Inferring Orthologs: Open Questions and Perspectives

Fredj Tekaia

Inferring Orthologs: Open Questions and Perspectives

Fredj Tekaia

Institut Pasteur, Unit of Structural Microbiology, CNRS URA 3528 and University Paris Diderot, Sorbonne Paris Cité, Paris, France.

ABSTRACT: With the increasing number of sequenced genomes and their comparisons, the detection of orthologs is crucial for reliable functional annotation and evolutionary analyses of genes and species. Yet, the dynamic remodeling of genome content through gain, loss, transfer of genes, and segmental and whole-genome duplication hinders reliable orthology detection. Moreover, the lack of direct functional evidence and the questionable quality of some available genome sequences and annotations present additional difficulties to assess orthology. This article reviews the existing computational methods and their potential accuracy in the high-throughput era of genome sequencing and anticipates open questions in terms of methodology, reliability, and computation. Appropriate taxon sampling together with combination of methods based on similarity, phylogeny, synteny, and evolutionary knowledge that may help detecting speciation events appears to be the most accurate strategy. This review also raises perspectives on the potential determination of orthology throughout the whole species phylogeny.

KEYWORDS: evolutionary processes, genome annotation quality, taxon sampling, synteny, phylogeny, HGT, multidomains, genome trees

CITATION: Tekaia. Inferring Orthologs: Open Questions and Perspectives. *Genomics Insights* 2016:9 17–28 doi:10.4137/GEI.S37925.

TYPE: Review

RECEIVED: November 18, 2016. **RESUBMITTED:** December 30, 2016. **ACCEPTED FOR PUBLICATION:** January 2, 2016.

ACADEMIC EDITOR: Gustavo Caetano-Anollés, Editor in Chief

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1211 words, excluding any confidential comments to the academic editor.

FUNDING: Author discloses no external funding sources.

COMPETING INTERESTS: Author discloses no potential conflicts of interest.

CORRESPONDENCE: tekaia@pasteur.fr

Paper subject to independent expert single-blind peer review. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Detection of orthologs is of fundamental importance in many fields of biology, particularly in the annotation of newly sequenced organisms, functional genomics, gene organization in species, evolutionary studies of biological systems, and phylogenomic analyses. Accurate determination of evolutionary relationships between gene (protein) families involving multiple species is of utmost importance for such goals.

Starting in 1995 with the first complete genome sequence of a free-living organism, *Haemophilus influenzae*,[1] genomics has opened a new research era on species evolution. The availability of full genome datasets raised the hope to decipher species relationships in terms of evolution. Only a few years later, the accumulation of complete genome sequences from three phylogenetic domains has revealed the existence of significant evolutionary processes such as gene exchange between species,[2,3] partial and whole-genome duplication,[4] and gene loss,[5,6] casting doubt on the established species tree topology. Moreover, significant incongruence was observed between species and gene tree topologies, especially in bacteria where horizontal gene transfers (HGTs) tend to blur the phylogeny of species[7–9] and, therefore, the accurate detection of orthologs. The tree representation of cell life was thus questioned, and its replacement by a net or ring of life[7,10] suggested to reflect gene exchange between species. With the exponential increase of available full genome sequences and their studies,[11] several new methods that take into account these evolutionary processes have been introduced for gene and species tree construction.[9,12–14]

Most methods for large-scale detection of orthologs are based on homology as inferred by sequence similarity. However, proper identification of orthologous genes is a major challenge because the accumulation of evolutionary dynamic events tends to blur the recognition of true orthologs among homologs.[15–18] Numerous methods were elaborated to solve this problem, with various advantages and limitations.[17,19,20] In general, most of these methods suffer from a common limitation, namely, the difficulty in constructing orthologous classes in the presence of paralogs. For distantly related species, assessing orthology can become quite difficult, typically due to low similarity between protein sequences and the likely increase of gene birth and death.[15] In contrast to closely related species, synteny conservation provides useful information to identify conserved chromosomal segments, in which orthologous genes can be more steadily searched for.

In this review, a reminder of basic definitions concerning homology, paralogy, and orthology relationships is first proposed. Then, some of the numerous computational methods designed to infer orthologs are introduced, along with a discussion on their advantages and limitations. Accurate construction of species trees is directly linked to inference of orthologs. Difficulties related to species tree constructions undermine orthology inference and inversely. Some of these difficulties will be discussed. Finally, some open questions about ongoing efforts in computational development for the detection of orthologous genes are raised.

Concepts of Homology, Orthology, and Paralogy

Figure 1 illustrates the concepts of homology, paralogy, and orthology.

Homology. Homology is a basic concept at the core of evolutionary genomics. Identifying homology relationships between sequences is the first fundamental step in many biological research domains, and more particularly so in inferring orthologs and paralogs. According to Fitch,[21,22] two genes are homologs if they share a common origin, ie, derived from a common ancestor. Sequence similarity[23] is almost the only criterion available to infer homology. This criterion introduces a limitation to the detection of homology, because sequences may diverge beyond statistical recognition as the evolutionary distances between species increase. Additional complications include the dynamics of gene duplication, loss, transfer or fusion/fission, and shuffling that occurred during genome evolution.

Orthologs—paralogs. Orthologs are homologous genes resulting from a speciation event, whereas paralogs are homologous genes resulting from a duplication event. The dynamic of duplication/loss during evolution is such that paralogy does not require paralogous genes to be in the same genome.[16] Depending upon taxon sampling, genes in single copies in two distinct genomes may result from duplicated copies in their ancestor after differential gene loss in the two derived lineages (Fig. 1). Hence, there is a distinction between in-paralogs, corresponding to paralogs issued from duplication post speciation, and out-paralogs, corresponding to duplication prior to speciation.[24] Out-paralogs are also sometimes referred to as pseudoorthologs[16] in configuration with differential gene loss where they appear to be orthologs, while two in-paralogs issued from duplication in one lineage are themselves co-orthologs to the only copy present in another lineage in which no duplication took place.

Specific situations are generated by events of whole-genome duplication (WGD) and by HGT. The numerous pairs of in-paralogs left after a WGD have been designated ohnologs, a term that helps distinguish them from other

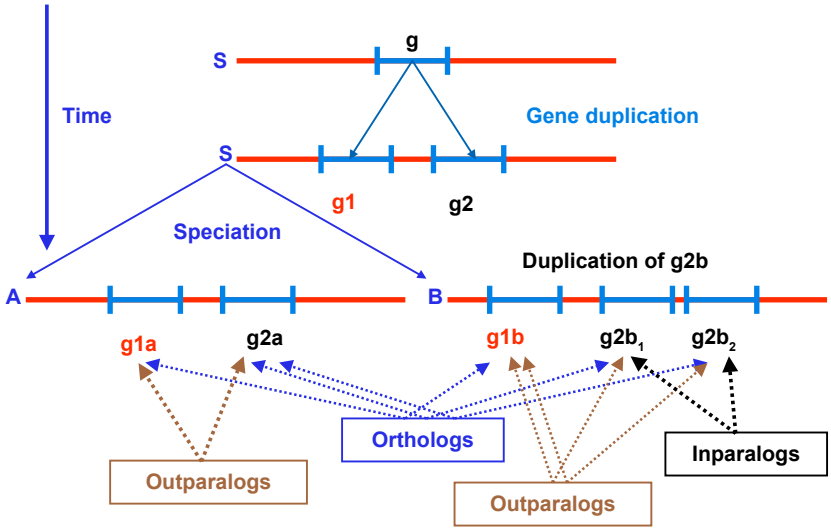

Figure 1. Homologs—paralogs—orthologs.

Notes: This figure illustrates speciation and duplication events and their resulting consequences on gene terminology.

The figure shows:

(1) an intraspecies duplication of gene g giving rise to two genes g1 and g2 (note that g is no more visible in species S);

(2) a speciation event giving rise to two species A and B with identical contents as S; in particular, g1 and g2 are denoted as g1a and g2a in A and g1b and g2b in B;

(3) we assume that in B, g2b is duplicated and gives rise to $g2b_1$ and $g2b_2$;

(Note that g2b is no more visible in B).

In this scheme and considering solely the last speciation event:

- g1 and g2 are homologs because they descend from g. Similarly, g1a and g1b are homologs because they descend from g1;
- g1 and g2 are in-paralogs, because they are duplicated in S;
- Similarly, $g2b_1$ and $g2b_2$ are in-paralogs because they are duplicated in B;
- g1a and g2a are out-paralogs because their ancestors are duplicated in S;
- Similarly, g1b and each of $g2b_1$ and $g2b_2$ are out-paralogs, because their ancestors are duplicated in S;
- g1a and g1b are orthologs because they are in distinct species A and B, respectively, with a common ancestor g1;
- g2a and $g2b_1$ and g2a and $g2b_2$ are orthologs because they are in distinct species A and B, respectively, with the same ancestor g2. $g2b_1$ and $g2b_2$ are also called co-orthologs to g2a.

Dashed arrows with different colors highlight pairs of orthologs, out-paralogs, and in-paralogs.

paralogs resulting from other ancestral duplication.[25] Xenologous genes are genes that appear falsely as orthologs because at least one of the pair is acquired via HGT from another species. Usually it is difficult to identify xenologs in pairwise genome comparisons. Xenologs and true orthologs might be distinguished in multiple genome comparisons if the origin of each gene can be identified.[16]

As originally recognized by Ohno,[26] gene duplication creates a novel evolutionary paradigm as the selective functional pressures act at different regimes when genes are single or multiple copy. Thus, paralogous genes tend to rapidly diverge in function.

Evolutionary Processes and Consequences on Ortholog Inference

Large-scale genome comparisons showed that genes and genomes are subject to strong evolutionary dynamics. These evolutionary processes (Fig. 2) include HGT between species,[2] gene loss and acquisition,[27] protein domain emergence, gain and loss[28,29] events, and, at the genome level, partial/whole-genome duplications and introgression events.[30] WGD events can take place in one round, as in yeast,[4] or in multiple rounds, as in plants, fishes, and other vertebrates.[31] Additional complications might arise by loss of genes in some descent species obtained after rounds of speciation and WGD events. All these events, together with gene transfer, accumulation, and loss, tend to blur the recognition of true orthologs among a set of homologs.[15–18,32]

Horizontal gene transfer—introgression events. HGT (also called lateral gene transfer) is the transmission of genes from a species to another one through processes distinct from ancestral inheritance (or vertical transfer). HGT has emerged as a major evolutionary process that has shaped genomes in all three domains of life. It has been recognized as one of the major evolutionary forces driving prokaryote evolution.[33] Recent estimates by Dagan et al suggest that on average 81% of prokaryote gene have been involved in HGT at some point in their history.[34] In eukaryotes, HGT occurs on a previously unsuspected scale.[35,36]

Introgression is another process of lateral transfer that concerns the transmission of regions of a species' genome to the genome of another species, occurring within closely as well as distantly related species.[18,37,38]

As a result, HGT and introgression events imply different evolutionary histories in the content of the host species,[39] affecting the concepts of evolutionary relationships between species,[40] and hence the detection of paralogs and orthologs and also phylogenetic tree constructions.

Protein domain emergence and gain and loss events. A domain is a structural constituent formed by a distinct region in a specific protein. A domain in a protein sequence might be unique or associated with other domains. Domains are evolutionarily well conserved across taxa[41] and are frequently rearranged (due to duplication, fusion, fission, as well as terminal domain loss) between and within proteins and genomes.[42] Emerging domains (ie, previously unreported) are more likely disordered in structure and spread more rapidly within their genomes than established domains.[29] A significant number of domains are lost along every lineage,[43] and insertions or deletions are more common than substitutions of domains. Domain insertions are significantly more common than domain deletions.[42] In the protein universe, the growth of a single-domain architecture is slow, whereas the growth of a multidomain architecture results from the combination of a single-domain architecture.[44]

Protein domains may give rise to supplementary difficulties in orthology inference due to the possible different ancestral origins of the corresponding gene(s) particularly when resulting from fission or fusion events. Different domains may have different ancestors. Proteins with multiple domains may have multiple significant hits with different proteins (each hit might be based on a given domain in the query protein sequence) with different relative positions and orientations. Consequently inferring an ancestral origin of such a protein is a difficult and challenging task. The need for clear evolutionary definitions is particularly acute for multidomain proteins, as their underlying coding sequences often have distinct, and even conflicting, evolutionary histories. Such proteins may share an inserted domain but are in fact unrelated.[45]

A supplementary difficulty, known as chaining effect, may arise when constructing families of orthologs, as proteins including multiple domains may attract, via auxiliary domains, unrelated proteins (including different combination of domains).

Further difficulties are related to isofunctional genes and novel gene creation, which may lead to erroneously inferred orthology. Isofunctional genes are likely to share

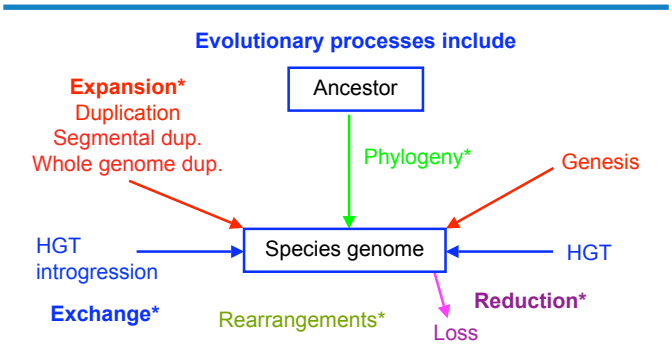

Figure 2. Evolutionary processes.

Notes: This figure illustrates some significant evolutionary processes as revealed by large-scale comparative analyses of predicted proteomes: phylogeny, expansion, exchange, and reduction. Phylogeny is the direct descent from ancestor to actual genome. Expansion (in red) includes gene duplication, segmental and whole-genome duplication, and genesis. Exchange (in blue) includes mainly HGT and introgression. Reduction is represented by gene loss. Rearrangements include inversions, translocations, fusion, and fissions.

high sequence similarity, particularly at the protein level, and thus might be considered as homologs by sequence similarity methods,[46] leading to the erroneous detection of orthologs. The creation of novel genes is possible from noncoding sequences, as well as through domain shuffling, incorporation of mobile elements, or gene fission and fusion.[47–49] Because homology is based on similarity criteria, new genes derived from noncoding sequences might be inferred as orthologs to true genes, whereas it is obvious that they are not: they do not share a common ancestor. To avoid this pitfall, such genes should be filtered out before applying any procedure for the search of orthologs.

Methods for Orthology Inference

There are two main approaches to determine orthologous gene classes, based either on sequence similarity or on phylogeny (Table 1).[50] Comparative studies assessing the performances of different strategies by these methods have already been reported.[17,19,51–53] Here, the salient features of some methods and their corresponding advantages and limitations are highlighted.

Similarity approaches. Similarity-based approaches rely on genome comparisons and clustering of highly similar genes to identify orthologous groups. These approaches include the following: COG,[54] InParanoid,[55] OrthoMCL,[56] TribeMCL,[57] eggNOG,[58] OrthoFocus,[59] OrthoInspector,[60] SuperPartitions of Ortholog (SPO),[61] and OrthFinder.[62] The clustering of the set of inferred orthologs is generally based on classification criteria such as the single linkage method or the Markov Cluster Algorithm.[57] Other similarity methods based on evolutionary distance metrics criteria include roundup[63] (based on reciprocal smallest distance [RSD]), RSD,[64] OMA,[65] and a minimum evolution method.[66] A combination of sequence similarity, genome rearrangements, and duplication events is implemented in MSOAR[67] to identify pairs of orthologs in closely related species.

Similarity approaches relying mainly on reciprocal best hits (RBHs) are recognized to perform well in terms of comparative accuracy. Benchmark studies[19,51,68] concluded that RBH methods outperform other approaches. However, the RBH approach was criticized for underappreciation of orthology in the presence of paralogy. More importantly, this simple

Table 1. Methods for orthology inference.

METHOD	ALGORITHM
COG[54]	Similarity—Single linkage clustering + Constraints
InParanoid/MultiParanoid[55]	Similarity (pair-wise species)/Extends to multiple species
OrthoMCL[56]	Similarity—MCL clustering algorithm
TribeMCL[57]	Similarity—MCL clustering algorithm
eggNOG[58]	Similarity—Detects false RBH due to gene fusion and protein domain shuffling
OrthoFocus[59]	Similarity—extended RBH to handle many-to-one and many-to-many relationships
OrthoInspector[60]	Smilarity
SPO[61]	Similarity (RBH)—Partition of orthologs includes Intra-species Partition and MCL clustering.
OrthoFinder[62]	Similarity—Clustering
Roundup[63]	Reciprocal Smallest Distance
RSD[64]	Reciprocal Smallest Distance (evolutionary distance = estimated number of amino acid substitutions)
OMA[65]	Similarity—Global sequence alignment
ME[66]	Minimum Evolution Method
MSOAR[67]	Similarity—Genome rearrangement—duplication
Orthostrapper[69]	Phylogeny—bootstrap
RIO[70]	Similarity (HMMER)—bootstrap—Phylogeny
PhIGs[71]	Similarity—Multiple sequence alignments—Phylogenetic trees
PhyOP[72]	Similarity (overlapping limits)—phylogeny based on d_S (synonymous substitution rates)
TreeFam[73]	Infer orthologs—paralog from the phylogenetic tree
LOFT[74]	Assigns hierarchical orthology numbers to genes based on a phylogenetic tree
EnsemblCompara GeneTrees[75]	Clustering—multiple alignment—tree generation based on TreeBeST method
SYNERGY[76]	Sequence similarity—species phylogeny—reconstruction of underlying gene evolutionary histories
PHOG[77]	Precomputed phylogenic trees followed by identification of orthologs as sequences from different species that are each others reciprocal nearest neighbors
COCO-CL[78]	Similarity—Correlation between sequences—single linkage clustering

Note: This table shows some orthology inference methods with corresponding reference and a short description of their underlying algorithm.

approach was reported to suffer from conceptual drawbacks: (i) RBH analyses are restricted to the class of 1:1 orthologs, failing thus in the detection of many-to-(one and/or many) orthologs; (ii) the RBH approach may lead to overinclusiveness particularly when gene losses are involved in some of the considered genomes.[68] In such cases, a gene is erroneously considered as best hit because the real counterpart is lost; and (iii) the classification of orthologs detected with RBH methods is suspected to be prone to chaining effects in the motif and domain compositions. Such effects could be associated with gene fusion and domain shuffling events in multidomain proteins that evolved through different speciation and duplication events.[59,61]

Phylogeny approaches. Phylogeny-based approaches use candidate gene families determined by similarity and then rely on merging gene and species phylogeny to determine the subset of orthologs: Orthostrapper,[69] RIO,[70] PhIGs,[71] PhyOP,[72] TreeFam,[73] LOFT,[74] EnsemblCompara GeneTrees based on TreeBeST method,[75] SYNERGY,[76] and PHOG.[77] COCO-CL[78] is based on a hierarchical clustering algorithm of correlated genes guided by phylogenetic relationships. Species phylogeny was originally based on rRNA genes, and with the availability of complete genomes, it is now based on sets of shared genes. It should be noted that species phylogeny construction is still subject to debate, as discussed below.

Obtaining the correct phylogenetic gene tree and performing accurate reconciliation is crucial for the detection of orthologs. Phylogeny-based methods are deemed to be more reliable than similarity approaches,[79] but are difficult to automate because of the intrinsic weakness of the multiple alignment and phylogenetic tree construction methods that underlie gene phylogenies. Intrinsic weaknesses of multiple alignments include sequences of different lengths that require the introduction of gaps, and reshuffled sequences that lead to inaccurate alignments.[80] Moreover, the complexity of the phylogeny-based methods grows with the number of taxa,[81] particularly in large families where orthology is more difficult to assess. Such approaches are also subject to controversy on conceptual grounds,[82] as species phylogeny is not always straightforwardly established,[9,16,61,83] with gene and species phylogenies not necessarily coinciding.

In practice, combination of the similarity and phylogeny-based approaches, together with manual annotation, helps to a rather reliable identification and clustering of orthologs.[84] This combination has been successfully illustrated by the reconstruction of gene histories in Ascomycota fungi.[85]

Synteny conservation approaches. Although synteny describes the colocalization of loci on the same chromosome, synteny conservation is defined as the conservation of gene order and orientation along chromosomes[86] and constitutes a substantial source of evidence for determining gene ancestry. The adjacency of orthologous genes in different species provides reliable information to identify orthology relationships, because the comparison of closely related species revealed an extensive, quasi-integral conservation of gene arrangements along chromosomes.[87,88] However, synteny conservation suffers from large interfamily evolutionary distances, asynchronous to the sequence divergence between orthologous genes, as shown for example in both drosophilids[89] and yeast species.[90]

Similarity approaches were associated with conservation of synteny[84,91,92] and neighborhood of genes between species[93,94] to reliably detect orthologs in closely related species. The conservation of chromosomal environments has been used in specific methods to refine the identification of orthologous groups, as applied for example on the specific group of hemiascomycetous yeasts[95,96] and vertebrates.[31] The information and conservation of gene arrangements help in improving the accuracy of orthology assignment and, in some cases, constitute the unique possibility to check for the reliability of detected orthologs. Figure 3 illustrates an example of a difficult scenario where a species S including a gene g that has been duplicated into g_1 and g_2, is followed by a speciation event, giving rise to two species S_1 and S_2, and by a gene duplication solely in S_2, of g_{12} (resulting in g_{12a} and g_{12b}) and g_{22} (resulting in g_{22a} and g_{22b}). Two neighboring genes descendant of g_0 and g_3 are conserved throughout. In this situation, if genes g_{11}, g_{22a}, and g_{22b} are lost (see the dashed lines under genes in Fig. 3), most existing methods (based on similarity and/or phylogeny) will erroneously consider the remaining actual genes as co-orthologs despite the fact that they are not [the two pairs (g_{21} and g_{12a}) and (g_{21} and g_{22b}) are not derived from the same ancestral gene after the speciation event]. Thus, it is solely the conservation of the neighboring genes in S_1 and S_2 that may help to hypothesize on the speciation and gene duplication events, and consequently on the nonorthology of (g_{21} and g_{12a}) and (g_{21} and g_{12b}). A similar, albeit more complex, example involving two rounds of WGD and speciation events is shown in Ref. 92.

Beyond *sequence comparisons, synteny, and phylogeny*-based methods, there are other, less traditional, methods that attempt to improve the prediction of orthologs and their functional analyses.[97] Some of these are based on large-scale analysis of protein–protein interactions and gene coexpression networks.[98,99] A recent database, IsoBase,[100] resulting from function-oriented ortholog identification, seeks the integration of sequence data and protein–protein interaction networks to help identifying functionally related proteins.

Other methods based on shared protein domains have been suggested, but these methods are implicitly based on sequence similarity.[101–103] Multidomain proteins pose a challenging question because of their ancestral origin particularly when they have undergone domain shuffling, and consequently on inferring their orthology.[45]

SuperPartitions of orthologs. In large-scale proteome comparisons, a method called SuperPartitions[61] was introduced for ortholog inference and clustering into families called SPOs. The procedure is based on the partitioning of RBHs, with the further merging of partitions, including members

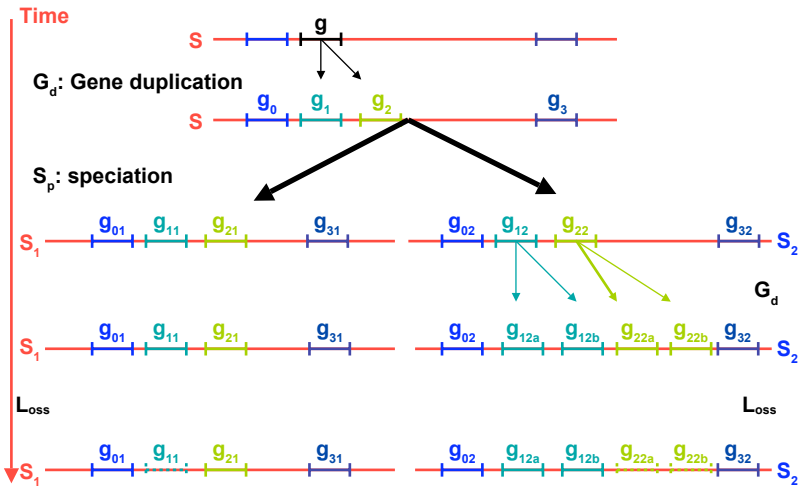

Figure 3. Example of a misleading situation in orthology inference.
Notes: A species S is shown including a gene g that has been duplicated (Gd) into g_1 and g_2. A speciation event (Sp) gave rise to two species S_1 and S_2, followed by a duplication (Gd) solely in S_2 of g_1 (resulting in g_{1a} and g_{1b}) and of g_2 (resulting in g_{2a} and g_{2b}). The neighboring genes g_0 and g_3 are conserved. If genes g_1 in S_1 and g_{2a} and g_{2b} in S_2 are lost, most similarity and phylogenetic methods for orthology detection will assign erroneously orthology to g_2, g_{1a}, and g_{2b}. Indeed, these are not orthologous, because g_2, g_{1a}, and g_{2b} do not result from the same ancestral gene after the speciation event. Conservation of their neighboring genes and synteny may help to suspect speciation and gene duplication events and therefore conclude for the nonorthology of these genes.

of the same paralogous classes. This procedure is detailed in Ref. 61; here, only the main steps are summarized in the following section.

Distinct proteomes by pairwise comparisons lead to sets of RBH proteins that are considered orthologs. The set of all RBH proteins (deduced from the considered proteomes) is partitioned and each part is denoted as Pn.m (n being the number of distinct proteins in the part and m is an arbitrary index used to differentiate partitions containing the same number of proteins).

For each species, intraspecies comparisons lead to a set of nonunique proteins (proteins that have at least one hit in their own proteome) that are considered paralogs. These paralogous proteins are clustered according to their similarity using the mcl programme.[104] Each cluster is denoted as Cp.q (p being the number of proteins in the cluster and q is an arbitrary index used to differentiate clusters with identical number of elements). In each species, unique proteins, ie, with no hit in their own proteome are denoted as *single*.

Considering the whole set or orthologs obtained from all pairwise comparisons, each protein is part of an RBH partition denoted as Pn.m and of a cluster of paralogs in its own species denoted as Cp.q (or *single* if it is unique). Both classifications are joined to form the grouping category of each protein in the set of orthologs: Pn.m.Cp.q, ie, corresponding orthologous part and paralogous cluster.

RBHs partitions were further processed, by merging partitions including proteins belonging to the same intraspecies class Cp.q (labeled Pn.m.Cp.q), into a SuperPartition denoted as SPOr.s with r the number of proteins in the SPO and s an arbitrary order for indexing.

In order to facilitate the in-depth checking and study of predicted ortholog families (SPOs), the conservation profiles of the obtained clusters were used, which allow simple detection of SPOs containing duplicated members. For each SPO, the *meme*[104] suite of programs was used to search for shared motifs by all (or a subset of) protein sequences in the SPO, thus allowing the evolutionary structure of the members to be revealed (a practical example of this coding scheme is shown in Fig. 4).

As a result, members of a Superpartition (SPOr.s) were characterized (see Ref. 61 for the coding) by: (a) their corresponding RBH partition (Pn.m); (b) their intraspecies mcl cluster Cp.q (with the attached intraspecies partition Pn.m.Cp.q); and (c) their shared motifs as detected by *meme*. These detailed descriptions help in the checking and the assessment of orthology between the SPO members. In the example of Figure 4, four shared motifs appear in the same order within all members of the SPO, thus enhancing the validation of the predicted SPO members.

Perspectives and Open Questions

The search for orthologs is a milestone in the genome era, as its proper detection is crucial for evolutionary studies of genes and species. Comparative genomic analyses represent a powerful approach to recognize similarities between species, notably with the exponentially increasing amount of data generated by genome projects. Consequently, one of the primary tasks of evolutionary genomics is the determination of gene families from sets of taxa. The reconstruction of the evolutionary histories of genes and species relies critically on the accurate identification of orthologs. Such identifications

SpecCode_ProtId	Partition_RBH	Paralogs	Motifs SPO29.1: map methionine aminopeptidase
MYTC_MT2929	P12.799	P2.101.C2.346	12 5 11 1 9 3 6 2 4
MYTU_mapB	P12.799	P2.184.C2.449	12 5 11 1 9 3 6 2 4
MYBO_Mb2886c	P12.799	P2.99.C2.133	12 5 11 1 9 3 6 2 4
MYMA_MMAR1842	P12.799	P2.240.C2.328	12 5 11 1 9 3 6 2 4
MYAV_MAV3721	P12.799	P2.66.C2.150	12 5 11 1 9 3 6 2 4
MYAP_MAP2934c	P12.799	P2.16.C2.65	12 5 11 1 9 3 6 2 4
MYLE_ML1576c	P12.799	P2.34.C2.46	12 5 11 1 9 3 6 2 4
MYSM_MSMEG2587	P12.799	P10.11.C4.47	12 5 11 1 9 3 6 2 4
MYJL_Mjls2033	P12.799	P3.78.C3.153	12 5 11 1 9 3 6 2 4
MYVA_Mvan2272	P12.799	P3.106.C3.184	12 5 11 1 9 3 6 2 4
MYUL_MUL2091	P12.799	P2.183.C2.244	13 7 1 9 3 6 2 4
MYAB_MAB3164c	P12.799	P3.29.C3.59	13 7 1 9 3 6 2 4
MYTC_MT0758	P12.977	P2.101.C2.346	10 7 1 9 3 6 14 2 4
MYTU_mapA	P12.977	P2.184.C2.449	10 7 1 9 3 6 14 2 4
MYBO_Mb0755	P12.977	P2.99.C2.133	10 7 1 9 3 6 14 2 4
MYMA_MMAR1072	P12.977	P2.240.C2.328	10 7 1 9 3 6 14 2 4
MYAV_MAV4432	P12.977	P2.66.C2.150	10 7 1 9 3 6 14 2 4
MYAP_MAP4200	P12.977	P2.16.C2.65	10 7 1 9 3 6 14 2 4
MYLE_ML1831c	P12.977	P2.34.C2.46	10 7 1 9 3 6 2 4
MYSM_MSMEG1485	P12.977	P10.11.C4.47	10 7 1 9 3 6 14 2 4
MYJL_Mjls1076	P12.977	P3.78.C3.153	10 7 1 9 3 6 14 2 4
MYVA_Mvan1349	P12.977	P3.106.C3.184	10 7 1 9 3 6 14 2 4
MYUL_MUL0830	P12.977	P2.183.C2.244	10 7 1 9 3 6 14 2 4
MYAB_MAB3782c	P12.977	P3.29.C3.59	7 1 9 3 6 14 2 4
MYSM_MSMEG5050	P4.215	P10.11.C4.47	7 15 8 3 6 2 4
MYJL_Mjls3082	P4.215	P3.78.C3.153	7 15 8 3 6 2 4
MYVA_Mvan0391	P4.215	P3.106.C3.184	7 15 8 3 6 2 4
MYAB_MAB0094c	P4.215	P3.29.C3.59	7 15 8 3 6 2 4
MYSM_MSMEG5683	P+	P10.11.C4.47	7 8 3 6 2 4

Figure 4. Assessment of members of orthologs in an SPO cluster by detecting motifs and their distribution.
Notes: Motifs in SPOs are illustrated with the example of SPO29.1, from the considered 12 mycobacterial species. This SPO contains proteins corresponding to mapA and mapB (methionine aminopeptidase). Column headings are as follows: (a) SpecCode_ProtID: species code (see coding conventions below) followed by the protein identification; (b) Partition_RBH: partition of RBHs in pairwise proteome comparisons of considered species) denoted Pl.r where l is the number of proteins in the partition and r is an arbitrary index; (c) paralogs: paralogous class Pn.m is a partition of intraspecies RBHs and Cp.q is the cluster obtained by the mcl programme (see Ref. 62 for more details on the coding scheme of Pn.m.Cp.q classes); and (d) motifs: distributions of motifs as obtained with the meme/mast programs. The distributions highlight motifs shared by all proteins (ancestral motifs: 3,6,2,4) and motifs shared by subsets of proteins. Checking of the detailed description of paralogs allowed adding the last line (MYSM_MSMMEG5683) because only three from the P10.11.C4.47 cluster were found by the RBH procedure.

Code	Species	Code	Species
MYTU	*Mycobacterium tuberculosis* H37R	MYAV	*Mycobacterium avium*
MYBO	*Mycobacterium bovis*	MYAP	*Mycobacterium avium paratuberculosis*
MYTC	*Mycobacterium tuberculosis* CDC 1551	MYJL	*Mycobacterium* JLS
MYUL	*Mycobacterium ulcerans*	MYVA	*Mycobacterium vanbaalenii* PYR-1
MYMA	*Mycobacterium marinum*	MYSM	*Mycobacterium smegmatis* MC2 155
MYLE	*Mycobacterium leprae*	MYAB	*Mycobacterium abscessus* ATCC 19977T

are crucial for addressing a series of fundamental evolutionary questions concerning the determination of shared genes by different species, genes that share common core evolutionary history, or yet genes subjected to duplication, deletion, or transfer. The accumulation of these evolutionary events makes the reliable identification of orthologous genes a major challenge, particularly for distantly related species where traces of similarity are hardly recognizable.

Even though the manual detection of orthologs may be efficient for a small number of genes, automatic approaches are needed to deal with the large amount of genome data currently available. However, despite great efforts devoted to the

development of orthology detection methods, the situation appears unsettled and to a large extent the *quest for orthologs*[17] is still an ongoing task in large-scale genome comparisons. The motivations of "The Quest for Orthologs" (http://questfororthologs.org/) consortium that is an open community are to benchmark, to improve accuracy and standardize orthology inference through collaboration, use of shared reference datasets (http://www.ebi.ac.uk/reference_proteomes), and evaluation of emerging new methods.

Species tree and ancestral sequence reconstructions may help to determine accurate orthologs. The species tree is crucial to delineate speciation events, whereas ancestral sequence reconstruction of actual sequences provides further information about their evolution over time. In the following section, some hints that may help methodological developments in inferring orthologs are pointed out.

Problematic Quality and Completeness of Whole-genomic Data

Genome assembly and annotation. The starting point for orthologs detection is the availability of completely annotated genomes with their corresponding complete sets of genes; otherwise, it is unlikely to recover the correct full set of orthologs. For technical and methodological reasons, available genome sequences are of different quality: some are complete, while others are incomplete or in draft state (see statistics on: http://www.genomesonline.org/cgi-bin/GOLD/index.cgi). Genome data annotations are also of highly diverse quality. Apart from a few genomes, particularly from yeast,[69,105] many of the available genome data are automatically annotated and have not been adequately checked. Moreover, annotations greatly depend on the quality of the sequences themselves. This is particularly serious for draft sequences, for which the lack of additional sequencing efforts to complete the assembly of chromosomes may mislead on the actual number of genes and the corresponding sequences, as quantified in Ref.[106] High-throughput sequencing technologies are now producing large amount of genomic data, raising fear of the low quality of assembly and annotation data. Consequently, an important emerging question is how accurate and useful are these data regarding orthology detection at a large scale?

This issue has been partially addressed in the case of model organisms. For example, a set of genomes have been extensively annotated, with the aim to study genes showing specific functions involved in the adaptation of species in specific environments (see Ref.[107] for the Archaea example) and to be used as reference for homology annotation of genes from closely related species. Only a few distantly related species received particular attention concerning this matter, from the human and teleostan fish genomes to the bacteria *Escherichia coli* and *Bacillus subtilis*, including the yeast *Saccharomyces cerevisiae* and the ciliate *Paramecium tetraurelia*. These efforts confidently allowed the automatic annotation of a number of closely related species, but rapidly have shown their limits when large-scale comparisons have been considered. As an illustration, the rate of erroneous annotations has been measured to be >30% in the UniProt/Swiss-Prot database.[108]

Taxon sampling and species coverage. With the exception of the model organisms mentioned earlier, the accurate selection of sequenced genomes is still a pending question. While a number of scientific communities focus on very specific groups of monophyletic species as for example the 1000[109] and 1001[110] genome projects for *Homo sapiens* and the plant *A rabidopsis thaliana*, respectively, the great majority was attached to sequence and characterize distantly related species, from known or unknown taxonomic groups, with the aim of extending our knowledge on the tree of life. This is the case for example in yeast for the Dikaryome project that aims to extend the known phylogeny to the Dykarya phylum.[111] This bias is important in orthology identification using phylogenetic methods, and obviously penalizes automatic discrimination between true orthologs and horizontally acquired genes.[112] False bacterial hits are favored in the case of eukaryotic genes, and erroneous *most similar* hits in the case of bacterial genes. Indeed, the biased distribution of proteins toward bacteria and model organisms in current databases artificially increases the chance to find bacterial (rather than eukaryotic) genes as the most similar sequences to the one under study.

The reconstruction of syntenic blocks, chromosomal segments in which the order and orientation of the genes are conserved through the evolution of corresponding species, is also limited by the correct taxon sampling of the species under study. In closely related species, a large part of genes is still found in short syntenic blocks as for example in fly and in yeast.[89,90] However, at large evolutionary distances, it remains very difficult to accurately define conserved segments, due to frequent events of synteny breakage. Consequently, one has to choose an appropriate set of closely related species for the analysis of synteny conservation, and hence for the reconstruction of ancestral genomes.

Therefore, the taxon sampling issue has a decisive impact on both phylogenetic tree and synteny reconstruction, thus affecting the identification of orthologous sets of genes. This could be partly solved by the correct choice of sequenced species, in this way filling the gap of uncovered phylogenetic regions and decreasing the span between compared species.

Ancestral sequence reconstructions and species tree. Reconstructions of ancestral gene sequences and evolutionary events traditionally reported by species trees are key to orthology detection and validation. In silico reconstruction of ancestral gene, protein, and chromosome sequences provides information about evolutionary events. For example, conserved genomic islands identify features that are protected from variation by natural selection. When an ancestral sequence is predicted for a set of actual sequences (genes, chromosomes, or genomes), multiple alignments of the predicted ancestor and the actual sequences should provide hints about mutations (substitutions), deletions, and insertions that

took place during the evolutionary time in these sequences. Such reconstructions might give indications about gene gain or loss, as well as genome reduction. Consequently, ancestral prediction represents one of the expected results that may validate orthology detection as shown by ongoing efforts in algorithmic development[6,85,113–116] and particularly when such reconstructions are performed in association with species tree topologies, shedding light on specific evolutionary events at a given node of the tree. Optimal strategies for ancestral sequence and genome reconstructions are still under development, evaluation, and debate.[30,117–120]

Building the tree of life genome by genome,[13,14,121] and tracing the tree of life,[122] together with advances in methodological tree of life construction[9] will allow the setup of an accurate species tree topology. Because of the observed incongruence between gene tree and species tree topologies, specific methodological developments to reconcile these topologies have been initiated to delineate the degree of confidence in such tree topologies.[123–128] The reconciliation processes usually involve hypotheses on gene duplications and losses in order to account for the topological incongruences.

The assessment of such theoretical models in constructing species trees can benefit from the availability of recently discovered data and new resources. In this regard, the availability of ancient DNA sequences might be a major resource in filling gaps between distantly related species on the species tree and allows for better estimation of speciation events.[129] Other significant resources are shown by an interesting example,[130] which traced the first step to speciation in a salamander population.

Experimental validation of predicted sets of orthologs?. While the vast majority of published orthology sets only relies on computational analysis of gene sequences without further experimental validation, functional similarities can be assessed assuming the absence of paralogs in the considered groups of genes.[131] One approach is the heterologous replacement of a particular gene by its ortholog from another species. This experimental setup allows measuring the complementation of the function of a gene by its orthologous gene, by complementing mutant cells or restoring a particular phenotype. Only a few cases of orthologous relationships have been validated using this type of isofunctionality testing in *S. cerevisiae* mutants.[132,133]

Isofunctionality can also be assessed by analyzing the similarity of interacting partners within protein interaction networks. In vitro two-hybrid experiments could reveal the biological proof for the existence of orthologous genes. In this case, pairs of purified gene products are tested for interaction, measured by two-hybrid essays. This method has long been used to establish interaction maps within species[134] and could be adapted to cross-species comparisons. Once the interaction between two proteins is verified within one particular species, the positive match after replacement by an orthologous protein from another species could provide direct evidence of functional complementation, whereas the negative match does not necessarily imply absence of orthology.

With the exception of few cases, the proofs for reliability and completeness of predicted groups of orthologs in a set of species are still pending questions. Indeed, a list of orthologs in a set of species is trusted as long as the corresponding species phylogeny is trusted.

At the structure level, relationship between sequence similarity and structural similarity has long been established, but little is known about the impact of orthology on the relationship between protein sequence and structure.[135] It has been shown that orthologous proteins exhibit a greater similarity of domain architectures (ie, domains structure along a sequence) as compared with paralogous proteins at the same level of similarity.[135–137] This result is confirmed by the comparison of orthologs and paralogs with the available crystal structures.[135]

In this regard, it is interesting to note that the conservation of sequence, structure, or genomic context is not implicit in the definition of orthology.[137]

A significant issue on reliability is the distribution of predicted orthologs (or conservation profile)[61] in a given cluster, conveying important information about the content, expansion, and reduction of such a cluster among the surveyed species. In this regard, only a few reports have established a possible gain or loss of members in a given cluster of orthologs by considering systematic studies in specific situations.[6,85] The work of Trachana et al[53] focused on the assessment of accuracy of the methods in ortholog assignments by considering 70 manually annotated protein families, and hence the possibility of comparing the ability of each method to detect orthologs in a given family.

For practical reasons, there is a need for a public standard set of genomes that can be used to compare methods in predicting orthology. This set should include the typical difficulties discussed earlier and could be used to test the sensitivity and selectivity of orthology detection methods. This concept has already been discussed,[138,139] and sets of reference proteomes (http://www.ebi.ac.uk/reference_proteomes/) and orthology-curated databases (http://questfororthologs.org/orthology_databases and http://eggnog.embl.de/orthobench2) to be considered for ortholog assignments and evaluation have been suggested. Unfortunately, while these sets are useful to show the differences between orthology detection methods, they are insufficient to estimate their respective accuracy with regard to the aforementioned difficulties (including HGT, duplication, loss, and proteins with multidomain ancestry, among others).

An optimal standard set should include both closely and more distantly related species sampled from the species tree. The selected species should correspond to the pointed difficulties in detecting orthologs, including gene loss, HGT, duplication, WGD (one or several rounds), and genome reduction. Analyses of this set by orthology prediction methods should illustrate their corresponding ability to detect a given difficulty

and their accuracy in identifying the correct set of orthologs and their clustering.

Why do so many orthology predicting methods exist?. At this point, one may wonder why so many methods exist (although only a few of them are cited here). The simple answer is that there is still no evidence on how to deal with the complex evolutionary events that have been mentioned in this work and that hamper correct orthology detection. Current methods have been developed to overcome particular difficulties, but none appears to be reporting *universal* solutions.

Cross-comparison studies of orthology detection methods[17,19,51–53,68,97] show significant differences in their corresponding sets of orthologs, and even contradictory results between large-scale studies that evaluate the relative algorithms.[68] It has been reported[97] that popular methods show <50% of concordance in establishing overlapping sets of orthologs, and even <30% when distantly related species are taken into account. These authors further suggested that a deeper understanding of the error-prone steps in the algorithms could trigger developments toward better ortholog detection and clustering, by focusing on inconsistent sets of orthologs predicted by different methods. They conclude that "challenges for RBH-based approaches center around how to reduce false positives. In contrast, phylogeny-based approaches have many more aspects to consider including: selection of genes to build the tree and the accuracy of the tree reconciliation with known phylogeny."

In practice, the combination of methods (similarity and phylogeny based) together with the known organization of the considered genomes (particularly concerning synteny conservation, introgression, and segmental and whole-genome duplication) are currently the best procedure to enhance the validity of the inferred set of orthologs and paralogs.

Concluding Notes

This review has reported some of the many available methods in inferring orthologs and their clustering. The evolutionary dynamics involving duplication, loss, fusion/fission of genes, and segmental and whole-genome duplication are some of the difficulties that hamper the detection and clustering of orthologs in a large set, including distantly related species. The review has also mentioned the contribution of methodological developments and resources that may help establishing a species tree that is at the core of reliable orthology detection.

Appropriate sequencing and annotation efforts in sets of sampled species should provide reliable sets of orthologs when using a combination of similarity and phylogeny methods together with prior knowledge related to evolutionary dynamic events. Large-scale orthology detection from distantly related species can be approached in two steps, first on locally (at the taxon level) reliably defined clusters of orthologs that can then be assembled in a second step to cover the whole large set of species.

Finally, it is suggested to filter out, prior to performing large-scale orthology inference, the considered proteomes from proteins resulting from HGT and to fix possible difficulties related to domain shuffling, gain, and loss.

Key Points

- This review describes some available methods in detecting orthologs between species, with brief indications about their advantages and limitations.
- Orthology detection is hindered by the evolutionary dynamics, including duplication, transfer and loss of genes, and shuffled multidomain proteins, as well as by the questionable quality of available genome data in terms of completeness and annotation.
- It is suggested that orthology detection at large scale should be first performed locally at the taxon level, where existing methods and manual validation generally result in reliable detection and clustering of orthologs, and then assembled following the species tree as a template.
- Methodological developments are still needed in the automatic inference and clustering of orthologs in conjunction with species tree construction, as both concepts are tightly linked.

Acknowledgments

I thank Bernard Dujon for constant support, Pedro Alzari for his careful reading of the manuscript and his suggestions, and Thomas Rolland for constructive discussions.

Author Contributions

Conceived the concepts: FT. Analyzed the data: FT. Wrote the first draft of the manuscript: FT. Developed the structure and arguments for the paper: FT. Made critical revisions: FT. The author reviewed and approved of the final manuscript.

REFERENCES

1. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995;269:496–512.
2. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*. 1999;96:3801–3806.
3. Choi IG, Kim SH. Global extent of horizontal gene transfer. *Proc Natl Acad Sci U S A*. 2007;104:4489–4494.
4. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*. 1997;387:708–713.
5. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000;290:1151–1155.
6. Hahn MW, Han MV, Han SG. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet*. 2007;3:e197.
7. Doolittle WF. Phylogenetic classification and the universal tree. *Science*. 1999; 284:2124–2129.
8. Galtier N, Daubin V. Dealing with incongruence in phylogenomic analyses. *Philos Trans R Soc Lond B Biol Sci*. 2008;363:4023–4029.
9. House CH. The tree of life viewed through the contents of genomes. *Methods Mol Biol*. 2009;532:141–161.
10. Rivera MC, Lake JA. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*. 2004;431:152–155.
11. 2015. Available at: http://www.ncbi.nlm.nih.gov/genome/
12. Eisen JA, Fraser CM. Phylogenomics: intersection of evolution and genomics. *Science*. 2003;300:1706–1707.
13. Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons. *Genome Res*. 1999;9(6):550–557.

14. Tekaia F, Yeramian E. Genome trees from conservation profiles. *PLoS Comput Biol*. 2005;1(7):e75.
15. Lynch M, Conery JS. The evolutionary demography of duplicate genes. *J Struct Funct Genomics*. 2003;3:35–44.
16. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 2005;39:309–338.
17. Kuzniar A, van Ham RC, Pongor S, Leunissen JA. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet*. 2008;24: 539–551.
18. Novo M, Bigey F, Beyne E, et al. Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proc Natl Acad Sci U S A*. 2009;106:16333–16338.
19. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*. 2009;5:e1000262.
20. Kristensen DM, Kannan L, Coleman MK, et al. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*. 2010;26:1481–1487.
21. Fitch WM. Distinguishing homologous from analogous proteins. *Syst Zool*. 1970;19:99–113.
22. Fitch WM. Homology a personal view on some of the problems. *Trends Genet*. 2000;16:227–231.
23. Tekaia F, Dujon B. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J Mol Evol*. 1999;49(5):591–600.
24. Sonnhammer EL, Koonin EV. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*. 2002;18:619–620.
25. Wolfe K. Robustness—it's not where you think it is. *Nat Genet*. 2000;25:3–4.
26. Ohno S. *Evolution by Gene Duplication*. Berlin: Springer-Verlag; 1970.
27. Babenko VN, Krylov DM. Comparative analysis of complete genomes reveals gene loss, acquisition and acceleration of evolutionary rates in Metazoa, suggests a prevalence of evolution via gene acquisition and indicates that the evolutionary rates in animals tend to be conserved. *Nucleic Acids Res*. 2004;32(17): 5029–5035.
28. Nasir A, Kim KM, Caetano-Anollés G. Global patterns of protein domain gain and loss in superkingdoms. *PLoS Comput Biol*. 2014;10(1):e1003452.
29. Moore AD, Bornberg-Bauer E. The dynamics and evolutionary potential of domain loss and emergence. *Mol Biol Evol*. 2012;29(2):787–796.
30. Geneva AJ, Muirhead CA, Kingan SB, Garrigan D. A new method to scan genomes for introgression in a secondary contact model. *PLoS One*. 2015;10(4):e0118621.
31. Nakatani Y, Takeda H, Kohara Y, Morishita S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res*. 2007;17:1254–1265.
32. Dalquen DA, Altenhoff AM, Gonnet GH, Dessimoz C. The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. *PLoS One*. 2013;8(2):e56925. doi: 10.1371/journal.pone.0056925.
33. Worning P, Jensen LJ, Nelson KE, Brunak S, Ussery DW. Structural analysis of DNA sequence: evidence for lateral gene transfer in *Thermotoga maritima*. *Nucleic Acids Res*. 2000;28:706–709.
34. Dagan T, Artzy-Randrup Y, Martin W. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A*. 2008;105:10039–10044.
35. Fitzpatrick DA. Horizontal gene transfer in fungi. *FEMS Microbiol Lett*. 2012; 329(1):1–8.
36. Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol*. 2015;16:50.
37. Galeote V, Bigey F, Beyne E, et al. Amplification of a *Zygosaccharomyces bailii* DNA segment in wine yeast genomes by extrachromosomal circular DNA formation. *PLoS One*. 2011;6(3):e17872.
38. Liu KJ, Steinberg E, Yozzo A, Song Y, Kohn MH, Nakhleh L. Interspecific introgressive origin of genomic diversity in the house mouse. *Proc Natl Acad Sci U S A*. 2015;112(1):196–201.
39. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring horizontal gene transfer. *PLoS Comput Biol*. 2015;11(5):e1004095.
40. Syvanen M. Evolutionary implications of horizontal gene transfer. *Annu Rev Genet*. 2012;46:341–358.
41. Forslund K, Sonnhammer EL. Evolution of protein domain architectures. *Methods Mol Biol*. 2012;856:187–216.
42. Weiner J III, Beaussart F, Bornberg-Bauer E. Domain deletions and substitutions in the modular protein evolution. *FEBS*. 2006;273:2037–2047.
43. Bornberg-Bauer E, Huylmans AK, Sikosek T. How do new proteins arise? *Curr Opin Struct Biol*. 2010;20(3):390–396.
44. Levitt M. Nature of the protein universe. *PNAS*. 2009;106(27):11079–11084.
45. Song N, Joseph JM, Davis GB, Durand D. Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput Biol*. 2008;4:e1000063.
46. Omelchenko MV, Galperin MY, Wolf YI, Koonin EV. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol Direct*. 2010;5:31.
47. Capra JA, Pollard KS, Singh M. Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol*. 2010;11:R127.
48. Long M, Betran E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*. 2003;4:865–875.
49. Giacomelli MG, Hancock AS, Masel J. The conversion of 3′ UTRs into coding regions. *Mol Biol Evol*. 2007;24:457–464.
50. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene Orthology inference. *Brief Bioinform*. 2011;12(5):379–391.
51. Hulsen T, Huynen MA, de Vlieg J, Groenen PM. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol*. 2006;7:R31.
52. Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One*. 2007;2:e383.
53. Trachana K, Larsson TA, Powell S, et al. Orthology prediction methods: a quality assessment using curated protein families. *Bioessays*. 2011;33(10):769–780.
54. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science*. 1997;278:631–637.
55. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*. 2001;314:1041–1052.
56. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13:2178–2189.
57. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30:1575–1584.
58. Jensen LJ, Julien P, Kuhn M, et al. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res*. 2008;36(Database issue):D250–D254.
59. Ivliev AE, Sergeeva MG. OrthoFocus: program for identification of orthologs in multiple genomes in family-focused studies. *J Bioinform Comput Biol*. 2008;6: 811–824.
60. Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*. 2011;12:11.
61. Tekaia F, Yeramian E. SuperPartitions: detection and classification of orthologs. *Gene*. 2012;492(1):199–211.
62. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 2015;16:157.
63. Deluca TF, Wu IH, Pu J, et al. Roundup: a multi-genome repository of orthologs and evolutionary distances. *Bioinformatics*. 2006;22:2044–2046.
64. Wall DP, Deluca T. Ortholog detection using the reciprocal smallest distance algorithm. *Methods Mol Biol*. 2007;396:95–110.
65. Roth AC, Gonnet GH, Dessimoz C. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*. 2008;9:518.
66. Kim KM, Sung S, Caetano-Anollés G, Han JY, Kim H. An approach of orthology detection from homologous sequences under minimum evolution. *Nucleic Acids Res*. 2008;36:e110.
67. Shi G, Zhang L, Jiang T. MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement. *BMC Bioinformatics*. 2010; 11:10.
68. Salichos L, Rokas A. Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One*. 2011;6(4):e18755.
69. Storm CE, Sonnhammer EL. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*. 2002;18:92–99.
70. Zmasek CM, Eddy SR. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*. 2004;3:14.
71. Dehal PS, Boore JL. A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics*. 2006;7:201.
72. Goodstadt L, Ponting CP. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol*. 2006;2:e133.
73. Li H, Coghlan A, Ruan J, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 2006;34(Database issue):D572–D580.
74. van der Heijden RT, Snel B, van Noort V, Huynen MA. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*. 2007;8:83.
75. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 2009;19:327–335.
76. Wapinski I, Pfeffer A, Friedman N, Regev A. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. 2007;23:i549–i558.
77. Datta RS, Meacham C, Samad B, et al. PHOG: phylofacts orthology group prediction web server. *Nucleic Acids Res*. 2009;37(Web Server issue):W84–W89.
78. Jothi R, Zotenko E, Tasneem A, Przytycka TM. COCO-CL: hierarchical clustering of homology relations based on evolutionary correlations. *Bioinformatics*. 2006;22:779–788.
79. Gabaldón T. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol*. 2008;9:235.
80. Thompson JD, Linard B, Lecompte O, Poch O. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PLoS One*. 2011;6(3):e18093.
81. Poptsova MS, Gogarten JP. BranchClust: a phylogenetic algorithm for selecting gene families. *BMC Bioinformatics*. 2007;8:120.

82. Dessimoz C, Gil M. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biol*. 2010;11:R37.
83. Bapteste E, Boucher Y, Leigh J, Doolittle WF. Phylogenetic reconstruction and lateral gene transfer. *Trends Microbiol*. 2004;12:406–411.
84. Cannon SB, Young ND. OrthoParaMap: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics*. 2003;4:35.
85. Wapinski I, Regev A. Reconstructing gene histories in Ascomycota fungi. *Methods Enzymol*. 2010;470:447–485.
86. Passarge E, Horsthemke B, Farber RA. Incorrect use of the term synteny. *Nat Genet*. 1999;23:387.
87. Fischer G, James SA, Roberts IN, Oliver SG, Louis EJ. Chromosomal evolution in *Saccharomyces*. *Nature*. 2000;405:451–454.
88. Lemoine F, Lespinet O, Labedan B. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC Evol Biol*. 2007;7:237.
89. Zdobnov EM, Bork P. Quantification of insect genome divergence. *Trends Genet*. 2007;23:16–20.
90. Souciet JL, Dujon B, Gaillardin C, et al. Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res*. 2009;19:1696–1709.
91. Zheng XH, Lu F, Wang Z, Zhong F, Hoover J, Mural R. Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics*. 2005;21:703–710.
92. Catchen JM, Conery JS, Postlethwait JH. Automated identification of conserved synteny after whole-genome duplication. *Genome Res*. 2009;19:1497–1505.
93. Byrne KP, Wolfe KH. Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser. *Nucleic Acids Res*. 2006;34(Database issue):D452–D455.
94. Seret ML, Diffels JF, Goffeau A, Baret PV. Combined phylogeny and neighbourhood analysis of the evolution of the ABC transporters conferring multiple drug resistance in hemiascomycete yeasts. *BMC Genomics*. 2009;10:459.
95. Kellis M, Patterson N, Birren B, Berger B, Lander ES. Methods in comparative genomics: Genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol*. 2004;11:319–355.
96. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*. 2005;15:1456–1461.
97. Fang G, Bhardwaj N, Robilotto R, Gerstein MB. Getting started in gene orthology and functional analysis. *PLoS Comput Biol*. 2010;6:e1000703.
98. Yosef N, Sharan R, Noble WS. Improved network-based identification of protein orthologs. *Bioinformatics*. 2008;24:i200–i206.
99. Towfic F, VanderPlas S, Oliver CA, et al. Detection of gene orthology from gene co-expression and protein interaction networks. *BMC Bioinformatics*. 2010;11(suppl 3):S7.
100. Park D, Singh R, Baym M, Liao CS, Berger B. IsoBase: a database of functionally related proteins across PPI networks. *Nucleic Acids Res*. 2011;39(Database issue):D295–D300.
101. Sjolander K, Datta RS, Shen Y, Shoffner GM. Ortholog identification in the presence of domain architecture rearrangement. *Brief Bioinform*. 2011;12:413–422.
102. Chiba H, Uchiyama I. Improvement of domain-level ortholog clustering by optimizing domain-specific sum-of-pairs score. *BMC Bioinformatics*. 2014;15:148.
103. Bitard-Feildel T, Kemena C, Greenwood JM, et al. Domain similarity based orthology detection. *BMC Bioinformatics*. 2015;16:154.
104. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology; 1994; AAAI Press, Menlo Park, CA:28–36.
105. Dujon B. Yeast evolutionary genomics. *Nat Rev Genet*. 2010;11:512–524.
106. Milinkovitch MC, Helaers R, Depiereux E, Tzika AC, Gabaldón T. 2x genomes-depth does matter. *Genome Biol*. 2010;11:R16.
107. Leigh JA, Albers SV, Atomi H, Allers T. Model organisms for genetics in the domain archaea: methanogens, halophiles, thermococcales and sulfolobales. *FEMS Microbiol Rev*. 2011;35(4):577–608.
108. Artamonova II, Frishman G, Gelfand MS, Frishman D. Mining sequence annotation databanks for association patterns. *Bioinformatics*. 2005;21(suppl 3): iii49–iii57.
109. 1000 Genomes Project. 2015. Available at: http://www.1000genomes.org/
110. 1001 Genomes Project. 2015. Available at: http://www.1001genomes.org/
111. Dujon B. *Genome Evolution in Yeasts*. Chichester: John Wiley & Sons, Ltd; 2015.
112. Rolland T, Neuvéglise C, Sacerdot C, Dujon B. Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. *PLoS One*. 2009;4:e6515.
113. Gordon JL, Byrne KP, Wolfe KH. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet*. 2009;5:e1000485.
114. Csurös M, Miklós I. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol Biol Evol*. 2009;26: 2087–2095.
115. Chauve C, Gavranovic H, Ouangraoua A, Tannier E. Yeast ancestral genome reconstructions: the possibilities of computational methods II. *J Comput Biol*. 2010;17:1097–1112.
116. Jiang T. Some algorithmic challenges in genome-wide ortholog assignment. *J Comput Sci Technol*. 2010;25:42–52.
117. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol*. 2006;2(6):e69.
118. Arenas M, Posada D. The effect of recombination on the reconstruction of ancestral sequences. *Genetics*. 2010;184(4):1133–1139.
119. Butzin NC, Lapierre P, Green AG, Swithers KS, Gogarten JP, Noll KM. Reconstructed ancestral myo-inositol-3-phosphate synthases indicate that ancestors of the Thermococcales and Thermotoga species were more thermophilic than their descendants. *PLoS One*. 2013;8:e84300.
120. Risso VA, Gavira JA, Gaucher EA, Sanchez-Ruiz JM. Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins*. 2014;82:887–896.
121. Pennisi E. Evolution. Building the tree of life, genome by genome. *Science*. 2008; 320:1716–1717.
122. Pennisi E. Human genome 10th anniversary. Tracing the tree of life. *Science*. 2011; 331:1005–1006.
123. Durand D, Halldorsson BV, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*. 2005;13:320–335.
124. Vernot B, Stolzer M, Goldman A, Durand D. Reconciliation with non-binary species trees. *J Comput Biol*. 2008;15:981–1006.
125. Conte MG, Gaillard S, Droc G, Perin C. Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. *BMC Genomics*. 2008;9:183.
126. Sennblad B, Lagergren J. Probabilistic orthology analysis. *Syst Biol*. 2009;58: 411–424.
127. Thomas PD. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*. 2010;11:312.
128. Than CV, Rosenberg NA. Consistency properties of species tree inference by minimizing deep coalescences. *J Comput Biol*. 2011;18:1–15.
129. Kuraku S. Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy on hagfish and lamprey gene phylogeny. *Integr Comp Biol*. 2010;50: 124–129.
130. Steinfartz S, Weitere M, Tautz D. Tracing the first step to speciation: ecological and genetic differentiation of a salamander population in a small forest. *Mol Ecol*. 2007;16:4550–4561.
131. Studer RA, Robinson-Rechavi M. How confident can we be that orthologs are similar, but paralogs differ? *Trends Genet*. 2009;25:210–216.
132. Agnan J, Korch C, Selitrennikoff C. Cloning heterologous genes: problems and approaches. *Fungal Genet Biol*. 1997;21:292–301.
133. Thierry A, Khanna V, Créno S, et al. Macrotene chromosomes provide insights to a new mechanism of high-order gene amplification in eukaryotes. *Nat Commun*. 2015;6:6154.
134. Yu H, Braun P, Yildirim MA, et al. High-quality binary protein interaction map of the yeast interactome network. *Science*. 2008;322:104–110.
135. Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Andrej Sali A. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci*. 2009;18:1306–1315.
136. Forslund K, Pekkari I, Sonnhammer EL. Domain architecture conservation in orthologs. *BMC Bioinformatics*. 2011;12:326.
137. Gabaldón T, Koonin EV. Functional and evolutionary implications of gene orthology. *Nat Rev Genet*. 2013;14(5):360–366.
138. Gabaldón T, Dessimoz C, Huxley-Jones J, Vilella AJ, Sonnhammer EL, Lewis S. Joining forces in the quest for orthologs. *Genome Biol*. 2009;10:403.
139. Trachana K, Forslund K, Larsson T, et al. A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS One*. 2014;9(11):e111122.