ESCI upf.
School of International Studies

MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

nf-core/reportho: a pipeline for comparative analysis of ortholog predictions

AUTOR/A: Igor Trujnara	**NIA:** 106743
GRAU: Bachelor's Degree in Bioinformatics	
CURS ACADÈMIC: tercer	
DATA: 18.06.2024	
TUTOR/S: Cedric Notredame	

FULL DE RESUM DEL TREBALL DE FI DE GRAU DEL BDBI (ESCI-UPF)

TÍTOL DEL PROJECTE: nf-core/reportho: a pipeline for comparative analysis of ortholog predictions	
AUTOR/A: Igor Trujnara	**NIA:** 106743
CURS ACADÈMIC: tercer	
DATA: 18.06.2024	
TUTOR/S: Cedric Notredame	

PARAULES CLAU (mínim 3)

- Català: ortologia, bases de dades públiques, analisi comparativa, *pipelines*
- Castellà: ortología, bases de datos públicas, análisis comparativa, *pipelines*
- Anglès: orthology, public databases, comparative analysis, pipelines

RESUM DEL PROJECTE (extensió màxima: 100 paraules per llengua)

- Català:

Els gens ortòlegs són crítics per a l'estudi de la funció i evolució de les proteïnes. S'han elaborat múltiples mètodes per predir ortòlegs. *Quest for Orthologs* ha fet un *benchmark* dels mètodes, però no ha creat una comparació completa de les prediccions. Proposem nf-core/reportho, un *pipeline* que obté prediccions públiques d'ortòlegs, realitza comparacions sistemàtiques, calcula la similitud i la presenta en un format llegible. El *pipeline* demostra bon rendiment i escalabilitat. Una execució amb una mostra representativa de proteïnes humanes mostra acord limitat entre les fonts i destaca els reptes per al camp, especialment en l'aspecte d'integració de dades.

- Castellà:

Los genes ortólogos son críticos para el estudio de la función y evolución de las proteinas. Se han elaborado multiples métodos para predecir ortólogos. *Quest for Orthologs* hizo un *benchmark* de los métodos, pero no creó una comparación completa de las predicciones. Presentamos nf-core/reportho, un *pipeline* que obtiene predicciones públicas de ortólogos, realiza comparaciones sistemáticas, calcula la similitud y la presenta de forma legible. El *pipeline* demuestra buen rendimiento y escalabilidad. Una ejecución con una muestra representativa de

proteinas humanas demuestra acuerdo limitado entre fuentes y destaca los retos del campo, especialmente en el aspecto de integración de datos.

- Anglès:

Orthologous genes are crucial for the study of protein function and evolution. Multiple methods have been created to predict orthologs. Quest for Orthologs has benchmarked those methods but has not created a comprehensive prediction comparison. We propose nf-core/reportho, a pipeline that retrieves public ortholog predictions, performs systematic comparisons, calculates agreement statistics, and presents them in a human-readable format. The pipeline shows satisfactory performance and strong scalability. A run on a representative sample of human proteins demonstrates limited agreement between sources and highlights challenges for the field, especially in the aspect of data integration.

ESCI-UPF
Bachelor's Degree in Bioinformatics (UPF-UB-UAB-UPC)
Final Degree Project

nf-core/reportho: a pipeline for comparative analysis of ortholog predictions

Igor Trujnara[1,*] **Luisa Santus**[1,‡] **and Cedric Notredame**[1,†]

[1]Centre for Genomic Regulation, Carrer del Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

*Corresponding author. igor.trujnara@crg.eu †As project supervisor. ‡As co-supervisor.

Abstract

Orthology is a highly relevant aspect of genomics, as orthologous genes allow functional inference, identify certain evolutionary constraints, and are used for reconstructing the tree of life. This is especially important in light of new massive sequencing initiatives, most importantly the Earth Biogenome Project, which will require multiple efficient and robust analysis methods to exploit the vast amount of sequence data it will generate. There is a large variety of publicly available orthology prediction methods, but the results they provide are highly varied and agreement is limited. Significant effort is made to assess the performance of those methods, most notably through the ongoing Quest for Orthologs, which created a comprehensive benchmark for orthology prediction. However, there is still a need for more universal, reference-free orthology benchmarks. We propose nf-core/reportho, a Nextflow pipeline for comparative analysis of ortholog predictions. Given a protein, the pipeline retrieves and integrates the ortholog predictions from public sources, performs comparative analyses, calculates agreement statistics, and creates summary visual representations. It also provides basic downstream analysis in the form of multiple sequence alignment and phylogenetic reconstruction. We envision that nf-core/reportho will enable and accelerate new research projects involving specific proteins, as well as systematic investigation of orthology databases. In benchmarks, nf-core/reportho demonstrates strong scalability. When tested with a representative sample of the human proteome, it indicates limited agreement between the databases, highlighting both the open nature of the ortholog prediction problem and the challenges of data interoperability.

Key words: orthology, comparative analysis, public databases, pipelines

Introduction

Although there are numerous definitions of a gene, and the concept has varied through time, a gene is certainly an inheritable unit of information [1]. However, genes are not merely inherited. They are highly dynamic and undergo changes between generations [2]. This process is crucial for the long-term maintenance of life, as it creates biodiversity and thus enables adaptation to the spatial and temporal diversity of the environment [3]. Although the processes that lead to the creation of new genes are diverse and complex, they can be classified into two main classes: de novo gene creation, where a gene is created in a previously non-coding region of the genome, and modification of existing genes, such as duplication, speciation, fission or fusion [2]. These processes give rise to homologous genes, or homologs, i.e. pairs of genes that originate from a single ancestral gene.

Orthologs are a special case of homologs. They are genes in different species that arose from their common ancestor through a speciation event [4]. The identification and study of orthologs is essential for the reconstruction of the tree of life [5]. The history of orthologous genes should recapitulate the evolutionary history of the species involved. Although duplications are possible in a gene tree of orthologs, they are exclusively species-specific, and thus introduce no ambiguity. Furthermore, orthologs exhibit more similar functions than other types of homologs. It should be stressed that the definition of orthologs, as stated above, does not require nor imply any functional similarity. Nonetheless, empirical observations suggest that orthologs exhibit significantly higher functional similarity than paralogs [6]. Finally, analysis of orthologs can reveal evolutionary constraints acting on the genes and resulting proteins. In particular, the preserved features of both the sequence and the structure should to an extent be relevant to the function, even though sequence conservation and function are not entirely equivalent [7].

Currently, there are multiple large-scale sequencing efforts, most notably the Earth Biogenome Project [8], which aims at sequencing all eukaryotic species. The massive amount of data that will be produced will create the need for accurate and efficient annotation methods, which strongly rely on a good understanding of ortholog history. Given the high demand and usefulness of accurate orthology identification, various research endeavors have targeted the orthology prediction problem [9, 10, 11, 12]. Given

that ancient genomic sequences are highly uncommon [13], and thus universal access to any ancestral sequence is infeasible, all of the resulting methods rely mostly on the genomes of currently existing species.

There are two main approaches to predicting orthologous relationships. The distance-based approach reconstructs orthology through the analysis of sequence divergence (distance) between genes. One common rule underlying a vast majority of distance-based methods is the reciprocal best hits heuristic – the notion that if two genes are each other's closest match in their respective species, then they are orthologous. Each method applies different additional corrections to this heuristic.

For instance, in the case of OMA, a pseudo-ortholog filter is applied to the pairs of best hits. In this filter, each putative ortholog pair is tested against a pair of closely matched outgroup sequences. If the putative orthologs form a single branch on the resulting phylogeny, they are considered true orthologs. Otherwise, they are rejected as pseudo-orthologs. This operation accounts for the problem of differential gene loss, a major confounding factor in orthology prediction [9]. EggNOG uses another similar approach to verify its predictions. It is based on the triangulation concept proposed by Tatusov and Koonin [14]. It requires that any pair of putative orthologs should be additionally supported by another gene, which is a reciprocal best hit with both of them. This approach naturally leads to the identification of the so-called clusters of orthologs (COGs) [11]. Finally, OrthoInspector uses an approach based on in-paralog groups. Putative in-paralogs in a genome are identified as such if their BLAST score is higher than that of the reciprocal best hit. Those in-paralogs are then verified by testing whether their best hit in the other genome is a putative in-paralog for the pair, and eliminated if not [15].

On the other hand, tree-based methods compare the gene tree to a reference species tree and reconstruct a sequence of duplication and gene loss events that could lead to the particular gene tree [16]. Due to the high computational cost of this operation, especially with larger inputs, the methods use different heuristics to minimize the required computation. One notable example of such a heuristic is species overlap. It is based on the simple assumption that any non-leaf node in the gene tree is a duplication if any species is found in both descendant branches, and a speciation otherwise [17]. This approach is still used in new orthology prediction methods [18].

Even though all the methods aim to reconstruct the same biological reality, the agreement of their predictions remains very limited. This is partially the result of the methods themselves, as the approaches are diverse and prioritize different types of signals [19]. Additionally, despite ongoing standardization efforts [20, 21], the data used for creating the precomputed ortholog databases is not uniform. There have been notable efforts to assess the performance of different orthology prediction methods, most notably the Quest for Orthologs, which has created a standard benchmark for ortholog predictors [20]. However, as of this manuscript being written, there is no standard method for compiling orthology predictions across different sources.

Thus, we propose nf-core/reportho, an open-source pipeline for the systematic retrieval and analysis of ortholog predictions. nf-core/reportho obtains ortholog predictions for a given protein or a list of proteins, performs automated integration and analysis of the obtained predictions, calculates agreement statistics, and generates a single final ortholog list. We envision that nf-core/reportho will provide new insight into the characteristics of ortholog prediction methods, as well as the specifics of predictions for different protein families.

Objectives

The core objective of this project is to develop a pipeline that performs a comprehensive comparative analysis of orthology predictions for a particular query protein or a list of proteins. The main facets of this objective include integration of data from different sources, calculation of statistics providing a concise numerical description of the agreement between the sources, as well as informative graphical representation. An additional aim of the project is to use current best practices of bioinformatics method development, including reproducible Nextflow pipelining and adhering to good software development practices enforced by nf-core.

Methods

Pipeline implementation

Due to differences in system configurations and software versions, workflows executed manually or created with simple tools like Bash scripting might lead to instability, posing a significant challenge to research reproducibility. Containerization systems in combination with workflow management tools are commonly used to mitigate this issue. Nextflow [22] is a Groovy-based workflow management framework. It ensures full reproducibility through the use of containers and is compatible with most modern container software, including Docker [23] and Singularity [24]. It is also compatible with many HPC systems, such as SLURM and Grid Engine, enabling task parallelization in compute clusters. nf-core [25] is a community effort focused on establishing high-quality Nextflow pipelines for bioinformatics analyses. The quality of the pipeline is ensured by providing developers with guidelines and tools for pipeline creation, as well as through internal peer review. nf-core uses GitHub for convenient and consistent version management and utilizes GitHub's continuous integration capabilities to ensure correct pipeline function throughout the development process by executing a test suite after every change. nf-core/reportho is implemented in Nextflow and is an official nf-core pipeline available on the nf-core website under: `https://nf-co.re/reportho`.

Selected Ortholog Repositories

For the first version of the pipeline, we chose to include data from OMA, PANTHER, OrthoInspector, and EggNOG. Those sources were selected by analyzing all the methods included in the Quest for Orthologs [20, 21]. To be included, a method had to fulfill two core criteria: have a public precomputed database, and provide output accessions in UniProt format, or one that easily maps to it (Ensembl, RefSeq). Additionally, each method had to provide at least one mode of programmatic access: either an API that allows queries in UniProt format ("online access"), or an FTP service that provides ortholog predictions as a single file ("offline access"). Where possible, databases were included for both online and offline access.

Input and Pre-processing

The pipeline receives as input one protein or a list of proteins for which the orthologs need to be retrieved. A protein can be provided in the form of a FASTA file with the corresponding sequence or a UniProt [26] accession number. We chose UniProt

as the main programmatic protein identifier due to its widespread use in existing orthology resources and the straightforward access to UniProt data. If the protein is provided as a FASTA file, the corresponding UniProt identifier is identified using the OMA API with the hybrid search strategy, i.e. first searching for an exact match and then performing a BLAST [27] search if none is found. It is reported whether the retrieved UniProt ID corresponds to the exact match. Once the UniProt ID is identified, OMA is used to determine the NCBI taxon identifier [28]. This is necessary for the PANTHER API.

Ortholog Fetching

The fetching of the ortholog entries can be performed via API calls or using local database snapshots, when available (local snapshots currently available for OMA, PANTHER, and EggNOG; API access available for OMA, PANTHER, and OrthoInspector). A combination of online and local searches is also supported. Certain databases provide ortholog predictions in non-Uniprot identifiers. In these cases, identifier mapping is performed using the UniProt mapping service or with a local mapping script. The script is only used if the user specifies an offline run, and uses official identifier maps provided by the respective databases. Unmapped identifiers are retained in their original form for subsequent analysis. Finally, for each database, a list with the ortholog predictions with the unified identifiers is provided.

Integrative Analysis

The retrieved ortholog lists are combined into a single CSV file, which reports the predicted orthologs, along with the databases that identified them as such. The score column reports the number of supporting databases. Additional statistics about the agreement of the predictions across databases are calculated, including pairwise agreement between sources (Jaccard index, see Equation 1 and Supplementary Figure S1), percentage consensus (the size of the intersection between all the sources), percentage of privates (predictions identified only in one database), as well as goodness, a custom statistic reflecting the full distribution of scores, described in Equation 2 and Supplementary Figure S2. The collected metrics are then visualized in graphical representations (Figure 2). Finally, a single list of orthologs is generated according to the selected criteria. Currently, this can be either the minimum number of databases supporting an ortholog (score threshold), or the predictions of the most concordant source, i.e. the database with the highest percentage intersection with the other databases.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{1}$$

where A and B are sets of predictions from different databases.

$$G = \frac{\sum s_i}{n \times d}, \tag{2}$$

where s_i are scores (numbers of supporting databases) of predicted orthologs, n is the total number of predicted orthologs, and d is the number of databases.

Downstream Analysis

The pipeline provides common analyses given a list of orthologs. Currently, this includes multiple sequence alignment (MSA) and rendering of a phylogenetic tree. Sequences for MSA are obtained using the OMA API where possible and the Uniprot API for IDs not found in the OMA API. If structure-based alignment is requested, structures are obtained from the AlphaFold Protein structure database [29]. It is an easily searchable resource, and it has been shown that predicted structures yield an MSA quality comparable to experimental structures [30]. Sequence-based MSAs are computed with T-Coffee [31], a commonly used progressive aligner. Structure-based MSAs are computed with 3D-Coffee [32]. nf-core/reportho currently supports phylogenetic reconstruction using maximum likelihood (IQ-TREE [33]) or minimum evolution techniques (FastME [34]). In either case, by default bootstrap support values are computed on 100 replicates, with the possibility of adjustment by the user.

Reporting

Optionally, an HTML report is generated for each input protein. It is created using React [35], a JavaScript framework for dynamic and modular web applications. The report summarizes relevant information from the run, including the predictions from each database, the distribution of scores and statistics, the final ortholog list, and the results of the MSA and phylogeny reconstruction. A run script is provided along with the report to enable its correct visualization. The report and the associated files are provided in a single folder per input protein, enabling convenient distribution of pipeline results. Additionally, a summary report is created using MultiQC [36]. It contains statistics (see above) and the number of orthologs found for each query in the run, as well as information about the run, including software versions used.

Output

nf-core/reportho produces a large number of distinct result files for each query. The key outputs, including some optional ones, are the following:

- score table – a CSV file containing all predicted orthologs for a query protein, the databases that predicted them, and the score (number of those databases),
- final ortholog list – a list of orthologs chosen based on user-defined criteria, used for alignment and phylogeny,
- agreement plots – plots representing the agreement between databases, in the form of a bar plot colored by score, Venn diagram, and tile plot representing the Jaccard index,
- MSA – a sequence alignment of the orthologs created with T-Coffee, in Clustal format,
- phylogeny – a phylogenetic tree created with ML or ME methods,
- per-query report – a detailed HTML report containing information about the orthologs of a single query,
- summary report – an HTML report summarizing the results for multiple queries, created with MultiQC.

Results

nf-core/reportho Workflow Overview

nf-core/reportho is a high-throughput workflow for systematic retrieval and analysis of ortholog predictions for a protein or a list of proteins. The key steps of the pipeline include orthologs prediction fetching from public databases, comparison and integration of the predictions, downstream analyses, and reporting (see Figure 1 for detailed steps). The prediction sources

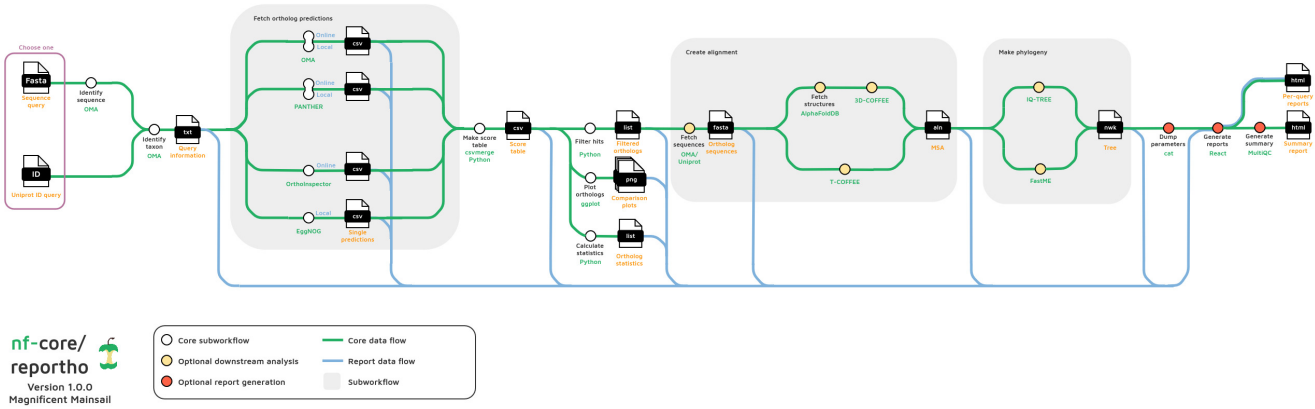

Fig. 1: Graphical representation of the nf-core/reportho pipeline. Lighter grey boxes highlight the subworkflows. Dots represent the steps of the main subworkflows (white) and the optional subworkflows for downstream analysis and report generation (yellow and red correspondingly). File icons indicate key outputs.

currently supported by the pipeline are OMA [37], PANTHER [10], OrthoInspector [15, 12], and EggNOG [11]. The sources used in the analysis, as well as other parameters, can be provided by the user through a command line interface or a configuration file, as extensively described in the pipeline documentation (see https://nf-co.re/reportho/dev/docs/usage) The input for the pipeline is a UniProt identifier or a FASTA sequence, which will be automatically converted to a UniProt identifier. UniProt [26] is a common reference resource for proteins, provides stable identifiers, and is used by multiple orthology predictors, making it the ideal input format for the pipeline. nf-core/reportho's key outputs include a CSV file per input protein with the corresponding ortholog predictions, summary plots, and a clean, human-readable HTML report of the results. In a nutshell, nf-core/reportho is a Nextflow pipeline that handles the fetching and subsequent integration and analysis of ortholog predictions from multiple public databases. We envision that it will provide valuable initial orthology information in the study of novel proteins. In the next sections, we will first demonstrate the detailed function of the pipeline with a single query, and then perform a large-scale analysis on a representative sample of the human proteome.

BicD2 Ortholog Predictions Across Databases

Proteins involved in essential cellular processes tend to be highly conserved and are therefore good targets for the analysis of ortholog predictions. Bicaudal D cargo adaptor 2 (BicD2) is a fibrous protein that participates in microtubule transport [38] and is therefore expected to have an ortholog in every eukaryotic species, making it a reasonable target for orthology predictions. To demonstrate the functionalities of the pipeline for a real-world example, we perform an example run with human BicD2 as the input protein.

The input for the pipeline in this run is the UniProt accession of the protein (`Q8TD16`).

A total of 904 BicD2 orthologs were identified across databases: 347 in EggNOG, 165 in OMA, 35 in PANTHER, and 480 in Orthoinspector (Table 1). It is noticeable that the different sources highly differ in the number of predictions from different sources (Table 1, Figure 1A) and the extent of the support of the predictions, ranging from high agreement in PANTHER (77% intersection with other sources) to high disagreement in EggNOG (<0.1% intersection).

A natural question is whether these predictions are largely concordant or are mostly unique to each database. Although there is only one true evolutionary history and orthology databases should in theory contain, at least partially, the same information, disagreement between different orthology predictors is a known issue [19]. In the case of BicD2, we do indeed observe that many of the orthologs identified are identified by 1 database (Table 1, Figure 2A). Only 15 out of the 904 identified orthologs are found across at least 3 databases, which corresponds to less than 2% (Table 2). This set includes orthologs from primates (e.g. *Macaca mulatta*), farm animals (e.g. *Bos taurus*), model species (e.g. *Mus musculus*) as well as a few miscellaneous ones. Notably, all orthologs with a score of 3 and all but one with a score of 2 are identified with a UniProt accession, indicating that UniProt is indeed the most interoperable accession format, at least with this query, and the presence of other formats might aggravate disagreement.

Furthermore, the pairwise overlap between the sources is rather small, with 11.8% of all predictions being supported by more than 1 source (Table 1-2, Figure 2B) and Jaccard indices (pairwise agreement) between databases ranging from 0 to 17% (Figure 2C). This observation demonstrates that the issue of disagreeing predictions across sources is present in the case of BicD2. All of the reported numbers and plots (Table 1, Figure 2A-C) are automatically generated in nf-core/reportho and available to the user in the output report. Here, we showcase the usefulness of nf-core/reportho by investigating the landscape of ortholog predictions for the BicD2 protein across orthology databases and unsurprisingly observed a high pairwise and global disagreement of the retrieved ortholog predictions.

Source	Number of orthologs found	Number of private orthologs
EggNOG	347	345
OMA	165	70
OrthoInspector	480	374
PANTHER	35	8
Total	904	797

Table 1. Number of identified orthologs of BicD2 per source database in databases; private orthologs are orthologs that are only supported by a single database; the counts do not sume to the total, as some orthologs were predicted by multiple sources.

Number of supporting sources (score)	Count
1	798
2	92
3	15
4	0

Table 2. The distribution of scores (numbers of supporting databases) for predicted orthologs of BicD2.

Sampling the Orthology Landscape of the Human Proteome

One significant advantage of nf-core/reportho is the automated integration of multiple databases and the retrieval of the predictions for input proteins in a parallelized manner. This enables automated large-scale orthology analyses that would otherwise require considerable manual effort and computational resources to be executed. We reasoned that an interesting application of nf-core/reportho would be the exploration of the available orthology predictions across the full human proteome. However, due to the heavy computational cost of the process for thousands of samples, we could only carry out a subset of it in the time frame of this project.

We ran the pipeline on a sample of the single-isoform version of the reference human proteome from UniProt [39]. The full proteome contains 20,590 annotated proteins, out of which we randomly selected 1,000 assuming homogeneous subsampling. As this sample covers ~5% of the human proteome, it should reasonably reflect its diversity. Due to the size of the dataset, we performed the analysis fully offline, i.e. without any runtime access to remote servers or APIs thus limiting the predictions obtained to OMA, PANTHER, and EggNOG. The run took under 2 hours on an HPC cluster. We observe limited agreement across the prediction of the three inspected databases. Across all the queries, no ortholog was found in all 3 databases. This is unexpected, as at least the orthologs from primate species should have been identified by any method. We were not able to identify whether this is a result of the methods themselves or of identifier mismatch.

In about 45% of the queries, all the orthologs were reported by exclusively one database. In the remaining queries, the percentage of such orthologs was relatively high, above 0.9 in most cases, further indicating limited intersection. Accounting for the size of the intersections does not improve the result much, as demonstrated by the distribution of goodness. In virtually all cases, this statistic is between its minimum of 0.33 (due to 3 databases in the run) and 0.4, indicating that the support of the predictions is very far from the theoretical maximum.

The distribution of hit counts provides additional insight into the behavior of the predictions. The total counts are mostly in the hundreds, with a median of 308, although there are zeros, as well as some exceptionally high values, up to tens of thousands (Supplementary Figure S3). The databases also vary in the number of predictions they provide. OMA and PANTHER exhibit similar behavior, providing tens to hundreds of predictions per query, with a small number of zeros and a few outliers in the thousands. In the case of EggNOG, the distribution is different, with a very significant number of zeros and substantially larger non-zero counts (Supplementary Figure S4).

As orthologs supported by a single source might be noise or unmappable identifiers, it is interesting to observe how hit counts change when those are removed. As expected, a major decrease in total ortholog counts is observed, with most of the counts in the scale of tens (Figure 3). The change is not uniform across databases. OMA and PANTHER both move from several hundred to tens, and the number of zero counts increases slightly. For EggNOG, the counts decrease from thousands to tens, and the number of zeros increases sharply (Supplementary Figure S5). It can be concluded that while low-confidence hits are found in all databases, the extent is highly variable.

Discussion

Despite its importance to evolutionary biology, ortholog prediction remains a remarkably challenging task. As we have reported, the degree of disagreement across databases is highly significant. Previous efforts outlined the biases and the sensitivity-specificity tradeoff of each method [21], but it remains unclear whether a single optimal methodology for the orthology prediction problem currently exists. As we have outlined in the introduction, the knowledge of orthologs is fundamental for multiple lines of evolutionary research. Further prediction method investigation, and the development of robust predictors, remain a priority for the field. This is especially relevant in the context of ongoing large-scale sequencing efforts, including the Earth Biogenome Project [8], as high-quality orthology prediction will be crucial to the correct functional inference of proteins [40] and phylogenetic placement of species [5].

nf-core/reportho cannot currently provide a definitive solution to the problem. However, it delivers a systematic approach to programmatically comparing existing prediction methods. This will enable new evolutionary studies by providing high-confidence

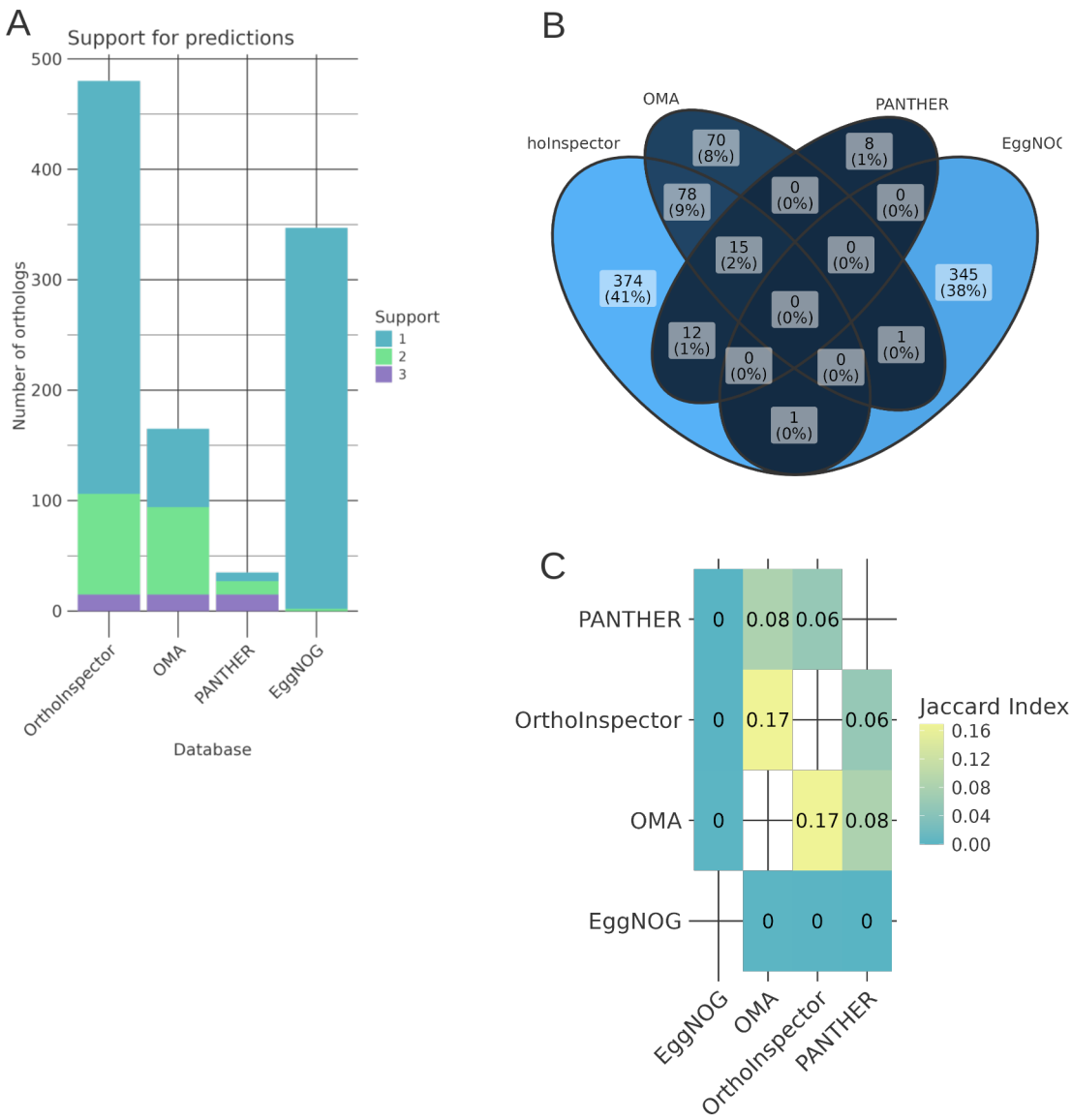

Fig. 2: The plots automatically produced by the pipeline. **A:** The number of predictions per database and the scores (number of supporting sources) of the predictions. **B:** Venn diagram showing the intersection across databases. **C:** Tile plot with the Jaccard index for each pair of databases.

predictions. At the same time, it will provide a novel insight into the specifics of each prediction method, as well as identify how specific classes of proteins behave in the prediction process.

After investigating the landscape of orthology predictions across a large sample of the human proteome, we notice limitations of interoperability of the available orthology prediction methods. Although the agreement found is in the general range found in prior studies [19], it is definitely on the lower end of this range. This is likely caused by two main factors. The first is the inherent difference between the methods. Even though the databases were built with access to a set of benchmark proteomes [20], the input data was certainly not identical, leading to limitations on the set of predictions that can be provided. Furthermore, the technical differences between the methods inevitably lead to disagreement.

However, this alone cannot justify the magnitude of the differences. The second factor is data integration. During development, we observed that all the databases use Uniprot, Ensembl, and Refseq identifiers, as well as custom formats. The lack of a universal vocabulary, or a comprehensive identifier mapping system limits the possibility of full comparison between the public ortholog prediction databases. This highlights a major issue for the field, as a shared vocabulary is crucial to the correct comparison and combination of data from different sources. Benchmarks circumvent this issue by recomputing the predictions

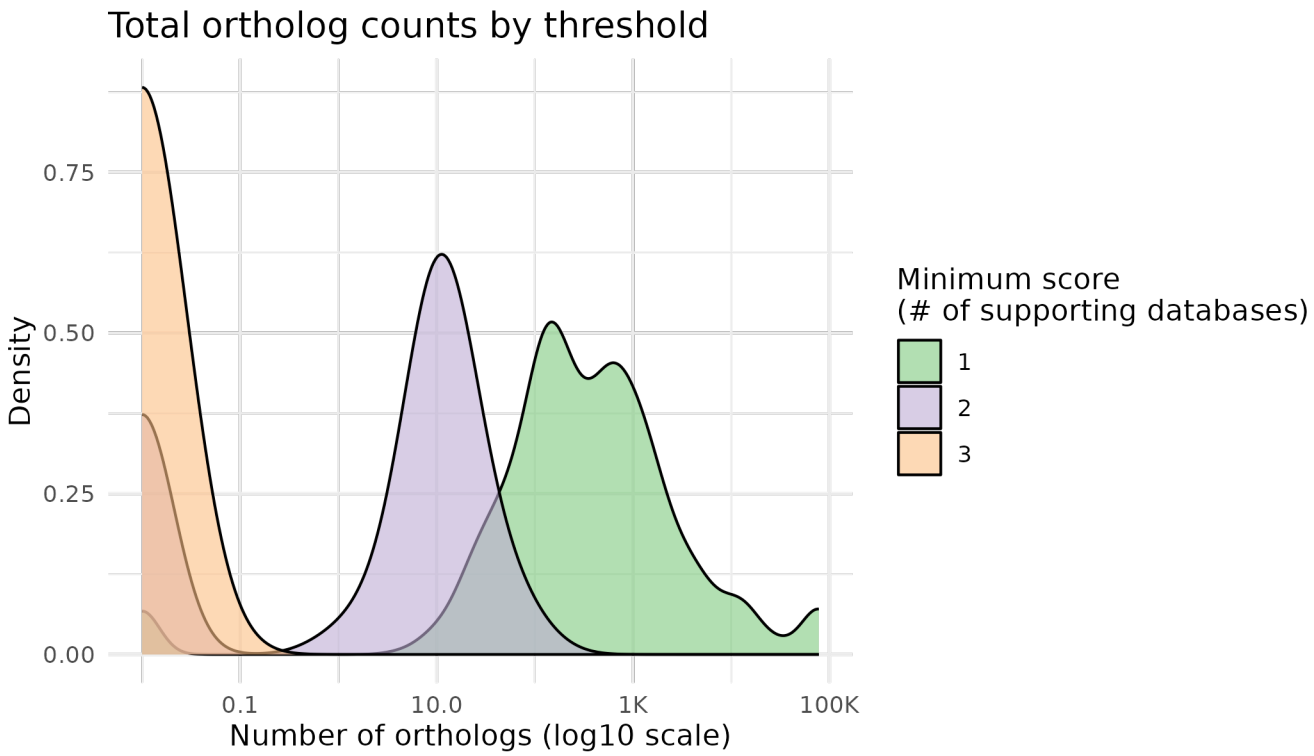

Fig. 3: Change in total ortholog count distribution after filtering by score in the sample of 1000 human proteins; note that for minimum score 3 all values are equal to 0.

with coherent reference proteomes [20], but this recomputation is a costly process and is only feasible in some studies. The observations from this project indicate that interoperability, one of the four pillars of the FAIR data principles [41], remains a significant challenge in the field of orthology research, and protein research in general.

Conclusion

As the main predictors of taxonomic relationships and protein function, orthologs are a key component in our understanding of life. However, the identification of orthologs remains an open problem, and there is much work ahead in the field. With nf-core/reportho, we propose a tool to integrate ortholog predictions for evolutionary studies, as well as large ortholog datasets for big data and machine learning purposes. Furthermore, the tool has the potential to introduce new insights and increase understanding of the various approaches currently available. In our tests, nf-core/reportho has underlined the challenges of public orthology predictions, and we envision that it will support further research in this area.

We are planning to expand nf-core/reportho with new capabilities. Currently, we want to include taxonomic analysis, as well as the possibility to compare results between multiple sets of queries (e.g. compare prediction quality between different proteomes). We will also expand the pipeline with new databases should they become sufficiently compatible.

Code and data availability

The code of the pipeline is available under the nf-core organization at `https://github.com/nf-core/reportho` and released under the MIT license. Additional code for the report rendering is available at `https://github.com/itrujnara/orthologs-report`. Minimal test datasets for the pipeline are available at `https://github.com/nf-core/test-datasets/tree/reportho`. Additional scripts used to analyze the output are available at `https://github.com/itrujnara/reportho-extra-scripts`.

Acknowledgments

The author thanks:

- Cedric Notredame for scientific supervision,
- Luisa Santus for support in design and writing,
- Jose Espinosa-Carrasco for technical support and code review,
- Alessio Vignoli for design support,
- Júlia Mir Pedrol for code review,
- the nf-core community for creating tools and standards that guided the project, as well as the supervision and support throughout.

References

1. Mark B. Gerstein, Can Bruce, Joel S. Rozowsky, Deyou Zheng, Jiang Du, Jan O. Korbel, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, and Michael Snyder. What is a gene, post-ENCODE? History and updated definition.

Genome Research, 17(6):669–681, June 2007. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

2. Sidi Chen, Benjamin H. Krinsky, and Manyuan Long. New genes as drivers of phenotypic evolution. *Nature reviews. Genetics*, 14(9):645–660, September 2013.
3. Torsten Nygaard Kristensen, Tarmo Ketola, and Ilkka Kronholm. Adaptation to environmental stress at different timescales. *Annals of the New York Academy of Sciences*, 1476(1):5–22, September 2020.
4. Eugene V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, 39:309–338, 2005.
5. Frédéric Delsuc, Henner Brinkmann, and Hervé Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361, 2005.
6. Toni Gabaldón and Eugene V. Koonin. Functional and evolutionary implications of gene orthology. *Nature Reviews. Genetics*, 14(5):360–366, May 2013.
7. Chris P. Ponting. Biological function in the twilight zone of sequence conservation. *BMC Biology*, 15(1):71, August 2017.
8. Harris A. Lewin, Gene E. Robinson, W. John Kress, William J. Baker, Jonathan Coddington, Keith A. Crandall, Richard Durbin, Scott V. Edwards, Félix Forest, M. Thomas P. Gilbert, Melissa M. Goldstein, Igor V. Grigoriev, Kevin J. Hackett, David Haussler, Erich D. Jarvis, Warren E. Johnson, Aristides Patrinos, Stephen Richards, Juan Carlos Castilla-Rubio, Marie-Anne van Sluys, Pamela S. Soltis, Xun Xu, Huanming Yang, and Guojie Zhang. Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333, April 2018. Publisher: Proceedings of the National Academy of Sciences.
9. Alexander C. J. Roth, Gaston H. Gonnet, and Christophe Dessimoz. Algorithm of OMA for large-scale orthology inference. *BMC bioinformatics*, 9:518, December 2008.
10. Paul D. Thomas, Dustin Ebert, Anushya Muruganujan, Tremayne Mushayahama, Laurent-Philippe Albou, and Huaiyu Mi. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science*, 31(1):8–22, 2022.
11. Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, Lars J Jensen, Christian von Mering, and Peer Bork. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, January 2019.
12. Yannis Nevers, Arnaud Kress, Audrey Defosset, Raymond Ripp, Benjamin Linard, Julie D Thompson, Olivier Poch, and Odile Lecompte. OrthoInspector 3.0: open portal for comparative genomics. *Nucleic Acids Research*, 47(D1):D411–D418, January 2019.
13. Ludovic Orlando, Robin Allaby, Pontus Skoglund, Clio Der Sarkissian, Philipp W. Stockhammer, María C. Ávila Arcos, Qiaomei Fu, Johannes Krause, Eske Willerslev, Anne C. Stone, and Christina Warinner. Ancient DNA analysis. *Nature Reviews Methods Primers*, 1(1):1–26, February 2021. Publisher: Nature Publishing Group.
14. R. L. Tatusov, E. V. Koonin, and D. J. Lipman. A genomic perspective on protein families. *Science (New York, N.Y.)*, 278(5338):631–637, October 1997.
15. Benjamin Linard, Julie D. Thompson, Olivier Poch, and Odile Lecompte. OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, 12(1):11, January 2011.
16. Adrian M. Altenhoff and Christophe Dessimoz. Inferring orthology and paralogy. *Methods in Molecular Biology (Clifton, N.J.)*, 855:259–279, 2012.
17. Jaime Huerta-Cepas, Hernán Dopazo, Joaquín Dopazo, and Toni Gabaldón. The human phylome. *Genome Biology*, 8(6):R109, June 2007.
18. Xavier Grau-Bové and Arnau Sebé-Pedrós. Orthology Clusters from Gene Trees with Possvm. *Molecular Biology and Evolution*, 38(11):5204–5208, November 2021.
19. Feng Chen, Aaron J. Mackey, Jeroen K. Vermunt, and David S. Roos. Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. *PLOS ONE*, 2(4):e383, April 2007. Publisher: Public Library of Science.
20. Adrian M. Altenhoff, Brigitte Boeckmann, Salvador Capella-Gutierrez, Daniel A. Dalquen, Todd DeLuca, Kristoffer Forslund, Jaime Huerta-Cepas, Benjamin Linard, Cécile Pereira, Leszek P. Pryszcz, Fabian Schreiber, Alan Sousa da Silva, Damian Szklarczyk, Clément-Marie Train, Peer Bork, Odile Lecompte, Christian von Mering, Ioannis Xenarios, Kimmen Sjölander, Lars Juhl Jensen, Maria J. Martin, Matthieu Muffato, Toni Gabaldón, Suzanna E. Lewis, Paul D. Thomas, Erik Sonnhammer, and Christophe Dessimoz. Standardized benchmarking in the quest for orthologs. *Nature Methods*, 13(5):425–430, May 2016. Publisher: Nature Publishing Group.
21. Yannis Nevers, Tamsin E M Jones, Dushyanth Jyothi, Bethan Yates, Meritxell Ferret, Laura Portell-Silva, Laia Codo, Salvatore Cosentino, Marina Marcet-Houben, Anna Vlasova, Laetitia Poidevin, Arnaud Kress, Mark Hickman, Emma Persson, Ivana Piližota, Cristina Guijarro-Clarke, the OpenEBench team the Quest for Orthologs Consortium, Wataru Iwasaki, Odile Lecompte, Erik Sonnhammer, David S Roos, Toni Gabaldón, David Thybert, Paul D Thomas, Yanhui Hu, David M Emms, Elspeth Bruford, Salvador Capella-Gutierrez, Maria J Martin, Christophe Dessimoz, and Adrian Altenhoff. The Quest for Orthologs orthology benchmark service in 2022. *Nucleic Acids Research*, 50(W1):W623–W632, July 2022.
22. Paolo Di Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, April 2017.
23. Dirk Merkel. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2:2, March 2014.
24. Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):e0177459, May 2017. Publisher: Public Library of Science.
25. Philip A. Ewels, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 38(3):276–278, March 2020.
26. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, January 2023.

27. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.
28. Scott Federhen. The Taxonomy Project. In *The NCBI Handbook [Internet]*. National Center for Biotechnology Information (US), August 2003.
29. Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, January 2022.
30. Athanasios Baltzis, Leila Mansouri, Suzanne Jin, Björn E. Langer, Ionas Erb, and Cedric Notredame. Highly significant improvement of protein sequence alignments with AlphaFold2. *Bioinformatics (Oxford, England)*, 38(22):5007–5011, November 2022.
31. C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217, September 2000.
32. Orla O'Sullivan, Karsten Suhre, Chantal Abergel, Desmond G. Higgins, and Cédric Notredame. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *Journal of Molecular Biology*, 340(2):385–395, July 2004.
33. Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, May 2020.
34. Vincent Lefort, Richard Desper, and Olivier Gascuel. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, 32(10):2798–2800, October 2015.
35. React.
36. Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, 32(19):3047–3048, October 2016.
37. Adrian M Altenhoff, Clément-Marie Train, Kimberly J Gilbert, Ishita Mediratta, Tarcisio Mendes de Farias, David Moi, Yannis Nevers, Hale-Seda Radoykova, Victor Rossier, Alex Warwick Vesztrocy, Natasha M Glover, and Christophe Dessimoz. OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. *Nucleic Acids Research*, 49(D1):D373–D379, January 2021.
38. Daniël Splinter, David S. Razafsky, Max A. Schlager, Andrea Serra-Marques, Ilya Grigoriev, Jeroen Demmers, Nanda Keijzer, Kai Jiang, Ina Poser, Anthony A. Hyman, Casper C. Hoogenraad, Stephen J. King, and Anna Akhmanova. BICD2, dynactin, and LIS1 cooperate in regulating dynein recruitment to cellular structures. *Molecular Biology of the Cell*, 23(21):4226–4241, November 2012.
39. Homo sapiens (Human) | Proteomes | UniProt.
40. Mengfei Cao and Lenore J. Cowen. WHEN SHOULD WE NOT TRANSFER FUNCTIONAL ANNOTATION BETWEEN SEQUENCE PARALOGS? *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22:15–26, 2017.
41. Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018, March 2016.