A phylogeny aware analysis of gene function for the biodiversity genomics era

Gaurav D. Diwan[1,2,*], Paschalis Natsidis[3], Mu-en Chung[1,2], Juan-Carlos Gonzalez Sanchez[1,2], Roderic Guigó[4,5], John K. Colbourne[6,7], Maximilian J. Telford[3] and Robert B. Russell[1,2,*]

[1]BioQuant, Heidelberg University; Im Neuenheimer Feld 267, 69120 Heidelberg, Germany

[2]Heidelberg University Biochemistry Center (BZH); Im Neuenheimer Feld 328, 69120 Heidelberg, Germany

[3] Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Ecology, University College London; London WC1E 6BT, United Kingdom

[4]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology; C/ Dr. Aiguader 88, Barcelona 08003 Catalonia, Spain

[5]Department of Medicine and Life Sciences, Universitat Pompeu Fabra (UPF); C/ Dr. Aiguader 80, Barcelona 08003, Catalonia, Spain

[6]School of Biosciences, University of Birmingham; Birmingham, United Kingdom

[7]Centre for Environmental Research and Justice (CERJ), University of Birmingham; Birmingham, United Kingdom

*Corresponding authors Email:

Gaurav D. Diwan (gaurav.diwan@bioquant.uni-heidelberg.de);

Robert B. Russell (robert.russell@bioquant.uni-heidelberg.de)

Abstract:

In the biodiversity genomics era, vast gene catalogs have rapidly accumulated, but how gene functions map onto major evolutionary transitions remains vague. We integrated phylogenomics with functional annotations to trace the emergence of ~4.5 million genes and their associated functions in 508 representative species across the tree of life. Using this top-down, function-centric approach, we mapped the evolutionary history of well-annotated biological functions, revealing key insights and widespread parallelism. For instance, major founder events—bursts of gene and function gain—aligned with key evolutionary transitions, including the origins of eukaryotes, animals, plants, and vertebrates. Independent transitions to multicellularity and terrestrial life showed convergent enrichment of analogous functions, often drawing on ancient gene families. Our framework enables scalable interpretation of new genomes in an evolutionary context and offers a roadmap for exploring the genomic basis of biodiversity.

Introduction

We are now in the biodiversity genomics era, marked by massive international projects cataloguing life's genetic diversity at unprecedented scale (*1–3*). Building on landmark efforts like the Human (*4*) and Mouse (*5*) Genome Projects, these initiatives have driven major advances in genome assembly, annotation, and comparative analysis (*6, 7*). While whole-genome sequencing and comparative analyses are now routine, major challenges remain in this era (*8*), especially in assigning and interpreting gene function in an evolutionary context.

To fully utilise these expanding genomic resources, we must move beyond cataloguing genes to understanding what they reveal about the biology and evolution of organisms. How can we use the depth and breadth of sequenced genomes to reconstruct the deep past of life on Earth? And how can we use that perspective to identify which genes and gene functions are linked to particular phenotypic (clade-specific) traits, and when in evolutionary history they arose? Answers to such questions would deepen our understanding of major evolutionary transitions (*9*), such as the emergence of eukaryotes, transitions to new environments, multicellularity, and cell type diversity, and illuminate core biological processes like development, regeneration, and even human-specific conditions including cancer and aging. Realizing the potential contained within the data we now have requires a robust framework for tracing functional annotations across the tree of life.

Phylogenomic analyses provide an avenue for this by enabling inference of the gene content and phenotypic traits of ancestral species (*10–12*). While it is straightforward to determine the genomic content of ancestors either for a handful of genes/functions or for a few broad taxa ((*13–22*) and OMA (*23*), EggNOG (*24*), Matreex (*25*)), being able to consider the evolutionary history of the entire set of functions associated with all genes in a diverse set of organisms promises far more profound insights. A much broader perspective might show whether certain functions are phylogenetically restricted, and such discoveries could allow annotation rules in the future to reflect these expectations for newly sequenced species from these clades.

To explore these opportunities, we asked whether combining phylogenetic relationships with functional annotations could identify the evolutionary emergence of widely understood traits and functions. We analyzed genomic annotations as currently available in public databases, in the context of a deeply rooted phylogeny of life comprising 508 representative species (Table S1). Using OrthoFinder (*26*), we sorted ~4.5 million proteins into hierarchical orthologous groups (HOGs); built a species tree from the most conserved genes (see Materials and Methods); and traced the evolutionary origin of each gene and its associated functions using Gene Ontology terms (*27, 28*), protein domains (Pfam (*29*)), and biological pathways (KEGG (*30*); Reactome (*31, 32*)). All of these inferences can be explored via a web application (https://funcevol.russelllab.org).

The largest founder events correspond to specific biological traits

Of the largest founder (gain) events we can detect (Fig. 1, Table S2), most are confined to the eukaryotic superkingdom: in the ancestors of eukaryotes, slime molds (Evosea), land plants (Embryophyta and Magnoliopsida), animals (Metazoa) and vertebrates (Chordata and Euteleostomi). Among the prokaryotes, two large founder events are apparent in Cyanobacteria and one in Archaea (Halobacteria). For most founder events, genes (i.e. HOGs; see Materials and Methods), protein domains and pathways were gained in large quantities simultaneously (Fig. S2).

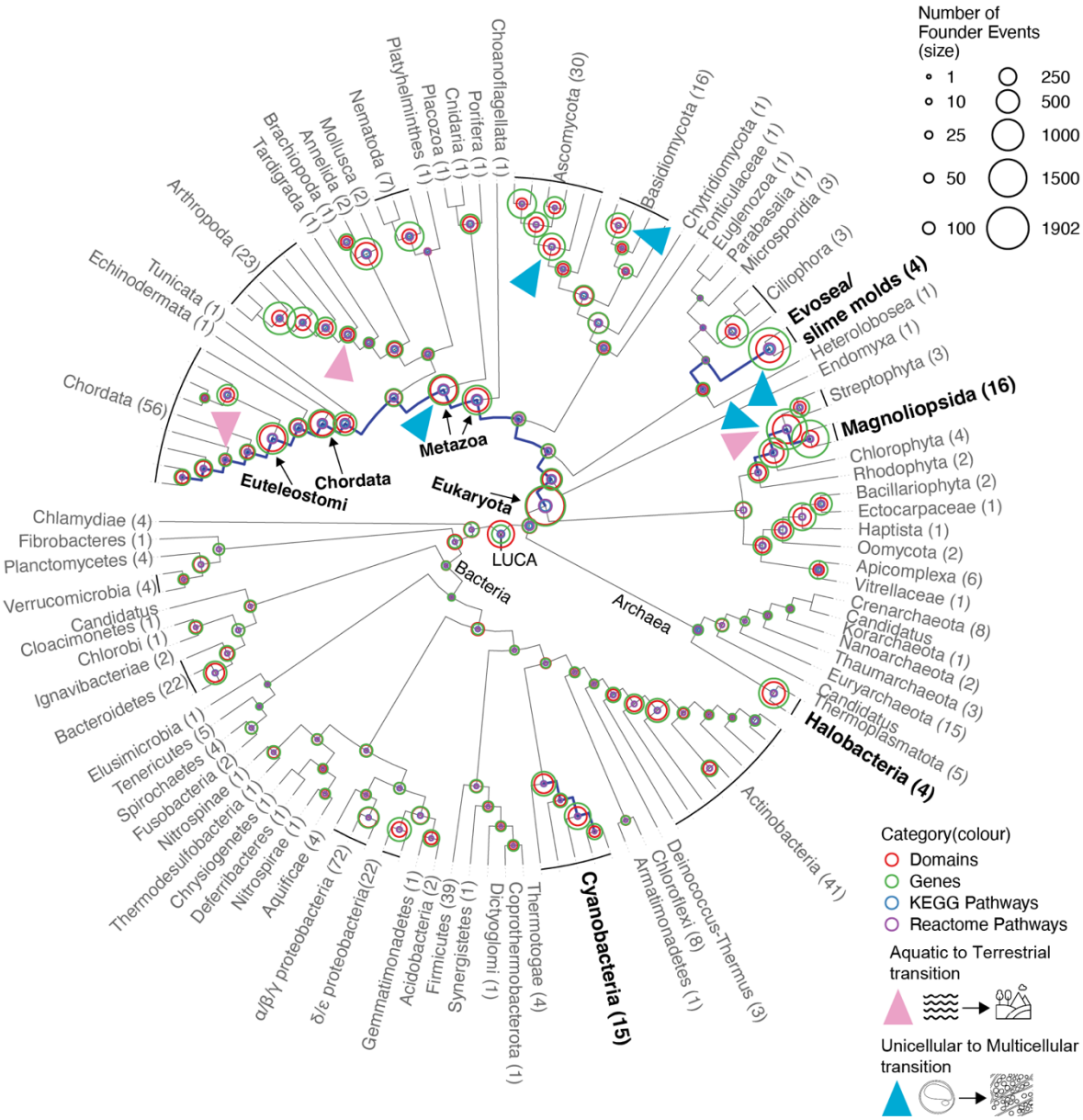

Fig. 1: Ancestral reconstruction of genes, domains and pathways reveal nodes with the largest founder events. *Phylogenetic tree (cladogram) showing the founder events of biological entities in the ancestors of the clades in this study (number of species in parentheses; the complete tree with support values is shown in Fig. S1). The size of the coloured circles represents the number of each category that was gained at a given node. N.B. KEGG/Reactome pathways are depicted as the sum of the proportion of pathways gained at the node. Lineages with large and simultaneous emergence of multiple biological categories have been emphasized (dark blue lines). These are further detailed in Fig. 2. LUCA = Last Universal Common Ancestor*

Reassuringly, the integration of phylogenomic and functional information readily identified many functions known to be associated with the origin of major evolutionary groups. For example, the eukaryotic root gained a large proportion of basic pathways such as

transcription (on average 54.6% of KEGG pathways in this category) and translation (37.4%), and showed an enrichment of broad GO terms such as cell cycle, mRNA processing and RNA splicing (Fig. 2A). The metazoan root gained pathways including development and regeneration (30.9%), circulatory system (30.1%) and cell motility (29.1%), and terms such as cell differentiation and embryo development (Fig. 2A). Further specificity of function was observed in the vertebrate ancestor with large gains in pathways for signalling processes (27.7%) and the immune (24.6%), sensory (18.7%) and digestive (18.7%) systems, and in terms such as 'immune response' and signaling pathways (Fig. 2A). The ancestor of angiosperms showed an enrichment of GO terms such as 'response to abscisic acid', 'auxins' and 'bacteria' and pathways such as Brassinosteroid synthesis (92.8%), plant hormone signal transduction (79.5%), plant circadian rhythm (70.5%) and zeatin biosynthesis (69.2%) (Fig. 2C) whereas the slime molds (*Evosea*) showed an enrichment for sorocarp development (Fig. 2B).

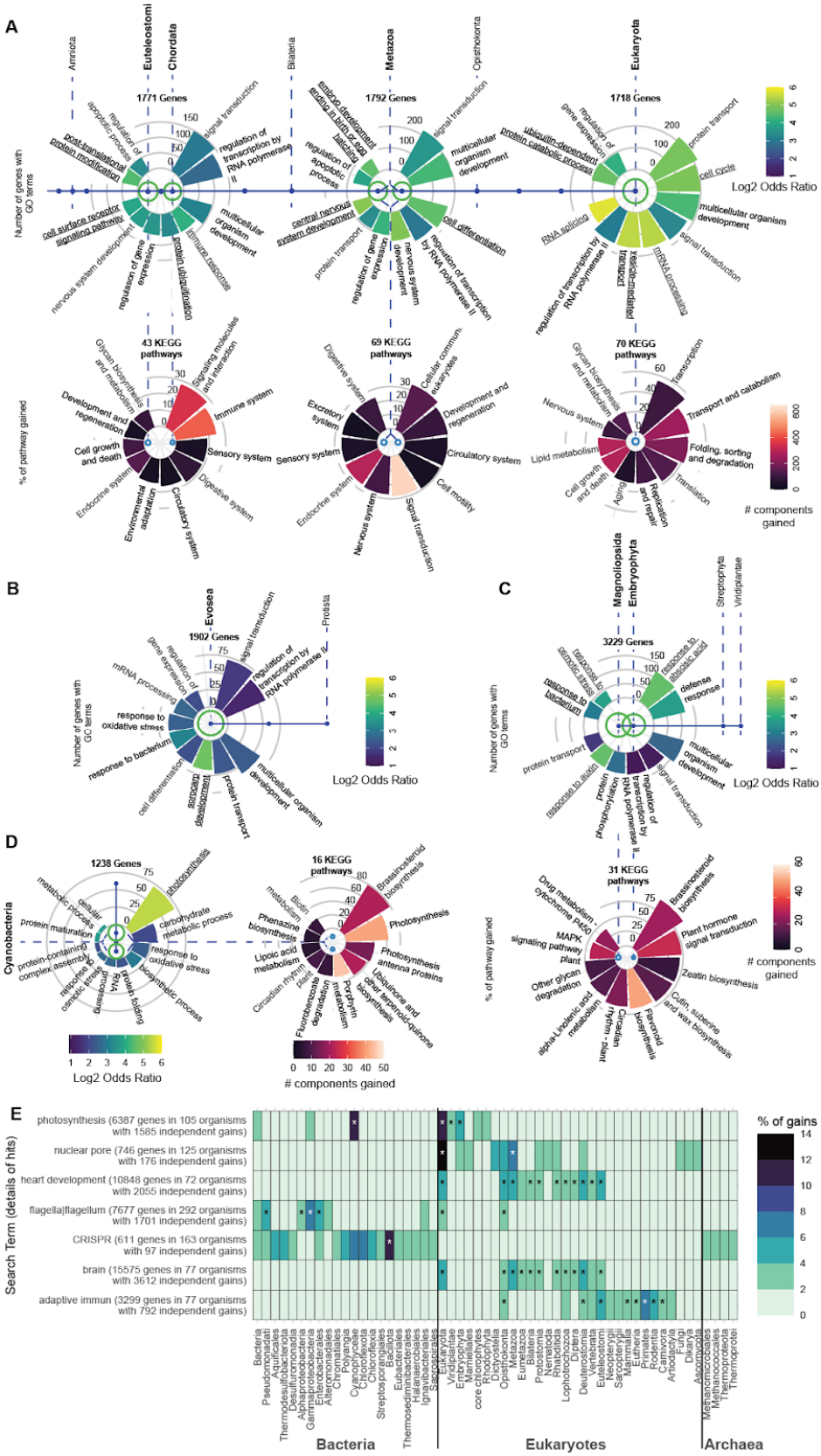

Fig. 2: Combining phylogeny with functional and pathway annotations shows gains of well-known traits at different nodes in the tree. A-D) *Figures showing a detailed look at the*

largest founder events from Fig. 1. The highlighted lineages (dark blue lines) in Fig. 1 have been depicted in a linear manner and have been labelled with the relevant taxonomic information wherever possible (dashed lines). In each panel, bar plots show the ten enriched GO terms with the highest number of genes in each category and the ten KEGG pathways with the highest % of the pathway gained at the respective node in the tree. The GO term bars are coloured by the Log2 Odds Ratio from a Fisher's exact test and the KEGG pathway bars are coloured by the number of components (i.e., proteins) gained for each pathway. Unique GO terms from each clade are also underlined. ***E)*** *Heatmap showing the % of gains for the genes captured by a search term occurring at each of the nodes on the x-axis. Asterisks indicate the taxa that have significantly higher % gains than the mean (z-score > 1.96) calculated for each search term independently.*

To generalize this approach beyond curated functional sets, we used keyword-based text queries across annotations (Fig. 2E). For each term we identified clades where genes specific to the term originated. For instance, the genes associated with the term "photosynthesis" originate mostly in eukaryotes and plants (Viridiplantae and Embryophyta, Fig. 2E). A few genes originated in bacteria, with the largest gains in cyanobacteria (Cyanophyceae). Terms such as "nuclear pore," "heart development," "brain," and "adaptive immun(e)" showed gains in eukaryotes and their descendant clades, often peaking in vertebrates (e.g., adaptive immunity genes in Euteleostomi and Primates). Genes linked to "flagella|flagellum" originated independently in bacteria and eukaryotes, reflecting convergent functional roles in bacterial motility and ciliary structures such as intraflagellar transport complexes. As expected, genes associated with "CRISPR" were confined to bacteria and archaea.

Convergent genomic patterns in multicellularity and terrestrial transitions

To investigate the genomic basis of major evolutionary transitions, we focused on two well-characterized shifts in eukaryotic biology: the transition from aquatic to terrestrial environments, and the transition from unicellularity to multicellularity. Our goal was twofold. First, we asked whether genes previously described as playing a role in these transitions within specific clades are also present in other clades that underwent the same transitions independently. Second, we aimed to determine when the broader homologous gene families underlying these genes first emerged. To do this, we examined both the presence of orthologs for the known genes and the timing of founder events in the corresponding hierarchical orthologous groups (HOGs), using the emergence of a HOG as a proxy for the origin of a gene family.

For the aquatic-to-terrestrial transition, we identified three clades in our phylogeny that independently colonized land: Embryophyta (land plants) *(33)*, Arthropoda, and Amniota (tetrapods excluding amphibians) *(34)* (Fig. S3, Table S2). Although fungi are also known to have undergone a terrestrial transition, many of the fungi in our dataset were host-associated or pathogenic, preventing us from confidently assigning a transition node. For the three clades, we tested whether known adaptive genes—previously described in the context of adaptation to living on land—were gained near or prior to their respective transition nodes (Table S3). These included transcription factors and signaling molecules such as BMP4, Noggin, Pax2, Eyeless/Pax6, KEAP1/NRF2, Hox genes, and Aquaporins *(34–36)*. In all three clades, we observed presence of orthologs close to the relevant transition nodes (Fig. 3A). In tetrapods and arthropods, many of these genes originated from founder events (HOGs)

inferred to have occurred before or at the transition node (gray bars in Fig. 3A, Fig. S5), suggesting their availability for recruitment into land-adaptive roles. Notably, Eyeless *(37)* and Aquaporins also appeared in land plants, but as later ortholog gains from earlier-emerging HOGs, highlighting parallel use of ancient gene families across lineages.

We applied a similar approach to the emergence of multicellularity, identifying independent transitions in Metazoa, Dictyostelia (slime molds), Embryophyta, and multicellular fungi (Agaricomycotina and Pezizomycotina) (Fig. S4, Table S2). For each clade, we examined the presence and timing of orthologs previously associated with multicellular development, including Beta-catenins, Cadherins, the Integrin Adhesion Complex *(38–40)*, NMD machinery and SR splicing factors *(41)*, as well as lineage-specific regulators such as flocculation genes in yeasts *(42)*, cyclins in *Chlamydomonas* *(43)*, and multicellularity-associated genes in fungi *(44)* (Table S4). Many of these were found not only in the clades where they had been originally described, but also in others that independently evolved multicellularity. For instance, Dictyostelia and Pezizomycotina gained orthologs of nearly all multicellularity-associated genes (except Beta-catenins), while Embryophyta gained all but the animal-specific Cadherins *(39)* and the Integrin Adhesion Complex (Fig. 3B).

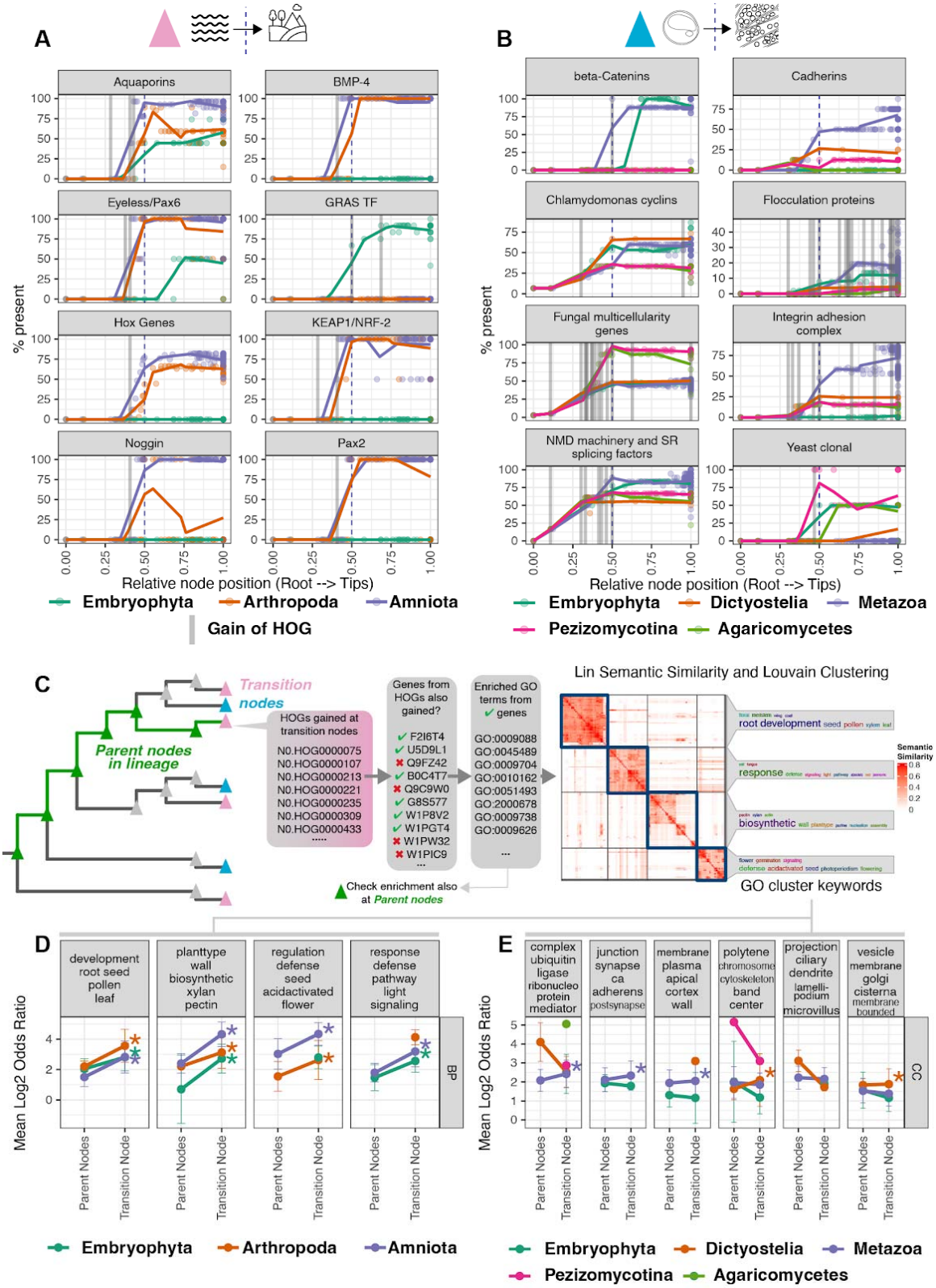

Fig. 3: Known genes and functional clusters adaptive for large phenotypic transitions are present in parallel at multiple nodes. A-B) *Plot showing the % of orthologs present at each*

node of a transition lineage. The distance to the root of each node in a lineage was scaled between 0 and 1 on the x-axis, while the transition node was forced to be at 0.5 (dashed vertical lines). Multiple linear models were fitted to the underlying points for each lineage (one model each for four equal sections of the x-axis, see Materials and Methods). Vertical gray bars indicate the earliest gain of HOGs associated with the known genes. ***C)*** *Schematic representation of the GO term analysis done to elucidate the general functions gained at each transition node. We used the HOGs gained at each node and took the GO terms of the gene members from these HOGs that were also gained at this node. We carried out an enrichment analysis of these terms at the transition nodes as well as the parent nodes for each lineage. The enriched terms were then clustered using Lin Semantic Similarity and Louvain Clustering and we determined the top keywords in each cluster to label them.* ***D-E)*** *Mean (and Standard deviation) of Log2 Odds Ratio of GO terms in the different clusters for transition nodes compared with parent nodes. For points only seen in transition nodes, we did not find an enrichment of terms for that cluster in the parent nodes. Asterisks indicate a significantly greater mean of Log2 Odds Ratio at the transition node. We focused on GO:BP (biological process) clusters for aquatic to terrestrial and GO:CC (cellular component) clusters for unicellular to multicellular.*

In both phenotypic transitions, we observed that the HOGs containing these orthologs often originated well before the actual phenotypic transition nodes. This pattern suggests that ancient gene families provided a shared genomic background that was independently co-opted for similar adaptive functions. For example, the presence of Hox genes in arthropods and tetrapods, or Aquaporins in plants, arthropods and tetrapods, reflects expansions or lineage-specific deployments of older gene families. Indeed, a recent study showed that functions associated with more than half of all human protein-coding genes have ancient origins, particularly at the root of all eukaryotes (*45*). Thus, while individual transitions and species may recruit different gene sets, they frequently draw from a common pool of conserved functions.

Together, these results show that genes implicated in key evolutionary transitions in one lineage are often present—as part of older gene families—in other, independently evolving lineages. This supports the idea that parallel phenotypic outcomes can arise through repeated use or expansion of shared genomic components and underscores the utility of combining phylogenomic inference with functional annotation to detect such patterns.

Convergent enrichment of functional categories across phenotypic transitions

Beyond examining known genes involved in adaptive transitions, we sought to determine whether broader functional signatures could be identified at the transition nodes. Specifically, we investigated which general biological functions were enriched in each lineage undergoing either a water-to-land or unicellular-to-multicellular transition. To achieve this, we identified enriched Gene Ontology (GO) terms associated with HOGs gained at each transition node, clustered these terms by semantic similarity (using Lin similarity (*46*)), and compared their enrichment to that of their ancestral nodes (see Materials and Methods; Fig. 3C). An equivalent exercise using pathway information proved uninformative largely because of limited (or even entirely flat) hierarchies, in contrast to the deep hierarchical structure of GO.

Each GO cluster represents a set of functionally related terms linked through shared parent categories in the GO hierarchy. Several clusters revealed biologically coherent patterns that, in retrospect, match known adaptive pressures associated with these transitions.

For the aquatic-to-terrestrial transition, enrichment was observed in functional categories related to environmental adaptation and structural innovation. One major GO cluster included terms associated broadly with anatomical structure development—such as "seed," "pollen," "roots," and "development"—and was significantly enriched across all three water-to-land transition nodes. This enrichment was greater than in the corresponding ancestral nodes (Fig. 3D; Table S5). While the genes contributing to these clusters differed across clades (Fig. S6), the encoded functions converged on common adaptive strategies. For instance, in tetrapods, we found osteopontin (SPP1) and dentin matrix acidic phosphoprotein 1 (DMP1), both involved in bone development; in plants, we observed regulators like NAC transcription factors involved in seed, leaf and flower development (e.g., NAC034, NAC051, NAC098), lignification-related CYP73A5, and vascular patterning genes such as class III HD-Zip genes (e.g., Revoluta); in arthropods, gains included components of the open tracheal system such as Protein Vhl, Zinc Finger Protein Elbow, and Noc.

Another major cluster linked to water-to-land transitions involved stress responses, light perception, and defense. This cluster was significantly enriched at all terrestrial transition nodes (Fig. 3D; Table S6). In plants, this included genes involved in red/far-red light response such as *Empfindlicher im dunkelroten Licht* protein 1 (EID1), as well as abscisic acid signaling genes (e.g., ABI5, ABF1–4), which mediate responses to desiccation and other land-specific stresses. In land animals, this cluster encompassed expansions in components of the adaptive immune system, including interferons and LY96, reflecting increased pathogen defense complexity.

In the unicellular-to-multicellular transition, we expected functional enrichment related to cell–cell communication and coordination. Indeed, we observed a GO cluster involving terms like "synapse," "adherens," and "junction" in both land plants and Metazoa (Fig. 3E). In animals, this cluster included beta- and delta-catenins and Ephrins, which mediate intercellular adhesion and signaling (Fig. 3B, Fig. S7). In plants, the same cluster included genes involved in plasmodesmata formation, the plant-specific channels for intercellular communication.

Together, these results show that large phenotypic transitions are accompanied not only by lineage-specific gene innovations but also by convergent enrichment of functional categories. These shared functional signatures, detectable only when phylogenetic context is combined with functional annotation, point to common adaptive solutions despite underlying genomic differences.

Discussion

The genomic era has ushered in automated systems that assign gene annotations—GO terms, protein domains, and gene descriptions—at the time of sequence deposition into databases like UniProt *(47, 48)*, NCBI *(49)*, and ENSEMBL *(50)*. These pipelines typically rely on sequence homology and rule-based systems such as UniRule *(51)* and GO taxonomic constraints *(52)*. Over time, annotations are refined through community curation and experimental validation, notably through resources like UniProt-SwissProt *(53)* and ENSEMBL-HAVANA. As annotation systems and functional genomics improve, we edge closer to a more comprehensive understanding of all gene functions.

But just how informative are current annotations? Given the apparent biases of genome annotations towards humans, and a handful of other well-studied model organisms, it is reassuring that we can nevertheless capture common broad functions converging across vast evolutionary distances. For example, functional profiles gained at key evolutionary nodes correspond strikingly with phenotypes historically linked to those clades—developmental processes in animals, light and stress responses in land plants, or multicellular-specific adhesion and signaling in metazoans and fungi. These results not only validate the general reliability of functional annotations, but also point to their untapped potential as predictive tools in comparative genomics.

Crucially, we introduce a top-down approach to evolutionary functional genomics. Instead of selecting a clade or function of interest and tracing its evolutionary path, we infer the emergence of all annotated biological functions across a comprehensive, deeply rooted phylogeny. This framework makes it possible to detect parallel functional evolution—even when the underlying genes differ—by capturing convergent gains of broader functional categories. For instance, organisms that transitioned independently to terrestrial habitats show gains in functions related to environmental stress tolerance and anatomical structures, while multicellularity is associated with cell communication and structural coordination, despite involving entirely different gene families.

These findings demonstrate that such phylogenomic analyses can reveal evolutionary convergences, adaptive innovations, and latent parallels that traditional gene- or clade-centric methods may miss. Any newly sequenced genome can now be easily placed into this functional phylogenetic framework, revealing both its unique and conserved functional innovations. This can serve as a launch point for hypothesis generation in functional biology, development, and evolutionary studies of newly sequenced species.

While orthology inference and phylogenetic placement are not without pitfalls (e.g., errors in gene presence/absence or species misplacement (*54, 55*)), our results show that large-scale inferences when carefully made without tweaking the underlying species tree or orthology inferences can yield meaningful biological insights. Researchers focusing on specific lineages or functions can refine this approach further by deepening taxonomic sampling or incorporating curated annotations.

Looking ahead, this framework can be made even more powerful when coupled with codified phenotypic data. Descriptive traits, often scattered across textbooks, databases, or museums, are rarely digitized in a standardized way. Harmonizing and encoding phenotypic knowledge would unlock the possibility of building a comprehensive genotype–phenotype map of life, linking functional genomic innovations to organismal form and behavior on evolutionary timescales. Combined with modern phylogenomics, such a resource would be transformative for understanding evolution, development, and even biomedical questions rooted in fundamental biology.

References

1. H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, M. M. Goldstein, I. V. Grigoriev, K. J. Hackett, D. Haussler, E. D. Jarvis, W. E. Johnson, A. Patrinos, S. Richards, J. C. Castilla-Rubio, M.-A. van Sluys, P. S. Soltis, X. Xu, H. Yang, G. Zhang, Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4325–4333 (2018).

2. A. Rhie, S. A. McCarthy, O. Fedrigo, J. Damas, G. Formenti, S. Koren, M. Uliano-Silva, W. Chow, A. Fungtammasan, J. Kim, C. Lee, B. J. Ko, M. Chaisson, G. L. Gedman, L. J. Cantin, F. Thibaud-Nissen, L. Haggerty, I. Bista, M. Smith, B. Haase, J. Mountcastle, S. Winkler, S. Paez, J. Howard, S. C. Vernes, T. M. Lama, F. Grutzner, W. C. Warren, C. N. Balakrishnan, D. Burt, J. M. George, M. T. Biegler, D. Iorns, A. Digby, D. Eason, B. Robertson, T. Edwards, M. Wilkinson, G. Turner, A. Meyer, A. F. Kautt, P. Franchini, H. W. Detrich 3rd, H. Svardal, M. Wagner, G. J. P. Naylor, M. Pippel, M. Malinsky, M. Mooney, M. Simbirsky, B. T. Hannigan, T. Pesout, M. Houck, A. Misuraca, S. B. Kingan, R. Hall, Z. Kronenberg, I. Sović, C. Dunn, Z. Ning, A. Hastie, J. Lee, S. Selvaraj, R. E. Green, N. H. Putnam, I. Gut, J. Ghurye, E. Garrison, Y. Sims, J. Collins, S. Pelan, J. Torrance, A. Tracey, J. Wood, R. E. Dagnew, D. Guan, S. E. London, D. F. Clayton, C. V. Mello, S. R. Friedrich, P. V. Lovell, E. Osipova, F. O. Al-Ajli, S. Secomandi, H. Kim, C. Theofanopoulou, M. Hiller, Y. Zhou, R. S. Harris, K. D. Makova, P. Medvedev, J. Hoffman, P. Masterson, K. Clark, F. Martin, K. Howe, P. Flicek, B. P. Walenz, W. Kwak, H. Clawson, M. Diekhans, L. Nassar, B. Paten, R. H. S. Kraus, A. J. Crawford, M. T. P. Gilbert, G. Zhang, B. Venkatesh, R. W. Murphy, K.-P. Koepfli, B. Shapiro, W. E. Johnson, F. Di Palma, T. Marques-Bonet, E. C. Teeling, T. Warnow, J. M. Graves, O. A. Ryder, D. Haussler, S. J. O'Brien, J. Korlach, H. A. Lewin, K. Howe, E. W. Myers, R. Durbin, A. M. Phillippy, E. D. Jarvis, Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).

3. M. Blaxter, N. Mieszkowska, F. Di Palma, P. Holland, R. Durbin, T. Richards, M. Berriman, P. Kersey, P. Hollingsworth, W. Wilson, A. Twyford, E. Gaya, M. Lawniczak, O. Lewis, G. Broad, K. Howe, M. Hart, P. Flicek, I. Barnes, Sequence locally, think globally: The Darwin Tree of Life Project. *Proc. Natl. Acad. Sci. U. S. A.* **119** (2022).

4. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. Levine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, M. L. Hong, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. De La Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G.

Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

5. Mouse Genome Sequencing Consortium, R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigó, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S.-P. Yang, E. M. Zdobnov, M. C. Zody, E. S. Lander, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

6. M. R. Brent, Genome annotation past, present, and future: How to define an ORF at each locus. *Genome Res.* **15**, 1777–1786 (2005).

7. Y. Wang, Y. Zhao, A. Bollas, Y. Wang, K. F. Au, Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).

8. F. Langschied, N. Bordin, S. Cosentino, D. Fuentes-Palacios, N. Glover, M. Hiller, Y. Hu, J. Huerta-Cepas, L. P. Coelho, W. Iwasaki, S. Majidian, S. Manzano-Morales, E. Persson, T. A. Richards, T. Gabaldón, E. Sonnhammer, P. D. Thomas, C. Dessimoz, I. Ebersberger, Quest for Orthologs in the Era of Biodiversity Genomics. *Genome Biol. Evol.* **16**, evae224 (2024).

9. R. Guigó, Genome annotation: From human genetics to biodiversity genomics. *Cell Genom.* **3**, 100375 (2023).

10. E. V. Koonin, Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136 (2003).

11. T. Domazet-Lošo, J. Brajković, D. Tautz, A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).

12. L. J. Revell, L. J. Harmon, *Phylogenetic Comparative Methods in R* (Princeton University Press, Princeton, NJ, 2022).

13. M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 1–8 (2016).

14. J. Paps, P. W. H. Holland, Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. *Nat. Commun.* **9**, 1–8 (2018).

15. G. D. Diwan, D. Agashe, Wobbling Forth and Drifting Back: The Evolutionary History and Impact of Bacterial tRNA Modifications. *Mol. Biol. Evol.* **35**, 2046–2059 (2018).

16. C. Guijarro-Clarke, P. W. H. Holland, J. Paps, Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat. Ecol. Evol.* **4**, 519–523 (2020).

17. R. Fernández, T. Gabaldón, Gene gain and loss across the metazoan tree of life. *Nat. Ecol. Evol.* **4**, 524–533 (2020).

18. J. Barrera-Redondo, J. S. Lotharukpong, H.-G. Drost, S. M. Coelho, Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. *Genome Biol.* **24**, 54 (2023).

19. Z. Merényi, K. Krizsán, N. Sahu, X.-B. Liu, B. Bálint, J. E. Stajich, J. W. Spatafora, L. G. Nagy, Genomes of fungi and relatives reveal delayed loss of ancestral gene families and evolution of key fungal traits. *Nat. Ecol. Evol.* **7**, 1221–1231 (2023).

20. N. Konno, W. Iwasaki, Machine learning enables prediction of metabolic system evolution in bacteria. *Sci. Adv.* **9**, eadc9130 (2023).

21. P. O. Mulhair, R. J. Moran, J. S. Pathmanathan, D. Sussfeld, C. J. Creevey, K. Siu-Ting, F. J. Whelan, D. Pisani, B. Constantinides, E. Pelletier, P. Lopez, E. Bapteste, J. O. McInerney, M. J. O'Connell, Bursts of novel composite gene families at major nodes in animal evolution. *bioRxivorg*, 2023.07.10.548381 (2023).

22. M. Domazet-Lošo, T. Široki, K. Šimičević, T. Domazet-Lošo, Macroevolutionary dynamics of gene family gain and loss along multicellular eukaryotic lineages. *Nat. Commun.* **15**, 2663 (2024).

23. M. Zahn-Zabal, C. Dessimoz, N. M. Glover, Identifying orthologs with OMA: A primer. *F1000Res.* **9** (2020).

24. A. Hernández-Plaza, D. Szklarczyk, J. Botas, C. P. Cantalapiedra, J. Giner-Lamia, D. R. Mende, R. Kirsch, T. Rattei, I. Letunic, L. J. Jensen, P. Bork, C. von Mering, J. Huerta-Cepas, eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* **51**, D389–D394 (2023).

25. V. Rossier, C. Train, Y. Nevers, M. Robinson-Rechavi, C. Dessimoz, Matreex: Compact and interactive visualization for scalable studies of large gene families. *Genome Biol. Evol.* **16** (2024).

26. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).

27. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

28. S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. Van Auken, J. Ramsey, D. A. Siegele, R. L. Chisholm, P. Fey, M. C. Aspromonte, M. V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. dos Santos, S. Marygold, V. Strelets, C. J. Tabone, J. Thurmond, P. Zhou, S. H. Ahmed, P. Asanitthong, D. Luna Buitrago, M. N. Erdol, M. C. Gage, M. Ali Kadhum, K. Y. C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S. C. C. Saverimuttu, R. Su, K. E. Thurlow, R. C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. Sitnikov, C. Smith, A. Cuzick, J. Seager, L. Cooper, J. Elser, P. Jaiswal, P. Gupta, P. Jaiswal, S. Naithani, M. Lera-Ramirez, K. Rutherford, V. Wood, J. L. De Pons, M. R. Dwinell, G. T. Hayman, M. L. Kaldunski, A. E. Kwitek, S. J. F. Laulederkind, M. A. Tutaj, M. Vedi, S.-J. Wang, P. D'Eustachio, L. Aimo, K. Axelsen, A. Bridge, N. Hyka-Nouspikel, A. Morgat, S. A. Aleksander, J. M. Cherry, S. R. Engel, K. Karra, S. R. Miyasato, R. S. Nash, M. S. Skrzypek, S. Weng, E. D. Wong, E. Bakker, T. Z. Berardini, L. Reiser, A. Auchincloss, K. Axelsen, G. Argoud-Puy, M.-C. Blatter, E. Boutet, L. Breuza, A. Bridge, C. Casals-Casas, E. Coudert, A. Estreicher, M. Livia Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, C. Hulo, N. Hyka-Nouspikel, F. Jungo, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, I. Pedruzzi, L. Pourcel, S. Poux, C. Rivoire, S. Sundaram, A. Bateman, E. Bowler-Barnett, H. Bye-A-Jee, P. Denny, A. Ignatchenko, R. Ishtiaq, A. Lock, Y. Lussi, M. Magrane, M. J. Martin, S. Orchard, P. Raposo, E. Speretta, N. Tyagi, K. Warner, R. Zaru, A. D. Diehl, R. Lee, J. Chan, S. Diamantakis, D. Raciti, M. Zarowiecki, M. Fisher, C. James-Zorn, V. Ponferrada, A. Zorn, S. Ramachandran, L. Ruzicka, M. Westerfield, The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).

29. J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, A. Bateman, Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

30. M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, M. Ishiguro-Watanabe, KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).

31. A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P.

D'Eustachio, The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).

32. M. Milacic, D. Beavers, P. Conley, C. Gong, M. Gillespie, J. Griss, R. Haw, B. Jassal, L. Matthews, B. May, R. Petryszak, E. Ragueneau, K. Rothfels, C. Sevilla, V. Shamovsky, R. Stephan, K. Tiwari, T. Varusai, J. Weiser, A. Wright, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.* **52**, D672–D678 (2024).

33. A. M. C. Bowles, U. Bechtold, J. Paps, The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. *Curr. Biol.* **30**, 530–536.e2 (2020).

34. N. Shubin, *Your Inner Fish* (Random House, New York, NY, 2009).

35. P. W. H. Holland, Evolution of homeobox genes: Evolution of homeobox genes. *Wiley Interdiscip. Rev. Dev. Biol.* **2**, 31–45 (2013).

36. G. I. Martínez-Redondo, C. Simón Guerrero, L. Aristide, P. Balart-García, V. Tonzo, R. Fernández, Parallel duplication and loss of aquaporin-coding genes during the "out of the sea" transition as potential key drivers of animal terrestrialization. *Mol. Ecol.* **32**, 2022–2040 (2023).

37. Y. Takeda, S. Hatano, N. Sentoku, M. Matsuoka, Homologs of animal eyes absent (eya) genes are found in higher plants. *Mol. Gen. Genet.* **262**, 131–138 (1999).

38. N. Ros-Rocher, A. Pérez-Posada, M. M. Leger, I. Ruiz-Trillo, The origin of animals: an ancestral reconstruction of the unicellular-to-multicellular transition. *Open Biol.* **11**, 200359 (2021).

39. D. J. Richter, P. Fozouni, M. B. Eisen, N. King, Gene family innovation, conservation and loss on the animal stem lineage. *Elife* **7**, e34226 (2018).

40. M. R. Chastney, J. R. W. Conway, J. Ivaska, Integrin adhesion complexes. *Curr. Biol.* **31**, R536–R542 (2021).

41. X. Grau-Bové, G. Torruella, S. Donachie, H. Suga, G. Leonard, T. A. Richards, I. Ruiz-Trillo, Dynamics of genomic innovation in the unicellular ancestry of animals. *Elife* **6**, e26036 (2017).

42. T. Brunet, N. King, The origin of animal multicellularity and cell differentiation. *Dev. Cell* **43**, 124–140 (2017).

43. E. R. Hanschen, T. N. Marriage, P. J. Ferris, T. Hamaji, A. Toyoda, A. Fujiyama, R. Neme, H. Noguchi, Y. Minakuchi, M. Suzuki, H. Kawai-Toyooka, D. R. Smith, H. Sparks, J. Anderson, R. Bakarić, V. Luria, A. Karger, M. W. Kirschner, P. M. Durand, R. E. Michod, H. Nozaki, B. J. S. C. Olson, The Gonium pectorale genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. *Nat. Commun.* **7**, 11370 (2016).

44. E. Kiss, B. Hegedüs, M. Virágh, T. Varga, Z. Merényi, T. Kószó, B. Bálint, A. N. Prasanna, K. Krizsán, S. Kocsubé, M. Riquelme, N. Takeshita, L. G. Nagy, Comparative genomics reveals the origin of fungal hyphae and multicellularity. *Nat. Commun.* **10**, 4080 (2019).

45. M. Feuermann, H. Mi, P. Gaudet, A. Muruganujan, S. E. Lewis, D. Ebert, T. Mushayahama, Gene Ontology Consortium, P. D. Thomas, A compendium of human gene functions derived from evolutionary modelling. *Nature* **640**, 146–154 (2025).

46. D. Lin, “An Information-Theoretic Definition of Similarity” in *Proceedings of the Fifteenth International Conference on Machine Learning* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998)*ICML ’98*, pp. 296–304.

47. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

48. UniProt Consortium, UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).

49. N. A. O’Leary, E. Cox, J. B. Holmes, W. R. Anderson, R. Falk, V. Hem, M. T. N. Tsuchiya, G. D. Schuler, X. Zhang, J. Torcivia, A. Ketter, L. Breen, J. Cothran, H. Bajwa, J. Tinne, P. A. Meric, W. Hlavina, V. A. Schneider, Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* **11**, 732 (2024).

50. P. W. Harrison, M. R. Amode, O. Austine-Orimoloye, A. G. Azov, M. Barba, I. Barnes, A. Becker, R. Bennett, A. Berry, J. Bhai, S. K. Bhurji, S. Boddu, P. R. Branco Lins, L. Brooks, S. B. Ramaraju, L. I. Campbell, M. C. Martinez, M. Charkhchi, K. Chougule, A. Cockburn, C. Davidson, N. H. De Silva, K. Dodiya, S. Donaldson, B. El Houdaigui, T. E. Naboulsi, R. Fatima, C. G. Giron, T. Genez, D. Grigoriadis, G. S. Ghattaoraya, J. G. Martinez, T. A. Gurbich, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, M. Kay, V. Kaykala, T. Le, D. Lemos, D. Lodha, D. Marques-Coelho, G. Maslen, G. A. Merino, L. P. Mirabueno, A. Mushtaq, S. N. Hossain, D. N. Ogeh, M. P. Sakthivel, A. Parker, M. Perry, I. Piližota, D. Poppleton, I. Prosovetskaia, S. Raj, J. G. Pérez-Silva, A. I. A. Salam, S. Saraf, N. Saraiva-Agostinho, D. Sheppard, S. Sinha, B. Sipos, V. Sitnik, W. Stark, E. Steed, M.-M. Suner, L. Surapaneni, K. Sutinen, F. F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T. A. Walsh, D. Ware, E. Wass, N. L. Willhoft, J. Allen, J. Alvarez-Jarreta, M. Chakiachvili, B. Flint, S. Giorgetti, L. Haggerty, G. R. Ilsley, J. Keatley, J. E. Loveland, B. Moore, J. M. Mudge, G. Naamati, J. Tate, S. J. Trevanion, A. Winterbottom, A. Frankish, S. E. Hunt, F. Cunningham, S. Dyer, R. D. Finn, F. J. Martin, A. D. Yates, Ensembl 2024. *Nucleic Acids Res.* **52**, D891–D899 (2024).

51. A. MacDougall, V. Volynkin, R. Saidi, D. Poggioli, H. Zellner, E. Hatton-Ellis, V. Joshi, C. O’Donovan, S. Orchard, A. H. Auchincloss, D. Baratin, J. Bolleman, E. Coudert, E. de Castro, C. Hulo, P. Masson, I. Pedruzzi, C. Rivoire, C. Arighi, Q. Wang, C. Chen, H. Huang, J. Garavelli, C. R. Vinayaka, L.-S. Yeh, D. A. Natale, K. Laiho, M.-J. Martin, A. Renaux, K. Pichler, UniProt Consortium, UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics* **36**, 4643–4648 (2020).

52. J. I. Deegan née Clark, E. C. Dimmer, C. J. Mungall, Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC Bioinformatics* **11**, 530 (2010).

53. S. Poux, C. N. Arighi, M. Magrane, A. Bateman, C.-H. Wei, Z. Lu, E. Boutet, H. Bye-A-Jee, M. L. Famiglietti, B. Roechert, T. UniProt Consortium, On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* **33**, 3454–3460 (2017).

54. P. Natsidis, P. Kapli, P. H. Schiffer, M. J. Telford, Systematic errors in orthology inference and their effects on evolutionary analyses. *iScience* **24**, 102110 (2021).

55. P. Kapli, T. Flouri, M. J. Telford, Systematic errors in phylogenetic trees. *Curr. Biol.* **31**, R59–R64 (2021).

56. R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing [Preprint] (2023). https://www.R-project.org/.

57. H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

58. H. Wickham, M. Girlich, E. Ruiz, dbplyr: A "dplyr" Back End for Databases. [Preprint] (2024). https://CRAN.R-project.org/package=dbplyr.

59. H. Wickham, J. Ooms, K. Müller, RPostgres: C++ Interface to PostgreSQL. [Preprint] (2023). https://CRAN.R-project.org/package=RPostgres.

60. L. J. Revell, phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

61. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

62. G. Yu, Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).

63. C. O. Wilke, cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2." [Preprint] (2024). https://CRAN.R-project.org/package=cowplot.

64. T. L. Pedersen, patchwork: The Composer of Plots. [Preprint] (2024). https://cran.r-project.org/package=patchwork.

65. E. Neuwirth, RColorBrewer: ColorBrewer Palettes. [Preprint] (2022). https://CRAN.R-project.org/package=RColorBrewer.

66. J. D. Zhang, S. Wiemann, KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* **25**, 1470–1471 (2009).

67. D. Greene, S. Richardson, E. Turro, ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* **33**, 1104–1106 (2017).

68. Y. Nevers, T. E. M. Jones, D. Jyothi, B. Yates, M. Ferret, L. Portell-Silva, L. Codo, S. Cosentino, M. Marcet-Houben, A. Vlasova, L. Poidevin, A. Kress, M. Hickman, E. Persson, I. Piližota, C. Guijarro-Clarke, W. Iwasaki, O. Lecompte, E. Sonnhammer, D. S. Roos, T. Gabaldón, D. Thybert, P. D. Thomas, Y. Hu, D. M. Emms, E. Bruford, S. Capella-Gutierrez, M. J. Martin, C. Dessimoz, A. Altenhoff, The Quest for Orthologs orthology benchmark service in 2022. *Nucleic Acids Res.* **50**, W623–W632 (2022).

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

70. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

71. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

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

73. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).

74. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

75. H. Shimodaira, M. Hasegawa, Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).

76. S. Sherrill-Mix, taxonomizr: Functions to Work with NCBI Accessions and Taxonomy. [Preprint] (2023). https://CRAN.R-project.org/package=taxonomizr.

77. S. Ohno, *Evolution by Gene Duplication* (Springer Berlin Heidelberg, Berlin, Heidelberg, 1970; http://dx.doi.org/10.1007/978-3-642-86659-3).

78. M. Csurös, Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).

79. D. L. Swofford, W. P. Maddison, Reconstructing ancestral character states under Wagner parsimony. *Math. Biosci.* **87**, 199–229 (1987).

80. J. P. Bollback, SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7**, 88 (2006).

81. Z. Gu, simona: a comprehensive R package for semantic similarity analysis on bio-ontologies. *BMC Genomics* **25**, 869 (2024).

82. Z. Gu, D. Hübschmann, SimplifyEnrichment: A bioconductor package for clustering and visualizing functional enrichment results. *Genomics Proteomics Bioinformatics* **21**, 190–202 (2023).

Acknowledgments: We would like to thank Arne Sebé-Pedrós, Jan Korbel, Supriya Khedkar and the Russell group for feedback and suggestions. We thank Damien DeVienne and David Emms for help with the LifeMap and OrthoFinder tools. We thank Denise Duma for introducing semantic similarity. We acknowledge the use of chatGPT for checking grammar, overall clarity and flow of the manuscript text.

Funding:

Wellcome Trust UK 210585/A/18/Z (GDD)

German Federal Ministry of Research, Technology and Space (BMBF) PreDiCT, FKZ-01GM2306 (GDD)

European Union Horizon 2020 PrecisionTox project, Grant Agreement Number 965405 (JCGS, MEC)

Swedish Research Council Grant VR2018-05882, UmU project number 322045306 (JCGS)

Author contributions:

Conceptualization: RBR, GDD, MJT, RG, JKC

Methodology: GDD, RBR, PN, MJT

Data Curation: GDD, JCGS, MEC, PN, RBR

Software: GDD, JCGS, PN

Investigation: GDD, RBR

Visualization: GDD, RBR

Funding acquisition: RBR

Project administration: RBR

Supervision: RBR

Writing – original draft: GDD, RBR

Writing – review & editing: GDD, JCGS, MJT, RG, JKC, RBR

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data are available in the main text or the supplementary materials, an online web application and scripts via a GitHub repository.

Supplementary Materials

Materials and Methods

Figs. S1 to S8

Tables S1 to S7

References (*56–82*)

Materials and Methods

All analyses were done using the R programming language (56) and the tidyverse set of packages (57) unless noted otherwise. Most of our large data tables were added to a PostgreSQL database that was queried in R using the dbplyr (58) and RPostgres (59) packages. For tree manipulations and visualizations, we used packages such as phytools (60), ape (61), ggtree (62), cowplot (63), patchwork (64), and RColorBrewer (65).

Proteomes, pathways and functional annotations

We sampled 508 species from across the tree of life (superkingdom Archaea, Bacteria and Eukaryota) from the full list of Uniprot reference proteomes by trying to keep at least one representative from a taxonomic class while proportionately sampling larger clades (proteomes downloaded April 2021, Table S1). We downloaded the canonical protein sequences from each proteome (4,543,007 proteins in total) and used the Uniprot accessions as the anchor point for all subsequent analyses. Next, we downloaded the Uniprot accessions of all protein components for every pathway in KEGG (last updated July 2023) and in Reactome (downloaded July 2022). For each KEGG pathway, we also downloaded the "kgml" graph files associated with every pathway. Using the KEGGgraph R package (66), we then extracted the components (nodes) connections (egdes) in each pathway, be it protein-protein or protein-chemical and all the reactions catalyzed by different enzymes. For GO terms, we extracted all the three GO category terms for nearly every protein in our dataset (4.53 million out of 4.54 million proteins) from the Uniprot knowledgebase file "uniprot_sprot.dat.gz" (August 2021) using the Bio.SwissProt module in Biopython. For each term detected, we also captured the complete lineage using the GO obo file (https://purl.obolibrary.org/obo/go.obo) and the ontologyIndex R package (67).

Inferring orthologs and phylogenetic tree building

We used the OrthoFinder (versions 2.5.2 – 2.5.5) (26) program with the option -og to first infer the Orthogroups (OGs) for all of our proteomes. OrthoFinder has been shown as one of the highest performing orthology programs (68), it uses a phylogenetic framework and definition of orthology and is scalable for a large number of genomes. Overall, 91% of the 4.54 million proteins in our dataset were assigned to 154k OGs (Table S7). We determined the OGs that had <= 10 gene copies in at least 50% of the species. We built the multiple sequence alignments of each of the fasta sequences of the OGs using MAFFT-linsi (69) with the option --anysymbol and built gene trees for each one of them using IQ-tree (70) with the option of model testing with the LG, WAG and JTT models (71).

Using the gene trees, we removed paralogs using a custom python script; briefly, we checked if the paralogs from a species are monophyletic. If so, we only kept the paralog with the shortest branch length, else, we dropped all paralogs. We then resampled the OG fasta files by keeping one paralog per species and determined the OGs that had a sequence from at least 249 species (roughly 50% of all species). We aligned the 239 selected OG fasta files using MAFFT-linsi, removed positions with large number of gaps using trimAL (72) with option –gappyout, concatenated the alignments and used FastTree (73, 74) with the options -lg (for the LG model) and -cat 10 (for 10 rate categories) to get the species tree. We chose the LG model and 10 rate categories based on the model testing done for each OG gene tree mentioned above. FastTree uses the Shimodaira-Hasegawa test (75) to calculate local branch support values by testing all possible alternative topologies and 1000 resamplings of the alignments (Fig. S1). Finally, the phylogenetic species tree was used to split orthologs from paralogs by using the -fg, -s, and -y options in OrthoFinder.

Taxonomy of tree nodes

For every node, we sought to assign taxonomic information and obtained the entire NCBI taxonomy lineage for each species using the taxonomizr R package (76) (and the "names.dmp" and "nodes.dmp" files from NCBI taxonomy FTP; updated February 2025). We then determined the name from the broadest taxonomic level that encompassed all of the descendant species. In some cases, we also captured names from lower taxonomic levels that encompassed >75% of the descendant species. This approach worked pretty well, except for some nodes where the descendant lineages led to an equal proportion of species from multiple groups (Table S2). In such cases, the node usually had a very broad classification (e.g., superkingdom Eukaryota or Bacteria).

Ancestral reconstruction of genes/pathways/domains

For genes, OrthoFinder assigns them to Phylogenetically Hierarchical Orthogroups (HOGs - https://github.com/davidemms/OrthoFinder?tab=readme-ov-file#phylogenetic-hierarchical-orthogroups-directory) and we assumed that each HOG was a single gene at some point in evolutionary history and duplicated/diverged further (77). We used the HOGs from OrthoFinder and their presence/absence in every species in our study to carry out ancestral reconstruction. We also created similar presence/absence matrices for the orthologs of every gene calculated above. For KEGG pathways, we used the API to download the KEGG orthology groups associated with all pathways and hierarchies (https://rest.kegg.jp/link/path/ko) and the Uniprot accessions present in every KEGG orthology group. To make a presence/absence matrix for each KEGG orthology group in every pathway and hierarchy, we also extended the KEGG orthology using our own orthologs and then calculated the species where each group was present/absent. We needed to be careful here and not imply the unknown presence of a KEGG pathway in some organisms as a result of using our orthology. Hence, we forced the KEGG orthology groups associated with a pathway to be absent in an organism if the pathway also did not exist for that organism in KEGG.

In the case of Pfam domains, we determined the Pfam domain content for all Uniprot accessions in our dataset and calculated the presence/absence of each domain in our species. For ancestral reconstruction in each case, we used the Count program (78) with the Wagner parsimony (79) option and gain/loss penalty equal to 1 meaning both are equally possible. From the output that the Count program provided, we extracted the states of each entity at every node in the tree and noted the descendant nodes where the state changed from absent (in the parent node) to present (in the descendant node) as "gain nodes" and "loss nodes" for the opposite.

Biological entity counts

We counted the number of genes and Pfam domains, gained at every node in the tree by checking if the state of the parent node was "absent" (0) and the node of interest was "present" (1). For KEGG/Reactome pathways, we determined the proportion of components "present" at every node of the tree and used the difference between the proportion at a node and the proportion at the parent node as the proportion gained at each node (Table S1, Table S2).

Ancestral reconstruction of phenotypes

For phenotypes, we decided to focus only on eukaryotes as the encoding of phenotypes was distinct from bacteria. While bacterial habitats could be inferred for most species in our study, their habitats are very dynamic and non-discrete in many cases. Conversely, in eukaryotes we could use a broad classification. For instance, we considered only the primary

habitat of the adult where it spends the most time in nature. Only frogs and the duck-billed platypus were considered amphibians. All eukaryotes that were pathogens or commensals or endocytic were called "Host-associated". The terrestrial label also included airborne organisms. The cellularity phenotype was relatively straightforward to encode (Table S1).

Using the trait matrix for the tips of the eukaryotic subtree in our study (199 tips), we carried out stochastic character mapping *(80)* using the "make.simmap" function in the "phytools" package of R. In each case we used the Bayesian "mcmc" option to calculate the transition matrix with the equal rates (ER) model and the number of simulations being 100. We forced the eukaryotic root to be aquatic and unicellular and allowed the function to estimate the probabilities of each trait category at every node of the tree. To assign an ancestral state to the nodes in the tree, we used a posterior probability cutoff of 0.5 for habitat data and 0.75 for the cellularity data i.e. if any category had a probability >= the cutoff, we assigned that as the state. Otherwise, the state of the parent node was transferred to the descendant node. We then determined the nodes that had at least three descendant species where the states switched from aquatic to terrestrial or unicellular to multicellular.

GO term enrichment and clustering

Throughout the manuscript, we have carried out enrichment of GO terms using a modified Fisher's test. For any node of interest, we made a 2x2 contingency table as follows:

	At Node	Not at Node
With GO term	# unique HOGs	# unique HOGs
Without GO term	# unique HOGs	# unique HOGs

By considering the number of unique HOGs to represent genes that contain a term, we were able to ensure independence for the gene numbers since each HOG can be thought of as an independent entity. We believe this approach was extremely powerful and helped capture the results in our study. The p-values were corrected using the Benjamini-Hochberg method and enrichment was filtered by keeping GO terms with an Odds Ratio > 1 and an adjusted p-value < 0.05. To get the top terms in Fig. 2, we had to consider the hierarchical structure of the GO term data and hence we removed some parent terms in the filtered significant lists by keeping only the descendant terms if a broader parent term was also seen in the list.

For the GO term clustering in Fig. 3, we analyzed each GO category (BP, MF, CC) separately. We focused on the enriched terms at each of the transition nodes which were determined by taking the terms having p-value < 0.05, a finite odds ratio > 1 and at least five HOGs with the term. We used the "term_sim" function from the simona package *(81)* in R with the "Sim_Lin_1998" method *(46)*. We clustered the terms using the "cluster_terms" function from the simplifyEnrichment package *(82)* with the "louvain" method and a resolution parameter of 1. We determined the top five keywords in each cluster using the "keyword_enrichment_from_GO" function from the simplifyEnrichment package with default parameters. This function carries out a Fisher's test for the words in a GO term list and arranges them by adjusted p-values. These keywords were used to label each cluster (Table S5, Table S6). Within each cluster, we calculated the mean of the Log2 Odds Ratio for the terms at the transition node and compared them with the mean of the Log2 Odds ratio of all the parent nodes per lineage. Here, we used a t-test with an alternate hypothesis that the ratio at the transition node was greater than the parent nodes.

Term of interest searches

We exploited the hierarchical nature of GO terms to enhance the searching of a function of interest. Each keyword was used as a regular expression to search a SQL table containing the entire hierarchy of all terms associated with all genes in our study. As a result, even genes that may have the function of interest as a broader parent term can be included to understand the evolutionary history of the keyword i.e. function itself. This is especially useful when a specific GO term assigned to a gene may not necessarily contain the keyword of interest in the term name. For all the genes captured for a search, we took the nodes where they were gained and to reduce the noise in the taxonomic assignments, assigned nodes with taxonomy levels lower than order up to the order level and combined the calculations for each taxonomic name. This approach was in general conservative and sometimes captured minute presences of GO terms in spurious taxa (Fig. S8). To circumvent these, we used coarse intervals for the % genes gained at a node (2%) and also calculated the z-scores of the % gained for a taxon in relation to the mean for a search term and determined the taxa which had a z-score > 1.96 (asterisks in Fig. 2E).

Transition node orthologs

We sourced the known genes in the phenotypic transitions and made some choices that are detailed in Tables S3-4. For the unicellular to multicellular transition, there were other known genes in the animal ancestor that have been attributed to be involved in the unicellular to multicellular transition such as Wnt, TGF-ß, SMADs, Frizzled, dishevelled, Hemicentins, Fermitins and Dystroglycans, however these exclusively emerged in the Metazoa lineage in our analysis and were not shown in the main figure (Fig. S7). In a few cases, we searched the gene descriptions in our database of genes to get a set of starting genes.

In each case, we checked the gene/HOG/protein domain status at the nodes in each transition lineage from the ancestral reconstruction output described above. For the gene status at the tips, we also considered only the genes from species that were terrestrial or multicellular as applicable. The % of orthologs present at each node/tip was calculated as the sum of genes in a set that were present divided by the number of genes.

We converted our tree into a cladogram i.e. ignored branch lengths so as to make all the tips equidistant from the root of the tree (using the “chronos” function in the phytools package). This allowed us to scale the distance of each node/tip between 0 and 1 where 0 represents the root and 1 represents any tips. We also forced all transition nodes to be at 0.5 for better visual representation and analysis. For the % orthologs present values along a lineage, we fitted four linear models for scaled distance intervals of [0-0.25], (0.25-0.5], (0.5-0.75] and (0.75-1]. We took the fitted residuals to draw the lines in Fig. 3A and 3B.

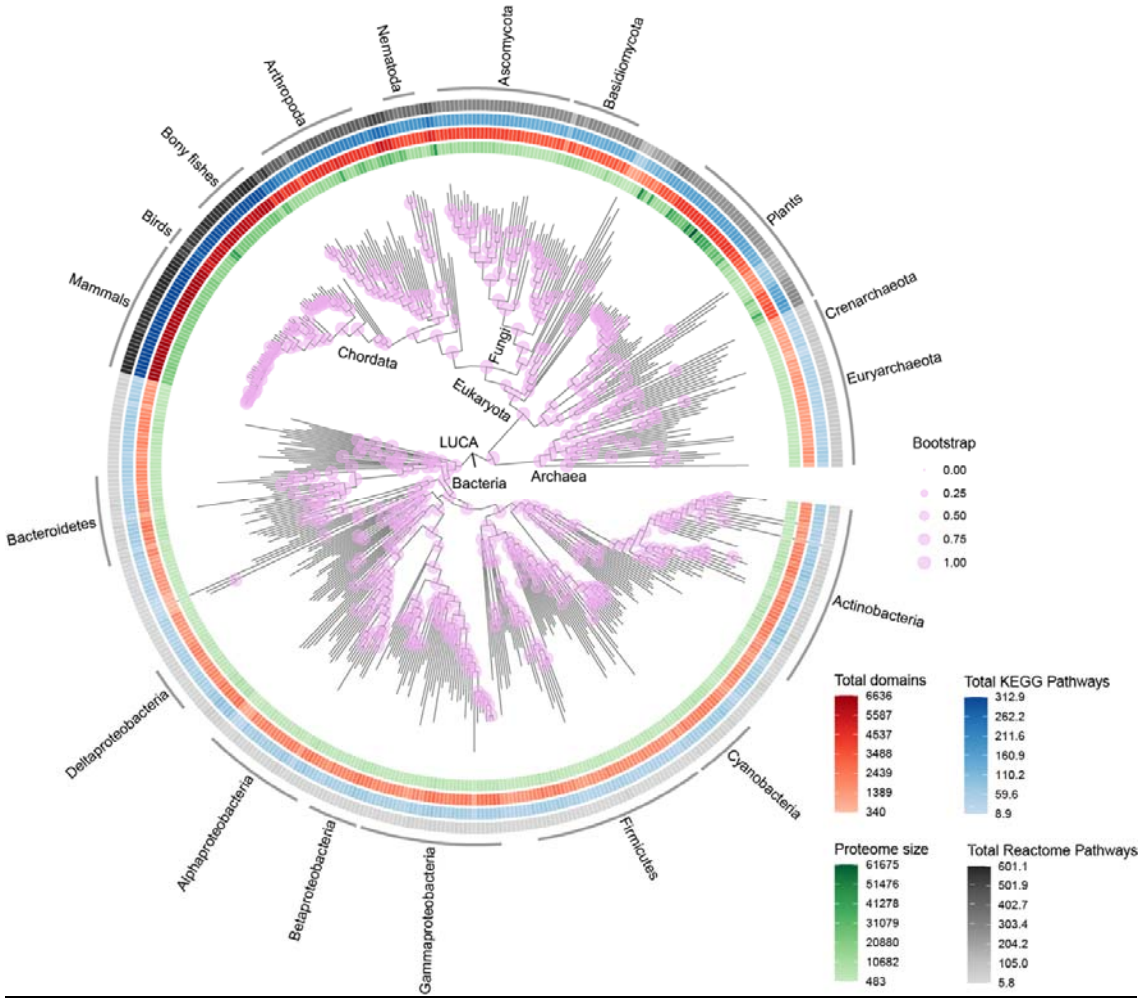

Fig. S1: Phylogenetic tree showing the relationships between 508 species from across the tree of life. The nodes have been annotated with points where the size is proportional to the bootstrap values (purple). The colored rings around the tree indicate the number of proteins (shades of green), number of unique Pfam domains (shades of red), total sum of the proportion of all KEGG pathways (shades of blue) and the total sum of the proportion of broad level Reactome pathways (shades of gray).

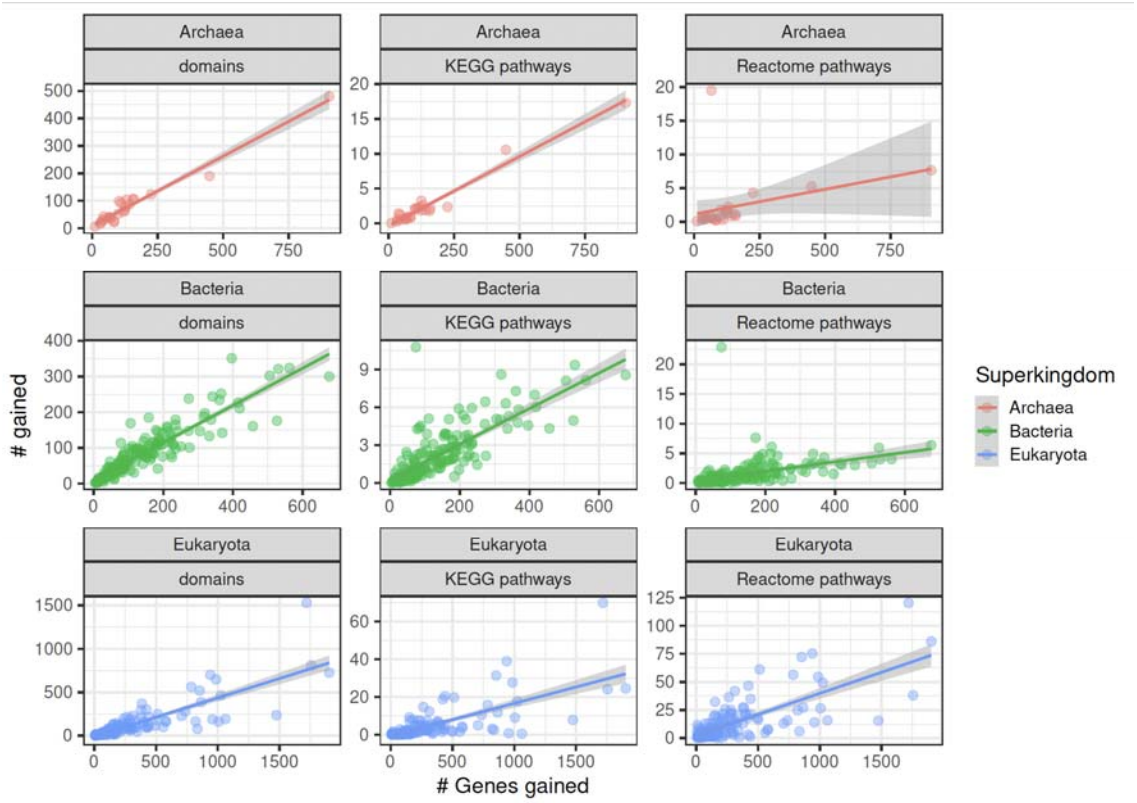

Fig. S2: Number of gained Pfam domains, KEGG pathways and Reactome pathways versus number of genes gained at every node in the phylogenetic species tree (nodes with >2 descendant species). Linear models were fitted to each category across superkingdoms and the grey regions indicate the standard error of the fit.

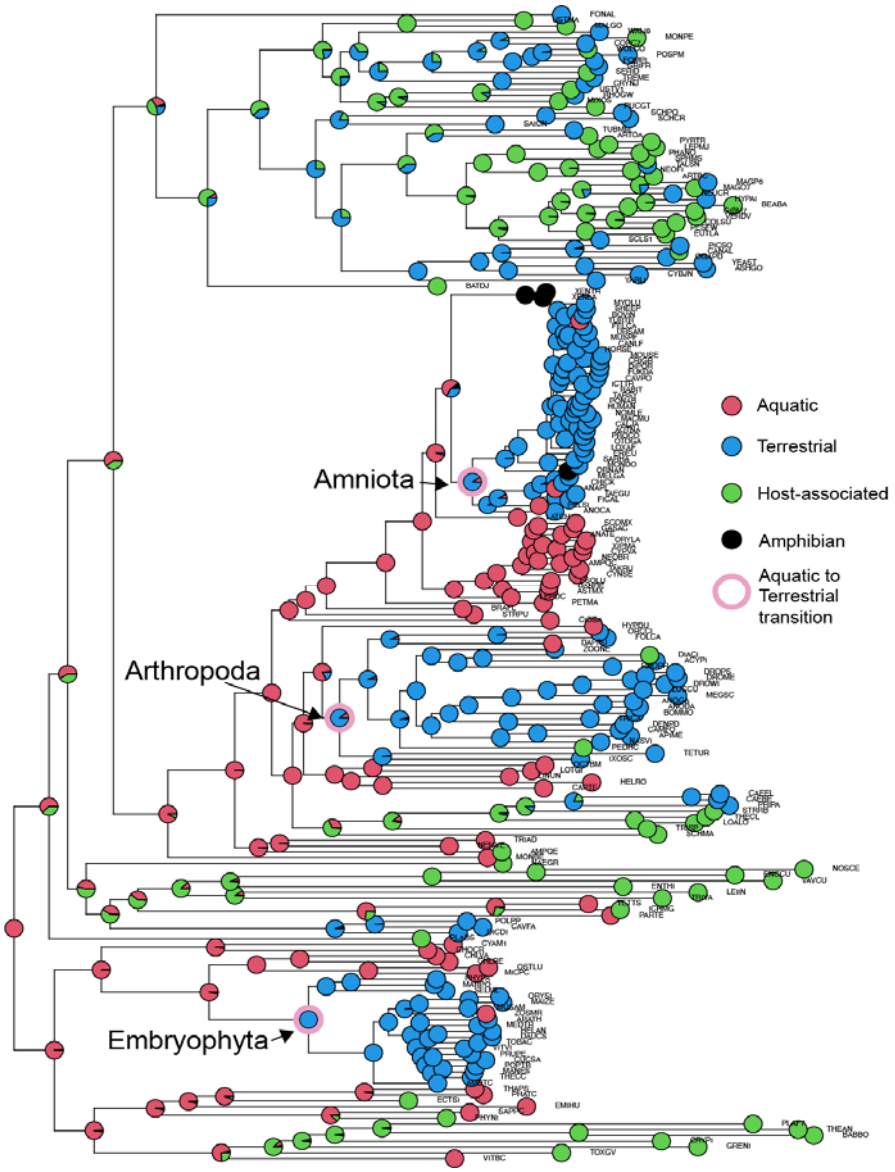

Fig. S3: Phylogenetic tree annotating the broad-level habitat of 199 eukaryotic species in this study. We used stochastic character mapping to determine the posterior probabilities of each habitat type at every node in the tree (pie charts at every node). The highlighted nodes indicate the nodes where a predominantly aquatic state transitioned to the terrestrial state. N.B. Host-associated indicates pathogens and symbionts.

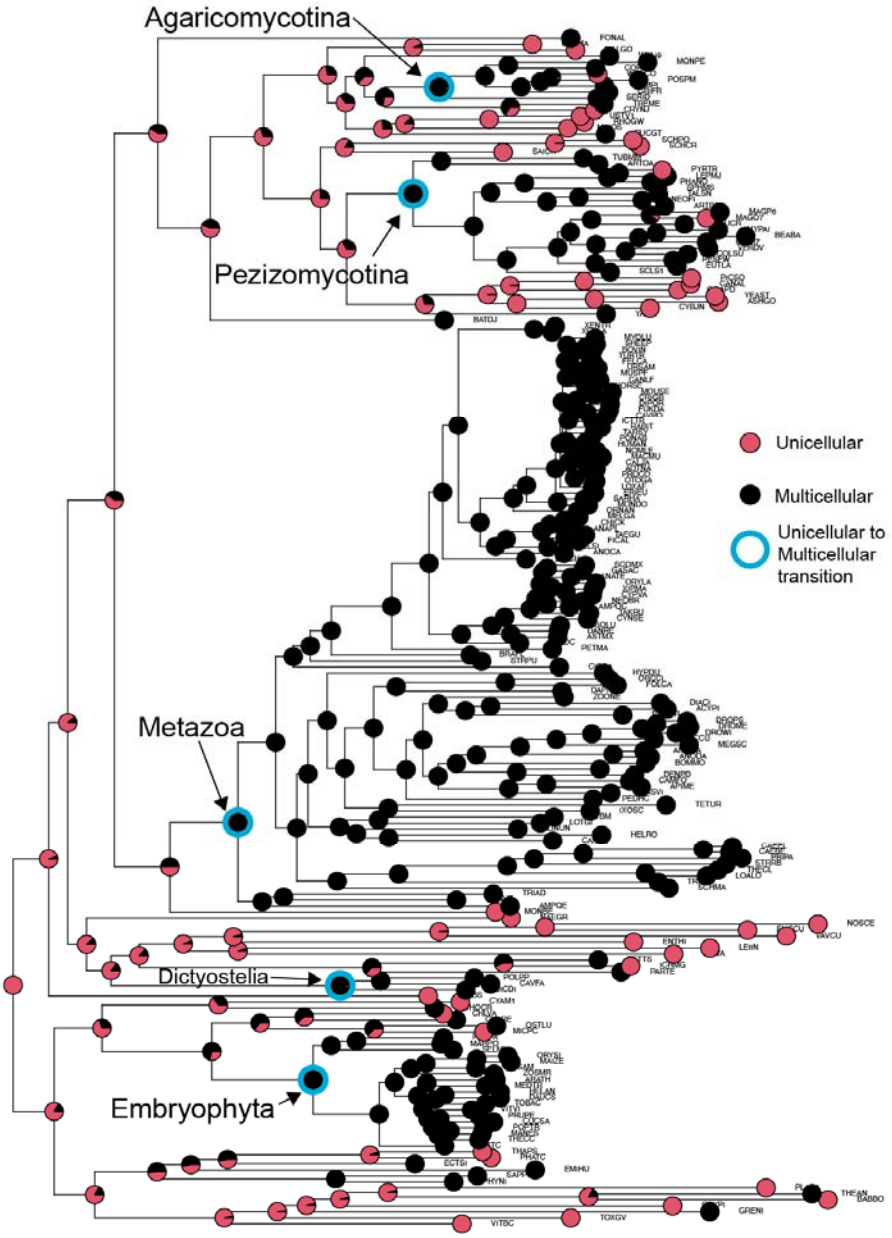

Fig. S4: Phylogenetic tree annotating the cellularity of 199 eukaryotic species in this study. We used stochastic character mapping to determine the posterior probabilities of each trait at every node in the tree (pie charts at every node). The highlighted nodes indicate the nodes where a predominantly unicellular state transitioned to the multicellular state.

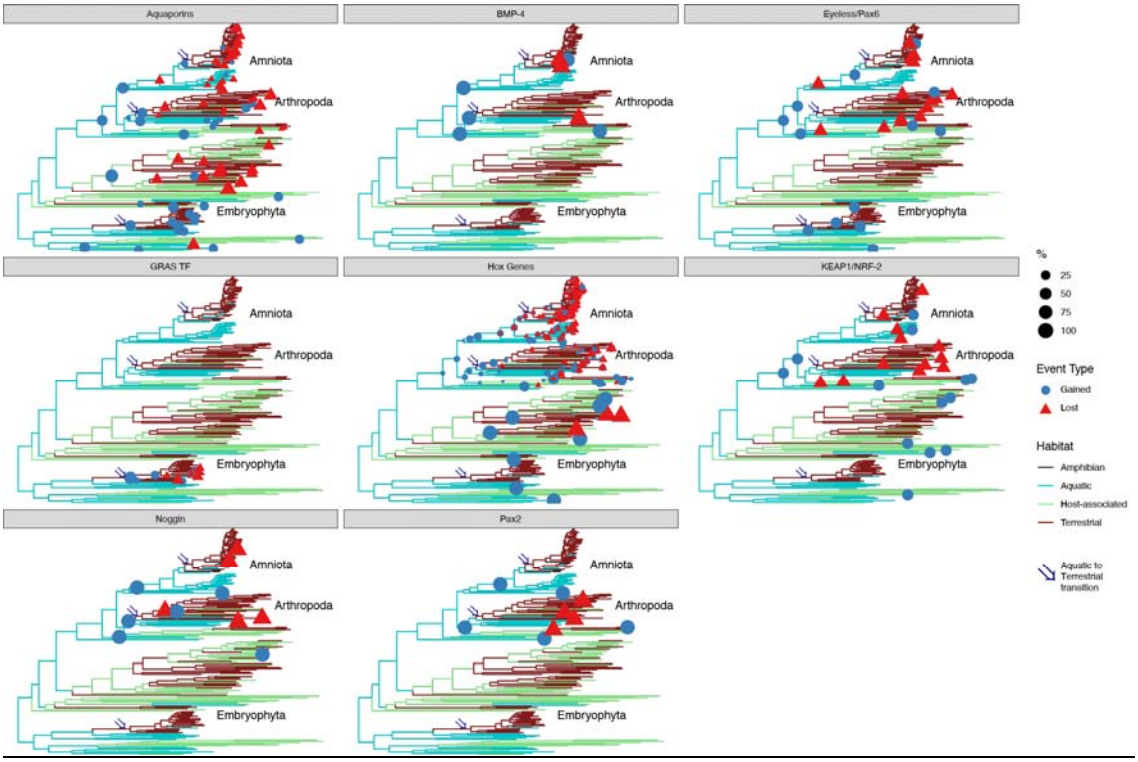

Fig. S5: Phylogenetic tree of the eukaryotes in our study with branches coloured by the inferred habitat of the parental nodes. The three transitions from aquatic to terrestrial habitats are marked with an arrow. Events depicting the gains or losses of ***orthologs*** for known genes (Fig. 3A) are marked by scaled blue filled circles or red filled triangles.

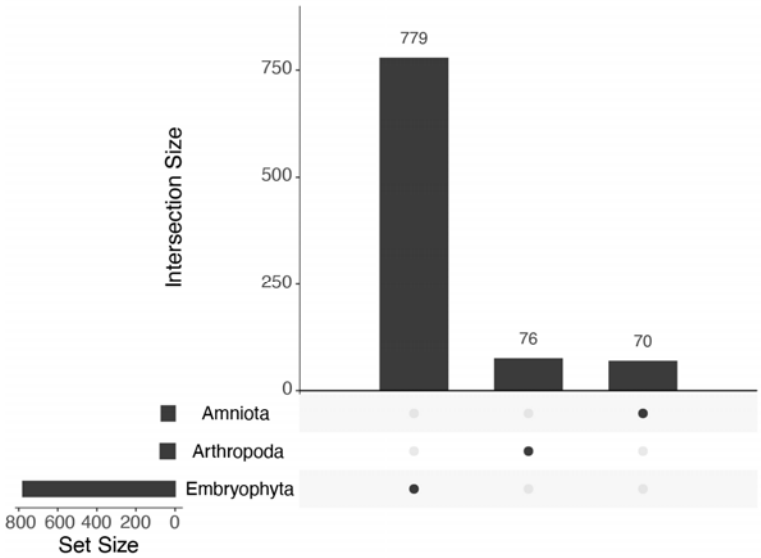

Fig. S6: Upset plot showing the number of HOGs emerging at each of the aquatic to terrestrial transition nodes and if any are shared between the nodes. The HOGs were taken for the GO cluster pertaining to anatomical structures (i.e., “root”, “seed”, “pollen”) and showed no overlap across the nodes indicating the independent gain of genes involved in the anatomical structures for each clade. GO clustering enables these genes to be analyzed together.

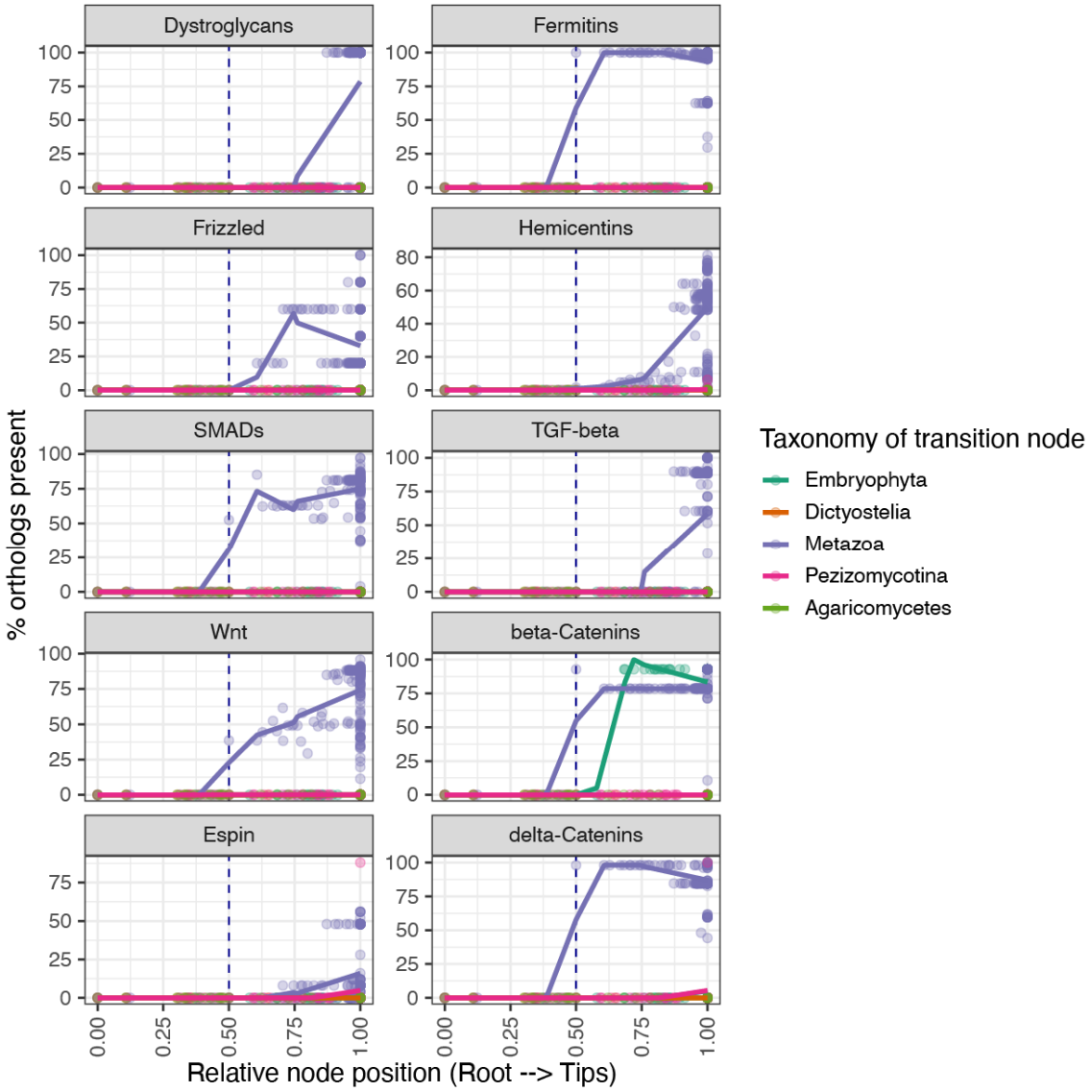

Fig. S7: Plot similar to Fig. 3B showing the % of orthologs for genes thought to be present in the animal ancestor at each node of a transition lineage. The distance to the root of each node in a lineage was scaled between 0 and 1 on the x-axis, while the transition node was forced to be at 0.5. Multiple linear models were fitted to the underlying points for each lineage.

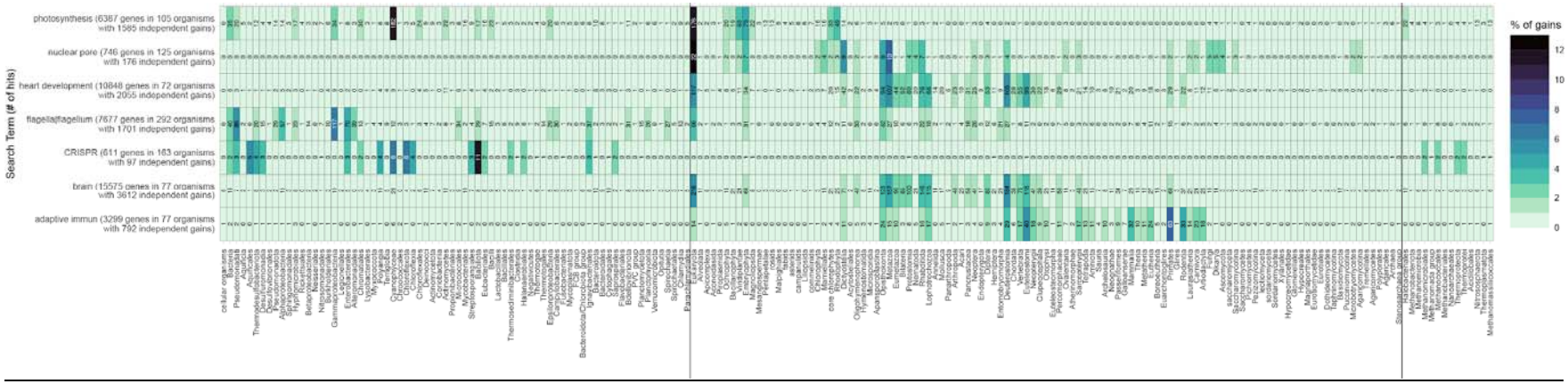

Fig. S8: A detailed heatmap, similar to Fig. 2E, showing the % of gains for the genes captured by each search term occurring at each taxon on the x-axis. The numbers in each cell indicate the number of gain events occurring at that node. There is no filtering of taxa in this heatmap and shows the small amount of spurious gain events being inferred in unexpected clades (discussed in Materials and Methods).

Data S1. (separate file)
Excel file containing all supplementary tables.