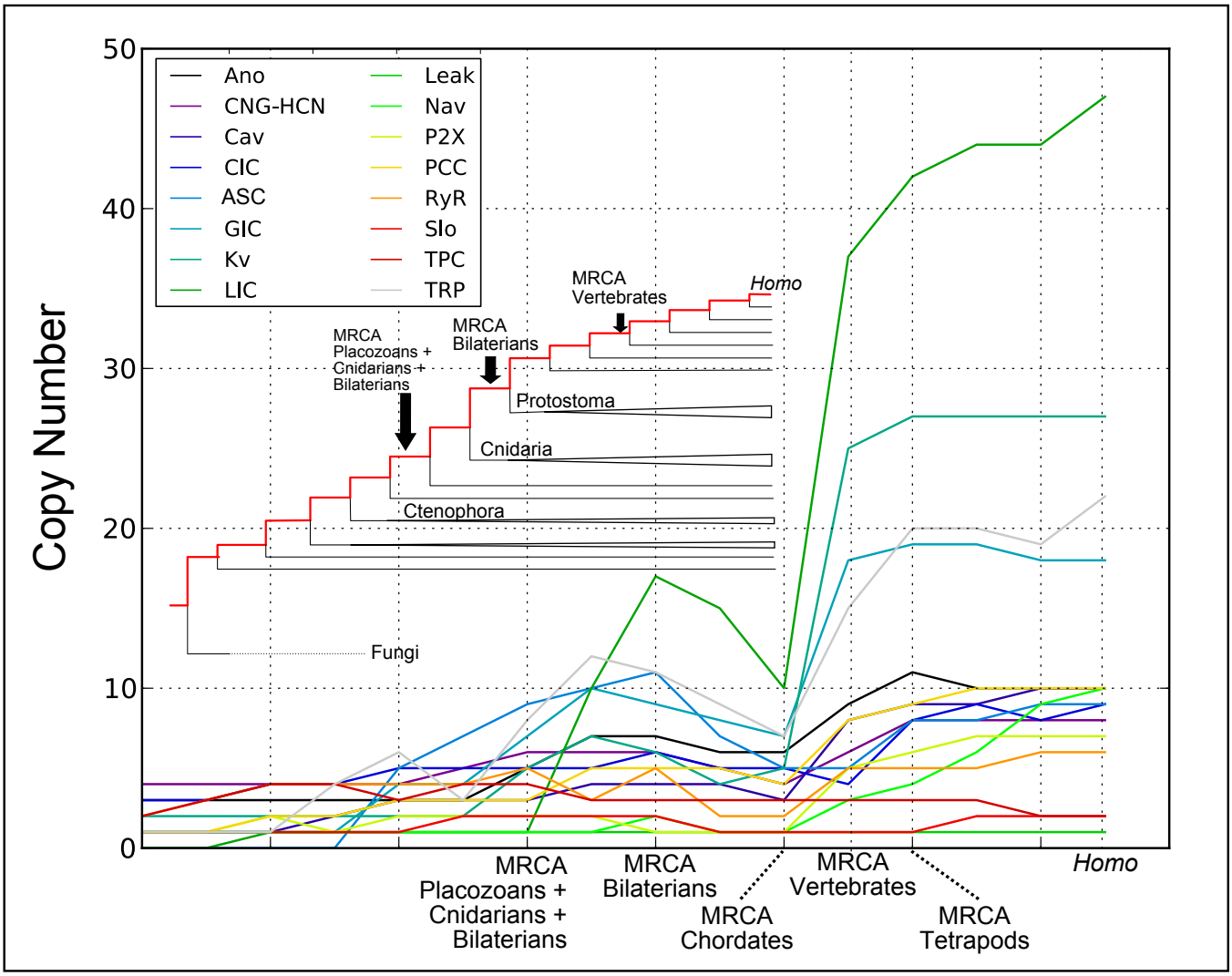


**Fig. S1:** Gain and loss of ion channel genes in the lineage leading to humans**.** Gene expansion in the MRCA of vertebrates was directly preceded by a large loss event in the MRCA of chordates. The gene families LIC, $K_v$, GIC, and TRP underwent the largest reductions. An inset tree is shown to illustrate the path of the lineage leading to humans (red branches).

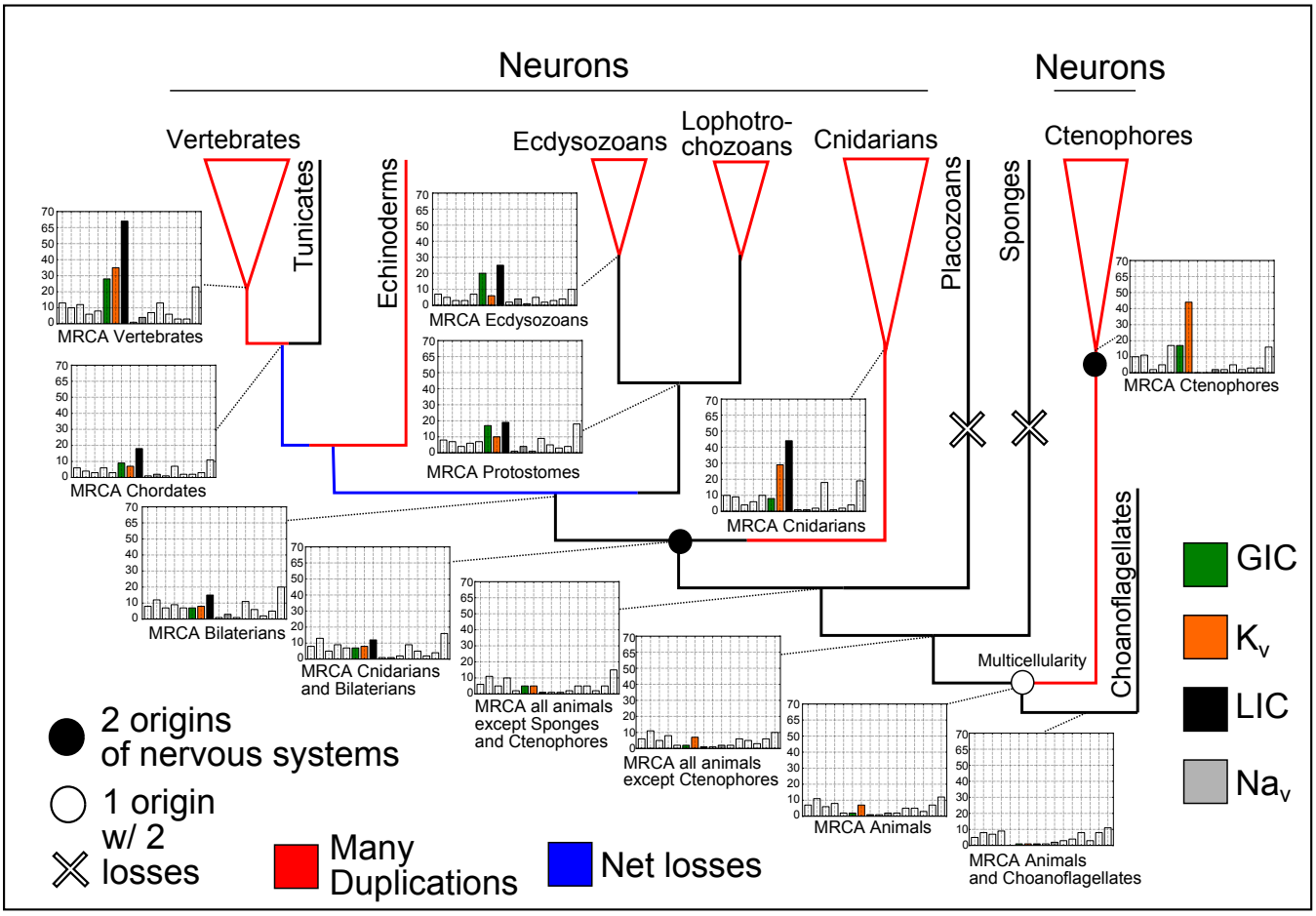


**Fig. S2:** Ion channel counts using majority-rule consensus gene tree/species tree reconciliation in Notung (3). Ancestral nodes have larger counts than those calculated by EvolMap (1), probably due to bias arising from rogue taxa (4). But the large gains in GICs, $K_v$s, and LICs can still be clearly seen in the MRCAs of vertebrates, cnidarians, and ctenophores. However, the signal for ion channel reductions in the MRCA of ecdysozoans and the gains in the MRCA of bilaterians are not found in this analysis.

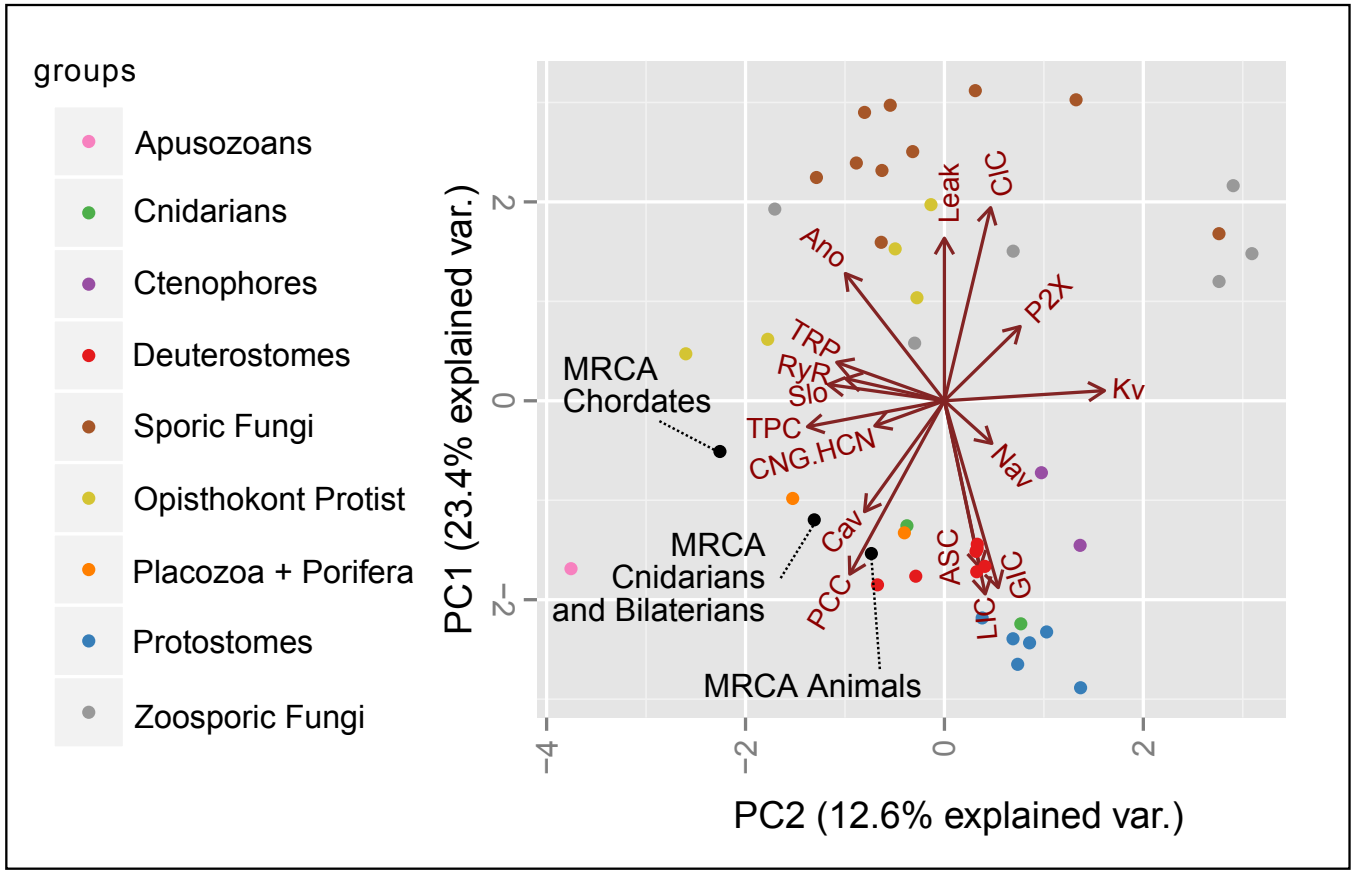


**Fig. S3:** Principal component analysis of normalized ion channel gene content using consensus gene tree/species tree reconciliation in Notung, as in Figure 3C. Results do not differ qualitatively from those acquired with EvolMap. Key ancestral nodes are still found to be distant from the terminal taxa with nervous systems that are their immediate descendents.

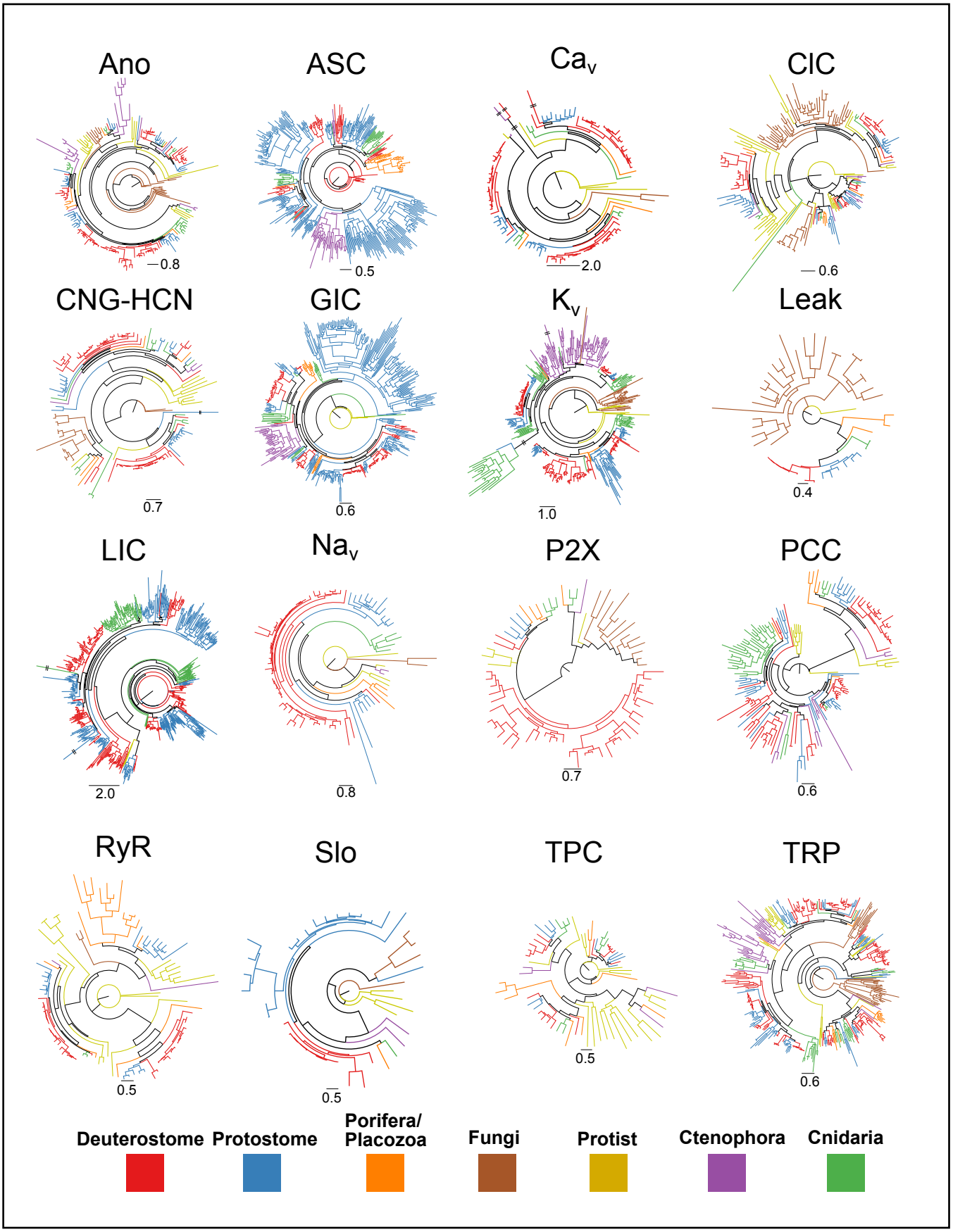


**Fig S4:** Maximum likelihood trees for all families. Large single-color clusters can be seen in the ASC, $K_v$, GIC and LIC families, denoting independent radiations in these lineages. Branches with two hashes have been shortened for ease of display.

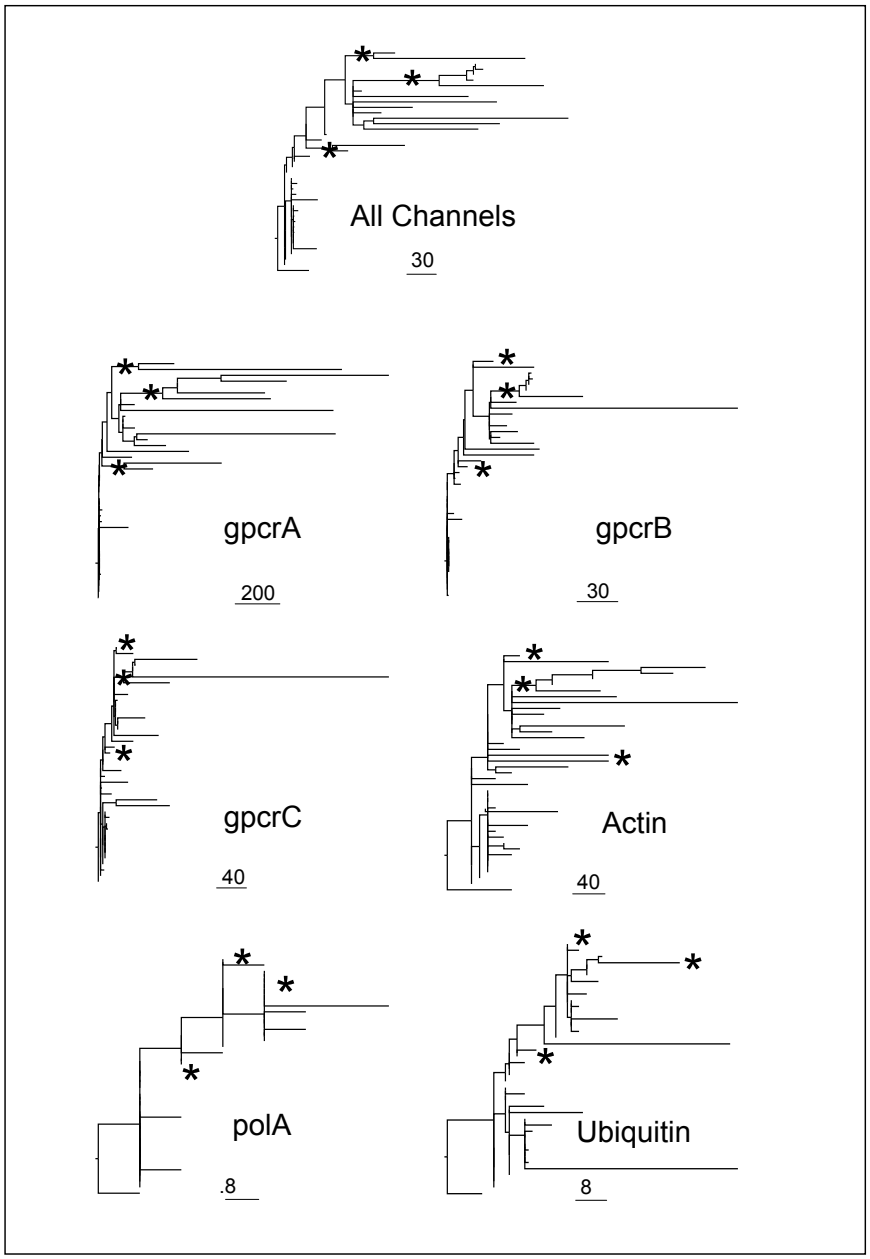


**Fig. S5:** Net gain trees of G protein-coupled receptors (GPCRs) and three other protein domains not strongly associated with nervous sysytems: Actin, DNA polymerase, and Ubiquitin. The tree of channels from Figure 1 is shown at top for comparison. Stars are placed on, from top to bottom, the MRCA of cnidarians, the MRCA of vertebrates, and the MRCA of ctenophores. On trees where the branches don't exist or are too small, the stars are placed next to these lineages. The pattern of gains in GPCRs is similar to that of ion channels, with many more gains in animals than in fungi. The patterns in the other genes are different. The different scales should also be noted.

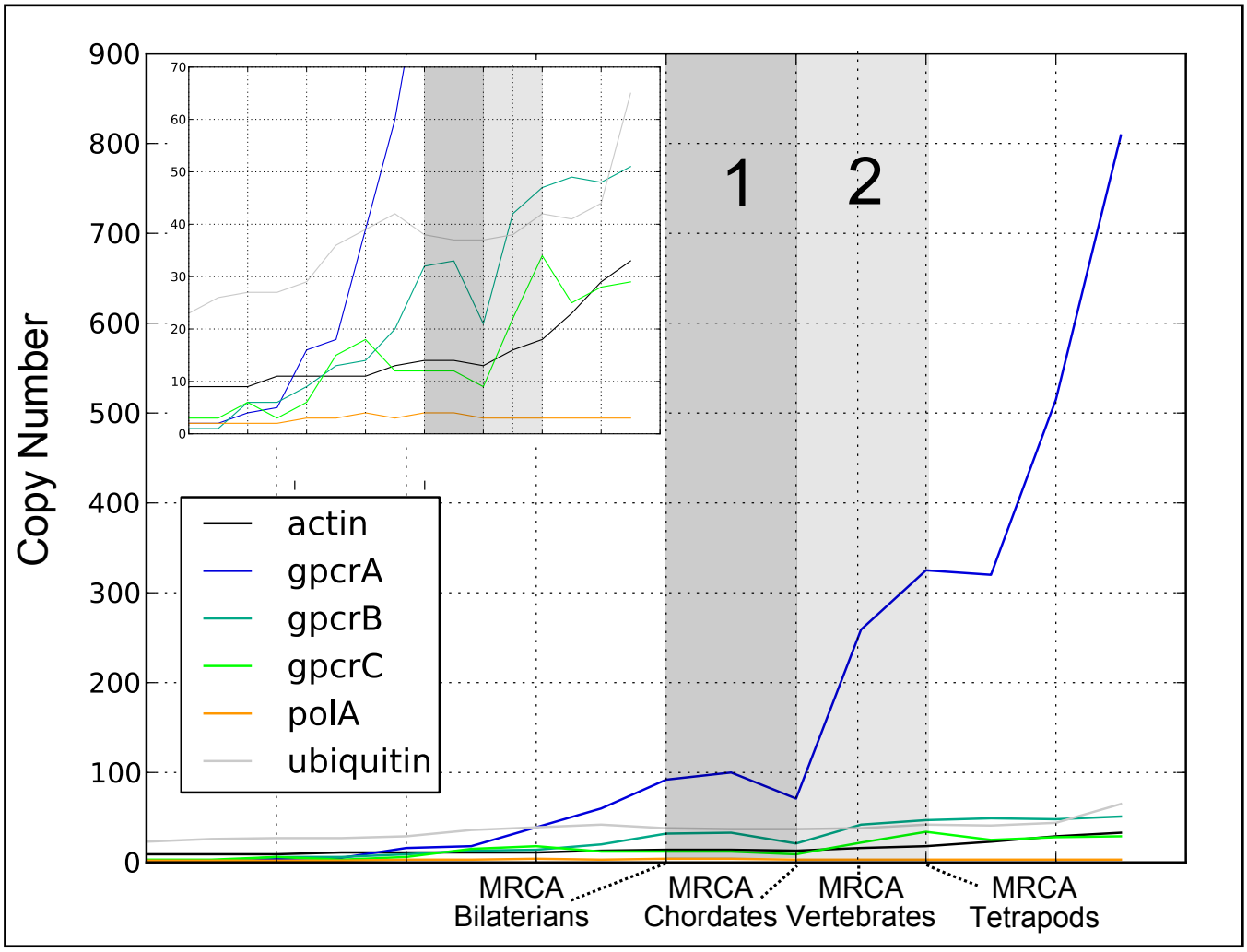


**Fig. S6:** Gene gain and loss of GPCRs and three non-nervous system genes in the human lineage. The inset is shown on a smaller scale so that the pattern of duplication and loss can be seen for genes families other than A-type GPCRs, which are a much larger family. GPCRs resemble ion channels in their pattern of gain and loss whereas the other genes do not. In particular, GPCRs underwent loss events in the common ancestor of chordates followed by a period of gain, primarily in the ancestor of vertebrates. The shaded regions highlight the periods of loss (1) and gain (2).

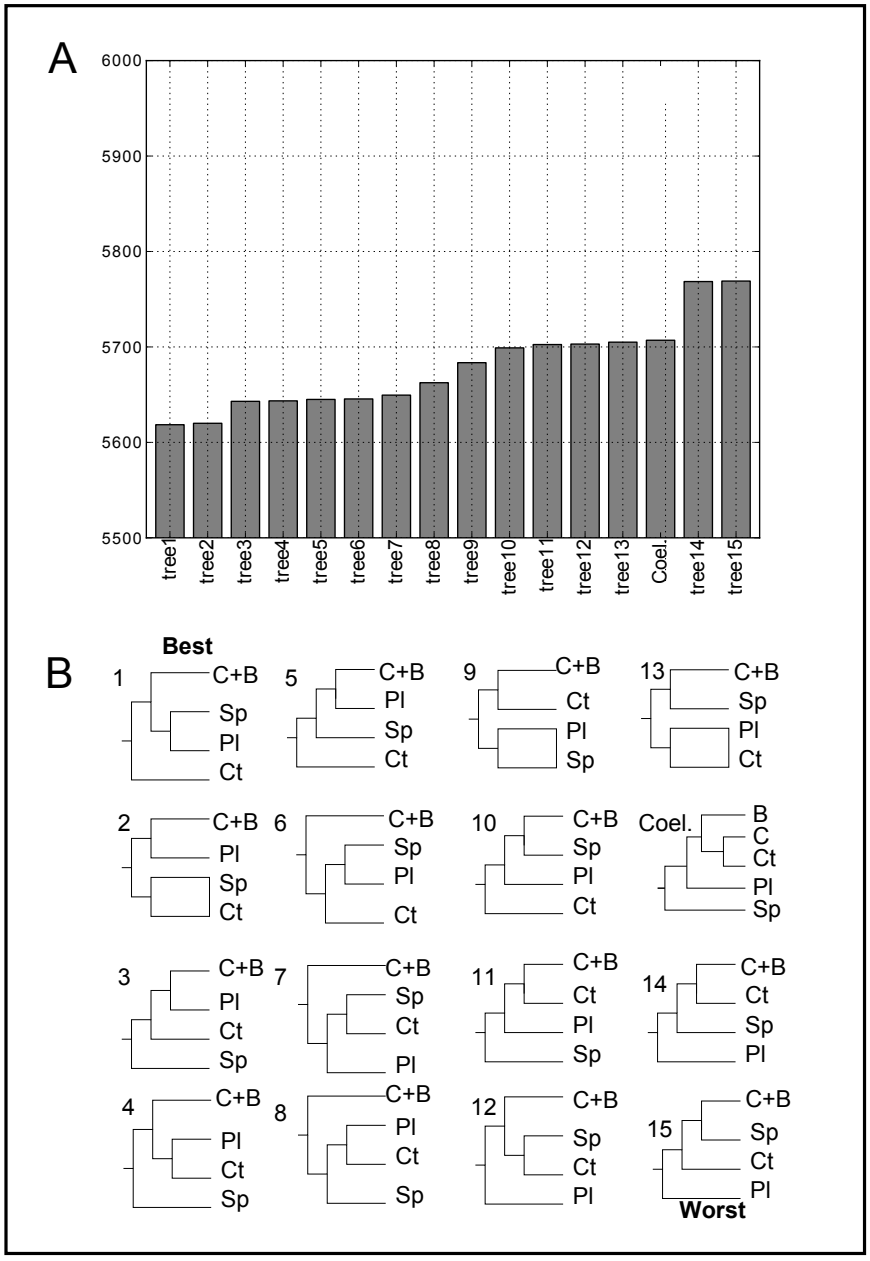


**Fig S7:** Support for different species tree topologies based on parsimony scores of gain and loss events calculated from gene tree/species tree reconciliation in Notung. A.) Scores for different resolutions ordered from best (left) to worst (right). B.) Topologies ordered from best (top left) to worst (bottom right) and numbered as in A. Taxa are labeled as follows: bilaterians (B), cnidarians (C), ctenophores (Ct), placozoans (Pl), and sponges (Sp). Scores generally support ctenophores as the sister group of remaining animals, but do not strongly favor this placement over other scenarios with the exception of Coelenterate hypothesis ("Coel", Philippe 2009). The latter tree yields highly unparsimonious patterns of gain and loss.

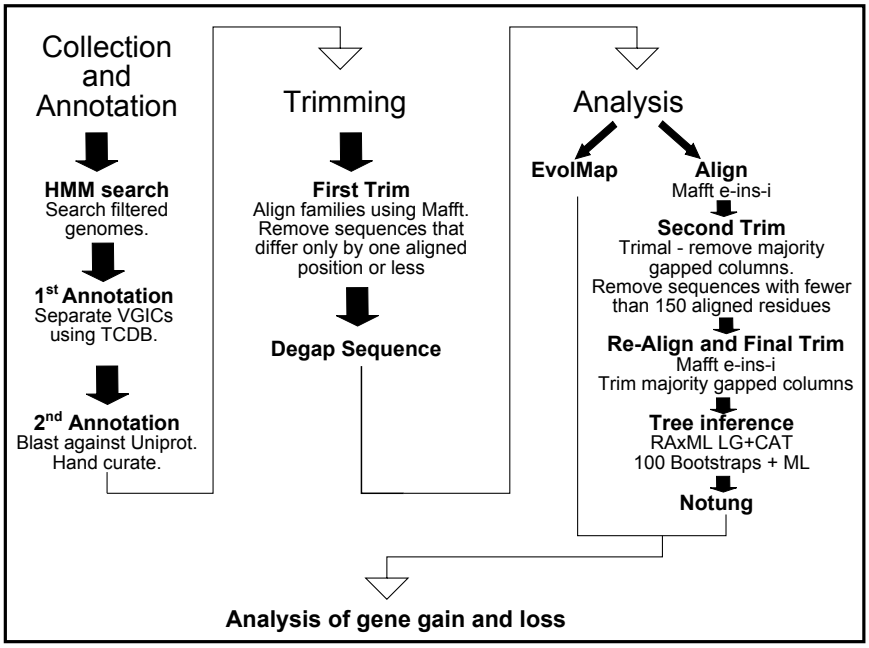


**Fig. S8:** Bioinformatics pipeline used for analysis.

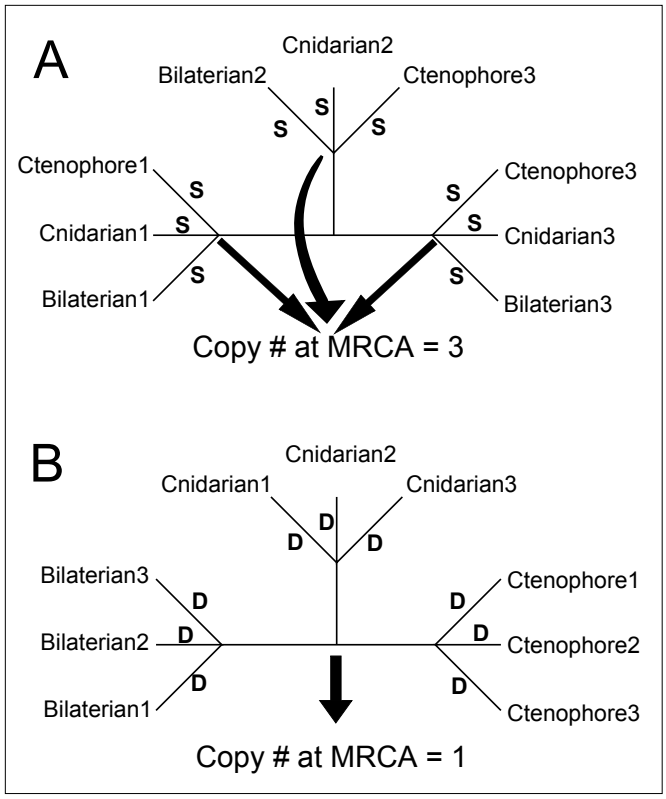


**Fig. S9:** Inferring ancestral copy number from gene trees. Two gene-tree topologies are shown comprising three genes each (1-3) from a bilaterian, a cnidarian, and a ctenophore. When different species group together (A), the terminal branches are interpreted as speciation events (S) and the copy number at the most recent common ancestor of bilaterians, cnidarians, and ctenophores (MRCA) is high. When the same species cluster together (B), the terminal branches are inferred as independent gene duplications (D), and the MRCA copy number is lower.

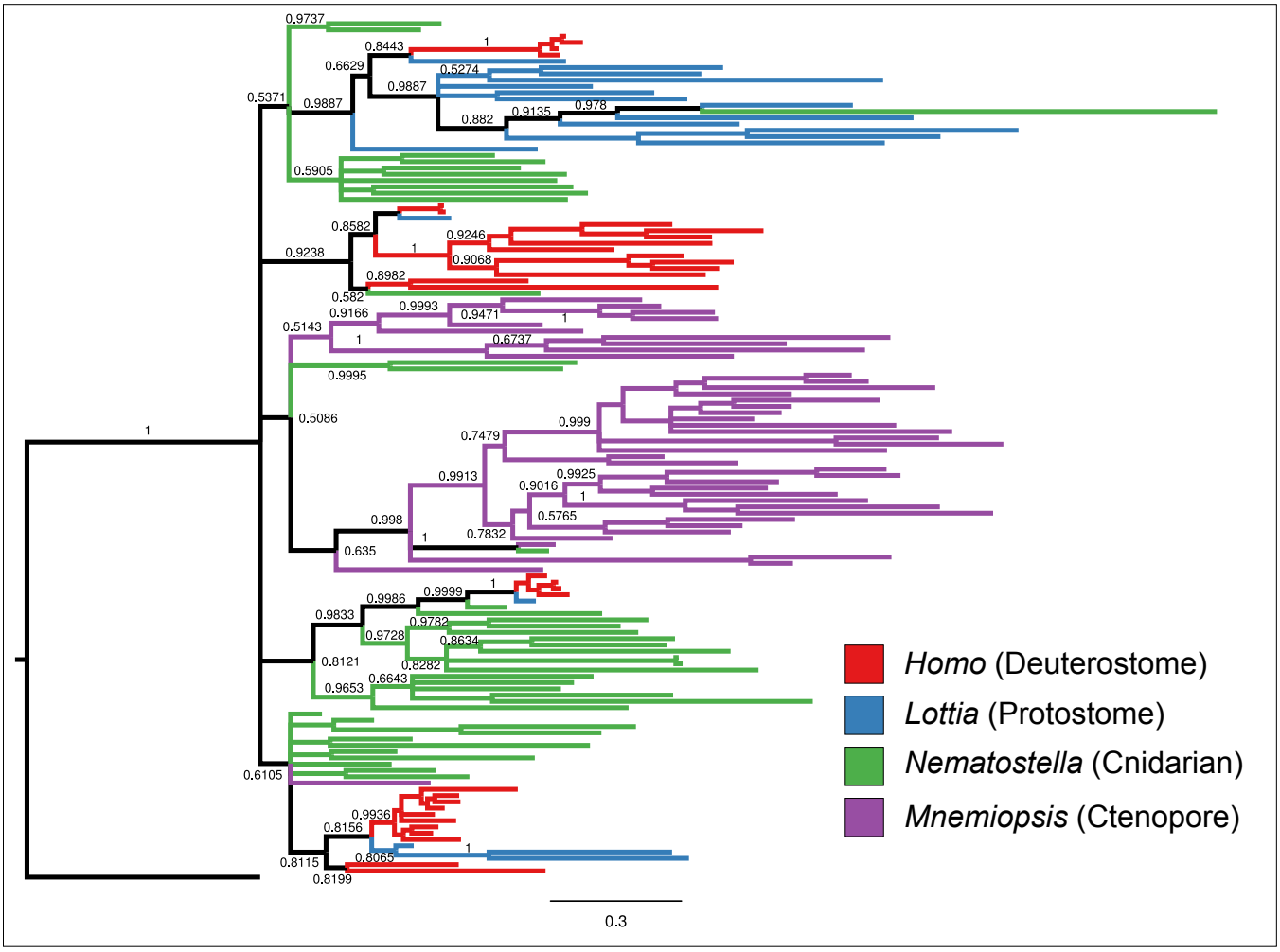


**Fig S10:** A reduced tree of the $K_v$ family. Only genes from four animal species are shown, plus a fungal outgroup. Bayesian posterior probabilities are shown on interior nodes. Large within-species groups suggest many independent bouts of gene duplication and a low ancestral copy number.

| Species | Data Source |
|---|---|
| Acropora | Matz Lab |
| Allomyces | Origins of Multicellularity |
| Amphimedon | Ensembl |
| Aspergillus | JGI |
| Capitella | Ensembl |
| Capsaspora | Origins of Multicellularity |
| Catenaria | JGI |
| Ciona | Ensembl |
| Coemansia | JGI |
| Conidiobolus | JGI |
| Danio | Ensembl |
| Drosophila | Ensembl |
| Fonticula | Origins of Multicellularity |
| Gallus | Ensembl |
| Gonapodya | JGI |
| Helobdella | Ensembl |
| Homo | Ensembl |
| Ixodes | Ensembl |
| Lottia | Ensembl |
| Melampsora | JGI |
| Mnemiopsis | Mnemiopsis Genome Project |
| Monodelphis | Ensembl |
| Monosiga | Origins of Multicellularity |
| Mortierella | Origins of Multicellularity |
| Mucor | JGI |
| Nematostella | Ensembl |
| Neurospora | JGI |
| Phycomyces | JGI |
| Piromyces | JGI |
| Pleurobrachia | Pleurobrachia Genome Project |
| Rhizophagus | JGI |
| Rozella | JGI |
| Saccharomyces | JGI |
| Salpingoeca | Origins of Multicellularity |
| Sphaeroforma | Origins of Multicellularity |
| Spizellomyces | Origins of Multicellularity |
| Strigamia | Ensembl |
| Strongylocentrotus | Ensembl |
| Thecamonas | Origins of Multicellularity |
| Trichoplax | Ensembl |
| Xenopus | Ensembl |

**Table S1:** Source of genomic information. URLs can be found in main text Methods section.

| Gene Family | PFAM name | PFAM Accession | Date |
|---|---|---|---|
| Actin | Actin | PF00022.14 | 10/8/2012 |
| Ano | Anoctamin | PF04547.7 | 10/6/2012 |
| ASC | ASC | PF00858.19 | 10/6/2012 |
| CNG-HCN | ion_trans | PF00520.26 | 10/9/2012 |
| Cav | ion_trans | PF00520.26 | 10/9/2012 |
| ClC | Voltage_ClC | PF00654.15 | 10/6/2012 |
| GIC | Lig_chan | PF00060.21 | 10/7/2012 |
| gpcrA | 7tm_1 | PF00001.16 | 10/11/2012 |
| gpcrB | 7tm_2 | PF00002.19 | 10/6/2012 |
| gpcrC | 7tm_3 | PF00003.17 | 10/6/2012 |
| LIC | Neur_chan_memb | PF02932.11 | 10/6/2012 |
| Kv | ion_trans | PF00520.26 | 10/9/2012 |
| Nav | ion_trans | PF00520.26 | 10/9/2012 |
| Leak | ion_trans | PF00520.26 | 10/9/2012 |
| P2X | P2X_receptor | PF00864.14 | 10/5/2012 |
| PCC | PKD_channel | PF08016.7 | 10/7/2012 |
| RyR | Ins145_P3_rec | PF08709.6 | 10/6/2012 |
| Slo | ion_trans | PF00520.26 | 10/9/2012 |
| TPC | ion_trans | PF00520.26 | 10/9/2012 |
| TRP | ion_trans | PF00520.26 | 10/9/2012 |
| Ubitquitin | ubiquitin | PF00240.18 | 10/9/2012 |

**Table S2:** PFAM domains used for initial searches.

| **Pipeline Step** | **Program name** | |
|---|---|---|
| Genome trimming | remove_splices | |
| Sequence collection | hits_pickler | |
| Annotation | TCDBAnnotate | |
| Sequence trimming | Alignment | |
| Ancestral Reconstruction | AncGenome | |
| | | |
| **Program name** | **Repository** | **Commit** |
| remove_splices | PhyloPreprocessing | 7f13302470c1f75b91b228f314e77c28378a6be9 |
| hits_pickler | TCDBAnnotate | 70b0370f86048e76ea16fdadb75d85cd8babaf6a |
| TCDBAnnotate | TCDBAnnotate | 70b0370f86048e76ea16fdadb75d85cd8babaf6a |
| Alignment | Alignment | 9b10e9ce92509f98f733cf44be98ac1e38e9c7e3 |
| AncGenome | AncestralGeneContent | 86416f8b370a02a99a1c2cf29b0cc11931276605 |

**Table S3:** Script information and availability: https://github.com/bliebeskind