HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2016 October 04.

Published in final edited form as:

Nat Methods. 2016 May ; 13(5): 425–430. doi:10.1038/nmeth.3830.

Standardised Benchmarking in the Quest for Orthologs

Adrian M. Altenhoff[1,2], **Brigitte Boeckmann**[3], **Salvador Capella-Gutierrez**[4,5,6], **Daniel A. Dalquen**[7], **Todd DeLuca**[8], **Kristoffer Forslund**[9], **Jaime Huerta-Cepas**[9], **Benjamin Linard**[10], **Cécile Pereira**[11,12], **Leszek P. Pryszcz**[4], **Fabian Schreiber**[13], **Alan Sousa da Silva**[13], **Damian Szklarczyk**[14,15], **Clément-Marie Train**[1], **Peer Bork**[9,16,17], **Odile Lecompte**[18], **Christian von Mering**[14,15], **Ioannis Xenarios**[3,19,20], **Kimmen Sjölander**[21], **Lars Juhl Jensen**[22], **Maria J. Martin**[13], **Matthieu Muffato**[13], **the Quest for Orthologs consortium**, **Toni Gabaldón**[4,5,23], **Suzanna E. Lewis**[24], **Paul D. Thomas**[25], **Erik Sonnhammer**[26], and **Christophe Dessimoz**[7,20,27,28,29,*]

[1]Dept. of Computer Science, ETH Zurich, 8092 Zurich, Switzerland [2]Computational Biochemistry Research Group, Swiss Institute of Bioinformatics (SIB), Zurich, Switzerland [3]Swiss-Prot Group, Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland [4]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona, Spain [5]Universitat Pompeu Fabra (UPF), Barcelona, Spain [6]Yeast and Basidiomycete Research Group, CBS Fungal Biodiversity Centre, Utrecht, The Netherlands [7]Dept. of Genetics, Evolution, and Environment, University College London, London, UK [8]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA [9]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany [10]Dept. of Life Sciences, Natural History Museum, London, UK [11]Univ Paris-Sud, Laboratoire de Recherche en Informatique, Orsay, France [12]Univ Paris-Sud, Institute for Integrative Biology of the Cell (I2BC), Orsay, France [13]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK [14]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland [15]Bioinformatics / Systems Biology Group, Swiss Institute of Bioinformatics (SIB), Zurich, Switzerland [16]Germany Molecular Medicine Partnership Unit (MMPU), University Hospital Heidelberg and European Molecular Biology Laboratory, Heidelberg, Germany [17]Max Delbrück Centre for Molecular Medicine, Berlin, Germany [18]LBGI, Computer Science Department, ICube,

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

[*]Correspondence: ; Email: Christophe.Dessimoz@unil.ch

Author contributions

AMA and CD conceived the project, with contributions from the rest of the QfO benchmarking working group (ASdS, BB, CP, ES, FS, KF, KS, JHC, MJM, MM, PDT, SEL, TG); ASdS, MJM contributed the reference proteomes; AMA, BB, FS, KF contributed benchmarks; AMA, BL, DS, ES, KF, JHC, LPP, MM, SCG, PDT, TDL, TG contributed predictions; PB, OL, CvM, IX, LJJ supervised contributions of benchmarks or predictions. AMA and CMT designed and implemented the benchmark service; DAD and AMA assessed the impact of incomplete lineage sorting on orthology inference; AMA and JHC generated plots; AMA, BB, CD, ES, FS, KF, KS, MM, PDT, SEL, TG analysed the results; AMA, BB, CD, ES, LJJ wrote the manuscript, with feedback from all other co-authors; CD coordinated the project.

Competing financial interests

The authors declare no competing financial interests.

Code Availability

The source code is available under an open source license (Mozilla Public License Version 2.0) at https://github.com/qfo/benchmark-webservice.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

University of Strasbourg, Strasbourg, France [19]Vital-IT, Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland [20]Center for Integrative Genomics (CIG), University of Lausanne, Lausanne, Switzerland [21]Dept. of Bioengineering, University of California, Berkeley, California, USA [22]The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark [23]Institució Catalana de Recerca I Estudis Avançats (ICREA), Barcelona, Spain [24]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA [25]Division of Bioinformatics, Dept. of Preventive Medicine, University of Southern California, Los Angeles, California, USA [26]Stockholm Bioinformatics Center, Dept. of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Solna, Sweden [27]Dept. of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland [28]Dept. of Computer Science, University College London, UK [29]Swiss Institute of Bioinformatics, Biophore Building, Lausanne, Switzerland

Abstract

The identification of evolutionarily related genes across different species—orthologs in particular—forms the backbone of many comparative, evolutionary, and functional genomic analyses. Achieving high accuracy in orthology inference is thus essential. Yet the true evolutionary history of genes, required to ascertain orthology, is generally unknown. Furthermore, orthologs are used for very different applications across different phyla, with different requirements in terms of the precision-recall trade-off. As a result, assessing the performance of orthology inference methods remains difficult for both users and method developers. Here, we present a community effort to establish standards in orthology benchmarking and facilitate orthology benchmarking through an automated web-based service (http://orthology.benchmarkservice.org). Using this new service, we characterise the performance of 15 well-established orthology inference methods and resources on a battery of 20 different benchmarks. Standardised benchmarking provides a way for users to identify the most effective methods for the problem at hand, sets a minimal requirement for new tools and resources, and guides the development of more accurate orthology inference methods.

Introduction

Evolutionarily related genes (homologs) across different species are often divided into gene pairs that originated through speciation events (orthologs) and those that originated through duplication events (*paralogs*)[1]. This distinction is useful in a broad range of contexts, including phylogenetic tree inference, genome annotation, comparative genomics, and gene function prediction[2–4]. Accordingly, dozens of methods[5] and resources[6–8] for orthology inference have been developed.

Because the true evolutionary history of genes is typically unknown, assessing the performance of these orthology inference methods is not straightforward. Several indirect approaches have been proposed. Based on the notion that orthologs tend to be functionally more similar than paralogs (a notion now referred to as the ortholog conjecture[9–12]), Hulsen *et al.*[13] used several measures of functional conservation (co-expression levels, protein-protein interactions, protein domain conservation) to benchmark orthology inference

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

methods. Chen *et al.*[14] proposed an unsupervised learning approach based on consensus among the different orthology methods tested. Altenhoff and Dessimoz[15] introduced a phylogenetic benchmark measuring the concordance between gene trees reconstructed from putative orthologs and undisputed species trees. More recently, several recent benchmarks have used “gold standard” reference sets, either manually curated[16,17] or derived from trusted resources[18]. Finally, Dalquen *et al.*[19] used simulated genomes to assess orthology inference in the presence of varying amounts of duplication, lateral gene transfer, and sequencing artifacts.

However, this wide array of different benchmarking approaches poses considerable conceptual and practical challenges to orthology methods developers and users. Conceptually, the choice of an appropriate benchmark strongly depends on the application at hand. Practically, most methods are not available as standalone programs and thus cannot be easily compared on a common set of data. Likewise, some benchmarks rely on complex pipelines which may be difficult to implement. If public results are available as part of a publication or a resource, inconsistent genome releases or identifiers severely complicate comparisons. Finally, some methods or benchmarks can be computationally costly to run. As a result, users cannot easily identify appropriate tools and methodological progress is hampered.

Here, we report on a community effort to standardise and facilitate orthology benchmarking. This entailed establishing a shared reference dataset and developing a web-based service for automatic orthology benchmarking (http://orthology.benchmarkservice.org). Building upon these new resources, a community experiment was run to assess 15 well-established orthology inference methods and resources on a wide array of phylogenetic and functional benchmarks. This constitutes the most comprehensive survey of state-of-the-art orthology resources to date. Furthermore, by providing a way to automatically include new methods and disseminate results publicly, this effort holds the promise of remaining continuously up-to-date to the benefit of orthology users and developers.

Results

We first provide an overview of the benchmark service and of the orthology inference methods tested. Then, we present the results obtained on the benchmarks grouped into three main categories: species discordance tests, reference gene trees, and functional tests. The benchmark service alone required evaluation of 70,390,701 orthologous relationships and the inference of 233,000 phylogenetic trees.

Benchmark service

To automate ortholog benchmarking on a broad range of tests (detailed below), we developed a publicly accessible web service. A schematic overview of the workflow is depicted in Figure 1. First, an orthology method developer infers orthologs using the Quest for Orthologs (QfO) reference proteome dataset. Orthology inference methods vary in the kind of output they provide—e.g. labelled gene trees, orthologous groups—but it is usually possible to reduce these to orthologous pairs, which thus constitute a natural “common denominator” for benchmarking. The benchmark service accepts these pairwise orthologs

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

predictions in OrthoXML[20] or tab-delimited format. As the OrthoXML format also supports InParanoid-style clusters and hierarchical orthologous groups, the service can automatically convert these to pairwise relationships. Next, the service ensures that only predictions among valid reference proteomes are provided (with scoring implicitly assuming that the uploaded inferences are complete). Benchmarks are then selected and run in parallel—some may take up to several hours. Finally, statistical analyses of method accuracy on each benchmark dataset are performed. Where possible, the accuracy is measured in terms of precision (also known as positive predictive value—the proportion of ortholog predictions that are correct) and recall (also known as sensitivity, or true positive rate—the proportion of actual orthologs that are correctly predicted). The raw data and results are stored and provided to the submitter, who can choose to make the results publicly available. For the sake of transparency and to encourage improvement to the service, its source code is released under an open source license (Mozilla Public License Version 2.0) at https://github.com/qfo/benchmark-webservice.

Methods investigated

The methods investigated here include a broad array of well-established methods. Three of them are tree-based methods: Ensembl Compara[21], PANTHER 8.0[22], and PhylomeDB[23]. Seven are graph-based (i.e. based on pairwise comparisons): Best Reciprocal Hits (BRH)[24], Reciprocal Smallest Distance (RSD)[25], EggNOG[26], Hieranoid[27], InParanoid[28], OMA[29], OrthoInspector[30]. Finally, MetaPhOrs[31] is a meta-method incorporating both tree-based and graph-based methods. For some of the methods, multiple variants are included in the analysis. Details are provided in the Methods section. Each method inferred orthologs on the 754,149 protein sequences from 66 reference genomes, except for MetaPhOrs, which inferred orthologs on all but three prokaryotes (see *Methods*).

Generalised species tree discordance test

Orthology was first defined in the context of species tree inference, which requires genes related through speciation[1]. The species tree discordance test exploits this connection by assessing the accuracy of orthologs in terms of the accuracy of the species tree that can be reconstructed from them[15]. The two main limitations of the original protocol were the species tree “comb” topology (a specific type of tree in which all bifurcations occur along a single path) and the small number of taxa (6). Here we overcome these two limitations by generalising the orthology sampling procedure to any tree topology and by employing larger reference trees from the SwissTree initiative. Furthermore, to minimise the possibility of gene-species tree discordance due to incomplete lineage sorting, we avoided sampling orthologs among species separated by branches shorter than 10 myr (see *Methods* and Supplementary Fig. 1).

We observed different trade-offs between average discordance (Robinson-Foulds[32] distance, as a proxy for the false-discovery rate—the complement of precision) and the number of trees that can be sampled (proxy for recall) across all methods (Fig. 2). An ideal method would be placed in the lower right corner. When considering eukaryotes, results with highest precision and lowest recall were obtained with OMA groups. At the other extreme, PANTHER 8.0 (all) tended to yield the highest recall and lowest precision results. Among

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the more balanced methods, no method consistently obtained a better balance than the other methods across all datasets, but for instance Orthoinspector, InParanoid, and PANTHER (LDO only) performed well overall. In terms of broad categories of methods, the results are interlaced, with no obvious systematic difference in performance between tree-based orthology inference methods (Ensembl, PANTHER, PhylomeDB) and graph-based orthology inference methods (the rest), or between methods relying on species tree (Ensembl, PANTHER, PhylomeDB, OMA GETHOGs, Hieranoid, EggNOG) and those that do not. The latter point is perhaps unexpected, as one could expect knowledge of the species tree to provide an "unfair" advantage in this particular benchmark. If there is any such effect, our results indicate that it is small.

These trends held when we measured recall in terms of the number of inferred orthologs (Supplementary Fig. 2) or if we focused on other clades (Supplementary Fig. 3–5). Among vertebrates, the results were largely consistent, but we note minor differences in the ranking of individual methods, with InParanoid Core yielding the highest precision and MetaPhOrs the highest recall (Supplementary Fig. 3). We also benchmarked the methods for their ability to recover ortholog relationships among "universal" genes, by applying the species discordance test on a tree spanning across archaea, bacteria, and eukaryotes. Once again, there were slight variations in the precise ranking of methods, but the overall trends were very similar to what is observed for eukaryotes only (Supplementary Fig. 5). Finally, if we included (high-confidence) short branches as well, the average concordance of reconstructed trees substantially decreased—both because short branches tend to be harder to infer and because of potential incomplete lineage sorting around them; interestingly however, the relative position of the methods remained practically unchanged, which is a further indication of the robustness of the benchmark (Supplementary Fig. 6).

Reference gene trees

The second series of orthology benchmarks employs evolutionary relationships of gene pairs derived from annotated high-quality gene trees. Such reference trees are inferred through a careful combination of computational inference and expert curation: results obtained at each step of the tree inference pipeline (homolog identification, alignment, tree inference, gene-species tree reconciliation) are individually inspected, poor-quality sequences are excluded from the analysis, and results are typically assessed using multiple models. This manual oversight is expected to yield gene phylogenies with high statistical support and topological consistency.

Concordance of orthology predictions was assessed with two sets of trees. The first was SwissTree[16,33], a small collection of large and high-confidence gene family phylogenies with different types of challenges for orthology prediction and species from all domains. The second, TreeFam-A[34], consisted of a larger set of metazoan gene trees and thus covers a taxonomically restricted but wider range of protein families. At first glance, the results obtained with the two benchmarks look different (Fig. 3a and b), but on closer inspection the similarity of the outcome of the two tests becomes apparent. Strikingly, on these benchmarks, virtually no trade-off between precision and recall appears to be necessary. The best performing methods were the ones that adopt a balanced precision-recall strategy, with

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

MetaPhOrs doing particularly well. Methods with a more skewed precision-recall strategy (in particular stringent OMA groups and permissive PANTHER (all)) fare poorly in comparison. In part, this may be due to the nature of the reference gene tree dataset, which focuses on gene families with a tractable evolutionary history. On ambiguous phylogenies, mistakes would become unavoidable and a skewed strategy could become preferable, depending on the application.

Functional benchmarks

The third series of benchmarks evaluated orthology in terms of their functional similarity. Although orthology is an evolutionary and not a functional relationship, we chose to include functional benchmarks for two reasons. First, for similar levels of sequence divergence, orthologs have been shown to be moderately, but significantly, more conserved than paralogs in terms of Gene Ontology (GO) annotation similarity[11]. For a given evolutionary distance, more accurate orthology inference is thus likely to be correlated with functionally more similar gene pairs. Second, many users are interested in orthologs to identify functionally conserved genes; for them, functional benchmarks are thus directly relevant.

We assessed functional similarity based on experimentally backed annotations from the UniProt-Gene Ontology Annotation (GOA) database[35] and Enzyme Commission (EC) numbers from the ENZYME database[36]. Though the two benchmarks consider different aspects of gene function, the results were largely consistent: in both cases, orthology inference methods showed a clear trade-off between precision (measured as the average Schlicker semantic similarity[37] of functional annotations associated with orthologs) and recall (measured as the number of ortholog relationships predicted) (Fig. 4). The only exception was with the EC number benchmark, where MetaPhOrs falls beneath the “Pareto frontier” (the frontier defined by the methods that are not dominated by any other method in both precision and recall). However, MetaPhOrs is also the only method with missing taxa, and the three missing ones contain a substantial number of genes with EC annotations (827 in total). This has a negative effect on the recall.

Discussion and outlook

For orthology methods developers, the orthology benchmark service overcomes many of the practical complications previously associated with orthology benchmarking. It enables systematic comparison of a new method with state-of-the-art approaches on a wide range of benchmarks. This sets a vastly better standard than the current benchmarking practice, which typically includes fewer methods, fewer tests, and less empirical data.

By relying on a common set of data for all methods, the benchmark service ensures that the results obtained by different methods are directly comparable. This also constitutes a substantial improvement over previous benchmarking efforts, which required painstaking and error-prone mapping of proteins between different sources, releases, and choice of alternative splicing variants. The only caveat is that, since proteomes vary in quality and difficulty to analyse, the results on the benchmark dataset may not entirely reflect the quality of the orthology assignments provided by each resource. The choice of species included in the QfO reference proteomes (see *Methods*) requires a compromise between i) increasing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the number of proteomes to make the benchmark set more representative of current resources, and ii) keeping the number of proteomes low to facilitate and encourage new submissions to the benchmark.

Submissions performed on a subset of the proteomes are discouraged, as all missing predictions are counted as false negatives. This provides an incentive for submitters to analyse the entire reference proteome dataset. Instead of penalising for missing predictions, we also considered alternative ways of handling submissions on partial data but they had major flaws: one alternative was to extrapolate scores obtained on the subset of proteomes considered in a particular submission to all data. However, this can introduce a bias in the analyses (e.g. some methods only predict orthologs for “easy” pairs of proteomes). Another alternative was to restrict comparisons to the intersection of proteomes analysed by all methods. This is, however, excessively wasteful of information, as the intersection can only decrease with each additional method.

Overall, results obtained across multiple phylogenetic and functional tests corroborate previous observations that the main difference among the established orthology inference methods tested here lies in the trade-off they produce in terms of precision and recall[13,15,17]. This trade-off was however not present in the reference gene tree test, perhaps because sequences with ambiguous location are typically excluded from these hand-curated trees. On these reference trees, the meta-method MetaPhOrs performed particularly well. The analysis also confirms that the widely-used reciprocal best hit approach has a relatively high precision, but relatively low recall[38,39]. Other methods fill different niches, with OMA group and PANTHER (all) often lying at the two extremes of the precision-recall trade-off. Among the more balanced approaches, InParanoid, Hieranoid, and OrthoInspector showed solid performance in most benchmarks.

Whether to favour a skewed or balanced approach to the precision-recall trade-off strongly depends on the context of application. For instance, hypotheses-generating analyses may favour a high recall, while phylogenomic species tree inference typically require high precision. Because of this, we refrain from computing a combined score, which would necessarily entail a statement of preference with respect to this trade-off.

To be deemed competitive, a method should ideally reach or exceed the Pareto frontier in a least a subset of the benchmarks. If not, the benchmark service may help uncover bugs or deeper flaws. Analogous to unit testing in software engineering, benchmarking can also provide quality control for new releases of established resources. As it happens, in the course of the present community benchmarking effort, over a hundred datasets have been submitted to the service. Many submitters did not make their results publicly available, presumably after discovering poor outcome in some of the benchmarks. This clearly demonstrates the effectiveness of the benchmark service for quality control.

The bane of benchmarking is circularity. Unfortunately, despite our best efforts, not all of it could be avoided. As mentioned above, some methods use knowledge of the species tree in their inference—though judging by their performance as a group, any unfair advantage this may confer seems negligible. More generally, many methods were trained or fine-tuned

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

using some of the benchmarks considered here. For instance, parameters of the meta-method MetaPhOrs were in part trained using TreeFam-A[31]. Similarly, the latest version of InParanoid[28] and PhylomeDB[23] used the benchmark service for parameter fine-tuning. As for the functional benchmarks, although GO annotations derived from sequence comparisons were excluded, experiments are very often guided by sequence similarity to proteins with known function. Thus, even when restricting analyses to experimentally backed GO annotations, we cannot entirely avoid circularity. Taken together, however, because the benchmarks are collectively underpinned by a large amount of data from a broad range of species (tens of thousands of trees, hundreds of thousands of pairs of functional annotations), the risk of overfitting seems low. Notwithstanding that, this potential risk will be monitored by the QfO benchmark working group and new benchmarks may be introduced over time to detect and discourage overfitting.

Presently, the benchmark service uses orthologous gene pairs as "common denominator" among all the methods. However, one should remember that many resources provide richer outputs—such as reconciled gene trees or hierarchical orthologous groups—and may indeed be optimised for these. The performance on pairwise data are thus not entirely representative of what they offer. In the future, however, the benchmark service could be extended to evaluate these richer, more specific orthology formats as well. Similarly, the benchmark service could also be extended to take into account confidence score or posterior probabilities, which are particularly relevant to likelihood-based orthology inference methods[40,41].

Meanwhile, for end-users of orthology predictions, the benchmark service provides the most comprehensive survey of methods to date. Furthermore, because it can process new submissions automatically and continuously, it holds the promise of remaining current and relevant over time. The benchmark service thus enables users to gauge the quality of the orthology calls upon which they depend, and to identify the methods most appropriate to the problem at hand.

Methods

Quest for Orthologs Reference Proteomes and Species Tree

The QfO consortium has defined a consensus dataset of proteomes and common file formats[6,7] to be used by diverse orthology inference methods, allowing for standardised benchmarks and to aid integration of multiple ortholog sources. The QfO Reference Proteomes datasets were created as a collection of data providing a representative protein for each gene in the genome of selected species. Such datasets have been generated annually from the UniProt Knowledgebase (UniProtKB) database[42] for the past five years. To this end, a gene-centric pipeline has been developed and enhanced over these years by UniProt. The QfO Reference Proteomes are a manually compiled subset of the UniProt reference proteomes, comprising well-annotated model organisms and organisms of interest for biomedical research and phylogeny, with the intention to provide broad coverage of the tree of life. These complete, non-redundant reference proteomes are publicly available at ftp://ftp.ebi.ac.uk/pub/databases/reference_proteomes/QfO. The datasets are provided either in SeqXML[20] format or as a collection of FASTA files.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The benchmarking effort reported here uses the reference proteomes dataset released in 2011, which comprises 754,149 non-redundant protein sequences from 66 species (40 eukaryotes and 26 bacteria/archaea).

The reference species tree used in this study was produced by the QfO species tree working group, who surveyed the literature to establish a well-supported tree topology for the 66 species[43] (Supplementary Fig. 1). The internal nodes of this reference species tree have assigned confidence levels based on the agreement among the resources surveyed (L90: congruent, significant branch support; L70: congruent, L50: one alternative species tree topology, L30: default level, L10: two or more alternative species tree topologies have been reported; for more detail, see Boeckmann et al.[43]). The latest version of the tree can be retrieved from http://swisstree.vital-it.ch/species_tree. To minimize the chance of including cases of incomplete lineage sorting in the species tree discordance benchmark, we estimated the evolutionary times of all internal branches using the timetree resource[44] and collapsed branches that were shorter than 10 myr.

Orthology databases and methods

EggNOG[26] (http://eggnogdb.embl.de) is a database of Orthologous Groups (OGs) and functional annotation covering prokaryotic and eukaryotic species. Since version 4.1, the EggNOG method is also capable of producing fine-grained (e.g. pairwise) orthology predictions based on the automated analysis of phylogenetic trees. For this study, the complete set of 66 reference proteomes was independently analyzed using the EggNOG pipeline, which involved 1) joining proteins into in-paralogous groups from closely related species and 2) de novo reconstruction of 38,513 OGs by clustering the obtained in-paralogous groups based on triangles of their reciprocal best hits[45]. Phylogenetic analysis and automated tree interpretation for each OG was subsequently performed using the workflow described in PhylomeDB22 as implemented in the ETE Toolkit v2.3[46]. The phylogenetic approach used included testing three aligners (MAFFT[47] v6.861b, Muscle[48] v3.8.31 and Clustal Omega[49] v1.2.1) and five evolutionary models (LG, WAG, JTT, VT and MtREV); applying alignment consensus and soft trimming techniques (M-Coffee[50] v10, trimAl[51] v1.3); and using maximum likelihood tree inference (PhyML[52] v3). This workflow is labeled as eggnog41 when using the ETE-build command and was applied in a per OG basis. Pairwise orthology predictions were derived from each tree using the species overlap algorithm[53] after rooting trees to midpoint. The predictions were submitted to the benchmark service in July 2015.

Ensembl Compara[21] uses a gene-species tree reconciliation pipeline. The predictions were run using the code released in the version 81 of the Ensembl (July 2015). However, Treebest (the software used to build phylogenetic trees) had to be adapted to accept alignments of protein sequences. Treebest makes a consensus out of trees built with various phylogenetic methods and some of them required nucleotide sequences, which were not provided in the QfO dataset. The list of maximum-likelihood models and distance methods (used for neighbour-joining) was thus updated to: WAG, JTT and Dayhoff instead of WAG and HKY (for maximum-likelihood), and JTT, Kimura and mixed amino-acid models instead of dN, dS and mixed nucleotide models (for neighbour-joining). The predictions were submitted to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the benchmark service in June 2015. An older submission based on the version 66 of the Ensembl code (June 2011) is also present on the benchmark service.

Hieranoid[27] performs pairwise orthology analysis using InParanoid at each node in a guide (species) tree as it progresses from its leaves to the root. This concept reduces the total runtime complexity from a quadratic to a linear function of the number of species. We ran Hieranoid 2.0. Hieranoid outputs ortholog groups structured as species trees with orthologs at all levels, hence there can be many outparalogs within an ortholog group. The trees were therefore parsed to extract ortholog pairs only at the last common ancestor of two species, for all species pairs. The predictions were submitted to the benchmark service in April 2015.

InParanoid[28] is a graph-based algorithm, which aims to generate orthologous groups that include all inparalogs but no outparalogs between species pairs. Version 4.1 of the algorithm was run with default parameters. Two variants were tested in this study: the regular InParanoid output containing all predicted pairs of orthologs (labelled InParanoid in the plots) and a high-confidence set including only orthologs with InParanoid's maximum confidence score of 1.0 (labelled Inparanoid (core)). The predictions were submitted to the benchmark service in June 2011.

MetaPhOrs[31] (Meta Phylogeny-based Orthologs) is a repository of orthologs and paralogs that were computed using phylogenetic trees available in several databases or computed from graph-based orthologous groups. For each orthology/paralogy prediction, MetaPhOrs (http://orthology.phylomedb.org/) provides two reliability scores: Evidence Level (informing about number of repositories from which prediction is retrieved) and Consistency Score (defining overall agreement of source databases about given prediction). MetaPhOrs does not include predictions for the three reference genomes Streptomyces coelicolor, Thermotoga maritima and Pyrococcus kodakaraensis (strain KOD1). The predictions were submitted to the benchmark service in February 2013.

OMA[29] (Pairs, Groups, HOGs) is a publicly available resource (http://omabrowser.org/) that provides orthology predictions among thousands of proteomes from all domains of life. OMA uses evolutionary distance estimates from Smith–Waterman alignments to infer orthologs. A distinct feature among graph based methods is the witness of non-orthology step in its pipeline, where cases of differential gene losses get detected. OMA provides three different groupings of orthologs: (i) the raw pairwise ortholog relationships form the OMA Pairs, a gene centric view that lists all the orthologs for a given gene. (ii) OMA Groups, a very stringent type of grouping where all member proteins are orthologous to one another within group. OMA Groups have been designed mainly for species tree inference purposes, as gene trees built from them should be congruent with the species tree. (iii) Lastly, we construct hierarchical orthologous groups (OMA HOGs). These are nested groups which contain genes that descend from a single common ancestral gene within a given taxonomic range using the GETHOGs algorithm[54]. The predictions were submitted to the benchmark service in June 2011 (OMA pairs and groups) and in March 2013 (OMA HOGs).

OrthoInspector[30] is a database of precomputed orthology and inparalogy relationships and a stand-alone package allowing large-scale predictions of orthology between thousands of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

proteomes (http://lbgi.fr/orthoinspector/). The resource has recently undergone a major new release, with improved speed and visualisation tools, but the inference algorithm is unchanged from the initial graph-based method described in Linard et al.[55] The predictions were submitted to the benchmark service in June 2011.

PANTHER 8.0[22] is based on version 8.0 of the PANTHER database (http://pantherdb.org), released in 2012 (the current version is 10.0, released in 2015). Family membership of each sequence is based on HMM scoring to the PANTHER "library" of HMMs (at both the family and subfamily levels). Sequences were aligned with MAFFT[56] and the resulting alignment was used to construct phylogenetic trees with the GIGA program[57]. GIGA (version 1.1 was used for PANTHER version 8.0) uses a species tree to guide tree construction, and all nodes in the tree are labeled as speciation or gene duplication events; these labeled nodes are used to infer orthologs (pairs of genes with a speciation event as their common ancestor). PANTHER predicts two types of orthologs: least-diverged orthologs (LDO) and other orthologs (O). LDO pairs can be simplistically thought of as "the same gene" in two different species. Formally, the two genes created by each gene duplication event in the tree are treated asymmetrically: the least diverged duplicate (the one with the shortest branch immediately following the duplication) remains in the same LDO group as its ancestor, while the other duplicate founds a new LDO group. The benchmarking was performed on either LDO only, or all orthologs (including both LDO and O). The predictions were submitted to the benchmark service in February 2013.

PhylomeDB[23] (http://phylomedb.org/) is a publicly available repository of phylomes, i.e. the complete collection of phylogenies for all genes of a given species in a predefined evolutionary context. PhylomeDB is unique among other repositories in that it follows an approach that is both gene-centric and genome-wide. PhylomeDB uses its phylogenetic trees to infer orthology and paralogy relationships. For the Quest for Orthologs project, 42 phylomes were reconstructed using different combinations of the 66 species in the benchmark. A total of 458,108 phylogenetic trees were generated, which were later on combined to provide orthology predictions for all proteins included in the benchmark. Briefly, each tree was scanned and only the partition of up to 30 sequences including the seed protein was kept. Then, evolutionary relationships were computed for those protein sequences based on a species overlap approach. Redundant predictions across the 42 phylomes were unified using the Consistency Score (CS) as implemented in MetaPhOrs (see above). Only those predictions having a Consistency Score greater or equal to 0.5 across the whole dataset were called orthologs. The predictions were submitted to the benchmark service in June 2013.

RBH[24] (Reciprocal best hit) is a classic method consisting in identifying the pairs of genes with mutually highest alignment score between every pair of species. Here, we use reciprocal blastp hits as orthologs, with minimum E-value of 1e-2, and keep all hits that are ≥99% of the highest score. The predictions were submitted to the benchmark service in January 2016.

RSD[25] (Reciprocal smallest distance) infers orthology relationships by finding pairs of genes whose nearest gene, computed using PAML, is the other gene in the pair. Candidates

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

genes are also filtered using BLAST E-value and multiple-sequence alignment divergence thresholds. This method is implemented in the database RoundUp[58], a large-scale orthology database developed by the Wall Lab. The database is no longer maintained, but the source code is still available at https://github.com/todddeluca/reciprocal_smallest_distance/. To identify orthologs, we ran the algorithm with divergence and E-value cutoffs of 0.8 and 1e-5, respectively. The predictions were submitted to the benchmark service in February 2012.

Benchmarks

Generalised species tree discordance—The idea behind the species tree discordance test is simple. Two genes are orthologous if they started diverging through a speciation event. Therefore, if we sample putative orthologous genes such that all resulting genes are related through speciation events, the resulting tree should be congruent with the species tree. Previously, we presented a sampling strategy for fully imbalanced tree topologies[15]. Here, we extend this idea to arbitrary reference trees—including those with soft polytomies (unresolved nodes).

The following procedure is repeated a large number of times: we start with a random gene in a random genome. We then attempt to sample a maximal path along the tree by selecting an orthologous gene in the “next” species in the tour from the list of reported orthologs (Supplementary Fig. 7a). If there are multiple possibilities in the choice of the “next” species due to soft polytomies, or in the choice of the orthologous counterparts due to one-to-many or many-to-many orthology, a choice is made at random. If there is no predicted ortholog at any step along the path, the sample is deemed unsuccessful. Alternatively, if at least one orthologous counterpart is predicted at each step, this results in a set of n sequences. Assuming that i) the reference tree is correct, ii) the retrieved orthologs are all correct, and iii) all within-species variation are fixed (i.e. no incomplete lineage sorting), it is easy to prove that the unrooted evolutionary tree relating these sequences should only contain speciation nodes and should therefore be congruent with the reference species tree.

Proof: The n sequences sampled through the circular tour are sampled by starting from a random sequence and retrieving n−1 pairs of orthologs. By construction, these n−1 pairs of orthologs belong to pairs of species that have distinct last common ancestors and thus coalesce in different speciation nodes in the phylogenetic tree of these sequences. Therefore, that tree contains at least n−1 distinct speciation nodes. However, the rooted, fully-resolved evolutionary tree of n species has exactly n−1 internal nodes. Thus, all the internal nodes of the gene tree are speciation nodes. Since we assume that there is no incomplete lineage sorting, as long as the input orthologs are correct, the tree relating these sequences should therefore be congruent with the species tree.

A least-squares distance tree is reconstructed for each set of putative orthologous sequences. After aligning the sequences with MAFFT[47], maximum likelihood distances and their variances (using the inverse Fisher information) are estimated using the EstimatePam() function in the Darwin programming environment[59] for each pair of sequences. Next, the gene tree is estimated using Darwin’s MinSquareTree() function, which is a fast implementation of the weighted least squares trees[60] constrained to non-negative branch

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

lengths[61]. We have previously shown that orthology benchmarking results obtained with such distance trees are consistent with more computationally demanding Maximum likelihood trees[15]. The Robinson–Foulds[32] distance between this gene tree and the reference tree measures the false discovery rate, while the total number of trees is used as a proxy of recall. Due to the stochastic nature of the algorithm, repeated runs of the benchmark may lead to slightly (albeit non-significantly) different results.

Reference gene trees—Reference gene trees labeled with speciation and duplication events were downloaded from SwissTree on March 23, 2015 (http://swisstree.vital-it.ch/) and Treefam-A version 7 (http://www.treefam.org/). As sequences analysed in these two resources can differ from those of the QfO reference proteomes, sequences were mapped based on gene identifiers or sequence identifiers. After mapping, for each family the $n(n - 1) / 2$ induced pairwise evolutionary relationships were extracted and compared with the orthologous predictions from each orthology prediction method as follows. Let $G = \{g_i\}$ be the set of all genes in the reference gene tree and $R_O = \{(g_i, g_j)\} \mid g_i \in G, g_j \in G, g_i \neq g_j, label(g_i,g_j) = speciation$ the set of true orthologs according to the reference tree. Likewise, let R_P be the set of non-orthologous relations in that family and $P = \{(g_i, g_j)\}$, be the set of all predictions made by the orthology method. With $P_F = \{(g_i, g_j)\} \mid (g_i, g_j) \in P \cap g_i \in G \cap g_j \in G$, we denote the set of orthologs where both members are part of the reference gene family. Now, the true/false positives/negatives are simply $TP = P_F \cap R_O$, $FP = P_F \cap R_P$, $FN = R_O - P_F$ *and* $TN = R_P - P_F$. From these values we can compute positive predictive values (PPV) and true positive rate (TPR): $PPV = |TP| / (|TP| + |FP|)$, $TPR = |TP| / (|TP| + |FN|)$.

We can further estimate the uncertainties of these rates by treating them as binomially distributed random variables, e.g. $\sigma^2(\mathrm{PPV}) = \mathrm{PPV}(1 - \mathrm{PPV}) / (|\mathrm{TP}| + |\mathrm{FP}|)$. Finally, we combine all the families by building averages of the rates. As an example, for the positive predictive value this results in

$$\mathrm{avgPPV}=1/\mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{PPV}_{\mathrm{i}},\ \text{and}\ \sigma^2(\mathrm{avgPPV})=1/\mathrm{n}^2\sum_{\mathrm{i}=1}^{\mathrm{n}}\sigma^2(\mathrm{PPV})_{\mathrm{i}}.$$

Functional tests—We downloaded the Gene Ontology annotations[62] for all the genes in the reference genomes from the Nov 2014 release of UniProt-GOA[35] and excluded any annotation with a "NOT" qualifier from this set. For the analysis shown in here, we only use annotations with experimental evidence codes (EXP, IPI, IDA, IMP, IGI, IEP). Likewise, we collected the hierarchical EC number assignments of the ENZYME database[36], maintained by Swiss-Prot. The computation of the functional similarities between gene pairs is done in the same way for both types of data, using the approach of Schlicker et al[37]: the semantic similarity between annotations *sim*(i,j) is measured using Lin's metric[63]; between any two genes, the most similar pairs of annotations are identified and averaged, i.e.

$$\mathrm{GeneSim}_{\mathrm{Schlicker}}=1/(|\mathrm{p}_1|+|\mathrm{p}_2|)\left(\sum_{\mathrm{i}\in\mathrm{p}1}\max_{\mathrm{j}\in\mathrm{p}2}(\mathrm{sim}(\mathrm{i},\mathrm{j}))+\sum_{\mathrm{j}\in\mathrm{p}\mathit{2}}\max_{\mathrm{i}\in\mathrm{p}\mathit{1}}(\mathrm{sim}(\mathrm{i},\mathrm{j}))\right)$$

where p_i is the set of function annotations associated with protein i.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by Swiss National Science Foundation grant PP00P3_150654 (to CD), UK Biotechnology and Biological Sciences Research Council grant BB/L018241/1 (to CD), Spanish Ministry of Economy and Competitiveness grant BIO2012-37161 (to TG), Qatar National Research Fund NPRP 5-298-3-086 (to TG), European Research Council grant ERC-2012-StG-310325 (to TG), National Institutes of Health (NIH) grant R24 OD011883 (to SEL), U41 HG002273 (to SEL and PDT), U41 HG007822 (to MJM and IX), Swiss State Secretariat for Education, Research and Innovation (SERI) funding (to IX and CD), US National Science Foundation EAGER Award #1355632 (to KS), ANR project BIP-BIP ANR-10-BINF-03-02 (to OL). Furthermore, ASdS, JHC, MJM, MM, PB acknowledge support from European Molecular Biology Laboratory, MM also acknowledges support from the Wellcome Trust (WT095908), SEL acknowledges support from Lawrence Berkeley National Laboratory core funds (Office of Basic Energy Sciences and U.S. Department of Energy Contract No. DE-AC02-05CH11231), LJJ acknowledges support from the Novo Nordisk Foundation (Grant No. NNF14CC0001), and LPP acknowledges support from the La Caixa-CRG International Fellowship Program.

References

1. Fitch WM. Distinguishing homologous from analogous proteins. Syst. Zool. 1970; 19:99–113. [PubMed: 5449325]
2. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 2005; 39:309–338. [PubMed: 16285863]
3. Gabaldón T, Koonin EV. Functional and evolutionary implications of gene orthology. Nat. Rev. Genet. 2013; 14:360–366. [PubMed: 23552219]
4. Dessimoz C. Editorial: Orthology and applications. Brief. Bioinform. 2011; 12:375–376. [PubMed: 21949264]
5. Altenhoff, AM.; Dessimoz, C. Evolutionary Genomics. Anisimova, M., editor. Vol. 855. Humana Press; 2012. p. 259-279.
6. Gabaldón T, et al. Joining forces in the quest for orthologs. Genome Biol. 2009; 10:403. [PubMed: 19785718]
7. Dessimoz C, et al. Toward community standards in the quest for orthologs. Bioinformatics. 2012; 28:900–904. [PubMed: 22332236]
8. Sonnhammer ELL, et al. Big data and other challenges in the quest for orthologs. Bioinformatics. 2014; 30:2993–2998. [PubMed: 25064571]
9. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput. Biol. 2011; 7:e1002073. [PubMed: 21695233]
10. Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake Ja. On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. PLoS Comput. Biol. 2012; 8:e1002386. [PubMed: 22359495]
11. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS Comput. Biol. 2012; 8:e1002514. [PubMed: 22615551]
12. Chen X, Zhang J. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput. Biol. 2012; 8:e1002784. [PubMed: 23209392]
13. Hulsen T, Huynen Ma, de Vlieg J, a Groenen PM. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 2006; 7:R31. [PubMed: 16613613]
14. Chen F, Mackey AJ, Vermunt JK, Roos DS. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One. 2007; 2:e383. [PubMed: 17440619]
15. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput. Biol. 2009; 5:e1000262. [PubMed: 19148271]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

16. Boeckmann B, Robinson-Rechavi M, Xenarios I, Dessimoz C. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. Brief. Bioinform. 2011; 12:423–435. [PubMed: 21737420]

17. Trachana K, et al. Orthology prediction methods: A quality assessment using curated protein families. Bioessays. 2011; 33:1–12. [PubMed: 21157784]

18. Salichos L, Rokas A. Evaluating ortholog prediction algorithms in a yeast model clade. PLoS One. 2011; 6:e18755. [PubMed: 21533202]

19. 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:e56925. [PubMed: 23451112]

20. Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL. SeqXML and OrthoXML: standards for sequence and orthology information. Brief. Bioinform. 2011; 12:485–488. [PubMed: 21666252]

21. Vilella AJ, et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2008; 19:327–335. [PubMed: 19029536]

22. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013; 41:D377–D386. [PubMed: 23193289]

23. Huerta-Cepas J, Capella-Gutiérrez S, Pryszcz LP, Marcet-Houben M, Gabaldón T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 2014; 42:D897–D902. [PubMed: 24275491]

24. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. U. S. A. 1999; 96:2896–2901. [PubMed: 10077608]

25. Wall DP, Fraser HB, Hirsh AE. Detecting putative orthologs. Bioinformatics. 2003; 19:1710–1711. [PubMed: 15593400]

26. Powell S, et al. eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res. 2014; 42:D231–D239. [PubMed: 24297252]

27. Schreiber F, Sonnhammer ELL. Hieranoid: hierarchical orthology inference. J. Mol. Biol. 2013; 425:2072–2081. [PubMed: 23485417]

28. Sonnhammer ELL, Östlund G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res. 2015; 43:D234–D239. [PubMed: 25429972]

29. Altenhoff AM, et al. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic Acids Res. 2015; 43:D240–D249. [PubMed: 25399418]

30. Linard B, et al. OrthoInspector 2.0: Software and database updates. Bioinformatics. 2015; 31:447–448. [PubMed: 25273105]

31. Pryszcz LP, Huerta-Cepas J, Gabaldón T. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res. 2011; 39:e32. [PubMed: 21149260]

32. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math. Biosci. 1981; 53:131–147.

33. Zhang X, Krause K-H, Xenarios I, Soldati T, Boeckmann B. Evolution of the ferric reductase domain (FRD) superfamily: modularity, functional diversification, and signature motifs. PLoS One. 2013; 8:e58126. [PubMed: 23505460]

34. Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res. 2013; 42:D922–D925. [PubMed: 24194607]

35. Dimmer EC, et al. The UniProt-GO Annotation database in 2011. Nucleic Acids Res. 2012; 40:D565–D570. [PubMed: 22123736]

36. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res. 2000; 28:304–305. [PubMed: 10592255]

37. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics. 2006; 7:302. [PubMed: 16776819]

38. Wolf YI, Koonin EV. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol. Evol. 2012; 4:1286–1294. [PubMed: 23160176]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

39. Dalquen, Da; Dessimoz, C. Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades such as Plants and Animals. Genome Biol. Evol. 2013; 5:1800–1806. [PubMed: 24013106]

40. Sennblad B, Lagergren J. Probabilistic orthology analysis. Syst. Biol. 2009; 58:411–424. [PubMed: 20525594]

Methods-only References

41. Akerborg O, Sennblad B, Arvestad L, Lagergren J. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:5714–5719. [PubMed: 19299507]

42. The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2012; 40:D71–D75. [PubMed: 22102590]

43. Boeckmann B, et al. Quest for Orthologs (QfO) entails Quest for Tree of Life (QfToL): in Search of the Gene Stream. Genome Biol. Evol. 2015; 7:1988–1999. [PubMed: 26133389]

44. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. 2015; 32:835–845. [PubMed: 25739733]

45. Jensen LJ, et al. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 2008; 36:D250–D254. [PubMed: 17942413]

46. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010; 11:24. [PubMed: 20070885]

47. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 2008; 9:286–298. [PubMed: 18372315]

48. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

49. Sievers F, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 2011; 7:539. [PubMed: 21988835]

50. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 2006; 34:1692–1699. [PubMed: 16556910]

51. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25:1972–1973. [PubMed: 19505945]

52. Guindon S, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 2010; 59:307–321. [PubMed: 20525638]

53. Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. The human phylome. Genome Biol. 2007; 8:R109. [PubMed: 17567924]

54. Altenhoff AM, Gil M, Gonnet GH, Dessimoz C. Inferring hierarchical orthologous groups from orthologous gene pairs. PLoS One. 2013; 8:e53786. [PubMed: 23342000]

55. Linard B, Thompson JD, Poch O, Lecompte O. OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics. 2011; 12:11. [PubMed: 21219603]

56. Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. Bioinformatics. 2010; 26:1899–1900. [PubMed: 20427515]

57. Thomas PD. GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. BMC Bioinformatics. 2010; 11:312. [PubMed: 20534164]

58. DeLuca TF, Cui J, Jung J-Y, St Gabriel KC, Wall DP. Roundup 2.0: enabling comparative genomics for over 1800 genomes. Bioinformatics. 2012; 28:715–716. [PubMed: 22247275]

59. Gonnet GH, Hallett MT, Korostensky C, Bernardin L. Darwin v. 2.0: an interpreted computer language for the biosciences. Bioinformatics. 2000; 16:101–103. [PubMed: 10842729]

60. Wikipedia contributors. Least squares inference in phylogeny. Wikipedia: The Free Encyclopedia; 2013. at <https://en.wikipedia.org/w/index.php?title=Least_squares_inference_in_phylogeny&oldid=552325441> [last accessed 7 March 2016]

61. Felsenstein, J. Inferring Phylogenies. Palgrave Macmillan; 2004.

62. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015; 43:D1049–D1056. [PubMed: 25428369]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

63. Lin, D. An information-theoretic definition of similarity. Proc. 15th International Conf. on Machine Learning; Morgan Kaufmann; San Francisco, CA. 1998. p. 296-304.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

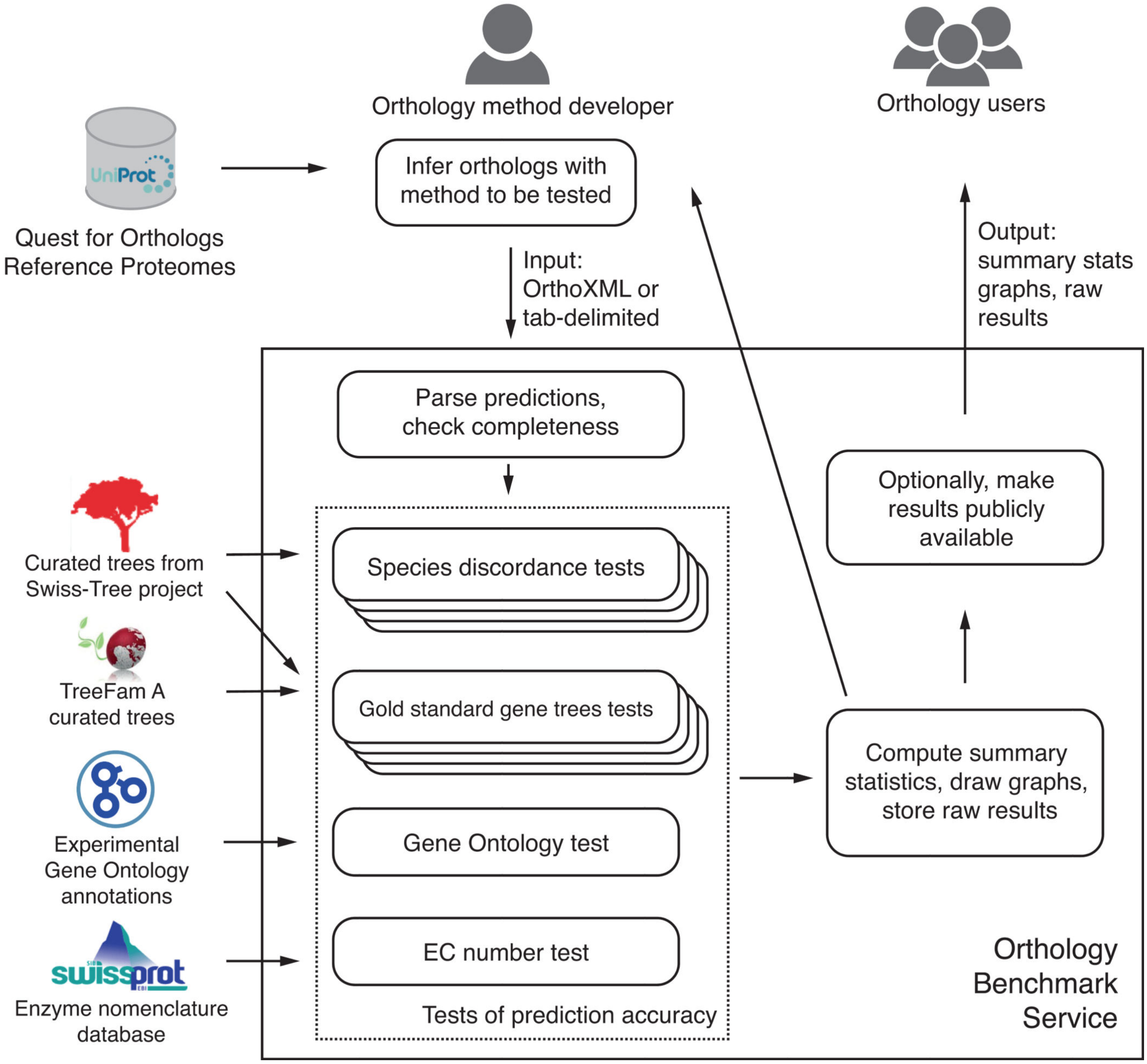
Orthology method developer
Orthology users
UniProt
Quest for Orthologs Reference Proteomes
Infer orthologs with method to be tested
Input: OrthoXML or tab-delimited
Output: summary stats graphs, raw results
Parse predictions, check completeness
Curated trees from Swiss-Tree project
Species discordance tests
TreeFam A curated trees
Gold standard gene trees tests
Experimental Gene Ontology annotations
Gene Ontology test
swissprot
Enzyme nomenclature database
EC number test
Tests of prediction accuracy
Optionally, make results publicly available
Compute summary statistics, draw graphs, store raw results
Orthology Benchmark Service

Figure 1.

The Orthology Benchmark service facilitates assessment and comparison of orthology inference methods. Orthology method developers run their methods on a reference proteome set and submit the inferred orthologs to the service. The predictions are subjected to a battery of phylogenetic and functional tests, and the results are returned to the method developer, who can choose to disclose them publicly.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

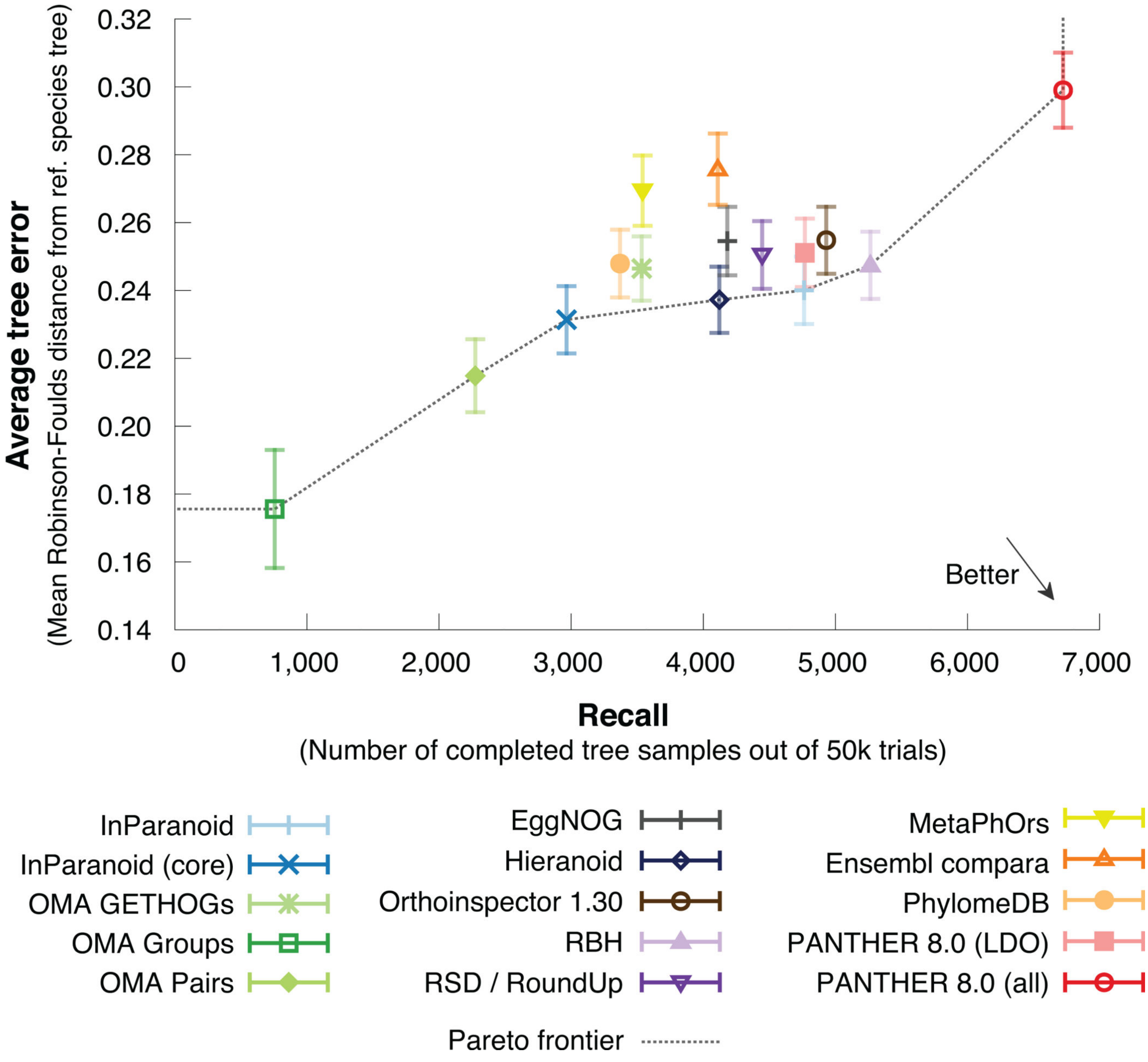

Figure 2.

The Generalised Species Tree Discordance test assesses the congruence of inferred orthologs with a trusted reference tree, focusing on eukaryotes. A trade-off between precision (measured in terms of tree error in the y-axis) and recall (measured in terms of completed tree samples in the x-axis; see Methods) can be observed. Only high-confidence branches of the reference tree (L90, see Methods), at least 10 myr long, are considered. Results for other clades are provided in Supplementary Fig. 3–5. Error bars indicate 95% confidence intervals and the line the “Pareto frontier”.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

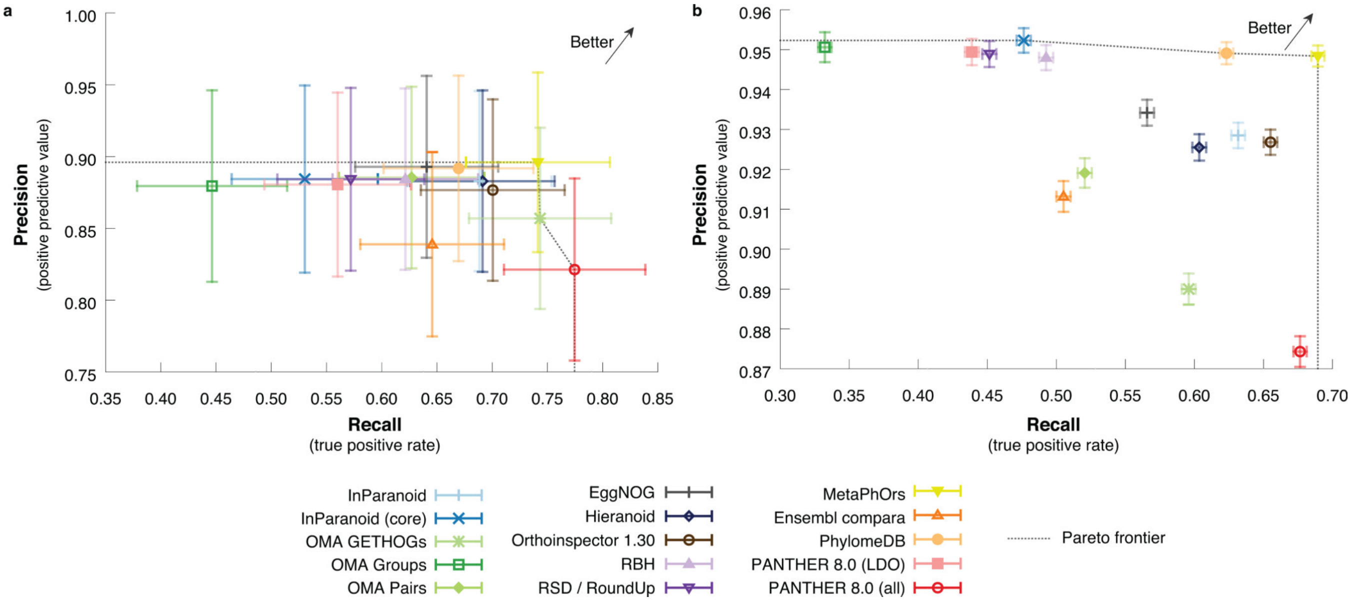

Figure 3.

Scatter plot of precision and recall for the benchmark with sets of reference gene trees. Evolutionary gene relationships are predicted for the QfO reference proteomes by 15 different methods. From the results, pairs of orthologous relationships are determined for each method and compared to those obtained from the reference gene trees of (a) SwissTree and (b) and TreeFam-A. In these tests, no trade-off between precision and recall can be observed and methods with a balanced trade-off perform especially well. Error bars indicate 95% confidence intervals.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

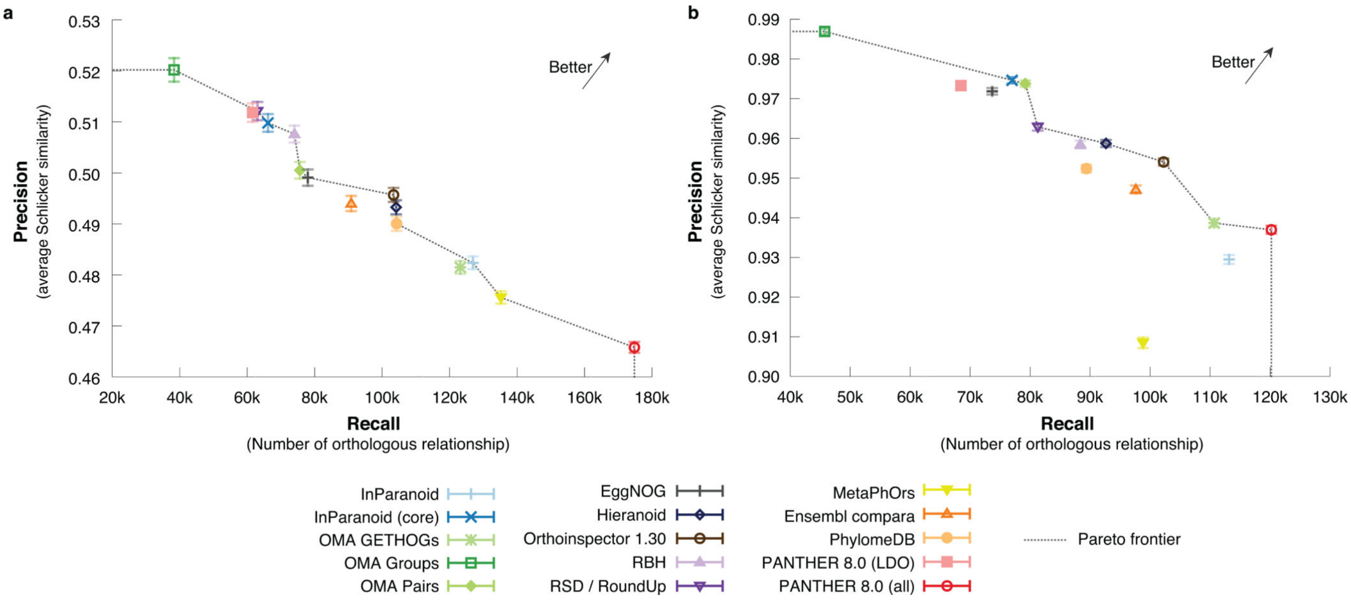

Figure 4.

Benchmarks of functional similarity between inferred orthologous gene pairs. A clear trade-off between precision (measured in Schlicker similarity[37]) and recall (measured in number of ortholog relationships involving proteins with functional annotations) can be observed with two different types of functional annotations: (a) experimentally supported GO annotations (b) Enzyme Commission (EC) numbers. Error bars indicate 95% confidence intervals.