Quest for Orthologs in the Era of Biodiversity Genomics

Felix Langschied [1,*], Nicola Bordin [2], Salvatore Cosentino [3], Diego Fuentes-Palacios [4,5], Natasha Glover [6,7], Michael Hiller [8], Yanhui Hu [9,10], Jaime Huerta-Cepas [11], Luis Pedro Coelho [12], Wataru Iwasaki [13], Sina Majidian [6,7], Saioa Manzano-Morales [4,5], Emma Persson [14], Thomas A. Richards [15], Toni Gabaldón [4,5,16,17], Erik Sonnhammer [14], Paul D. Thomas [18], Christophe Dessimoz [6,7], Ingo Ebersberger [1,19,20,*]

[1]Department for Applied Bioinformatics, Institute of Cell Biology and Neuroscience, Goethe University, Frankfurt, Germany
[2]Institute of Structural and Molecular Biology, University College London, WC1E 6BT, London, UK
[3]Department of Integrated Biosciences, The University of Tokyo, 277-0882 Tokyo, Japan
[4]Barcelona Supercomputing Center (BSC-CNS), 08034 Barcelona, Spain
[5]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028 Barcelona, Spain
[6]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
[7]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland
[8]Department of Comparative Genomics, Institute of Cell Biology and Neuroscience, Goethe University, Frankfurt, Germany
[9]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[10]Drosophila RNAi Screening Center, Harvard Medical School, Boston, MA 02115, USA
[11]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Campus de Montegancedo-UPM, Madrid, Spain
[12]Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, Woolloongabba, Queensland, Australia
[13]Department of Integrated Biosciences, University of Tokyo, 277-0882 Tokyo, Japan
[14]Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Solna, Sweden
[15]Department of Biology, University of Oxford, Oxford, OX1 3SZ UK
[16]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
[17]CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain
[18]Department of Population and Public Health Sciences, University of Southern California, Los Angeles, CA, USA
[19]LOEWE Centre for Translational Biodiversity Genomics, 60325 Frankfurt, Germany
[20]Senckenberg Biodiversity and Climate Research Centre (S-BIK-F), Frankfurt am Main, Germany

*Corresponding authors: E-mails: langschied@bio.uni-frankfurt.de; ebersberger@bio.uni-frankfurt.de.

Accepted: October 11, 2024

Abstract

The era of biodiversity genomics is characterized by large-scale genome sequencing efforts that aim to represent each living taxon with an assembled genome. Generating knowledge from this wealth of data has not kept up with this pace. We here discuss major challenges to integrating these novel genomes into a comprehensive functional and evolutionary network spanning the tree of life. In summary, the expanding datasets create a need for scalable gene annotation methods. To trace gene function across species, new methods must seek to increase the resolution of ortholog analyses, e.g. by extending analyses to the protein domain level and by accounting for alternative splicing. Additionally, the scope of orthology prediction should be pushed beyond well-investigated proteomes. This demands the development of specialized methods for the

identification of orthologs to short proteins and noncoding RNAs and for the functional characterization of novel gene families. Furthermore, protein structures predicted by machine learning are now readily available, but this new information is yet to be integrated with orthology-based analyses. Finally, an increasing focus should be placed on making orthology assignments adhere to the findable, accessible, interoperable, and reusable (FAIR) principles. This fosters green bioinformatics by avoiding redundant computations and helps integrating diverse scientific communities sharing the need for comparative genetics and genomics information. It should also help with communicating orthology-related concepts in a format that is accessible to the public, to counteract existing misinformation about evolution.

Key words: ortholog search, annotation transfer, domain architecture, protein structure, FAIR, noncoding RNA.

Significance

The identification and analysis of orthologs play a crucial role in evolutionary research, especially in the rapidly advancing field of biodiversity genomics. In this review, we highlight recent advancements in orthology-based research and propose strategies for addressing current limitations. A comprehensive understanding of existing methods and open challenges is essential for utilizing orthology assignments effectively in state-of-the-art biodiversity genomics research.

Introduction

Biodiversity loss is listed as one of the five most pressing global risks by the World Economic Forum (https://www.weforum.org/publications/global-risks-report-2023/). Estimated consequences include the alteration of food webs (Pilling et al. 2020), changes in the metabolic pan-network of an ecosystem through the extinction of species (Cannell et al. 2020), and, ultimately, the potential collapse of entire ecosystems. Investigating the causes and consequences of biodiversity loss has substantially intensified over the past years. Since the study of biodiversity loss increasingly involves the comparative analysis of biological sequences at the genome scale, it has become a focal area of "Biodiversity Genomics" (Supple and Shapiro 2018). As a common principle, Biodiversity Genomics investigates how sequences and the functions they convey have changed over time, which sequence variants are present in what taxa, and which evolutionary regimes have shaped this process. On this basis, the presence, the abundance, and the genetic diversity of species in an ecosystem can be monitored over evolutionary time (Leigh et al. 2019). Additionally, biological sequences help to mine the hitherto only marginally tapped wealth of natural products via, e.g. the identification of novel biosynthetic gene clusters (Marcet-Houben et al. 2023).

"Biodiversity Genomics" has significantly benefited from the ease with which sequences of even very large genomes can be assembled. This has resulted in the rapid accumulation of publicly available genome sequences that represent the tree of life with increasing coverage. The integration of these novel genomes into Biodiversity Genomics studies essentially depends on the accuracy with which the evolutionary relationships of the compared sequences can be determined. Biological sequences that share a common ancestry (homologs) can be categorized based on the evolutionary event they originated from (Fig. 1a). Orthologs emerge as a consequence of a speciation event, whereas paralogs emerge via gene duplications (Fitch 1970). Consequently, paralogous sequences from two species are evolutionarily more distantly related than the corresponding ortholog pairs. It is for this reason that orthologs are often loosely referred to as the "corresponding" genes in two species and often share the same or at least similar function (Gabaldón and Koonin 2013). This makes the identification of orthologs the basis of many Biodiversity Genomics analyses (Fig. 1b).

Since 2009, current challenges in the field of orthology inference have been discussed and addressed by the Quest for Orthologs (QfO) Consortium. Since its inception, the QfO Consortium successfully established a regularly updated benchmark service for orthology predictions (Nevers et al. 2022), curating and maintaining a set of reference proteomes (https://www.ebi.ac.uk/reference_proteomes). It has been also advancing the state of the art of orthology-related applications for more than 10 years (Linard et al. 2021). The QfO Consortium gathered for its seventh on-site meeting on 2022 September 17 to 18 in Sitges, Spain, in conjunction with the 21st European Conference on Computational Biology (ECCB 2022). Using key points from this meeting as anchor points, we will here review challenges and future directions for orthology inference. We will use the expanding datasets in the biodiversity genomics era to motivate the relevance of these issues (Fig. 2). The almost exponentially growing number of newly sequenced genomes results in an annotation bottleneck that needs to be addressed at an appropriate scale (Challenge 1). It is further essential to bridge the gap between potential and actual knowledge that can be derived from this flood of data. To do this, the resolution of orthology prediction must be increased to the protein domain level to better track the change of gene function through time (Challenge 2). Additionally, the analysis of

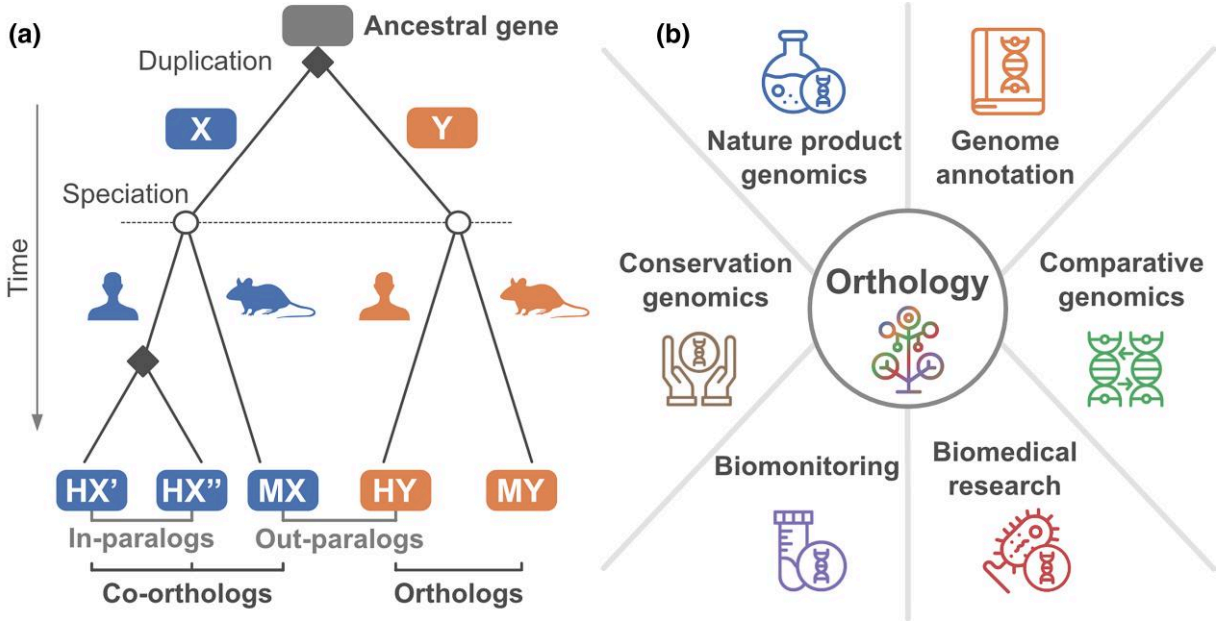

Fig. 1. Biodiversity genomics are rooted in ortholog identification. a) The evolutionary concepts of orthology and paralogy. The evolutionary lineages of genes sharing the same ancestry (homologs) are split either by gene duplication events (square nodes), giving rise to paralogs, or by speciation events (circle nodes), resulting in orthologs. The depicted tree represents a scenario where a gene duplication prior to the speciation event gives rise to the out-paralogous genes X and Y that reside in the human and mouse lineages, respectively. The split of the human (H) and mouse (M) lineages resulted in the orthologous groups HY and MY. In the case of gene X, a subsequent gene duplication in human formed the in-paralogous gene pair HX′ and HX″ where both human genes are co-orthologous to MX. Pictograms provided by PhyloPic. b) The identification of orthologs forms the foundation for gene annotation transfer between species, because they represent our best inferences of genes with corresponding functions. Together with de novo annotation methods, orthologs are needed to annotate comprehensive catalogues of genes present in an organism (Genome annotation). Consequently, any analysis that compares which functions (i.e. genes) are available in different genomes is also rooted in orthology (Comparative genomics). Identifying orthologs that are specific to a taxonomic group can help to identify pathogenicity-related factors (Biomedical research) or to define genetic markers hat help in sequence-based species identification (Biomonitoring). They inform how robustly a molecular function is represented in an ecosystem (Conservation genomics), and they can identify novel gene clusters that produce secondary metabolites (Natural product genomics). All symbols provided by Icon Market from Noun Project.

large genome collections should extend beyond standard protein-coding genes (Challenge 3). With machine learning-driven predictions of protein structures being readily available, it will be necessary to explore how the comparison of protein structures can help to identify distantly related orthologs whose sequences are no longer more similar than expected by chance (Challenge 4). Finally, efforts of the ortholog community should be made more accessible to scientists of all communities (Challenge 5) and redundant computations should be minimized to reduce the collective computational carbon footprint (Grealey et al. 2022). Each of these challenges will be expanded upon in separate sections below.

Expanding Datasets Create Opportunities and Challenges

Cataloguing Biodiversity With Targeted Sequencing

The scope of large-scale sequencing projects has expanded rapidly in recent years. One manifestation of this is the Earth Biogenome Project (EBP)—a "moonshot" project that aims to sequence all 1.6 million taxonomically classified eukaryotic species (Lewin et al. 2022). As of January 2024, 2,313 species from 24 phyla have been sequenced, assembled, and made publicly available under the umbrella of the EBP (https://goat.genomehubs.org/projects/EBP), and this number is growing rapidly. Many of these genome sequences have been contributed by associated projects, each of which targets a specific group of species. This includes the vertebrate genome project (https://vertebrategenomesproject.org/) which aims to sequence over 260 different vertebrate genomes in its first phase and the Moore Foundation Aquatic Symbiosis project (https://www.aquaticsymbiosisgenomics.org/) which targets hundreds of symbiotic organisms (e.g. corals, sponges, lichens, algae, and protists). The Darwin Tree of Life (DToL) project, as another example, aims to sequence all eukaryotic species in Britain and Ireland and increasingly tackles more challenging genome assembly projects. Recently, DToL celebrated its 1,000th genome sequence assembly, the European mistletoe (*Viscum album*) whose genome is 30 times larger than that of humans. Current large-scale sequencing efforts are committed to the production

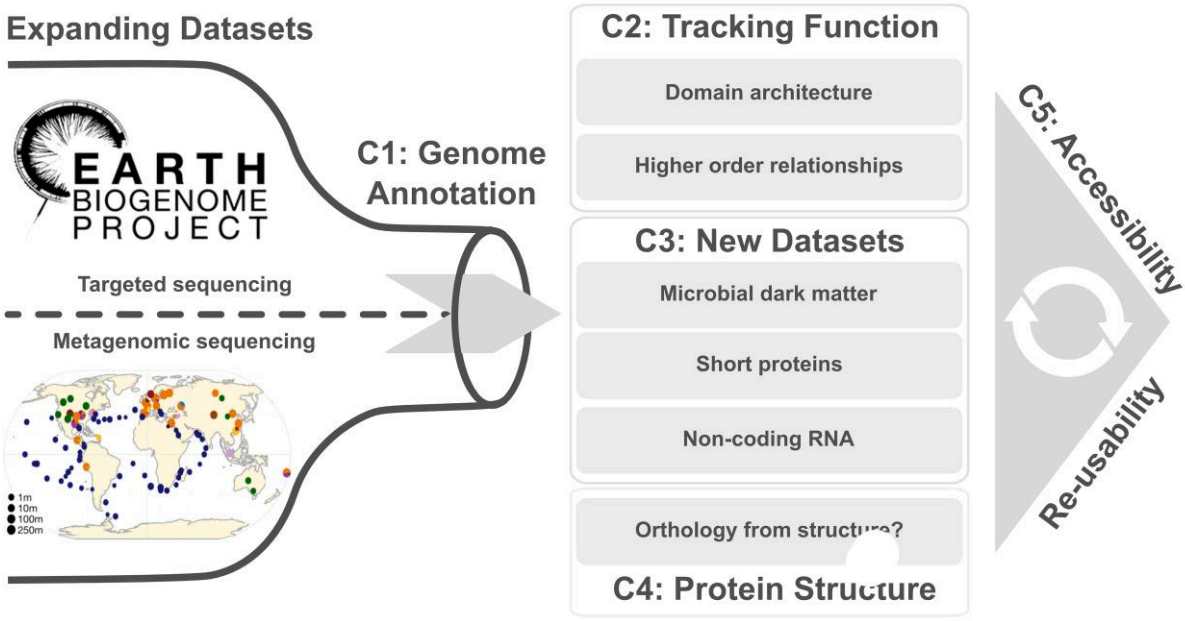

Fig. 2. Current challenges of ortholog-based research. The logo of the EBP is used to symbolize large, targeted sequencing projects. The world map in the bottom half shows the sampling distribution of 13,174 metagenomes from 14 habitats as an example of a large-scale metagenome project [Figure adapted from Coelho et al. (2022)]. The bottle shape represents the large number of available datasets which present several new challenges. As the first challenge, all new datasets must pass through the genome annotation bottleneck (C1). Challenges regarding technical aspects of orthology prediction are presented in boxes C2-4. Finally, orthology assignments should be FAIR (C5) (Wilkinson et al. 2016).

of chromosome-level, reference-grade assemblies—an important foundation for obtaining comprehensive gene catalogues with little missing data due to assembly gaps. This helps to minimize artifacts in downstream orthology prediction steps (see Challenge 1). With corresponding genome announcements written in a semiautomatic manner, the data are made publicly available and monitored in real time via the Genome on a Tree (GoaT) project (Challis et al. 2023). Thus, the status of the EBP provides a glimpse of the data scale and assembly quality that will become available in the upcoming years. In the long run, these projects will allow us to chart a large part of the genomic diversity on Earth.

Metagenomic Sequencing Unearths the Diversity of Prokaryotic Communities

Catalogues of prokaryotic diversity are expanding quickly, with huge datasets emerging from metagenomic sequencing of diverse environments (Laiolo et al. 2024). Metagenomic analyses are no longer used only for diversity tag or barcode sequencing, but to assemble genomes directly from metagenomic reads (Pasolli et al. 2019). Metagenome-assembled genomes (MAGs) are key for evaluating the ecosystem service that is provided by different microbial communities (Grossart et al. 2020). MAGs can also help to compare taxonomic and functional profiles in an ecosystem, for example, by identifying redundant metabolic functions present in multiple taxonomically distinct organisms (Louca et al. 2018).

One example for large-scale metagenomic sequencing efforts is the Global Microbial Gene Catalogue across 13,174 metagenomes from 14 habitats. To date, it includes 303 million nonredundant genes (Coelho et al. 2022). Functionally annotating such large-scale datasets requires specialized algorithms like the eggNOG-mapper v2 that is optimized for fast database queries and I/O operations (Cantalapiedra et al. 2021) or the MMseqs2 suite. MMseqs2 quickly assigns contigs with taxonomic labels by concentrating on sequence fragments that share a minimum level of similarity to sequences in the reference database and removes redundant protein sequences by linear scaling clustering (Mirdita et al. 2021). However, removing redundancy also means potentially losing information. This can become an issue for downstream analyses, especially since bacterial research communities increasingly focus their investigations on intraspecies variations (Iruegas et al. 2023).

The rapid accumulation of genomic data from both targeted and metagenomic sequencing poses a challenge that outpace strategies devised at a time when sequencing a single genome was a major feat. Consequently, databases, publications, and the way we investigate and report gene family evolution must be adjusted to cope with this flood of data.

Challenge 1: Ortholog Search Across Tens of Thousands of Genomes

Less than 20 yr ago, genome-based research was possible for only a few model organisms (e.g. *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*), but it became quickly obvious that both accuracy and resolution of comparative genomics analyses benefit substantially from a more comprehensive taxon sampling (Putnam et al. 2007). As the bottleneck imposed by lack of genome assemblies ceases to exist, comparing genomic sequences themselves is now often only of moderate interest. Instead, most analyses are concerned with lineage-specific changes in gene content (Ocaña-Pallarès et al. 2022), transposon activity (Martelossi et al. 2023), or horizontal gene transfer (Irwin et al. 2022), to name a few. All these analyses require accurately inferred protein-coding genes for the respective genome assemblies, and artifacts introduced during this process can generate spurious signals that confound the downstream conclusions (Bálint et al. 2024). The scarcity of genome availability, once a major bottleneck, has now been superseded by the challenge of functionally annotating tens of thousands of genome assemblies and inferring patterns of gene evolution—an endeavor for which understanding orthology relationships is essential.

State-of-the-art methods of gene prediction integrate intrinsic sequence information like coding-potential and splice site motifs with extrinsic data (Gabriel et al. 2023). Genes with homologs in other species can be inferred with information from related organisms (Bruna et al. 2024). Complementary to the evolutionary approach, genes can be predicted with the help of transcriptomic data (e.g. Hoff et al. 2016). This allows for the detection of evolutionarily young genes that lack homologs in public databases, and of genes that evolve so quickly that the homology inference fails (Jain et al. 2019). However, genes that are lowly expressed or have a particular spatial or temporal expression pattern are likely to be missed. Mitigating these limitations is necessary for cataloguing the full breadth of sequences that make up life on Earth, but generating these data for all assembled genomes is not feasible. Therefore, Guigó (2023) has recently proposed a hierarchical sequencing scheme. In brief, this scheme suggests assembling the genome of one organism per taxonomic class to high contiguity and additionally constructing a comprehensive, whole-body, long-read RNA-seq cell atlas. Further taxa nested in this class can then be sampled with decreasingly elaborate sequencing efforts. However, even if transcriptomic data become available for one representative species per genus, it will still be necessary to project gene predictions and their functional annotation to ~10 times more species.

Quality Control of Genome Annotation

In ortholog-based research, common sources of artifacts include genes situated in assembly gaps that can be mistaken for gene loss. Genes identified only over part of their length, or the failure to differentiate between intergenic regions and introns can falsely indicate changes in gene structure. Additionally, an incomplete masking of e.g. transposons can lead to artificially inflated sets of protein-coding genes. Future studies should therefore determine whether the choice of annotation methods has a significant impact on downstream orthology inference. To what extent a gene set can be considered complete has been most addressed with the BUSCO completeness score (Manni et al. 2021). In brief, BUSCO relies on curated sets of single-copy orthologs that are conserved in predefined taxonomic groups. Orthologs for each member of these sets are identified and are labeled as "complete," "fragmented," or "missing" according to a length criterion. This general idea was recently extended with two new approaches that are not restricted to single-copy orthologs and hence query more marker genes than BUSCO. fCAT (https://github.com/BIONF/fCAT) assigns completeness labels not only based on gene length but also takes the expected feature protein architecture similarity score into account (Dosch et al. 2023). In contrast to BUSCO, fCAT can utilize custom sets of core genes that can be curated from any source of orthology assignments and can cover any taxonomic range. OMArk evaluates the completeness of proteomes by rapidly assigning query protein sequences to a set of evolutionarily conserved hierarchical orthologous groups (HOGs) for a taxonomic level (Nevers et al. 2024). Additionally, OMArk identifies proteins with no known homologs or with taxonomically unexpected orthologs in the target proteomes. Those may derive from fast-evolving or hitherto undiscovered gene families but can also indicate spurious gene models—an aspect typically ignored by other tools. Proteins with an unexpected taxonomy distribution may also be caused by contamination and are flagged as a contaminant by OMArk if more orthologs to it are found outside of the target assembly's lineage than expected by chance. Nevers et al. (2024) identified 72 proteomes from UniProt that are contaminated with sequences mostly of bacterial or fungal origin. This finding highlights the need for more stringent quality control during the curation of reference proteomes, and the need for reference-grade assemblies, a stated ambition of many current large-scale biodiversity genome diversity projects.

While contaminations from nontarget organisms do increase spurious orthology assignments in comparative analyses (Bálint et al. 2024), these contaminants may not be devoid of useful information. Indeed, recent studies do not only explore contaminations in genome assemblies to find residues of sample preparation (Chrisman et al.

2022), but have used them, e.g. for reconstructing the microbiome of a target species (Foo et al. 2023). To harness this unused potential, scalable approaches for the identification and analysis of nontarget sequences in newly sequenced assemblies are needed.

Computational Limitations of the Ortholog Search

Graph-based orthology inference methods scale quadratically with the number of sequences investigated (Altenhoff et al. 2019), making computational time a bottleneck for analyzing large datasets. It is therefore not surprising that even the latest releases of ortholog databases cover only up to 2,000 out of over 15,000 eukaryotic assemblies in the International Nucleotide Sequence Database Collaboration (INSDC) (Arita et al. 2021; Kuznetsov et al. 2023). To close this gap, multiple approaches were recently developed. For example, SonicParanoid2 reduces the runtime of orthology inference using machine learning (Cosentino et al. 2024). In brief, given a proteome pair *A* and *B* in which orthologs can be identified using a reciprocal similarity-based search, the search direction (A vs. B or B vs. A) with a shorter execution time is predicted using Adaptive Boosting. The "faster" search direction is then performed first, and based on its results, only those proteins that are likely to have an ortholog partner are considered in the second search. A benchmark revealed a speedup of the ortholog search by up to 42% while maintaining high precision and recall. Another method involves targeted profile-based searches that reduce the runtime to linear complexity, facilitating the identification of orthologs across thousands of taxa (Birikmen et al. 2021). Additionally, k-mer distance-based preclustering can be used to remove allelic variants and duplicates (Derelle et al. 2020). Dealing with the growing amount of incoming data will require even more efforts to reduce the computational burden. In this context, the carbon footprint associated with computational analysis has become a serious concern (Grealey et al. 2022), and it is important not only to optimize algorithms but also to avoid redundant analyses (see Challenge 5).

Ortholog Search in Unannotated Genome Assemblies

Thus far, all ortholog search algorithms depend on the availability of a comprehensively identified gene sets for a genome assembly. However, the gene prediction process itself benefits from the information about orthology relationships as well, indicating that both tasks can be integrated. Tool to infer Orthologs from Genome Alignments (TOGA) facilitates a computationally efficient and comprehensive projection of genome annotation across species using whole-genome alignments (Kirilenko et al. 2023). Graph-based ortholog detection exploits that a sequence in one organism tends to look more like its ortholog in a second species than like a paralog. TOGA extends this similarity principle to the genomic context of a gene. For each annotated gene in a reference species, TOGA first infers the orthologous locus (or loci in case of co-orthologs). It then determines the positions of coding exons in each orthologous locus, which provides a comprehensive, high-quality annotation of genes conserved between reference and target species. TOGA scales linearly with the number of genomes and can therefore be used for hundreds of species. For example, TOGA was applied with human and mouse as references to 488 placental mammals and with chicken as a reference to 500 birds. However, the use of TOGA is limited to closely related species for which even intronic and intergenic parts of the genomes can still be aligned in a meaningful way.

Many research questions are centered around a specific set of genes and do not necessarily require an annotation of the complete genome. In such use cases, the novel tool fDOG assembly can perform orthology inference for genes of interest in unannotated genome assemblies (Collins et al. 2023). fDOG assembly extends existing orthologous groups by first identifying the region in the unannotated assembly that is most likely to harbor an ortholog. It then annotates genes in these regions using either AUGUSTUS which is guided by block profiles generated from existing orthologous groups, or MetaEUK. Any identified genes are then tested if they can be added to the orthologous group with a reciprocal best similarity search hit criterion.

In a time where sequencing of new genomes rapidly progresses, these methods mentioned above provide a strategy to cope with the gene prediction and orthology inference bottleneck. The decision of whether a full reference-based inference of all genes or a targeted ortholog search for a specific gene set is more fitting will then ultimately depend on the research question and the annotation status of a given dataset. Nevertheless, one could legitimately ask whether including all available genomic data is necessary to infer orthologous relationships. Instead, it might be more practical to down-sample available assemblies, trying to maximize the taxonomic diversity in the dataset (Bonnie et al. 2024). Future studies should therefore also aim at investigating more closely how much information is gained by the addition of data to different phylogenomic applications.

Challenge 2: Tracking Functional Changes of Orthologs

Orthologs are typically considered the best guess for identifying functionally equivalent genes in two species. However, the probability of functional diversification increases over evolutionary time. While it is hard to trace

the conservation of function in silico, information can be gathered that indicates a change of function.

Increasing the Resolution to Feature Architecture Level

Amino acid sequence divergence alone is a poor proxy for the functional divergence of orthologs (Laurent et al. 2020). This opens the case for adding additional layers of information that aid to infer functional divergence of orthologs. One option is to annotate protein sequences with features that have more direct links to molecular function such as Pfam and SMART domains, signal peptides, transmembrane domains, or even low complexity regions. By now, many orthology databases provide the resulting feature architectures of the identified orthologs as accessory information (Altenhoff et al. 2021; Kuznetsov et al. 2023; Persson and Sonnhammer 2023). However, their comparison is left to the individual user, and the tracing of feature architecture changes across orthologs as an indicator of functional change is tedious. Dosch et al. (2023) simplified this task with FAS, a software that captures the pairwise feature architecture similarity between proteins as a score between 0 (no similarity) and 1 (identical). As a key innovation, a score maximization algorithm identifies the highest scoring linear path through redundant parts in a protein's feature architecture, e.g. resulting from overlapping annotations of Pfam and SMART domains. The method was applied to identify different variants of a bacterial pilus tip, indicating that different strains within the same species vary in the way they interact with the environment and/or their human host (Iruegas et al. 2023).

While FAS is a significant step toward the automated comparison of protein feature architectures and inferring functional divergence of orthologs, the scoring scheme remains ad hoc. Changes in protein feature architecture over evolutionary time can now be modeled with the tool DomArchov that exploits nonrandom constraints of multi domain architecture evolution (Cui et al. 2022). Using this software, it is now possible to simulate a range of protein feature architectures, compare observed changes in architectures over time, and determine whether these changes are significantly larger than expected.

Both FAS and DomArchov indicate that the conventional view of multidomain proteins as a single evolutionary unit during the ortholog search is helpful but sometimes misleading. While it simplifies the data processing and analysis considerably, it will provide spurious results if not all domains of a protein have the same evolutionary history (Persson et al. 2019). A few domain-based approaches to orthology prediction have been developed over the years, for example, SonicParanoid2 (Cosentino et al. 2024) or the methods used during construction of the prokaryotic COGs (Galperin et al. 2021) and MBGD (Uchiyama et al. 2019) databases. Recently, the InParanoiDB database was redesigned to contain both full-length and domain orthologs (Persson and Sonnhammer 2023). For the latter, the Domainoid algorithm (Persson et al. 2019) was applied to 640 eukaryotic and prokaryotic proteomes. The orthologous domains can then be used, e.g. to detect orthologous relations that are not found at the full-length level, or to extract discordant domain orthologs, where different domains have different evolutionary histories. InParanoidDB also provides a graphical display of the domain architectures in an ortholog group that allows domain searching and switching between full-length and domain ortholog groups.

Higher-Order Relationships of Gene Families

Functional annotation transfer is a prime application of orthology assignments. Many studies accomplish this by connecting their proteins of interest to manually curated groups of functionally annotated orthologs provided, e.g., by KEGG (Kanehisa et al. 2023), COG (Galperin et al. 2021), or Panther (Thomas et al. 2022). However, it has recently been shown that the complementary use of paralogs for transferring functional annotation can increase the amount of transferable information (Stamboulian et al. 2020). This is one objective of the Phylogenetic Annotation using Gene Ontology (PAN-GO) project (Gaudet et al. 2011). PAN-GO integrates experimental knowledge of gene functions across entire gene family trees, including both orthologs and paralogs, to create precise and comprehensive descriptions of gene functions. Using extensive automated functional annotation integrated with expert curation, the PAN-GO project has created models of function evolution across nearly 10,000 gene families, covering over 82% of human genes (Aleksander et al. 2023). PAN-GO annotations are available as part of the PANTHER database. They serve as a knowledge base of protein function that has been continually expanded since its initial release (Thomas et al. 2022). In addition to providing a comprehensive catalogue of gene functions, this work highlights again that a large fraction of our knowledge derives from experimental studies of homologs in model organisms like the mouse, fruit fly, *C. elegans*, and yeast.

Even with manual curation in place, the accuracy of the functional annotation transfer increases with decreasing phylogenetic distances between the compared species. This is because orthologs having less time to functionally diverge. This calls for a comprehensive set of functional studies from experimental assays in taxonomically diverse species. Paired with more sophisticated computational methods to detect functional diversification between orthologs (see Challenge 2), this can help to alleviate the burden of manual curation. Generating these datasets will also add to the pool of training data that is available to existing tools for the prediction of gene function directly from sequence data (Kulmanov et al. 2018).

Coevolution as an Indication of Functional Interdependence

Phylogenetic profiling studies identify functionally interacting proteins by comparing presence–absence patterns across large taxon collections (Dembech et al. 2023). Traditional phylogenetic profiling methods have been limited by the assumption of uniform correlation in coevolving proteins across all species (Moi and Dessimoz 2023). This assumption, effective for interactions common to ancestral lineages, struggles to capture lineage-specific interactions. To mitigate this, a novel approach employs graph neural networks (GNNs) to incorporate the species tree's structure and provide insights into the temporal and taxonomic emergence of these interactions (Moi and Dessimoz 2022). Such an approach will be particularly relevant to mine the wealth of biodiversity genomes discussed above. The ability of GNNs to predict when and in which taxa certain interactions appeared complements the PAN-GO project's efforts in creating detailed interaction models of function evolution across gene families. Fusing evolutionary data with interaction and functional analysis promises to increase the accuracy and predictive power of functional annotations based on orthologs and paralogs.

Challenge 3: Orthology Inference off the Beaten Track

In the past decade, the focus of comparative genomics was directed on protein-coding genes that can be annotated with sequence-based methods. These genes are consequently well covered in established orthology databases, but compiling a catalogue of all genes must also include those that are more difficult to annotate (Amaral et al. 2023). One must therefore not only process the amount of data that is becoming available, but also adapt existing methods to infer orthologs beyond standard proteomes.

Orthology of Short Proteins

The length distribution of annotated proteins is surprisingly uniform across the tree of life, with a high proportion of sequences ranging between 50 and 500 AA (Nevers et al. 2023). There is no consistent terminology, but proteins shorter than 100 AA (and particularly those shorter than 50 AA) are referred to as small- or microproteins (Storz et al. 2014). Microproteins are abundant across all domains of life and account for 3% to 5% of a species' proteome (Pueyo et al. 2016). However, they have so far flown under the radar of most comparative analyses for two main reasons. First, the corresponding genes are hard to annotate. This results in a high number of spurious gene predictions that prohibitively inflate the computational burden of orthology assignments (Storz et al. 2014). Experimental data like ribosomal profiling or mass spectrometry can provide more direct evidence for the annotation of microprotein-encoding genes (Kute et al. 2021). Unfortunately, the generation of such data does currently not scale with the growing number of available genome sequences. Second, it is challenging to identify orthologs of microproteins. Many of them may be evolutionarily very young and are thus only present in a narrow taxonomic range because the corresponding genes are more likely to be created de novo from random noncoding DNA than larger genes (Montañés et al. 2023). In addition, they evolve more rapidly than genes that encode longer proteins (Slavoff et al. 2013). These factors, together with their short length, which limits the amount of phylogenetic signal that is necessary to resolve their evolutionary relationships, make the detection of orthologs challenging (Jain et al. 2019). It is therefore particularly hard to distinguish between failing to sample (recover) an ortholog of a microprotein and its genuine absence. Tracking the full genetic biodiversity on Earth will consequently require the development of specialized applications for detecting orthologs to small proteins. This will not only help with the annotation transfer of short genes but also with shedding light on their evolution.

Novel Bacterial Gene Families

Insights from a recent large-scale metagenomic effort on the global microbiome (Coelho et al. 2022) indicate that there are many novel gene family clusters whose biological relevance is still unknown. Recent estimates indicate that about 25% to 50% of all observed environmental genes have no significantly similar sequences in the current databases, leaving their function elusive (Coelho et al. 2022). A recent study discovered novel orthologous groups (OGs) of high functional evolutionary significance from 140,000 prokaryotic MAGs of uncultivated taxa (del Río et al. 2024). The authors identified more than 400,000 protein-coding gene families found exclusively in uncultivated taxa that are missing from current databases. Still, several criteria indicate their functional relevance: they all show evidence of strong purifying selection, span multiple species, and contain a conserved protein region of at least 20 amino acids. Many of these novel OGs were functionally annotated by using structural alignments and/or mapped to highly conserved genomic locations, and they could be linked to important biological processes such as central metabolism, defense systems, and cell motility (del Río et al. 2024). Furthermore, hundreds of these novel OGs were identified as synapomorphies for high taxonomic ranks, highlighting their potential as lineage-defining traits contributing to the evolution and diversification of these uncultivated lineages.

Notably, the same study argues that these novel OGs cover only ~5% of all unknown sequences, the rest being discarded due to quality filters or lack of supporting data.

Thus, this new set of OGs (https://novelfams.cgmlab.org) may only represent the tip of the iceberg and the comparative microbial genomics community will soon face the challenge of an exponential increase in new genes, families, and orthologs. One approach to tackle the challenge of assigning tentative functions exploits the embedding of genes into their genomic context composed by neighboring genes. For example, genes that tend to conserve their gene order more than expected over evolutionary timescales, for example, through consistent co-association within operons (Rocha 2006), have been used to extend functional assignments to previously undescribed genes (Djahanschiri et al. 2022). Moreover, deep learning approaches have successfully used word embedding algorithms borrowed from natural language processing for assigning functions to genes lacking a significant sequence similarity to any sequence with a known function (Miller et al. 2022). However, systematic approaches for this kind of intragenomic annotation transfer are still lacking, which limits our ability to annotate the "dark matter" of bacterial genomes at scale.

Orthology of Genes and Gene Products

Orthology is typically identified on the protein level, because the phylogenetic signal that is necessary to infer the precise evolutionary relationships between sequences decays slower for amino acid sequences than for nucleotide sequences. This raises the question of how to deal with genes that give rise to alternatively spliced transcripts encoding different protein isoforms. Traditionally, only one isoform per gene was considered in the ortholog search. For example, there is a one-gene-one-protein layout for the reference proteomes provided by UniProt that are used for the QfO benchmark (Nevers et al. 2022). However, any prior selection of representative isoforms can impair the outcome of the ortholog search. Orthologs may be missed if a short isoform is chosen as the representative in one species and a long isoform represents the gene in another species. Alternatively, differences in the feature architecture of two orthologs (see Challenge 2) may reflect alternative splice events instead of evolutionary change. One approach to ameliorate this issue is the selection of one or several representative isoforms. The Ortho2tree pipeline, for example, provides a consistent set of canonical isoforms for closely related species using multiple sequence alignments and phylogenetic clustering (Insana et al. 2024). Other approaches additionally consider proteomic and structural data and involve human curation. However, these are currently only available for model organisms (Morales et al. 2022; Rodriguez et al. 2022). As an alternative concept, individual ortholog search tools allow for a selection on the fly by identifying the isoform that provides the best pairwise alignment scores to orthologs from all other considered species (Altenhoff et al. 2021).

Irrespective of the precise selection procedure, restricting comparative analysis to only one isoform per gene neglects the functional complexity in a proteome conveyed by these isoforms (Manuel et al. 2023) and how this has evolved. To accommodate this aspect, orthologous isoforms have to be identified, which makes it necessary to extend the concept of orthology to also include the evolutionary history of alternative splice patterns. To better compare the true functional repertoire of different proteomes, isoform-aware orthology assignment methods are required. As the first step in this direction, SplicedFamAlign introduces the concept of transcript orthologous groups. Specifically, the software analyses exon structure preservation as well as a one-to-one correspondence between the exons of two transcripts (Jammali et al. 2019).

Pan-genome-based Orthology Inference

In the current surge of genomic sequencing, available data do not only become more broadly distributed across the tree of life. It also becomes deeper, with multiple and sometimes even thousands of individuals being sequenced per species. This wealth of data allows the focus of comparative analyses to extend from an individual genome to all genes observed in a taxonomic clade, the pan-genome. Using a single assembly as a representative for the genomic diversity of a species is common practice, but it introduces a reference bias (Eizenga et al. 2020). For example, each additional human genome sequence adds, on average, 23 Mb of euchromatic autosomal sequence to the GRCh38 assembly (Liao et al. 2023). Genes residing in these nonreference regions are consequently missed in reference-based analyses.

One powerful alternative to reference-based approaches are pan-genome graphs. They employ sequence alignments (Eizenga et al. 2020) or conservation of gene order (Gautreau et al. 2020) to represent the pan-genome in a compact data structure. This works well to capture evolutionary dynamics in the pan-genome as long as the evolutionary distances remain short. With longer distances, however, both sequence similarity and gene-order conservation decrease. Additionally, the number of gene duplications increases which introduces reticulations in the pan-genome graph. Replacing the currently employed unidirectional searches to identify corresponding genes across genomes with an orthology inference could resolve these reticulations; however, the additional computational complexity diminishes the benefit of pan-genome graphs.

Capturing the pan-genome using hierarchical orthologous groups is an alternative to the reference-based approach, which, however, ignores gene-order conservation. This idea is applied in the Microbial Genome Database for Comparative Analysis (MBGD) that constructs groups of orthologs on multiple taxonomic levels

for 15,397 assemblies from 4,747 species in 1,444 genera (Uchiyama et al. 2019). This approach can cover larger evolutionary distances, but its range remains limited by the computational complexity of all-versus-all ortholog searches across gene set collections whose numbers increase in broader taxonomic scopes.

Future studies are now needed to find ways to harness the potential of integrating both gene-order conservation and orthology inference over larger evolutionary distances, while keeping the computational overhead to a minimum.

Orthologs of noncoding RNAs

Proteins were long treated as the main agent of biological function, and consequently, the Quest for Orthologs has first started to unravel the evolutionary history of protein-coding genes. However, up to 98% of all human transcripts are noncoding RNAs (ncRNAs) (Alessio et al. 2020). Even though there has been a surge of publications that start to elucidate the functional role of these noncoding transcripts (Mattick et al. 2023), transferring new information between them via ortholog prediction has remained challenging.

The identification of orthologs to microRNAs (miRNAs) was mainly hindered by their small size (~22 nt) and because they can be in repeat-rich regions. High-quality annotations of miRNAs require densely sampled small RNA-seq datasets and expert knowledge (Fromm et al. 2022). With the development of MirMachine, annotations of 508 miRNA families can now be extended to genome assemblies without the help of transcriptomic data (Umu et al. 2023). This is done by training a covariance model for each family with high-confidence miRNAs from MirGeneDB (Fromm et al. 2022). These models can serve as the first step of an ortholog search, in principle, but are limited to conserved miRNA families and training sequences are restricted to species represented in MirGeneDB. These limitations have been recently overcome by ncOrtho, which exploits regions of conserved synteny to first identify a set of positional miRNA orthologs in any given set of genome assemblies (Langschied et al. 2023). These high-confidence orthologs are then used for training covariance models which form the basis for a subsequent model-based ortholog search. Ortholog assignments by ncOrtho match gold-standard annotations in precision and facilitate high-resolution studies on the evolution and taxonomic distribution of miRNA families.

Predicting orthologs of other classes of noncoding RNAs remains challenging because their sequences change rapidly over time, even if their function remains conserved (Ross et al. 2021). Orthologs between species as closely related as human and mouse can be identified by considering conservation of splice sites and sites of active transcription (Chen et al. 2016). However, this is only possible if high-quality annotations, whole-genome alignments as well as transcriptomic data are available for both species. Therefore, lncRNA databases either lack orthology assignments completely (Zhou et al. 2021), are restricted to manually curated models of a few lncRNA families (Kalvari et al. 2021), or cover only a small taxonomic clade like primates (Bryzghalov et al. 2020).

The development of methods for the detection of miRNA orthologs is a first step for integrating ncRNAs into orthology-based frameworks. For the first time, miRNAs are available as markers in phylogenomic analyses, and their evolutionary dynamics can now be studied with unprecedented resolution. However, developing new methods for identifying orthologs from other classes of ncRNAs is key for continuing the Quest for Orthologs.

Challenge 4: Integrating Orthology With Protein Structure

Multiple studies have shown that the evolutionary age of many genes is underestimated due to sensitivity limits of sequence similarity-based orthology assignments (Jain et al. 2019; Weisman et al. 2020). Given that protein structure is up to ten times more conserved than the amino acid sequence (Illergård et al. 2009), this additional level of information promises to extend orthology prediction to even deeper timescales.

In July 2021, DeepMind and the European Bioinformatics Institute released AlphaFold Database, which in its current version covers the vast majority of proteins in UniProt with over 200 million protein structure models (Varadi et al. 2024). Additionally, embeddings from protein language models have been applied in recent years for various bioinformatics tasks, including function and structure prediction, but also homology assignments (Heinzinger et al. 2022). Therefore, high-quality models of protein structures are suddenly easy to obtain and provide an excellent basis for finding evolutionary-related proteins. With the development of the ultra-fast structural aligner Foldseek (van Kempen et al. 2023), homology assignments across large evolutionary timescales are increasingly viable (Ruperti et al. 2023). Indeed, protein structure-based searches enable the discovery of homologs across the most distant branches of the tree of life. For example, the same arrangement of the spaH domain followed by a beta-propeller protein domain is observed in both eukaryotic nuclear pore complex proteins and bacterial protomembranes, despite their sequence similarity being $< 4\%$ (Santarella-Mellwig et al. 2010).

However, available strategies for structure-based homology assignments still struggle to distinguish between orthologs and paralogs. Nevertheless, predicted protein structures can identify cases in which orthologs have diverged on structure level (Iruegas et al. 2023). In this sense,

protein structure serves as an additional feature such as domain architecture that can be used to infer if the function of orthologs has changed (see Challenge 2).

Applied researchers are eager to leverage similar protein structures as a means for transferring annotations of protein function, which has traditionally been based on orthology. Since structure and orthology are two different concepts, their assignments of "corresponding" proteins may clash. This happens, for example, in the case of convergent evolution of protein structures from different origins or the subfunctionalization of paralogs that is not reflected in the structure of a protein (Bordin et al. 2021). Both these concepts should therefore be used in complement to each other wherever possible, but a framework for combined utilization of these data is yet to be developed.

Challenge 5: Making Orthology Inferences FAIR

The concept of "Green bioinformatics" emphasizes the need to avoid unnecessary computations and energy usage (Lannelongue et al. 2023). As an orthology research community, we have therefore the responsibility to be efficient and environmentally responsible. The computational and logistical effort to precompute orthology assignments across many taxa only becomes justified if a wide range of users has access to these data and uses it (Mendes de Farias et al. 2022). Making data accessible to users from different backgrounds, including software developers, researchers and the general public, is best achieved by adhering to the FAIR data principles (Wilkinson et al. 2016). The various dedicated public databases for the dissemination of orthology inferences have been a first step in this direction. Publicly available orthology relationships are summarized in DIOPT, a platform integrating the prediction results of 19 algorithms as well as the annotation effort from model organism databases, that allows users to filter orthologs based on votes and rankings, providing protein alignments, domain information, and species conservation data (Hu et al. 2011). It further gives individual users the option to adjust the strictness with which orthologs are assigned: do they want to take orthologs that are predicted by many tools and are therefore likely to be true, or do they want to cast a broad net and take all putative orthologs inferred by any given method? Being originally tailored toward the *Drosophila* community, the functionality of DIOPT was later extended to other model organisms (Hu et al. 2017) and was used mapping various datasets across species such as interaction data (Hu et al. 2018) and single-cell data (Hu et al. 2021). Other platforms like the Gene Nomenclature Committee of the Human Genome Organization (HGNC) Comparison of Orthology Predictions (HCOP) tool (Wright et al. 2005) and the Alliance of Genome Resources (The Alliance of Genome Resources Consortium 2020) also provide meta-predictions that integrate various sources of orthology assignments with functional and other additional information.

To improve accessibility, new ways of querying databases may prove useful. One possibility is to convert scientific questions to database queries. Natural language processing could enable users to query databases in a manner that more closely resembles natural language questions. This has the potential to make the accessibility of orthology databases similar to that of chatbots like ChatGPT. Additionally, federated SPARQL queries could facilitate the gathering of information between databases (Sima et al. 2019).

A further approach for reducing the environmental footprint of ortholog searches is to share resources, results, and knowledge to avoid redundant computations. The OrthoXML format for orthology data has been proposed as a standard format for sharing this information (Schmitt et al. 2011). Unfortunately, it still faces challenges in terms of compatibility and user-friendliness and more work is necessary to provide a relevant data exchange standard. To reduce the need for all-against-all computations, new tools like OMAmer (Rossier et al. 2021) and SHOOT (Emms and Kelly 2022) have been developed to assign new gene sets to orthologous groups and place query sequences onto phylogenetic trees, respectively. In this vein, the matter of cross-referencing between releases becomes increasingly important. This is especially of concern for the reconstruction of meta-databases like MetaPhOrs v2.5 (Chorostecki et al. 2020) or phylome databases like PhylomeDB v5 (Fuentes et al. 2022). Unless advancements occur in the methods employed for crosslinking information from disparate databases across different versions of a species' proteome, a necessity to recalibrate analyses using the latest proteomes and their corresponding annotations will persist. Not addressing this challenge jeopardizes the preservation of previously documented information but also imposes an extensive computational burden.

While considerable efforts have been made thus far to make orthology accessible to the scientific community, communicating orthology-related scientific concepts to the public has lagged. Without a basic understanding of scientific concepts, the public may fall prey to harmful misinformation and sensationalism. Evolution, a foundational pillar of biology, is often misunderstood or misrepresented. Science communication projects like "In the Light of Evolution" (https://lightofevolution.org/) aim to foster curiosity about evolution and bioinformatics among people of all ages (Blatter et al. 2023). These projects create engaging stories and activities based on genuine scientific publications, offering participants tangible learning experiences. Orthology data from databases like OMA play a key role in these projects, helping to educate the public about common ancestry, relatedness, and evolution in a fun and

engaging way. By providing opportunities that reflect real-world scientific practices, these initiatives provide valuable insights into the workings of scientists.

Accessibility and reusability of orthology information are essential for the orthology community. By understanding the needs of biologists, developing user-friendly tools, incorporating NLP for database querying, engaging in science communication with the public, and promoting resource sharing and green bioinformatics practices, we can make orthology more accessible and beneficial for all stakeholders.

Conclusion

Densely sampled collections of genome assemblies allow biodiversity genomics projects to operate at new scales. The inference of evolutionary relationships forms the basis of these projects but faces several challenges that need to be addressed in future studies. Different challenges connect to different areas of biodiversity genomics (Fig. 1). For example, conservation genomics benefits most from identifying complete gene catalogues in large collections of assemblies (Challenge 1) but might be less interested in the functional differences of the orthologs therein (Challenge 2). In contrast, finding differences in molecular functions will be one of the main focuses of biomedical research. Testing orthologs for conserved domain architecture or protein structure similarity promises to improve the accuracy of any fine-grained analysis in this field (Challenge 2). Other areas like natural product genomics are impacted by each challenge presented here. Finding genes that synthesize new or specific secondary metabolites benefits from a large body of sequenced and annotated genes (Challenge 1). Changes in domain architecture or predicted protein structure might indicate functional adaptations that have an impact on the produced metabolite (Challenges 2 and 4). Biosynthetic genes may produce short proteins or remain hidden in the unannotated "bacterial dark matter" (Challenge 3). Lastly, researchers that are interested in natural product genomics must be aware of recent advances in orthology prediction and know how to apply them (Challenge 5). Currently, many different communities rely on orthology prediction implicitly because they are interested in the annotation transfer between genes. Looking ahead, we must strive to connect these communities and create more awareness of the opportunities created by the Quest for Orthologs.

Funding

We wish to acknowledge the following support for attending the meetings and/or writing this manuscript: I.E. and M.H. were supported by the research funding program Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE) of the State of Hessen, Research Center for Translational Biodiversity Genomics (TBG), S.M. and C.D. were funded by the Swiss National Science Foundation (Grant no. 205085). N.G. was supported by the Swiss National Science Foundation (Grant no. 199775) and the Swiss Institute of Bioinformatics. T.A.R. was supported by a Royal Society University Research Fellowship (URF\R\191005). N.B. acknowledges funding from the Wellcome Trust (Grant 221327/Z/20/Z). E.S. acknowledges funding from the Swedish Research Council (Grant 2019-04095). T.G. group acknowledges support from the Spanish Ministry of Science and Innovation for grants PID2021-126067NB-I00, CPP2021-008552, PCI2022-135066-2, and PDC2022-133266-I00, cofounded by ERDF "A way of making Europe"; from the Catalan Research Agency (AGAUR) SGR01551; from the European Union's Horizon 2020 Research and Innovation Programme (ERC-2016-724173); from the Gordon and Betty Moore Foundation (Grant GBMF9742); from the "La Caixa" foundation (Grant LCF/PR/HR21/00737), and from the Instituto de Salud Carlos III (IMPACT Grant IMP/00019 and CIBERINFEC CB21/13/00061-ISCIII-SGEFI/ERDF). J.H.-C. was supported by the National Programme for Fostering Excellence in Scientific and Technical Research, MCIN/AEI/10.13039/501100011033/and FEDER, Una manera de hacer Europa (Grant no. PID2021-127210NB-I00).

Members of the Quest for Orthologs (QfO) Consortium

Adrian Altenhoff, Aida Ouangraoua, Alex Warwick Vesztrocy, Arnaud Kress, Christophe Dessimoz, Dannie Durand, David Emms, Diego Fuentes-Palacios, Emma Persson, Erik Sonnhammer, Felix Langschied, Ikuo Uchiyama, Ingo Ebersberger, Jaime Huerta-Cepas, Laetitia Poidevin, Luis Pedro Coelho, Maria J. Martin, Michael Hiller, Natasha Glover, Nicola Bordin, Odile Lecompte, Paul D. Thomas, Saioa Manzano-Morales, Salvador Capella-Gutierrez, Salvatore Cosentino, Silvia Prieto Baños, Sina Majidian, Sofia Kirke Forslund-Startceva, Stefano Pascarelli, Thomas A. Richards, Toni Gabaldón, Vinh Tran, Wataru Iwasaki, Yan Wang, Yannis Nevers.

Data Availability

No new data were generated or analyzed in support of this research.

Literature Cited

Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, Hill DP, et al. The gene ontology knowledgebase in 2023. Genetics. 2023:224(1):iyad031. https://doi.org/10.1093/genetics/iyad031.

Alessio E, Bonadio RS, Buson L, Chemello F, Cagnin S. A single cell but many different transcripts: a journey into the world of long non-coding RNAs. Int J Mol Sci. 2020:21(1):302. https://doi.org/10.3390/ijms21010302.

Altenhoff AM, Glover NM, Dessimoz C. Inferring orthology and paralogy. In: Anisimova M, editors. Evolutionary genomics: statistical and computational methods. New York (NY): Springer New York; 2019. p. 149–175.

Altenhoff AM, Train C-M, Gilbert KJ, Mediratta I, Mendes de Farias T, Moi D, Nevers Y, Radoykova H-S, Rossier V, Warwick Vesztrocy A, et al. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids Res. 2021:49(D1):D373–D379. https://doi.org/10.1093/nar/gkaa1007.

Amaral P, Carbonell-Sala S, De La Vega FM, Faial T, Frankish A, Gingeras T, Guigo R, Harrow JL, Hatzigeorgiou AG, Johnson R, et al. The status of the human gene catalogue. Nature. 2023:622(7981):41–47. https://doi.org/10.1038/s41586-023-06490-x.

Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. 2021:49(D1):D121–D124. https://doi.org/10.1093/nar/gkaa967.

Bálint B, Merényi Z, Hegedüs B, Grigoriev IV, Hou Z, Földi C, Nagy LG. ContScout: sensitive detection and removal of contamination from annotated genomes. Nat Commun. 2024:15(1):936. https://doi.org/10.1038/s41467-024-45024-5.

Birikmen M, Bohnsack KE, Tran V, Somayaji S, Bohnsack MT, Ebersberger I. Tracing eukaryotic ribosome biogenesis factors into the archaeal domain sheds light on the evolution of functional complexity. Front Microbiol. 2021:12:739000. https://doi.org/10.3389/fmicb.2021.739000.

Blatter M-C, Zahn-Zabal M, Moix S, Pichon B, Dessimoz C, Glover N. Bringing science to the public in the light of evolution. Biol Methods Protoc. 2023:8(1):bpad040. https://doi.org/10.1093/biomethods/bpad040.

Bonnie JK, Ahmed OY, Langmead B. Dandd: efficient measurement of sequence growth and similarity. iScience. 2024:27(3):109054. https://doi.org/10.1016/j.isci.2024.109054.

Bordin N, Sillitoe I, Lees JG, Orengo C. Tracing evolution through protein structures: nature captured in a few thousand folds. Front Mol Biosci. 2021:8:668184. https://doi.org/10.3389/fmolb.2021.668184.

Bruna T, Lomsadze A, Borodovsky M. GeneMark-ETP: Automatic Gene Finding in Eukaryotic Genomes in Consistency with Extrinsic Data. bioRxiv 524024. https://doi.org/10.1101/2023.01.13.524024, 03 January 2024, preprint: not peer reviewed.

Bryzghalov O, Szcześniak MW, Makałowska I. SyntDB: defining orthologues of human long noncoding RNAs across primates. Nucleic Acids Res. 2020:48:D238–D245. https://doi.org/10.1093/nar/gkz941.

Cannell N, Emms DM, Hetherington AJ, MacKay J, Kelly S, Dolan L, Sweetlove LJ. Multiple metabolic innovations and losses are associated with major transitions in land plant evolution. Curr Biol. 2020:30(10):1783–1800.e11. https://doi.org/10.1016/j.cub.2020.02.086.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021:38(12):5825–5829. https://doi.org/10.1093/molbev/msab293.

Challis R, Kumar S, Sotero-Caio C, Brown M, Blaxter M. Genomes on a tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. Wellcome Open Res. 2023:8:24. https://doi.org/10.12688/wellcomeopenres.18658.1.

Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, Hanna JH, Regev A, Garber M. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome Biol. 2016:17(1):19. https://doi.org/10.1186/s13059-016-0880-9.

Chorostecki U, Molina M, Pryszcz LP, Gabaldón T. Metaphors 2.0: integrative, phylogeny-based inference of orthology and paralogy across the tree of life. Nucleic Acids Res. 2020:48(W1):W553–W557. https://doi.org/10.1093/nar/gkaa282.

Chrisman B, He C, Jung J-Y, Stockham N, Paskov K, Washington P, Wall DP. The human "contaminome": bacterial, viral, and computational contamination in whole genome sequences from 1000 families. Sci Rep. 2022:12(1):9863. https://doi.org/10.1038/s41598-022-13269-z.

Coelho LP, Alves R, del Río ÁR, Myers PN, Cantalapiedra CP, Giner-Lamia J, Schmidt TS, Mende DR, Orakov A, Letunic I, et al. Towards the biogeography of prokaryotic genes. Nature. 2022:601(7892):252–256. https://doi.org/10.1038/s41586-021-04233-4.

Collins G, Schneider C, Boštjanči[?] LL, Burkhardt U, Christian A, Decker P, Ebersberger I, Hohberg K, Lecompte O, Merges D, et al. The MetaInvert soil invertebrate genome resource provides insights into below-ground biodiversity and evolution. Commun Biol. 2023:6(1):1241. https://doi.org/10.1038/s42003-023-05621-4.

Cosentino S, Sriswasdi S, Iwasaki W. SonicParanoid2: fast, accurate, and comprehensive orthology inference with machine learning and language models. Genome Biol. 2024:25(1):195. https://doi.org/10.1186/s13059-024-03298-4.

Cui X, Xue Y, McCormack C, Garces A, Rachman TW, Yi Y, Stolzer M, Durand D. Simulating domain architecture evolution. Bioinformatics. 2022:38(Suppl. 1):i134–i142. https://doi.org/10.1093/bioinformatics/btac242.

del Río ÁR, Giner-Lamia J, Cantalapiedra CP, Botas J, Deng Z, Hernández-Plaza A, Munar-Palmer M, Santamaría-Hernando S, Rodríguez-Herva JJ, Ruscheweyh H-J, et al. Functional and evolutionary significance of unknown genes from uncultivated taxa. Nature. 2024:626(7998):377–384. https://doi.org/10.1038/s41586-023-06955-z.

Dembech E, Malatesta M, De Rito C, Mori G, Cavazzini D, Secchi A, Morandin F, Percudani R. Identification of hidden associations among eukaryotic genes through statistical analysis of co-evolutionary transitions. Proc Natl Acad Sci U S A. 2023:120(16):e2218329120. https://doi.org/10.1073/pnas.2218329120.

Derelle R, Philippe H, Colbourne JK. Broccoli: combining phylogenetic and network analyses for orthology assignment. Mol Biol Evol. 2020:37(11):3389–3396. https://doi.org/10.1093/molbev/msaa159.

Djahanschiri B, Di Venanzio G, Distel JS, Breisch J, Dieckmann MA, Goesmann A, Averhoff B, Göttig S, Wilharm G, Feldman MF, et al. Evolutionarily stable gene clusters shed light on the common grounds of pathogenicity in the *Acinetobacter calcoaceticus-baumannii* complex. PLoS Genet. 2022:18(6):e1010020. https://doi.org/10.1371/journal.pgen.1010020.

Dosch J, Bergmann H, Tran V, Ebersberger I. FAS: assessing the similarity between proteins using multi-layered feature architectures. Bioinformatics. 2023:39(5):btad226. https://doi.org/10.1093/bioinformatics/btad226.

Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, et al. Pangenome graphs. Annu Rev Genomics Hum Genet. 2020:21(1):139–162. https://doi.org/10.1146/annurev-genom-120219-080406.

Emms DM, Kelly S. SHOOT: phylogenetic gene search and ortholog inference. Genome Biol. 2022:23(1):85. https://doi.org/10.1186/s13059-022-02652-8.

Fitch WM. Distinguishing homologous from analogous proteins. Syst Zool. 1970:19(2):99–113. https://doi.org/10.2307/2412448.
Foo A, Cerdeira L, Hughes GL, Heinz E. Recovery of metagenomic data from the *Aedes aegypti* microbiome using a reproducible snakemake pipeline: MINUUR. Wellcome Open Res. 2023:8:131. https://doi.org/10.12688/wellcomeopenres.19155.2.
Fromm B, Høye E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, Umu SU, Chabot PJ, Kang W, Aslanzadeh M, et al. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. Nucleic Acids Res. 2022:50(D1):D204–D210. https://doi.org/10.1093/nar/gkab1101.
Fuentes D, Molina M, Chorostecki U, Capella-Gutiérrez S, Marcet-Houben M, Gabaldón T. PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. Nucleic Acids Res. 2022:50(D1):D1062–D1068. https://doi.org/10.1093/nar/gkab966.
Gabaldón T, Koonin EV. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 2013:14(5):360–366. https://doi.org/10.1038/nrg3456.
Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv 544449. https://doi.org/10.1101/2023.06.10.544449, 29 February 2024, preprint: not peer reviewed.
Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res. 2021:49(D1):D274–D281. https://doi.org/10.1093/nar/gkaa1018.
Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the gene ontology consortium. Brief Bioinform. 2011:12(5):449–462. https://doi.org/10.1093/bib/bbr042.
Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, Perrin A, Médigue C, Calteau A, Cruveiller S, et al. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol. 2020:16:e1007732. https://doi.org/10.1371/journal.pcbi.1007732.
Grealey J, Lannelongue L, Saw W-Y, Marten J, Méric G, Ruiz-Carmona S, Inouye M. The carbon footprint of bioinformatics. Mol Biol Evol. 2022:39(3):msac034. https://doi.org/10.1093/molbev/msac034.
Grossart H-P, Massana R, McMahon KD, Walsh DA. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. Limnol Oceanogr. 2020:65(S1):S2–S20. https://doi.org/10.1002/lno.11382.
Guigó R. Genome annotation: from human genetics to biodiversity genomics. Cell Genom. 2023:3(8):100375. https://doi.org/10.1016/j.xgen.2023.100375.
Heinzinger M, Littmann M, Sillitoe I, Bordin N, Orengo C, Rost B. Contrastive learning on protein embeddings enlightens midnight zone. NAR Genom Bioinform. 2022:4(2):lqac043. https://doi.org/10.1093/nargab/lqac043.
Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016:32(5):767–769. https://doi.org/10.1093/bioinformatics/btv661.
Hu Y, Comjean A, Mohr SE; The FlyBase Consortium; Perrimon N. Gene2Function: an integrated online resource for gene function discovery. G3 (Bethesda). 2017:7(8):2855–2858. https://doi.org/10.1534/g3.117.043885.
Hu Y, Flockhart I, Vinayagam A, Bergwitz C, Berger B, Perrimon N, Mohr SE. An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics. 2011:12(1):357. https://doi.org/10.1186/1471-2105-12-357.
Hu Y, Tattikota SG, Liu Y, Comjean A, Gao Y, Forman C, Kim G, Rodiger J, Papatheodorou I, Dos Santos G, et al. DRscDB: a single-cell RNA-seq resource for data mining and data comparison across species. Comput Struct Biotechnol J. 2021:19:2018–2026. https://doi.org/10.1016/j.csbj.2021.04.021.
Hu Y, Vinayagam A, Nand A, Comjean A, Chung V, Hao T, Mohr SE, Perrimon N. Molecular interaction search tool (MIST): an integrated resource for mining gene and protein interaction data. Nucleic Acids Res. 2018:46(D1):D567–D574. https://doi.org/10.1093/nar/gkx1116.
Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. Proteins. 2009:77(3):499–508. https://doi.org/10.1002/prot.22458.
Insana G, Martin MJ, Pearson WR. Improved selection of canonical proteins for reference proteomes. NAR Genom Bioinform. 2024:6:lqae066. https://doi.org/10.1093/nargab/lqae066.
Iruegas R, Pfefferle K, Göttig S, Averhoff B, Ebersberger I. Feature architecture aware phylogenetic profiling indicates a functional diversification of type IVa pili in the nosocomial pathogen *Acinetobacter baumannii*. PLoS Genet. 2023:19(7):e1010646. https://doi.org/10.1371/journal.pgen.1010646.
Irwin NAT, Pittis AA, Richards TA, Keeling PJ. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nat Microbiol. 2022:7(2):327–336. https://doi.org/10.1038/s41564-021-01026-3.
Jain A, Perisa D, Fliedner F, von Haeseler A, Ebersberger I. The evolutionary traceability of a protein. Genome Biol Evol. 2019:11(2):531–545. https://doi.org/10.1093/gbe/evz008.
Jammali S, Aguilar J-D, Kuitche E, Ouangraoua A. SplicedFamAlign: CDS-to-gene spliced alignment and identification of transcript orthology groups. BMC Bioinformatics. 2019:20(S3):133. https://doi.org/10.1186/s12859-019-2647-2.
Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 2021:49(D1):D192–D200. https://doi.org/10.1093/nar/gkaa1047.
Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023:51(D1):D587–D592. https://doi.org/10.1093/nar/gkac963.
Kirilenko BM, Munegowda C, Osipova E, Jebb D, Sharma V, Blumer M, Morales AE, Ahmed A-W, Kontopoulos D-G, Hilgers L, et al. Integrating gene annotation with orthology inference at scale. Science. 2023:380(6643):eabn3107. https://doi.org/10.1126/science.abn3107.
Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018:34:660–668. https://doi.org/10.1093/bioinformatics/btx624.
Kute PM, Soukarieh O, Tjeldnes H, Trégouët D-A, Valen E. Small open reading frames, how to find them and determine their function. Front Genet. 2021:12:796060. https://doi.org/10.3389/fgene.2021.796060.
Kuzmin E, Taylor JS, Boone C. Retention of duplicated genes in evolution. Trends Genet. 2022:38(1):59–72. https://doi.org/10.1016/j.tig.2021.06.016.
Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Res. 2023:51(D1):D445–D451. https://doi.org/10.1093/nar/gkac998.
Laiolo E, Alam I, Uludag M, Jamil T, Agusti S, Gojobori T, Acinas SG, Gasol JM, Duarte CM. Metagenomic probing toward an atlas of

the taxonomic and metabolic foundations of the global ocean genome. Front Sci. 2024:1:1038696. https://doi.org/10.3389/fsci.2023.1038696.
Langschied F, Leisegang MS, Brandes RP, Ebersberger I. ncOrtho: efficient and reliable identification of miRNA orthologs. Nucleic Acids Res. 2023:51(13):e71. https://doi.org/10.1093/nar/gkad467.
Lannelongue L, Aronson H-EG, Bateman A, Birney E, Caplan T, Juckes M, McEntyre J, Morris AD, Reilly G, Inouye M. GREENER principles for environmentally sustainable computational science. Nat Comput Sci. 2023:3(6):514–521. https://doi.org/10.1038/s43588-023-00461-y.
Laurent JM, Garge RK, Teufel AI, Wilke CO, Kachroo AH, Marcotte EM. Humanization of yeast genes with multiple human orthologs reveals functional divergence between paralogs. PLoS Biol.2020:18(5):e3000627. https://doi.org/10.1371/journal.pbio.3000627.
Leigh DM, Hendry AP, Vázquez-Domínguez E, Friesen VL. Estimated six per cent loss of genetic variation in wild populations since the industrial revolution. Evol Appl. 2019:12(8):1505–1512. https://doi.org/10.1111/eva.12810.
Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, et al. The Earth BioGenome Project 2020: starting the clock. Proc Natl Acad Sci U S A. 2022:119(4):e2115635118. https://doi.org/10.1073/pnas.2115635118.
Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. A draft human pangenome reference. Nature. 2023:617(7960):312–324. https://doi.org/10.1038/s41586-023-05896-x.
Linard B, Ebersberger I, McGlynn SE, Glover N, Mochizuki T, Patricio M, Lecompte O, Nevers Y, Thomas PD, Gabaldón T, et al. Ten years of collaborative progress in the quest for orthologs. Mol Biol Evol. 2021:38(8):3033–3045. https://doi.org/10.1093/molbev/msab098.
Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, Connor O, Ackermann MI, Hahn M, Srivastava AS, Crowe DS, et al. Function and functional redundancy in microbial systems. Nat Ecol Evol. 2018:2(6):936–943. https://doi.org/10.1038/s41559-018-0519-1.
Manni M, Berkeley MR, Seppey M, Zdobnov EM. BUSCO: assessing genomic data quality and beyond. Curr Protoc. 2021:1(12):e323. https://doi.org/10.1002/cpz1.323.
Manuel JM, Guilloy N, Khatir I, Roucou X, Laurent B. Re-evaluating the impact of alternative RNA splicing on proteomic diversity. Front Genet. 2023:14:1089053. https://doi.org/10.3389/fgene.2023.1089053.
Marcet-Houben M, Collado-Cala I, Fuentes-Palacios D, Gómez AD, Molina M, Garisoain-Zafra A, Chorostecki U, Gabaldón T. EvolClustDB: exploring eukaryotic gene clusters with evolutionarily conserved genomic neighbourhoods. J Mol Biol. 2023:435(14):168013. https://doi.org/10.1016/j.jmb.2023.168013.
Martelossi J, Nicolini F, Subacchi S, Pasquale D, Ghiselli F, Luchetti A. Multiple and diversified transposon lineages contribute to early and recent bivalve genome evolution. BMC Biol. 2023:21(1):145. https://doi.org/10.1186/s12915-023-01632-z.
Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen L-L, Chen R, Dean C, Dinger ME, Fitzgerald KA, et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023:24(6):430–447. https://doi.org/10.1038/s41580-022-00566-8.
McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017:2(4):17040. https://doi.org/10.1038/nmicrobiol.2017.40.
Mendes de Farias T, Wollbrett J, Robinson-Rechavi M, Bastian F. Lessons learned to boost a bioinformatics knowledge base reusability, the Bgee experience. Gigascience. 2022:12:giad058. https://doi.org/10.1093/gigascience/giad058.
Miller D, Stern A, Burstein D. Deciphering microbial gene function using natural language processing. Nat Commun. 2022:13(1):5731. https://doi.org/10.1038/s41467-022-33397-4.
Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021:37(18):3029–3031. https://doi.org/10.1093/bioinformatics/btab184.
Moi D, Dessimoz C. Reconstructing protein interactions across time using phylogeny-aware graph neural networks. bioRxiv 501014. https://doi.org/10.1101/2022.07.21.501014, 22 July 2022, preprint: not peer reviewed.
Moi D, Dessimoz C. Phylogenetic profiling in eukaryotes comes of age. Proc Natl Acad Sci U S A. 2023:120(19):e2305013120. https://doi.org/10.1073/pnas.2305013120.
Montañés JC, Huertas M, Messeguer X, Albà MM. Evolutionary trajectories of new duplicated and putative De Novo genes. Mol Biol Evol. 2023:40(5):msad098. https://doi.org/10.1093/molbev/msad098.
Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, Cox E, Davidson C, Ermolaeva O, Farrell CM, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022:604(7905):310–315. https://doi.org/10.1038/s41586-022-04558-8.
Nevers Y, Glover NM, Dessimoz C, Lecompte O. Protein length distribution is remarkably uniform across the tree of life. Genome Biol. 2023:24(1):135. https://doi.org/10.1186/s13059-023-02973-2.
Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, Codo L, Cosentino S, Marcet-Houben M, Vlasova A, et al. The quest for orthologs orthology benchmark service in 2022. Nucleic Acids Res. 2022:50(W1):W623–W632. https://doi.org/10.1093/nar/gkac330.
Nevers Y, Warwick Vesztrocy A, Rossier V, Train C-M, Altenhoff A, Dessimoz C, Glover NM. Quality assessment of gene repertoire annotations with OMArk. Nat Biotechnol. 2024. https://doi.org/10.1038/s41587-024-02147-w.
Ocaña-Pallarès E, Williams TA, López-Escardó D, Arroyo AS, Pathmanathan JS, Bapteste E, Tikhonenkov DV, Keeling PJ, Szöllősi GJ, Ruiz-Trillo I. Divergent genomic trajectories predate the origin of animals and fungi. Nature. 2022:609(7928):747–753. https://doi.org/10.1038/s41586-022-05110-4.
Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A, Ghensi P, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell. 2019:176(3):649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.
Persson E, Kaduk M, Forslund SK, Sonnhammer ELL. Domainoid: domain-oriented orthology inference. BMC Bioinformatics. 2019:20(1):523. https://doi.org/10.1186/s12859-019-3137-2.
Persson E, Sonnhammer ELL. InParanoiDB 9: ortholog groups for protein domains and full-length proteins. J Mol Biol. 2023:435(14):168001. https://doi.org/10.1016/j.jmb.2023.168001.
Pilling D, Bélanger J, Hoffmann I. Declining biodiversity for food and agriculture needs urgent global action. Nat Food. 2020:1(3):144–147. https://doi.org/10.1038/s43016-020-0040-y.
Pueyo JI, Magny EG, Couso JP. New peptides under the s(ORF)ace of the genome. Trends Biochem Sci. 2016:41(8):665–678. https://doi.org/10.1016/j.tibs.2016.05.003.
Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and

genomic organization. Science. 2007:317(5834):86–94. https://doi.org/10.1126/science.1139158.
Rocha EPC. Inference and analysis of the relative stability of bacterial chromosomes. Mol Biol Evol. 2006:23(3):513–522. https://doi.org/10.1093/molbev/msj052.
Rodriguez JM, Pozo F, Cerdán-Vélez D, Di Domenico T, Vázquez J, Tress ML. APPRIS: selecting functionally important isoforms. Nucleic Acids Res. 2022:50(D1):D54–D59. https://doi.org/10.1093/nar/gkab1058.
Ross CJ, Rom A, Spinrad A, Gelbard-Solodkin D, Degani N, Ulitsky I. Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. Genome Biol. 2021:22(1):29. https://doi.org/10.1186/s13059-020-02247-1.
Rossier V, Warwick Vesztrocy A, Robinson-Rechavi M, Dessimoz C. OMAmer: tree-driven and alignment-free protein assignment to subfamilies outperforms closest sequence approaches. Bioinformatics. 2021:37(18):2866–2873. https://doi.org/10.1093/bioinformatics/btab219.
Ruperti F, Papadopoulos N, Musser JM, Mirdita M, Steinegger M, Arendt D. Cross-phyla protein annotation by structural prediction and alignment. Genome Biol. 2023:24(1):113. https://doi.org/10.1186/s13059-023-02942-9.
Santarella-Mellwig R, Franke J, Jaedicke A, Gorjánácz M, Bauer U, Budd A, Mattaj IW, Devos DP. The compartmentalized bacteria of the planctomycetes-verrucomicrobia-chlamydiae superphylum have membrane coat-like proteins. PLoS Biol. 2010:8(1):e1000281. https://doi.org/10.1371/journal.pbio.1000281.
Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. Brief Bioinform. 2011:12(5):485–488. https://doi.org/10.1093/bib/bbr025.
Sheikhizadeh Anari S, de Ridder D, Schranz ME, Smit S. Efficient inference of homologs in large eukaryotic pan-proteomes. BMC Bioinformatics. 2018:19(1):340. https://doi.org/10.1186/s12859-018-2362-4.
Sima AC, Mendes de Farias T, Zbinden E, Anisimova M, Gil M, Stockinger H, Stockinger K, Robinson-Rechavi M, Dessimoz C. Enabling semantic queries across federated bioinformatics databases. Database. 2019:2019:baz106. https://doi.org/10.1093/database/baz106.
Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A. Peptidomic discovery of short open reading frame–encoded peptides in human cells. Nat Chem Biol. 2013:9(1):59–64. https://doi.org/10.1038/nchembio.1120.
Stamboulian M, Guerrero RF, Hahn MW, Radivojac P. The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. Bioinformatics. 2020:36(Suppl. 1):i219–i226. https://doi.org/10.1093/bioinformatics/btaa468.
Storz G, Wolf YI, Ramamurthi KS. Small proteins can no longer be ignored. Annu Rev Biochem. 2014:83(1):753–777. https://doi.org/10.1146/annurev-biochem-070611-102400.
Supple MA, Shapiro B. Conservation of biodiversity in the genomics era. Genome Biol. 2018:19(1):131. https://doi.org/10.1186/s13059-018-1520-3.
The Alliance of Genome Resources Consortium. Alliance of genome resources portal: unified model organism research platform. Nucleic Acids Res. 2020:48(D1):D650–D658. https://doi.org/10.1093/nar/gkz813.
Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, MiH. PANTHER: making genome-scale phylogenetics accessible to all. Protein Sci. 2022:31(1):8–22. https://doi.org/10.1002/pro.4218.
Uchiyama I, Mihara M, Nishide H, Chiba H, Kato M. MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. Nucleic Acids Res. 2019:47(D1):D382–D389. https://doi.org/10.1093/nar/gky1054.
Umu SU, Paynter VM, Trondsen H, Buschmann T, Rounge TB, Peterson KJ, Fromm B. Accurate microRNA annotation of animal genomes using trained covariance models of curated microRNA complements in MirMachine. Cell Genom. 2023:3(8):100348. https://doi.org/10.1016/j.xgen.2023.100348.
van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2023:42(2):243–246. https://doi.org/10.1038/s41587-023-01773-0.
Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 2024:52(D1):D368–D375. https://doi.org/10.1093/nar/gkad1011.
Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020:18(11):e3000862. https://doi.org/10.1371/journal.pbio.3000862.
Wilkinson MD, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016:3(1):160018. https://doi.org/10.1038/sdata.2016.18.
Wright MW, Eyre TA, Lush MJ, Povey S, Bruford EA. HCOP: the HGNC comparison of orthology predictions search tool. Mamm Genome. 2005:16(11):827–828. https://doi.org/10.1007/s00335-005-0103-2.
Zhou B, Ji B, Liu K, Hu G, Wang F, Chen Q, Yu R, Huang P, Ren J, Guo C, et al. EVLncRNAs 2.0: an updated database of manually curated functional long non-coding RNAs validated by low-throughput experiments. Nucleic Acids Res. 2021:49(D1):D86–D91. https://doi.org/10.1093/nar/gkaa1076.

Associate editor: Laura Eme