Chapter 9
Orthology: Promises and Challenges

Yannis Nevers, Audrey Defosset, and Odile Lecompte

Abstract Orthology is a cornerstone of comparative genomics and has numerous applications in current biology. In this chapter, we first introduce the concepts of orthology and paralogy. We then present the currently available orthology inference methods and the community-led efforts of standardization and benchmarking accompanying these developments. The large panel of available orthology resources is compared in terms of species coverage, access, contextual data and tools proposed to end-users to facilitate the analysis and exploitation of orthology data. We then review the importance of orthology applications, ranging from the study of protein families and information transfer to the comparison of genomes and genotype/phenotype correlations. Finally, we discuss the current challenges in the orthology field, faced with an ever-increasing number of proteomes of particularly heterogeneous quality. We highlight the urgent need of considering orthology at the protein domain and transcript levels and the conceptual and practical difficulties that this raises.

Y. Nevers · A. Defosset · O. Lecompte (✉)
Complex Systems and Translational Bioinformatics, ICube UMR 7357,
Université de Strasbourg, Strasbourg, France
e-mail: odile.lecompte@unistra.fr

Y. Nevers
e-mail: yannis.nevers@unil.ch

A. Defosset
e-mail: adefosset@etu.unistra.fr

Y. Nevers
SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

P. Pontarotti (ed.), *Evolutionary Biology—A Transdisciplinary Approach*,
https://doi.org/10.1007/978-3-030-57246-4_9

9.1 Introduction

Homology is a central concept in biology and is essential for any intraspecies or interspecies sequence comparison. Originally employed to compare phenotypic traits, it is now mainly used to define relationships between genomic regions, genes and, by extension, between proteins or even sub-protein regions. In this context, homology describes the relationship between two molecular entities (usually genes or proteins) that descend from the same ancestor. Two main categories of homologs were distinguished in the early days of molecular biology (Fitch 1970): paralogs that derive from a common ancestor by a duplication event and orthologs that emerge after a speciation event (Fig. 9.1a). *Stricto sensu*, these definitions only refer to the evolutionary history of genes. However, it is commonly accepted that orthologs tend to retain a similar function, while paralogs may have different fates in the course of evolution. Indeed, the paralogous copies may develop more specialized functions compared to the ancestral gene (tissue/stage-specific expression, complementation of functions initially performed by a single gene) or one copy may evolve a new function under the reduced selection pressure or even degenerate into a pseudogene (Force et al. 1999). The 'orthology conjecture' states that orthologs frequently retain ancestral function while paralogs tend to diversify is widely used to transfer functional information between orthologs. Although this hypothesis is commonly accepted by the community, it has been challenged in some cases (Studer and Robinson-Rechavi 2009; Nehrt

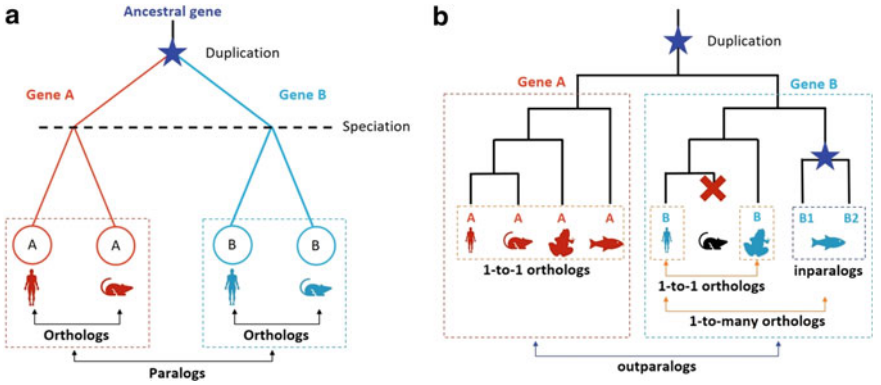

Fig. 9.1 Homology relationships. **a** Evolutionary history of a gene family with duplication and speciation events. Genes A (in red) present in humans and mouse emerged after a speciation event, they are orthologous to each other. The same is true for genes B (in blue). Genes A and B are paralogous between each other because they are separated by a duplication event in their evolutionary history. **b** Genes A (in red) are only separated by speciation events, they are 1-to-1 orthologs. The evolutionary history of genes B (in blue) is more complex with a lineage-specific loss in mouse and a 'recent' duplication in fish. Considering the evolutionary history of vertebrates, genes B1 and B2 are inparalogs to each other and co-orthologs to the human gene B. Thus, there is a 1-to-many orthology relation between the human gene B and the fish genes B1 and B2 genes. Considering Vertebrates, genes A and B are outparalogs between each other because they emerged after a duplication that occurred in the vertebrate ancestor, i.e., before speciations

et al. 2011), especially among highly similar genes. Nevertheless, it still generally holds (Altenhoff et al. 2012; Chen and Zhang 2012). Notably, it has been shown that the organization of introns (Henricson et al. 2010), the three-dimensional structure of proteins (Peterson et al. 2009) and domain architecture (Forslund et al. 2011) tend to be more conserved between orthologs than paralogs. In addition, orthologs are generally expressed in the same tissues in contrast to paralogs (Kryuchkova-Mostacci and Robinson-Rechavi 2015).

The debate around the orthology conjecture underlines the importance of taking into account the chronology of speciation and duplication events to establish functional links between homologous genes. Indeed, paralogs that derive from a 'recent' duplication event may still share the same function in contrast to distant paralogs separated over millions of years of evolution. Unfortunately, there is no objective threshold to define recent versus ancient paralogs, and in fact, it all depends on the evolutionary distance between compared species. This has been conceptualized with the terms 'outparalogs' and 'inparalogs' coined in 2002 (Sonnhammer and Koonin 2002). When comparing two species, paralogs deriving from a duplication event that occurred prior to the speciation event are called outparalogs, while paralogs originating from a duplication event subsequent to the speciation event are called inparalogs. Inparalogs are considered to be co-orthologs of genes descending from the speciation event in the other species (Fig. 9.1b). Hence, inparalogy and outparalogy are relative notions: The same paralogous sequences can be considered inparalogs or outparalogs depending on the speciation referred to. The co-orthology concept also introduces different orthology relationships: 1-to-1, 1-to-many and many-to-many orthologs (Fig. 9.1b).

The characterization of these intricate homology relationships is far from trivial since there is no direct record of past speciation or duplication events, and evolutionary scenarios can be further complicated by lineage-specific gene losses, whole genome duplications (WGD) and horizontal gene transfers (HGT). WGD or polyploidy can arise within a single species by the doubling of the chromosome set (autopolyploidy) or can result from the merging of the chromosome sets of two different species and subsequent genome doubling (allopolyploidy) (see Van de Peer et al. 2017 for a recent review). Homologs arising by autopolyploidy are called ohnologs (Wolfe 2000) and constitute a special case of paralogs, since both copies evolved originally in the same genomic context. Homeologs that result from an allopolyploidy event are more complex to define (reviewed in Glover et al. 2016) but are observed in many plants. Like orthologs, they originally emerge after a speciation event, but they are subsequently integrated in a single genome through autopolyploidization. Thus, homeologs experience a mosaic fate by initially evolving like orthologs and then after hybridization, undergoing an evolutionary pressure usually exerted on paralogs.

In HGT, the relationship does not rely on vertical transmission of genes but on acquisition of genetic material from another species. Genes whose history since their common ancestor involves an horizontal transfer are called xenologs (Gray and Fitch 1983; Fitch 2000). Xenology is especially prevalent in prokaryotes with HGT frequently involving mobile genetic elements, but it can also occur between

prokaryotes and eukaryotes (notably in the case of endosymbiosis or endoparasitism) or even between eukaryotes (reviewed in Soucy et al. 2015). Xenology relationships encompass a wide range of evolutionary histories, and xenolog classes have been proposed to reflect the events associated with the divergence of xenologs and the relative timing of transfer and speciation events (Darby et al. 2017).

The first step in the process of characterization of homology relations is based on sequence comparison. It is assumed that genes/proteins are homologous if they exhibit a higher similarity than would be expected by chance. Thus, homology detection usually relies on similarity searches, typically a BLAST search (Altschul et al. 1997; Camacho et al. 2009), with a fixed threshold of score, percentage identity, expect-value, etc. The distinction at the genome scale between the different types of homology (1-to-1 orthology, co-orthology, inparalogy, outparalogy, xenology) then requires dedicated approaches. The methods used to infer orthology and the corresponding available resources are presented in the first section of this chapter. We then review the main applications of orthology in biology. In the last section, we highlight the practical and conceptual challenges around the notion of orthology and its uses.

9.2 Orthology Inference and Resources

9.2.1 Orthology Inference Methods

An exhaustive description of the plethora of available programs is beyond the scope of this review (for a recent review on methods, see Altenhoff et al. 2019). However, these different programs can be classified into four main categories: graph-based, tree-based, hybrid and meta-prediction methods that are presented briefly below.

In graph-based methods, genes/proteins are represented by nodes and homology relationships by edges in the graph. The graph construction relies on all-against-all similarity searches between genes/proteins from two genomes. The simplest approach, called reciprocal best hit (RBH), will predict an orthology relationship between proteins A and B from two genomes if A is the genome-wide closest relative of B and vice versa (Overbeek et al. 1999). This approach only considers 1-to-1 orthology relationships, thus overlooking one-to-many and many-to-many orthologs. To circumvent this problem and offer a more comprehensive view of evolutionary relationships, other algorithms have been developed where inparalogy relations are inferred and included during graph construction. Examples of such methods include COG (Tatusov et al. 1997), Inparanoid (Remm et al. 2001), OrthoMCL (Li et al. 2003), OMA (Roth et al. 2008), EggNOG (Jensen et al. 2008), OrthoInspector (Linard et al. 2011) and OrthoFinder (Emms and Kelly 2015). The homology relationships predicted between a pair of genomes can then be extended to a set of species, in order to define groups of orthologs (also called orthogroups) present in these species. The groups are delineated on the basis of the structure of the graph by transitivity or clustering. For instance, OrthoMCL uses Markov clustering to partition the homology

graph into orthogroups containing highly connected orthologs and recent paralogs. OMA groups are based on cliques, i.e., fully connected subgraphs corresponding to genes that are all orthologs to each other, thus de facto excluding orthologs involved in 1-to-many or many-to-many relations.

Tree-based methods infer orthologs based on the gene's evolutionary history, which is reconstructed by reconciling the gene family tree with the species tree. First, a multiple alignment of homologous sequences is constructed to generate a phylogenetic tree of the gene family. Then, the nodes of this gene tree are labeled as duplication or speciation events by comparison to the species tree during the reconciliation step, allowing the prediction of orthology and paralogy relationships. This type of approach is implemented in numerous programs, including RIO (Zmasek and Eddy 2002), Orthostrapper (Storm and Sonnhammer 2002), PhylomeDB (Huerta-Cepas et al. 2007), Ensembl Compara (Vilella et al. 2009), PANTHER (Mi et al. 2010). These methods produce hierarchical ortholog groups, i.e., groups of orthologs and inparalogs deriving from a common ancestor, in the form of trees. These hierarchical groups are more informative than simple orthology relationships between pairs of species or flat groups of orthologs without evolutionary information about intra-group relations. Unfortunately, tree-based methods are highly dependent on the construction of correct multiple alignments and trees and are computationally demanding, preventing their application to very large datasets.

Although hierarchical groups are naturally produced by tree-based methods, they can also be generated by a post-processing of orthogroups obtained by graph-based methods. As an example, EggNog and OrthoDB explicitly delineate the hierarchy of ortholog groups by identifying orthogroups at different taxonomic levels of the species tree. Hybrid methods go further by using attributes of graph-based and tree-based methods in the inference of orthology relationships itself. The method of OMA Hierarchical Orthologous Groups (HOG) (Altenhoff et al. 2013) uses an orthology graph of pairwise relations to form groups, starting with the most specific taxonomic level and progressively merging groups toward the root of the species tree. Hieranoid (Schreiber and Sonnhammer 2013) progressively calculates pairwise orthology relationships using RBH at each node of a guide tree from the leaves to the ancestor. At each node, a consensus or a profile is built from the child nodes and used for subsequent pairwise comparisons, which considerably reduces the number of required pairwise comparisons. OrthoFinder 2 (Emms and Kelly 2019) first identifies orthogroups among a set of species using the OrthoFinder graph-based approach (Emms and Kelly 2015) and then uses the orthogroups to infer approximate gene trees and a species tree. Finally, each gene tree is compared to the species tree to infer duplication events and refine prediction of orthology and paralogy relations.

Meta-prediction methods are designed to exploit predictions generated by different programs and thus can potentially highlight false positives and negatives. As an example, DIOPT (Hu et al. 2011) assigns a score to each orthology relationship according to the number of independent methods predicting this relation. The MARIO program (Pereira et al. 2014) goes further by delineating a group of orthologs from predictions of several methods and constructing a hidden Markov model (HMM) profile of these orthologous sequences. This profile is then used

to evaluate the predictions made by each individual method. MetaPhOrs (Pryszcz et al. 2011) integrates phylogenetic trees constructed by several methods to predict orthology relations and assigns a score depending on the number of predictions. This filters unreliable results linked to poor resolution of phylogenetic trees. The WORMHOLE program (Sutphin et al. 2016) uses a classifier based on support vector machines (SVM) trained on a positive set of validated orthology relationships and a negative set of non-orthology gene pairs. The algorithm assigns a weight to each prediction method depending on its performance in different test cases (e.g., according to the proximity of the species under consideration). This weight is then used to combine predictions on a complete dataset and extract reliable orthology relations.

9.2.2 Standardization and Benchmarking

Given the multiplicity of orthology inference methods available, it is crucial to cross-reference, compare and evaluate their predictions in different biological contexts in order to choose the relevant program for a given biological question and to improve prediction methods. This requires a standardization of orthology prediction formats and an objective benchmarking. These topics are the central goals of the Quest For Orthologs (QFO) consortium (Gabaldón et al. 2009). QFO addresses both conceptual issues and technical challenges in orthology prediction. For example, community efforts led to the development of the standardized OrthoXML format (Schmitt et al. 2011) designed to represent orthology predictions for both graph- and tree-based methods. An ontology (Fernández-Breis et al. 2016) has also been developed to formalize the representation of orthology relationships. This ontology allows the representation of data according to a semantic Web standard, resource descriptions framework (RDF) that facilitates interoperability between resources.

The QFO consortium has also defined a QFO reference proteome dataset to allow the comparison of methods on a common set of species and proteins. The dataset is updated every year and currently comprises 78 UniProt Reference proteomes. It includes sequences from model organisms, species of interest for biomedical or agronomic research or species of interest from a phylogenetic point of view (Sonnhammer et al. 2014). In parallel, a variety of benchmarks have been developed to evaluate orthology prediction methods according to phylogenetic and functional criteria. A large-scale benchmarking study (Altenhoff et al. 2016) comparing 15 orthology methods highlighted a trade-off between sensitivity and specificity and clearly showed that the best approach is highly dependent on the biological context. Overall, the orthogroup predictions of OMA are characterized by high specificity, whereas the tree-based method used in PANTHER has high sensitivity. However, there is no systematic difference between tree-based and graph-based methods. Finally, Inparanoid, Hieranoid and OrthoInspector as well as OrthoFinder in the most recent version of the benchmark (results available at https://orthology.benchmarkser

vice.org) show a good balance between specificity and sensitivity over all benchmarks. Orthology predictions from the best methods identified by the benchmarking are now integrated in the Alliance of Genome Resources (Alliance) portal (Alliance of Genome Resources Consortium 2020). The Alliance aims to facilitate exploration of orthologous genes in human and well-studied model organisms in order to exploit the wealth of genetic and genomic studies available in these organisms.

9.2.3 Orthology Resources

Most orthology inference programs can be installed and executed locally on a user-defined set of proteomes, but many of them are also used to generate databases of orthology relationships. These resources are essential for the routine use of the orthology concept by non-experts. The databases differ in terms of number and diversity of represented species (Table 9.1), which determines the granularity with which orthology relationships can be exploited. Some generalist databases cover a large panel of species such as EggNog (Huerta-Cepas et al. 2016), HOGENOM (Penel et al. 2009), Inparanoid (Sonnhammer and Östlund 2015), MBGD (Uchiyama et al. 2019), OMA (Altenhoff et al. 2018), OrthoDb (Kriventseva et al. 2019) and OrthoInspector (Nevers et al. 2019). EggNog and OrthoDB also include viral genomes. Other resources are clade-specific, including TreeFam (for Metazoa) (Schreiber et al. 2014), FungiPath (for Fungi) (Grossetête et al. 2010), and GreenPhylDB (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018) that focus on plants. With the exception of MetaPhOrs (Pryszcz et al. 2011), the resources based on meta-predictions generally focus on a small number of model species (Table 9.1). In addition to the databases dedicated to orthology, orthology relationships are also provided in more general biological portals, such as PANTHER (Mi et al. 2019), Ensembl Compara (Herrero et al. 2016) and HomoloGene (NCBI Resource Coordinators 2016).

Orthology databases offer diverse access to information, via Web interfaces for manual exploration or using programmatic access through Web services or SPARQL (SPARQL Protocol and RDF Query Language) interfaces. Users can search for orthologs of a given gene using genes/proteins or orthogroup identifiers or perform a sequence similarity search. Information can also be accessed through functional annotation of the gene of interest (keywords, description or GO annotations). For instance, OrthoInspector (Nevers et al. 2019) allows users to retrieve all proteins of a given species associated with a given GO term and visualize their evolutionary histories. OrthoMCL (Chen et al. 2006) and GreenPhylDB (Rouard et al. 2011) propose searches for groups with a given protein domain. Genes can also be retrieved on the basis of their phylogenetic distribution, i.e., the presence or absence of an ortholog in different taxa. This phylogenetic profiling search is implemented in MBGD (Uchiyama et al. 2019), OrthoDb (Kriventseva et al. 2019), OrthoInspector (Nevers et al. 2019), OrthologeDB (Whiteside et al. 2013), OrthoMCL (Chen et al. 2006) and GreenPhylDB (Rouard et al. 2011). It can be used to perform genotype/phenotype studies as discussed in the applications section.

Table 9.1 Main orthology resources

Resource		Coverage					Exploration							Representation						
Type	Name	Genomes	Bacteria	Eukaryota	Archaea	Viruses	Gene Id	Group Id	Sequence	Function	Distribution	SPARQL	Webservice	Orthologues	Function	Domains	MSA	Tree	Synteny	Distribution
General	Inparanoid	273	/	/	/	0	✓	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
	OMA	2 327	1688	485	154	0	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓	✓
	EggNOG	2 031	1678	115	238	352	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✗	✓
	OrthoDb	7284	5609	1271	404	7963	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
	OrthoMCL	150	36	98	16	0	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✗	✗	✓
	Hieranoid	66	20	40	6	0	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗
	OrthoInspector	4753	3863	711	179	0	✓	✗	✓	✓	✓	✗	✓	✓	✓	✗	✓	✗	✗	✓
	MBGD	6318	5861	203	254	0	✓	✗	✓	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✓
	OtholugeDb	2069	/	0	/	0	✓	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗
	HOGENOM	13367	12326	593	224	0	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗
	PhylomeDb	1 862	/	/	/	0	✓	✗	✓	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗
Specific	TreeFam	109	0	109	0	0	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓	✓	✓	✗	✓
	FungiPath	165	0	165	0	0	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓	✗	✗
	Greenphyl	37	0	37	0	0	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✓
	PLAZA	119	0	119	0	0	✓	✗	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓
Meta-predictions	P-POD	12	1	11	0	0	✓	✗	✗	✓	✗	✗	✗	✓	✓	✗	✗	✓	✗	✓
	MetaPhOrs	2713	1 720	877	116	1	✓	✗	✓	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗
	WORMHOLE	6	0	6	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
	DIOPT	10	0	10	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
	YOGY	11	1	10	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓
	HCOP	19	0	19	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Other	Panther	142	35	99	8	0	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✓	✗	✗
	Ensembl	1191	123*	1068	/	0	✓	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓
	Homologene	21	0	21	0	0	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓

References: EggNog (Huerta-Cepas et al. 2016), HOGENOM (Penel et al. 2009), Inparanoid (Sonnhammer and Östlund 2015), MBGD (Uchiyama et al. 2019), OMA (Altenhoff et al. 2018), OrthoDb (Kriventseva et al. 2019), OrthoInspector (Nevers et al. 2019), OrthologueDB (Whiteside et al. 2013), OrthoMCL (Chen et al. 2006), PhylomeDB (Huerta-Cepas et al. 2014), TreeFam (Schreiber et al. 2014), FungiPath (Grossetête et al. 2010), GreenPhylDB (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018), PANTHER(Mi et al. 2019), Ensembl Compara (Herrero et al. 2016), HomoloGene (NCBI Resource Coordinators 2016)

*123 prokaryotic species (mainly Bacteria but also some Archaea) are included in the Pan-Compara resource which includes a selection of prokaryotic and eukaryotic species

All orthology databases provide orthology predictions in the form of a list of orthologs in the covered species, but many of them contextualize this minimum information by adding relevant data and tools to analyze and exploit the evolutionary information (Table 9.1). Hence, they frequently provide additional information about the function (GO term annotation, enzyme classification numbers…) or architecture (protein domains) of the predicted orthologs as illustrated in Table 9.1. This functional information most often comes from automatic annotations that must be handled with care. However, viewing the annotations for all the orthologs of a protein makes it easier to detect inconsistencies and spurious annotations. For example, OMA (Altenhoff et al. 2018) offers a synthetic representation of the GO annotations

of the detected orthologs with a color code that distinguishes between automatic annotation, annotation validated by an expert and annotation based on experimental data. Multiple sequence alignment (MSA) and phylogenetic trees also constitute an essential analytical tool for a more in-depth understanding of the relationships between orthologs and paralogs. As such, they are often made available, in particular by tree-based methods. They are either pre-calculated and available directly on the Web interface or can be constructed 'on the fly' for a selection of predicted orthologous sequences. In addition, some resources provide information about the genomic context of the query gene and its orthologs, allowing to detect syntenic stretches of genes that can be helpful for the validation of orthology relations and may be indicative of a functional link between syntenic genes. Finally, orthology resources can provide the taxonomic distribution of detected orthologs in each species represented in the orthology database. This is suitable for clade-specific resources such as GreenPhylDb (Rouard et al. 2011) and PLAZA (Van Bel et al. 2018). For generalist orthology resources, a synthetic view of distributions is required as exemplified by OrthoInspector (Nevers et al. 2019) that provides schematic representations of phylogenetic distributions at different granularity levels.

9.3 Orthology: The Swiss Army Knife of Genomics

9.3.1 Exploration of Gene and Protein Families

Since their definition in the early seventies, orthologs and paralogs have been traditionally used to study gene and protein families, in particular in the framework of multiple alignment analysis. By placing a gene or a protein sequence in its evolutionary context, the multiple alignment reveals selection pressure existing at particular sequence positions, allowing the straightforward detection of conserved motifs, localization signals or key functional residues for a considered family of orthologs or a superfamily regrouping several paralogous families (Lecompte et al. 2001). Such evolutionary analyses are essential for the determination of catalytic sites or residues involved in protein interactions for example. This can be exploited to decipher residues, motifs or domains involved in the specificity of paralogous families, for instance, to identify residues responsible for the enzyme substrate specificity in a multienzymatic family. In addition, alignments of orthologs or homologs are exploited in both 2D and 3D structure prediction methods by comparative protein modeling (reviewed in Khan et al. 2016). With the increase of experimentally determined structures, a wide range of accurate models are now available that can be used to predict protein binding sites, effects of protein mutations, and for structure-guided virtual screening (Liu et al. 2011; Leelananda and Lindert 2016).

Orthologous sequences are directly exploited by many mutation analysis tools, such as PolyPhen (Adzhubei et al. 2010) or SIFT (Vaser et al. 2016), to predict the phenotypic effects of variants. Pairwise or multiple alignments of orthologous

sequences are also used at the genomic level to highlight conserved regions that may reflect the existence of functional elements. Orthologs are also the cornerstone of phylogenetic studies aimed at deciphering the evolutionary history of a gene family or, more generally, phylogenetic relationships between species. The reconstruction of phylogenetic relationships between species has for a long time relied on a single family of genes, typically 16S/18S rRNA genes or well-conserved housekeeping protein genes. Today, species phylogenies can be built using comparisons of several protein families, including genome-wide comparisons (Crawford et al. 2012). These studies generally focus on widely conserved protein families exhibiting one-to-one orthology relationships. Orthofinder directly exploits orthogroups within a species set to construct a phylogenomics species tree using the species tree from all genes (STAG) algorithm (Emms and Kelly 2018). With the multiplication of available genomes and metagenomes, such phylogenomics analyses have renewed our vision of the tree of life, for instance, by highlighting the bacterial diversification (Hug et al. 2016), reshaping the eukaryotic tree (Burki et al. 2020) and revealing a new group of Archaea, the Asgard that questioned the position of Eukaryotes in the tree of life (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017).

Orthologous sequences and phylogenetic trees can also be exploited for ancestral sequence reconstruction. The leaves of the phylogenetic tree represent the extant sequences of the family, while the root corresponds to the extinct common ancestor. The ancestor can be synthesized to experimentally explore its biochemical properties. This approach allows to resurrect an ancestral precursor with selected properties, such as thermostability, in order to initiate synthetic evolution experiments (Gumulya and Gillam 2017). It can also be used to decipher past environmental conditions. For example, the reconstruction of translation elongation factors from organisms that lived 3.5 billion years ago revealed that the thermostability of these factors declines in the course of evolution and suggested a 30 °C decrease in environmental temperature (Gaucher et al. 2008). Ancestral sequence reconstruction methods also deduce the sequences present at each internal node of the tree. These intermediate states can help to elucidate evolutionary processes, in particular the main mutations involved in the distinct properties of extant proteins (Straub and Merkl 2019). Applied to whole genomes, ancestral reconstruction offers a partial view of ancestral gene repertoires, from the known repertoires of extant species. Such a resource is available on the ancestral genome portal, constructed from PANTHER inferences (Huang et al. 2019).

9.3.2 Information Transfer

As stated above, orthologous genes tend to retain equivalent functions across species and are thus widely used to transfer information from model species to poorly characterized ones. Typically, the functional annotation of genes in a newly sequenced genome is carried out by identifying annotated orthologs using similarity searches in protein databases such as UniProt (The UniProt Consortium 2019) or through the

Gene Ontology (The Gene Ontology Consortium 2019) and then transferring these annotations to genes of unknown function. Several protocols (compared in Amar et al. 2014) allow this automated annotation transfer. Although this approach is time efficient, it can also lead to bias since the orthology conjecture is not an absolute law and the ortholog/paralog distinction is not trivial, especially in superfamilies (Schnoes et al. 2009). The problem of misannotation is also particularly severe, with multidomain proteins exhibiting a differential conservation of some domains (discussed in Sect. 9.3.3 Beyond gene level orthology). In addition, automated transfer can propagate annotation errors. It is therefore wise to rely on closely related orthologs with expert-curated annotations, whenever possible, to avoid the 'percolation of annotation errors' modeled by Gilks and colleagues and its deleterious effects on database quality (Gilks et al. 2002).

More generally, orthology can be used to transfer experimentally evidenced information obtained from one species to another, provided that the organisms are sufficiently close. This principle is used by the Gene Ontology Consortium (Ashburner et al. 2000; The Gene Ontology Consortium 2019) to propagate standardized annotations not only on protein molecular function but also on their sub-cellular localization and the biological processes in which they are involved. The resulting annotations receive the IEA evidence code (Inferred from Electronic Annotation) in the case of an automatic transfer between orthologs. The Gene Ontology also integrates a semi-automated transfer protocol (Gaudet et al. 2011), taking into account annotations from several orthologs and the phylogenetic relationships between the corresponding species. These annotations are labeled with the IBA code (Inferred from Biological ancestry).

Information about protein–protein interactions (PPIs) can also be transferred from one species to another through the concept of interologs. The term 'interolog' (Walhout et al. 2000) refers to the conserved interaction between two pairs of proteins A1 and B1 from a first species and A2 and B2 from a second species. The A1/B1 interaction is considered as an interolog of the A2/B2 interaction if A1 and A2 are orthologs to B1 and B2, respectively. The concept of interology can be exploited in a predictive way: Orthologs of interacting proteins in one species are identified, and the PPI information is transferred to the pair of orthologs. To avoid false positive errors, interology inferences are usually combined with other data, as illustrated by the STRING interaction database (Szklarczyk et al. 2019) that relies on a large panel of diverse evidence (experiments, text mining, co-expression, synteny, etc.).

Finally, when working on human genes, orthology relationships are key elements to consider when choosing a relevant model species for experimental studies. In addition to practical considerations (duration, cost, etc.), the model species should be chosen to avoid 1-to-many or many-to-1 orthology relations between the human and the model species, since the existence of additional inparalogs in one species would considerably complicate the interpretation of experimental results.

9.3.3 Comparison of Genomes and Proteomes

Comparisons of complete genomes and proteomes are intrinsically linked to the proper delineation of orthologs and paralogs. Comparisons of orthologs at the sequence level are used to evaluate the selection pressure acting to model evolutionary rates in different species. One of the first examples of such genome-wide assessment of evolution rates was carried out in mammalian and nematode lineages (Castillo-Davis et al. 2004). This study showed that strong purifying selection seems to act on the same central cellular processes (such as translation and protein transport) in mammals and nematodes, whereas positive selection acts on different biological processes in each lineage (DNA-dependent transcriptional regulation in nematodes, signal transduction via receptors and host immune response in mammals). Such comparative analyses are also performed for non-coding RNA genes such as microRNA. For example, the study of microRNA substitution rates in human and chimpanzee genomes revealed that primate-specific microRNAs have twice as many substitutions as older microRNA families (Santpere et al. 2016).

Comparison of proteomes in terms of gene content has become a quasi-obligatory step when sequencing a new genome. It requires the prediction of orthology and paralogy relations between the proteomes under consideration and reveals the set of conserved protein families but also the acquisitions and losses that have taken place independently in each lineage. These comparisons have highlighted the extraordinary plasticity of the gene repertoire among species. This is particularly striking in the case of prokaryotic genomes. In a comparison of more than 500 bacterial species, Lapierre and Gogarten (2009) showed that the conserved bacterial core was reduced to about 250 gene families, with the notable exception of certain symbionts exhibiting a particularly reduced genome. This diversity of gene repertoire observed even among closely related species can be explained by lineage-specific expansion of gene families, acquisition of genes by horizontal transfer (xenologs) and differential gene losses. In some prokaryotes, the genomic versatility is so important that large differences in gene content can occur between different strains of the same species. This led to the definition of the pangenome concept, i.e., the set of all genes present in a given species, that can be divided into the conserved core and the accessory genome (reviewed in Brockhurst et al. 2019). In species with an 'open' pangenome, the core genome conserved in all strains represents only a small fraction of the pangenome, questioning the concept of species in Prokaryotes. For instance in *Escherichia coli*, the core genome is restricted to about 3000 gene families, while the pangenome reaches a total of about 90,000 families (Land et al. 2015).

Comparisons of orthologous genomic regions or complete chromosomes decipher the evolution of genome architecture by revealing differential gains/losses of genomic regions, segmental duplications and balanced rearrangements. These comparisons can be made at the nucleotide level using, for example, BLASTZ (Schwartz et al. 2003) or LASTZ and chaining/netting programs (Kent et al. 2003) to discriminate between orthologous and paralogous alignments. Alternatively, the comparison of genomic regions can be based on the comparison of genomic location of orthologs in

different genomes to identify conserved syntenic blocks, i.e., a stretch of genes with a conserved gene order in different species. Such comparisons delineate syntenic genes frequently linked by functional relations and allow the detection of elements involved in genomic plasticity at the syntenic regions boundaries. They are also used to reconstruct ancestral genomes with distance/event-based or homology/adjacency-based methods (reviewed in Feng et al. 2017).

9.3.4 Functional Inferences and Genotype/Phenotype Correlations

Comparisons of complete proteomes based on orthology relationships can be exploited to perform functional inferences between genes or to detect genes potentially involved in a phenotype. The rationale behind this approach is that functionally linked genes are preserved or lost in a correlated manner over the course of evolution and thus are found in the same species (Pellegrini et al. 1999). This assumption can be exploited in different ways. Subtractive analysis aims to identify genes restricted to species with a given phenotype. In practice, this means comparing the gene repertoire of at least two species (species A and B) possessing the phenotypic trait of interest and one or several related species (species C) lacking the considered phenotype. The set of genes with orthologs in species A and B but without orthologs in species C is likely to be enriched in genes associated with the phenotypic trait of interest. This approach was introduced by Huynen (Huynen et al. 1998) in the early days of comparative genomics in order to compare the genome of the pathogen *Helicobacter pylori* with that of another pathogen *Haemophilus influenzae* and a benign strain of *E. coli*. They identified 17 gene families restricted to the pathogenic species and potentially involved in virulence and host–pathogen interactions.

The subtractive method is applicable to the search for genes linked to a phenotypic trait or biological process that has been lost/acquired in some species during evolution. This approach can be extended to the comparison of tens or hundreds of genomes to allow a precise definition of the phenotypic distribution. The comparison of phylogenetically distinct lineages that have independently acquired (or lost) a given phenotype limits false positive predictions by eliminating genome differences simply due to random gains and losses of genes. For instance, Hecker and colleagues (Hecker et al. 2019) compared mammalian genomes to identify convergent gene losses associated with dietary adaptations in six independent herbivore lineages (16 species) and five independent carnivore lineages (15 species). Regarding the small evolutionary distances separating these placental mammals, they considered not only loss of entire genes or exons but also gene-inactivating mutations, using a genomic approach that combines the identification of orthologous regions and the CESAR program, a coding exon-structure aware realigner (Sharma et al. 2016).

At a larger evolutionary scale, another methodological framework is required. Phylogenetic profiles represent a generalization of subtractive analysis allowing the comparison of a large number of genomes that can be evolutionary distant. A phylogenetic profile of a gene represents the presence or absence of orthologs of that gene in the genomes of several species (Tatusov et al. 1997). Phylogenetic profiles were first used to infer the function of uncharacterized genes, and the method has been successfully applied to the annotation of genes, mainly prokaryotes (see Kensche et al. 2008 for examples). They are also exploited to predict functional links between genes, notably in the STRING (Szklarczyk et al. 2019) and OrthoInspector databases (Nevers et al. 2019).

Phylogenetic profiles can not only be compared to each other but also to all types of presence–absence distributions, including phenotypic traits. Phylogenetic profiling can thus be exploited to perform phenotype-genotype association studies. One of the first studies of this type was carried out on 86 prokaryotic genomes to identify genes associated with thermophily (Jim et al. 2004). Since then, many similar studies have been performed, notably to identify genes involved in human diseases thanks to the huge increase of available eukaryotic genomes that allows a detailed exploration of the distribution of human genes. For instance, Tabach et al. (2013) identified 54 clusters of phylogenetic profiles associated with a specific class of symptoms. More recently, the profiling of human genes in 100 eukaryotic species revealed 274 human genes exhibiting a phylogenetic distribution correlated with the distribution of cilia in eukaryotic lineages (Nevers et al. 2017). This set of predicted ciliary genes includes 87 new candidates. Among them, 21 have already been experimentally validated as ciliary genes.

9.4 Challenges

9.4.1 Keeping Up with the Data Flow

As seen above, orthology is the cornerstone of a plethora of applications in comparative genomics and biology, and orthology resources provide numerous contextual data and analytical tools to facilitate orthology exploitation. Coming into a new decade, they are now gearing up to adapt to new challenges, a data flow brought by the next generation sequencing and a need to assess orthology at different granularity levels. The last two decades have seen a massive increase in sequencing capacities, leading to the acquisition of numerous genomes from across the tree of life. These genomes have obvious usefulness for studying evolution at a broad scale and are increasingly incorporated into orthology resources. Nevertheless, they also lead to important challenges linked to the management and analysis of the ever-increasing volume of data and the heterogeneous data quality.

Genomic data, hence genome annotations, have been increasing at an exponential rate with the advent of high-throughput sequencing technologies. As of today, 19,163 complete genomes are registered in the Genome Online Database (Mukherjee et al. 2019), as well as 215,613 genomes in the permanent draft state. This increase in data generation represents a challenge for orthology resources. It is especially true for tree-based approaches, which are commonly more computationally intensive as they rely on phylogenetic tree inference tools and are traditionally limited in the number of species they can include. While less computationally intensive, the data increase is still onerous for graph-based approaches, as they rely on all-vs-all sequence comparisons, which grow quadratically with the number of sequences. The legacy tools for these kinds of comparison, namely BLAST (Altschul et al. 1990; Camacho et al. 2009) or Smith-Waterman (Smith and Waterman 1981) alignment, do not scale well, and resources that use them rely heavily on high-performance computing clusters. Other tools and resources use faster but generally less sensitive solutions: MMSeq2 (standard modes) (Steinegger and Söding 2017), DIAMOND (Buchfink et al. 2015) or *ad-hoc* methods as in SwiftOrtho (Hu and Friedberg 2019) for instance can perform all-vs-all comparisons with better performances.

Nonetheless, solutions bypassing computationally intensive all-vs-all computations are increasingly being investigated, in anticipation of an even bigger surge in data. These approaches such as EggNog-Mapper (Huerta-Cepas et al. 2017) aim to reduce the computation required to adding new proteomes by exploiting already precomputed ortholog groups that are assumed to be stable over time. Their goal is to use fast methods, e.g., hidden Markov models (Eddy 2011) or k-mer based sequence similarity searches, to identify likely existing orthologous groups in which each sequence fits. While fast, these methods rely on existing databases with sufficient clade coverage to be efficient.

Another aspect of data management, linked to computational time, is the size of databases produced. Storing a high number of orthologous relations or orthologous size implies storing Terabytes of data and induces longer access times to the data. Consequently, it is not necessarily optimal for orthology resources to include all available genomes, and a choice is often made concerning which data to select, with high variability of species chosen in each orthology resource. This is reflected by the number of species available in different resources and variable representation in terms of clades or domains of life. Notably, some resources specialize in specific clades such as Plaza (Van Bel et al. 2018) for plants or FungiPath for fungi (Grossetête et al. 2010). Even among the databases with a large number of species, a wide diversity of species is preferred rather than sheer number, as diversity is generally more important than number in comparative studies (Škunca and Dessimoz 2015). This can be achieved by limiting additional species to new taxa of interest or by limiting inter-clade computations to fewer species (Nevers et al. 2019) with several levels of taxonomic resolution. The decision to add or keep a species in an existing database is a product of multiple factors but may be informed by indicators of how the addition of one species affects the diversity. For example, the rarefaction curve proposed by the KinFin analysis tool (Laetsch and Blaxter 2017) (compatible with some orthology inference software suites) provides an objective measure of the novelties in terms

of orthogroups added by each included species. Favoring diversity is also beneficial for the fast-placing strategies mentioned above, moving toward resources with a limited number of species computed directly with the all-versus-all strategy, and other species added to existing groups using less computationally intensive strategies.

9.4.2 Addressing Proteome Quality

Another aspect of high-throughput data is the associated data quality issues, and the genomic data used in comparative genomics studies are no exception. Proteomes, i.e., the genome annotations of protein coding genes from genomic data result from a multi-step process ranging from genome sequencing to the actual annotation of the final assembled sequences, with multiple possible sources of error. Consequently, a proteome may have missing proteins (either being permanent draft or a misannotated complete genome) or contain proteins that are either fragmented or actually erroneous. All of these cases may in turn induce errors in orthology inference that rely heavily on sequence comparison and in comparative genomics approaches that assume data completeness.

Missing proteins, for instance, lead to missing orthology relationships between species with incomplete genomes and other species. Most orthology pipelines assume data completeness when inferring orthology, and while they are in principle robust to gene losses, incomplete gene sets may lead to errors in orthology inference and in orthogroup reconstruction. Some methods, e.g., Hamstr (Ebersberger et al. 2009) and OrthoGraph (Petersen et al. 2017) are designed to avoid this assumption by first excluding incomplete datasets (e.g., issued from RNA-seq data) during orthogroup construction. Sequences from the incomplete datasets are then mapped to the precomputed robust orthogroups. Even with correct orthology inference, incomplete genomes impact the phylogenetic placement of species, as fewer marker genes are available. This is particularly detrimental when relations between species are hard to resolve. More spectacularly, artificially missing proteins constitute a significant source of errors for comparative genomics methods relying on comparison of entire species gene repertoires, e.g., phylogenetic profiling.

Fragmented proteins are another matter and initially have an impact on orthology prediction via sequence similarity comparisons. For example, if a fragmented protein sequence corresponds to a single domain, reciprocal best hit methods may infer a false positive pairwise relation with a protein in another species having a homologous domain, although the full-length protein would not be identified as orthologous. Conversely, if the protein fragment corresponds to a low complexity, repeat-containing or divergent region, similarity based orthology prediction methods will miss it, leading to false negatives and in the worst case, may even be responsible for spurious relations (false positives). It is worth noting that issues caused by this kind of region, amplified in the presence of fragments, constitute a general limit of similarity-based orthology inference methods in any organism.

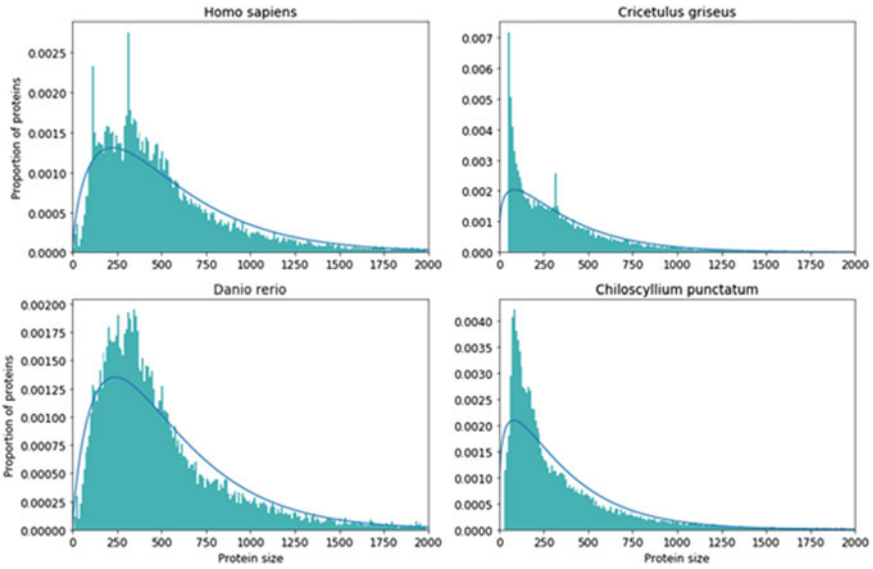

Fig. 9.2 Protein length distribution in four proteomes, from various vertebrate clades. On the left are examples of the distribution observed in well-studied species (*Homo sapiens* and *Danio rerio*), similar to the one observed in most proteomes. On the left, examples of atypical distributions with high number of small proteins for the rodent *Cricetulus griseus* and the chondrichthyes *Chiloscyllium punctatum*

A stark difference in proteome data quality is revealed by analysis of the distribution of protein length between publicly available proteomes. For example, Fig. 9.2 shows the protein length distribution, normalized for proteome size, in four vertebrate species. Most proteomes share a distribution centered on a peak in the range of 200–400 amino acids and a decreasing number of long proteins, as illustrated by *Homo sapiens* and *Danio rerio* (Fig. 9.2). In contrast, some proteomes present a peak for small proteins (less than 100 amino acid long), as exemplified by the other proteomes presented on Fig. 9.2. Strikingly, all manually curated proteomes of model species have the former distribution, and both distributions are distributed across the species tree, ruling out biological exceptions (Nevers et al. in prep). Instead, it indicates a high number of truncated or erroneous proteins.

One must thus be cautious when providing annotations of genomic data to public databases or using these data for orthology inference and comparative genomics. Quality measures exist to indicate the quality of genome assembly, N50 being a standard indicator of genome contiguity that is commonly provided with published genome assemblies. However, genome assembly quality does not necessarily correlate with proteome annotation quality. State-of-the-art tools exist that provide an indication of data completeness and fragmentation. For instance, CEGMA (Parra et al. 2007, 2009) and its successor, BUSCO (Waterhouse et al. 2018) make use of known conserved gene families, so-called core orthologs, in single-copy in most species for

the latter, to assess the completeness of the gene annotation for a given genome. The assumption being that the proportion of core orthologs found in a genome reflects the completeness of the gene annotation as a whole. BUSCO provides additional information about the state of the proteome, by indicating which proportion of core orthologs are found only in a fragmented state. Assessing BUSCO completeness is standard practice when publishing new genomes, and this information is now available in UniProt (The UniProt Consortium 2019) for most available proteomes.

However, empirical data show that BUSCO completeness assessment is not always correlated to the standard protein length distribution, suggesting that it does not capture all cases of genome misannotation. A better proxy of this bias can be obtained in the form of summary statistics, such as the proportion of extremely short proteins in the genome or the number of proteins annotated as not starting with a methionine (i.e., annotated genes for which no start codon was found by the annotation pipeline). These summary statistics can be used to filter genomes used in orthology analysis (Nevers et al. 2019), by setting thresholds under which proteomes are considered as not annotated. As these parameters are nearly orthogonal to core ortholog completeness, they can be used in parallel with methods like BUSCO and CEGMA to identify low quality proteomes. Despite these developments, work is still needed to further assess proteome quality and its impact on downstream applications, and this issue is an important target for future community efforts.

9.4.3 Beyond Gene-Level Orthology

While most orthology prediction methods are based on full-length gene or protein sequences, in certain cases, functional domains might be a more pertinent entity to consider. Indeed, the majority of known proteins consist of multiple domains, especially in the eukaryotic lineages, and it is known that multi-domain architectures tend to evolve over time as a result of different mechanisms, such as domain gains, losses and duplications, or gene fusion and fission (Buljan and Bateman 2009). The latter in particular can result in complex evolutionary histories for genes with domains of very different ancestral origins, which in turn makes orthology relations more complicated. In addition, domain architecture rearrangements have been observed several times between orthologs of species belonging to different phyla, possibly as a consequence of different organism complexity (Koonin et al. 2000, 2004). However, studies have shown that domain rearrangements can occur between relatively close species, such as mammals or members of the *Drosophila* genus, and it has been estimated that they could concern up to 50% of proteins (Forslund et al. 2011; Wu et al. 2012; Sonnhammer et al. 2014).

Divergences of domain content and/or order between orthologs can be challenging for traditional orthology inference methods. In some cases, parts of the protein sequence might be too highly divergent in some species to be properly detected as orthologs. In other cases, one protein might have significant similarity to multiple different protein families, each due to a different domain of the query protein, making

it hard to clearly establish orthologous relations. This shows a clear limitation of full-length analyses, as they ignore the natural tendency of proteins to be modular and to evolve not at the complete sequence level, but at the domain level. It would be beneficial to focus future improvements and developments on domain-aware orthology inference as a complement to full-length methods, in order to predict more precise ortholog relations and better understand architectural rearrangements in protein evolution. While it has been widely acknowledged that such methods are needed (Sjolander et al. 2011), very few currently take domains into account. Exceptions include the microbial genome database MBGD, which constructs ortholog groups at the domain level (Uchiyama et al. 2019), and Domainoid (Persson et al. 2019), a tool that uses Pfam (El-Gebali et al. 2019) defined domains to infer orthology relations at the single level domain. Domainoid has been shown to retrieve orthologs not detected by classical full-length approaches, thus showing the interest of combining both types of strategies.

Another hassle of focusing on gene-level orthology is that, in Eukaryotes, a single gene may be transcribed into several isoforms with different exons combinations. This process, called alternative splicing, is especially prominent in vertebrates (Keren et al. 2010). Its functional implication is debated, but it has been shown for particular genes that different isoforms may have different tissue expression and even sometimes produce proteins with antagonist cellular functions (Wang et al. 2008). This has direct implications on the way orthology is used to transfer function between genes, as two orthologous genes could display different splicing patterns and even two orthologous genes with orthologous exons may have substantially different transcripts. Integrating homology between alternative transcripts of orthologs will provide additional information on whether an evolutionary conserved isoform is more likely to be functional, and whether observations made in a model species on a particular isoform are likely to be applicable to other species.

Assessing orthology between alternative transcripts often relies on two conditions (Blanquart et al. 2016). Indeed, transcripts are orthologous if (1) they are transcripts of orthologous genes and (2) their exons are similar enough to assume they are orthologous and appear in the same order in the gene sequence. The first condition is a classical orthology inference problem. The second condition may be determined by spliced sequence alignment, using an exon-aware alignment method (Kapustin et al. 2008; Gotoh 2008; Sharma et al. 2016; Jammali et al. 2019). Transcript orthology prediction has been successfully employed to identify orthologous isoforms between the gene repertoires of mouse and human (Zambelli et al. 2010). Applying it to more species is trickier since it cannot be done with pairwise relations and requires the construction of gene trees, which is computationally demanding. Nonetheless, it has been used to study multiple gene families, mapping events of isoform gains and losses to the branches of the trees (Christinat and Moret 2012; Jammali et al. 2019). Nevertheless, one must still be cautious when using isoform orthology determination and ensure that expression of both isoforms can be detected through experimental means in the species of interest, to avoid the pitfalls of erroneous annotation transfer.

As can be seen, despite the major advances made in recent years in orthology inference and resources, there is still a long way to go in the quest for orthologs. The practical and conceptual challenges are numerous and will require the efforts of the entire comparative genomics community to invent new solutions. Substantial progress will be needed both in the development of new indicators of proteome quality and for the formal representation of orthology relationships at different granularity levels.

Acknowledgements The authors thank Julie Thompson for critical reading of the manuscript. The authors are also grateful to the anonymous referees for their useful suggestions.

References

Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249. https://doi.org/10.1038/nmeth0410-248

Alliance of Genome Resources Consortium (2020) Alliance of genome resources portal: unified model organism research platform. Nucleic Acids Res 48:D650–D658. https://doi.org/10.1093/nar/gkz813

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput Biol 8:e1002514. https://doi.org/10.1371/journal.pcbi.1002514

Altenhoff AM, Gil M, Gonnet GH, Dessimoz C (2013) Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS ONE 8:e53786. https://doi.org/10.1371/journal.pone.0053786

Altenhoff AM, Boeckmann B, Capella-Gutierrez S et al (2016) Standardized benchmarking in the quest for orthologs. Nat Methods 13:425–430. https://doi.org/10.1038/nmeth.3830

Altenhoff AM, Glover NM, Train C-M et al (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res 46:D477–D485. https://doi.org/10.1093/nar/gkx1019

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

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul SF, Madden TL, Schäffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amar D, Frades I, Danek A et al (2014) Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. BMC Plant Biol 14:329. https://doi.org/10.1186/s12870-014-0329-9

Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29. https://doi.org/10.1038/75556

Blanquart S, Varré J-S, Guertin P et al (2016) Assisted transcriptome reconstruction and splicing orthology. BMC Genomics 17:786. https://doi.org/10.1186/s12864-016-3103-6

Brockhurst MA, Harrison E, Hall JPJ et al (2019) The ecology and evolution of pangenomes. Curr Biol CB 29:R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176

Buljan M, Bateman A (2009) The evolution of protein domain families. Biochem Soc Trans 37:751–755. https://doi.org/10.1042/BST0370751
Burki F, Roger AJ, Brown MW, Simpson AGB (2020) The new tree of eukaryotes. Trends Ecol Evol 35:43–55. https://doi.org/10.1016/j.tree.2019.08.008
Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. BMC Bioinform 10:421. https://doi.org/10.1186/1471-2105-10-421
Castillo-Davis CI, Kondrashov FA, Hartl DL, Kulathinal RJ (2004) The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. Genome Res 14:802–811. https://doi.org/10.1101/gr.2195604
Chen X, Zhang J (2012) The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol 8:e1002784. https://doi.org/10.1371/journal.pcbi.1002784
Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34:D363–D368. https://doi.org/10.1093/nar/gkj123
Christinat Y, Moret BME (2012) Inferring transcript phylogenies. BMC Bioinform 13(Suppl 9):S1. https://doi.org/10.1186/1471-2105-13-s9-s1
Crawford NG, Faircloth BC, McCormack JE et al (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol Lett 8:783–786. https://doi.org/10.1098/rsbl.2012.0331
Darby CA, Stolzer M, Ropp PJ et al (2017) Xenolog classification. Bioinformatics 33:640–649. https://doi.org/10.1093/bioinformatics/btw686
Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol Biol 9:157. https://doi.org/10.1186/1471-2148-9-157
Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195
El-Gebali S, Mistry J, Bateman A et al (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995
Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. https://doi.org/10.1186/s13059-015-0721-2
Emms DM, Kelly S (2018) STAG: species tree inference from all genes. bioRxiv 267914. https://doi.org/10.1101/267914
Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238. https://doi.org/10.1186/s13059-019-1832-y
Feng B, Zhou L, Tang J (2017) Ancestral genome reconstruction on whole genome level. Curr Genomics 18:306–315. https://doi.org/10.2174/1389202918666170307120943
Fernández-Breis JT, Chiba H, Legaz-García MDC, Uchiyama I (2016) The orthology ontology: development and applications. J Biomed Semant 7:34. https://doi.org/10.1186/s13326-016-0077-x
Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113
Fitch WM (2000) Homology a personal view on some of the problems. Trends Genet TIG 16:227–231. https://doi.org/10.1016/s0168-9525(00)02005-9
Force A, Lynch M, Pickett FB et al (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545
Forslund K, Pekkari I, Sonnhammer ELL (2011) Domain architecture conservation in orthologs. BMC Bioinform 12:326. https://doi.org/10.1186/1471-2105-12-326
Gabaldón T, Dessimoz C, Huxley-Jones J et al (2009) Joining forces in the quest for orthologs. Genome Biol 10:403. https://doi.org/10.1186/gb-2009-10-9-403
Gaucher EA, Govindarajan S, Ganesh OK (2008) Palaeotemperature trend for precambrian life inferred from resurrected proteins. Nature 451:704–707. https://doi.org/10.1038/nature06510
Gaudet P, Livstone MS, Lewis SE, Thomas PD (2011) Phylogenetic-based propagation of functional annotations within the gene ontology consortium. Brief Bioinform 12:449–462

Gilks WR, Audit B, De Angelis D et al (2002) Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics 18:1641–1649. https://doi.org/10.1093/bioinformatics/18.12.1641

Glover NM, Redestig H, Dessimoz C (2016) Homoeologs: what are they and how do we infer them? Trends Plant Sci 21:609–621. https://doi.org/10.1016/j.tplants.2016.02.005

Gotoh O (2008) Direct mapping and alignment of protein sequences onto genomic sequence. Bioinformatics 24:2438–2444. https://doi.org/10.1093/bioinformatics/btn460

Gray GS, Fitch WM (1983) Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. Mol Biol Evol 1:57–66. https://doi.org/10.1093/oxfordjournals.molbev.a040298

Grossetête S, Labedan B, Lespinet O (2010) FUNGIpath: a tool to assess fungal metabolic pathways predicted by orthology. BMC Genomics 11:81. https://doi.org/10.1186/1471-2164-11-81

Gumulya Y, Gillam EMJ (2017) Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the "retro" approach to protein engineering. Biochem J 474:1–19. https://doi.org/10.1042/BCJ20160507

Hecker N, Sharma V, Hiller M (2019) Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. Proc Natl Acad Sci 116:3036–3041. https://doi.org/10.1073/pnas.1818504116

Henricson A, Forslund K, Sonnhammer ELL (2010) Orthology confers intron position conservation. BMC Genomics 11:412. https://doi.org/10.1186/1471-2164-11-412

Herrero J, Muffato M, Beal K et al (2016) Ensembl comparative genomics resources. Database J Biol Databases Curation. https://doi.org/10.1093/database/baw053

Hu X, Friedberg I (2019) SwiftOrtho: a fast, memory-efficient, multiple genome orthology classifier. GigaScience 8. https://doi.org/10.1093/gigascience/giz118

Hu Y, Flockhart I, Vinayagam A et al (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinform 12:357. https://doi.org/10.1186/1471-2105-12-357

Huang X, Albou L-P, Mushayahama T et al (2019) Ancestral genomes: a resource for reconstructed ancestral genes and genomes across the tree of life. Nucleic Acids Res 47:D271–D279. https://doi.org/10.1093/nar/gky1009

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T (2007) The human phylome. Genome Biol 8:R109. https://doi.org/10.1186/gb-2007-8-6-r109

Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP et al (2014) PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res 42:D897–D902. https://doi.org/10.1093/nar/gkt1177

Huerta-Cepas J, Szklarczyk D, Forslund K et al (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. https://doi.org/10.1093/nar/gkv1248

Huerta-Cepas J, Forslund K, Coelho LP et al (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol Biol Evol 34:2115–2122. https://doi.org/10.1093/molbev/msx148

Hug LA, Baker BJ, Anantharaman K et al (2016) A new view of the tree of life. Nat Microbiol 1:16048. https://doi.org/10.1038/nmicrobiol.2016.48

Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. FEBS Lett 426:1–5. https://doi.org/10.1016/s0014-5793(98)00276-2

Jammali S, Aguilar J-D, Kuitche E, Ouangraoua A (2019) SplicedFamAlign: CDS-to-gene spliced alignment and identification of transcript orthology groups. BMC Bioinform 20:133. https://doi.org/10.1186/s12859-019-2647-2

Jensen LJ, Julien P, Kuhn M et al (2008) eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res 36:D250–D254. https://doi.org/10.1093/nar/gkm796

Jim K, Parmar K, Singh M, Tavazoie S (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. Genome Res 14:109–115. https://doi.org/10.1101/gr.1586704

Kapustin Y, Souvorov A, Tatusova T, Lipman D (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct 3:20. https://doi.org/10.1186/1745-6150-3-20

Kensche PR, van Noort V, Dutilh BE, Huynen MA (2008) Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J Roy Soc Interface 5:151–170. https://doi.org/10.1098/rsif.2007.1047

Kent WJ, Baertsch R, Hinrichs A et al (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A 100:11484–11489. https://doi.org/10.1073/pnas.1932072100

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11:345–355. https://doi.org/10.1038/nrg2776

Khan FI, Wei D-Q, Gu K-R et al (2016) Current updates on computer aided protein modeling and designing. Int J Biol Macromol 85:48–62. https://doi.org/10.1016/j.ijbiomac.2015.12.072

Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. Cell 101:573–576. https://doi.org/10.1016/S0092-8674(00)80867-3

Koonin EV, Fedorova ND, Jackson JD et al (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7

Kriventseva EV, Kuznetsov D, Tegenfeldt F et al (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 47:D807–D811. https://doi.org/10.1093/nar/gky1053

Kryuchkova-Mostacci N, Robinson-Rechavi M (2015) Tissue-specific evolution of protein coding genes in human and mouse. PLoS ONE 10:e0131673. https://doi.org/10.1371/journal.pone.0131673

Laetsch DR, Blaxter ML (2017) KinFin: software for taxon-aware analysis of clustered protein sequences. G3 Bethesda Md 7:3349–3357. https://doi.org/10.1534/g3.117.300233

Land M, Hauser L, Jun S-R et al (2015) Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics 15:141–161. https://doi.org/10.1007/s10142-015-0433-4

Lapierre P, Gogarten JP (2009) Estimating the size of the bacterial pan-genome. Trends Genet TIG 25:107–110. https://doi.org/10.1016/j.tig.2008.12.004

Lecompte O, Thompson JD, Plewniak F et al (2001) Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene 270:17–30. https://doi.org/10.1016/s0378-1119(01)00461-9

Leelananda SP, Lindert S (2016) Computational methods in drug discovery. Beilstein J Org Chem 12:2694–2718. https://doi.org/10.3762/bjoc.12.267

Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189. https://doi.org/10.1101/gr.1224503

Linard B, Thompson JD, Poch O, Lecompte O (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinform 12:11. https://doi.org/10.1186/1471-2105-12-11

Liu T, Tang GW, Capriotti E (2011) Comparative modeling: the state of the art and protein drug target structure prediction. Comb Chem High Throughput Screen 14:532–547. https://doi.org/10.2174/138620711795767811

Mi H, Dong Q, Muruganujan A et al (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the gene ontology consortium. Nucleic Acids Res 38:D204–D210. https://doi.org/10.1093/nar/gkp1019

Mi H, Muruganujan A, Ebert D et al (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res 47:D419–D426. https://doi.org/10.1093/nar/gky1038

Mukherjee S, Stamatis D, Bertsch J et al (2019) Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res 47:D649–D659. https://doi.org/10.1093/nar/gky977

NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44:D7–D19. https://doi.org/10.1093/nar/gkv1290
Nehrt NL, Clark WT, Radivojac P, Hahn MW (2011) Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol 7:e1002073. https://doi.org/10.1371/journal.pcbi.1002073
Nevers Y, Prasad MK, Poidevin L et al (2017) Insights into ciliary genes and evolution from multi-level phylogenetic profiling. Mol Biol Evol 34:2016–2034. https://doi.org/10.1093/molbev/msx146
Nevers Y, Kress A, Defosset A et al (2019) OrthoInspector 3.0: open portal for comparative genomics. Nucleic Acids Res 47:D411–D418. https://doi.org/10.1093/nar/gky1068
Overbeek R, Fonstein M, D'Souza M et al (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96:2896–2901. https://doi.org/10.1073/pnas.96.6.2896
Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067. https://doi.org/10.1093/bioinformatics/btm071
Parra G, Bradnam K, Ning Z et al (2009) Assessing the gene space in draft genomes. Nucleic Acids Res 37:289–297. https://doi.org/10.1093/nar/gkn916
Pellegrini M, Marcotte EM, Thompson MJ et al (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96:4285–4288. https://doi.org/10.1073/pnas.96.8.4285
Penel S, Arigon A-M, Dufayard J-F et al (2009) Databases of homologous gene families for comparative genomics. BMC Bioinform 10(Suppl 6):S3. https://doi.org/10.1186/1471-2105-10-S6-S3
Pereira C, Denise A, Lespinet O (2014) A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics 15(Suppl 6):S16. https://doi.org/10.1186/1471-2164-15-S6-S16
Persson E, Kaduk M, Forslund SK, Sonnhammer ELL (2019) Domainoid: domain-oriented orthology inference. BMC Bioinform 20:523. https://doi.org/10.1186/s12859-019-3137-2
Peterson ME, Chen F, Saven JG et al (2009) Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci Publ Protein Soc 18:1306–1315. https://doi.org/10.1002/pro.143
Petersen M, Meusemann K, Donath A et al (2017) Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinform 18:111. https://doi.org/10.1186/s12859-017-1529-8
Pryszcz LP, Huerta-Cepas J, Gabaldón T (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res 39:e32. https://doi.org/10.1093/nar/gkq953
Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol 314:1041–1052. https://doi.org/10.1006/jmbi.2000.5197
Roth ACJ, Gonnet GH, Dessimoz C (2008) Algorithm of OMA for large-scale orthology inference. BMC Bioinform 9:518. https://doi.org/10.1186/1471-2105-9-518
Rouard M, Guignon V, Aluome C et al (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res 39:D1095–D1102. https://doi.org/10.1093/nar/gkq811
Santpere G, Lopez-Valenzuela M, Petit-Marty N et al (2016) Differences in molecular evolutionary rates among microRNAs in the human and chimpanzee genomes. BMC Genomics 17:528. https://doi.org/10.1186/s12864-016-2863-3
Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. Brief Bioinform 12:485–488. https://doi.org/10.1093/bib/bbr025
Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 5:e1000605. https://doi.org/10.1371/journal.pcbi.1000605

Schreiber F, Sonnhammer ELL (2013) Hieranoid: hierarchical orthology inference. J Mol Biol 425:2072–2081. https://doi.org/10.1016/j.jmb.2013.02.018
Schreiber F, Patricio M, Muffato M et al (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res 42:D922–D925. https://doi.org/10.1093/nar/gkt1055
Schwartz S, Kent WJ, Smit A et al (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103–107. https://doi.org/10.1101/gr.809403
Sharma V, Elghafari A, Hiller M (2016) Coding exon-structure aware realigner (CESAR) utilizes genome alignments for accurate comparative gene annotation. Nucleic Acids Res 44:e103. https://doi.org/10.1093/nar/gkw210
Sjolander K, Datta RS, Shen Y, Shoffner GM (2011) Ortholog identification in the presence of domain architecture rearrangement. Brief Bioinform 12:413–422. https://doi.org/10.1093/bib/bbr036
Škunca N, Dessimoz C (2015) Phylogenetic profiling: how much input data is enough? PLoS ONE 10:e0114701. https://doi.org/10.1371/journal.pone.0114701
Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197. https://doi.org/10.1016/0022-2836(81)90087-5
Sonnhammer ELL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet TIG 18:619–620. https://doi.org/10.1016/s0168-9525(02)02793-2
Sonnhammer ELL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43:D234–D239. https://doi.org/10.1093/nar/gku1203
Sonnhammer ELL, Gabaldón T, Sousa da Silva AW et al (2014) Big data and other challenges in the quest for orthologs. Bioinformatics 30:2993–2998. https://doi.org/10.1093/bioinformatics/btu492
Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. Nat Rev Genet 16:472–482. https://doi.org/10.1038/nrg3962
Spang A, Saw JH, Jørgensen SL et al (2015) Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521:173–179. https://doi.org/10.1038/nature14447
Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988
Storm CEV, Sonnhammer ELL (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18:92–99. https://doi.org/10.1093/bioinformatics/18.1.92
Straub K, Merkl R (2019) Ancestral sequence reconstruction as a tool for the elucidation of a stepwise evolutionary adaptation. Methods Mol Biol Clifton NJ 1851:171–182. https://doi.org/10.1007/978-1-4939-8736-8_9
Studer RA, Robinson-Rechavi M (2009) How confident can we be that orthologs are similar, but paralogs differ? Trends Genet TIG 25:210–216. https://doi.org/10.1016/j.tig.2009.03.004
Sutphin GL, Mahoney JM, Sheppard K et al (2016) WORMHOLE: novel least diverged ortholog prediction through machine learning. PLoS Comput Biol 12:e1005182. https://doi.org/10.1371/journal.pcbi.1005182
Szklarczyk D, Gable AL, Lyon D et al (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47:D607–D613. https://doi.org/10.1093/nar/gky1131
Tabach Y, Golan T, Hernández-Hernández A et al (2013) Human disease locus discovery and mapping to molecular pathways through phylogenetic profiling. Mol Syst Biol 9:692. https://doi.org/10.1038/msb.2013.50
Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637. https://doi.org/10.1126/science.278.5338.631
The Gene Ontology Consortium (2019) The gene ontology resource: 20 years and still going strong. Nucleic Acids Res 47:D330–D338. https://doi.org/10.1093/nar/gky1055
The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515. https://doi.org/10.1093/nar/gky1049

Uchiyama I, Mihara M, Nishide H et al (2019) MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. Nucleic Acids Res 47:D382–D389. https://doi.org/10.1093/nar/gky1054

Van Bel M, Diels T, Vancaester E et al (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res 46:D1190–D1196. https://doi.org/10.1093/nar/gkx1002

Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. Nat Rev Genet 18:411–424. https://doi.org/10.1038/nrg.2017.26

Vaser R, Adusumalli S, Leng SN et al (2016) SIFT missense predictions for genomes. Nat Protoc 11:1–9. https://doi.org/10.1038/nprot.2015.123

Vilella AJ, Severin J, Ureta-Vidal A et al (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19:327–335. https://doi.org/10.1101/gr.073585.107

Walhout AJ, Boulton SJ, Vidal M (2000) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. Yeast Chichester Engl 17:88–94. https://doi.org/10.1002/1097-0061(20000630)17:2%3c88::AID-YEA20%3e3.0.CO;2-Y

Wang ET, Sandberg R, Luo S et al (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476. https://doi.org/10.1038/nature07509

Waterhouse RM, Seppey M, Simão FA et al (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol 35:543–548. https://doi.org/10.1093/molbev/msx319

Whiteside MD, Winsor GL, Laird MR, Brinkman FSL (2013) OrthologueDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. Nucleic Acids Res 41:D366–D376. https://doi.org/10.1093/nar/gks1241

Wolfe K (2000) Robustness—it's not where you think it is. Nat Genet 25:3–4. https://doi.org/10.1038/75560

Wu Y-C, Rasmussen MD, Kellis M (2012) Evolution at the subgene level: domain rearrangements in the drosophila phylogeny. Mol Biol Evol 29:689–705. https://doi.org/10.1093/molbev/msr222

Zambelli F, Pavesi G, Gissi C et al (2010) Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC Genomics 11:534. https://doi.org/10.1186/1471-2164-11-534

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH et al (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358. https://doi.org/10.1038/nature21031

Zmasek CM, Eddy SR (2002) RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinform 3:14. https://doi.org/10.1186/1471-2105-3-14