A MOSAIC of methods: Improving ortholog detection through the integration of algorithmic diversity

M. Cyrus Maher[1] and Ryan D. Hernandez[2,3,4]

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco, 185 Berry Street, Lobby 5, Suite 5700 San Francisco, CA 94107 [2]Department of Bioengineering and Therapeutic Sciences, [3]Institute for Human Genetics, [4]Institute for Quantitative Biosciences (QB3), University of California, San Francisco, 1700 4th Street San Francisco, California 94158

Keywords:

Multiple sequence alignment, ortholog detection, evolution

Abstract:

Ortholog detection (OD) is a critical step for comparative genomic analysis of protein-coding sequences. There is a range of methods available for OD. However, relative performance varies by application, stymying attempts to identify a single best method. In this paper, we present a novel tool, MOSAIC, which is capable of integrating the entire swath of OD methods. We analyze the results of applying MOSAIC over four methodologically diverse OD methods. Relative to component and competing methods, we demonstrate large gains in the number of detected orthologs while simultaneously maintaining or improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality.

Background:

Orthologs are genes shared between organisms that derive from a common ancestral gene but have diverged from one another through speciation. This is in contrast to paralogs, which arise through gene duplication within a given genome. It is common in comparative genomics and phylogenetics to extract evolutionary information about a given gene through the alignment of orthologous sequences. To make this possible, orthologs must first be inferred, making ortholog detection (OD) an indispensible first step in a variety of phylogenetic inference tasks [1, 2].

In general, existing OD methods can be classified as tree-based, graph-based, or a hybrid of the two [3]. Tree-based methods may use reconciliation techniques between gene and species trees or may rely on the gene tree alone. Graph-based methods can employ a variety of metrics to quantify similarity between sequences, including BLAST scores or sequence identity. Information about the conserved gene neighborhood may also be included in this context. Techniques such as Markov clustering may then be applied to create orthologous groups, or one may simply define clusters based on the graph's existing connections [4].

Unfortunately, the few benchmarking studies that have sampled broadly from this methodological diversity have provided equivocal results. Although there are general trends in relative effectiveness between individual methods, performance is highly context-dependent and does not always favor more sophisticated approaches [5–7]. This is discouraging from the point of view of identifying a single best OD method, but it also suggests a new and relatively facile avenue for methodological improvement. By harnessing differences between OD methods, a wide variety of algorithms may play complementary roles within a cooperative inference framework.

Here, we chose to analyze four methodologically distinct OD methods: 1.) MultiParanoid, a reciprocal-BLAST plus Markov clustering method [8]; 2.) TBA, a synteny-based aligner used to produce UCSC's MultiZ alignments [9]; 3.) six-frame translated BLAT, a fast, approximately-scored protein query approach that does not rely on predicted proteomes [10]; and 4.) OMA, a well-established tree-based method [11]. We apply these methods to OD for a range of primates and closely related mammals and demonstrate that the relative quality of these methods varies widely by species and appear to depend critically on genome quality.

We then characterize the striking performance gains yielded by combining these methods using metrics based on sequence identity, phylogenetic tree concordance, and Hidden Markov Model-based functional agreement. Finally, we demonstrate that this integrative approach yields large improvements over metaPhOrs [12], a methodologically homogeneous OD database that reconciles data from seven phylogeny-based OD methods. The

implementation of this novel approach for the integration of diverse ortholog detection methods is presented as the software tool, MOSAIC, or **M**ultiple **O**rthologous **S**equence **A**nalysis and **I**ntegration by **C**luster optimization. MOSAIC offers striking gains in the number of orthologs detected relative to existing approaches, while simultaneously maintaining or improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality. MOSAIC is a python package that can be installed using easy_install bio-mosaic from the command-line. MOSAIC alignments, source code, and full documentation are available at http://pythonhosted.org/bio-MOSAIC.

Results and Discussion:

Ortholog detection methods frequently outperform one another

It is well-known in theory [13] and in practice [14] that the comparative performance of competing statistical inference algorithms often varies by context. Rather than search for a single best algorithm, researchers have sought to integrate a variety of methods in order to reap the benefits of methodological complementarity [15–17]. As might be expected, the gains yielded by this approach generally scale with the quality of the individual methods integrated, the number of methods included, and, importantly, the diversity of the comprised algorithms [18]. In the present study, we integrate results from four OD methods that were chosen for their high-quality and methodological diversity.

In Figure 1, we show the head-to-head performances of these different methods for a range of primates and closely related mammals. Performance is assessed on alignments generated between all human consensus coding sequences (CCDS) [19] and their

corresponding orthologs. More specifically, for each possible orthologous sequence, we examine the proportion of orthologs from all species for which the level of sequence identity to human is at least five percentage points higher for one particular method versus another. By this metric, we observe that one method significantly outperforms another 10% to 30% of the time. Importantly, no method uniformly outperforms all others, underlining the complementarity of the chosen methods.

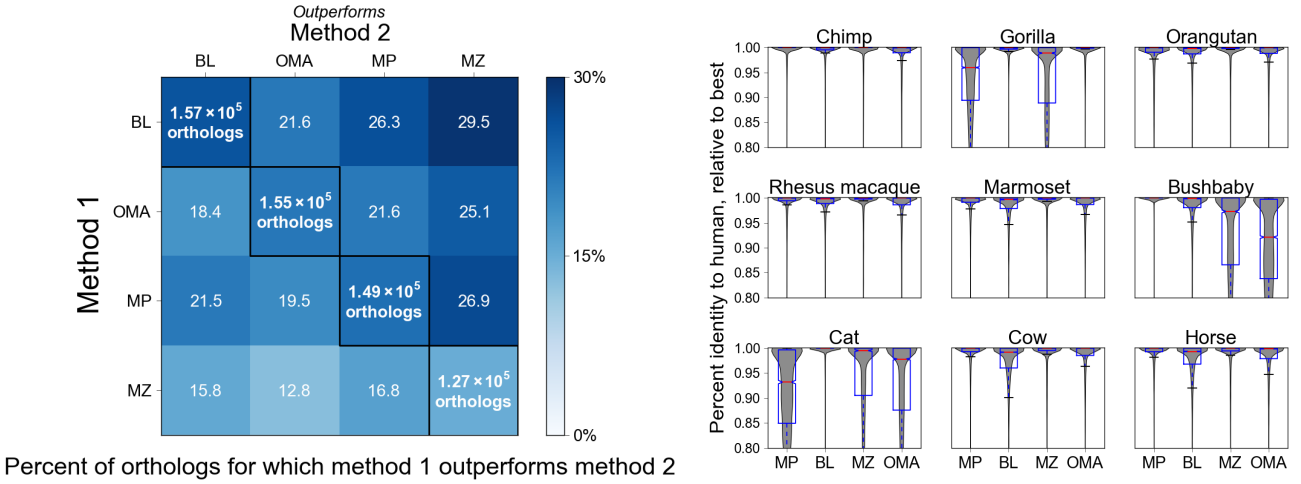

Figure 1. Comparison of sequence identity levels between methods *A.) Heat map of the percent of orthologs for which MultiParanoid (MP), OMA (OMA), BLAT (BL) and MultiZ (MZ) outperform one another. Performance is based on percent identity of each method's orthologs to the human sequence. One method is considered to outperform another method if it improves percent identity by at least five percentage points. Text in diagonal cells shows the number of orthologs identified by each method, colored by the percent of transcripts at which a given method outperforms all the others. B.) Distributions of percent identity relative to the highest scoring ortholog, stratified by species.*

We next evaluated percent identity to human for each ortholog proposed by each method relative to the highest scoring ortholog from all methods. Figure 2B demonstrates that relative performance is species-specific. In particular, we note that the performance disparities across methods are much more pronounced for gorilla, bushbaby, and cat, both in terms of the number and quality of obtained orthologs.

Examining each OD method in detail yields some hypotheses about the origin of these differences in performance. Errors in proteome prediction, both in terms of false-positives and false-negatives, are likely to have large effects on both MultiParanoid and OMA. Meanwhile, spurious syntenic information is expected to compromise the integrity of ortholog predictions produced by MultiZ. Finally, the lack of an assembled genome for bushbaby may negatively impact the quality of the one-way BLAT approach due to the segmentation of exon sets across multiple unordered scaffolds.

Combining multiple sequence alignments with MOSAIC

Having observed the complementarity between OD methods, we sought to develop a structure for the automatic integration of methodologically diverse OD methods such as those described above. We term this framework MOSAIC, or **M**ultiple **O**rthologous **S**equence **A**nalysis and **I**ntegration by **C**luster optimization. MOSAIC allows for the flexible integration of diverse OD methods through the application of standard or user-defined metrics of sequence divergence and ortholog cluster quality. Using specified divergence metrics, clusters of proposed orthologs are built. These orthologs are then adopted or rejected in order to optimize cluster completeness and quality (e.g., distance to a reference sequence or average pairwise distance).

For the examples presented here, we consider a protein set with relatively low levels of evolutionary divergence, and so choose percent identity as our metric for sequence divergence. However, for more distantly related species, the application of scoring matrices [20, 21] or Hidden Markov Models [22] may be preferable for measuring divergence. For each human sequence, each method may propose an ortholog from each species.

Corresponding putative orthologs are then evaluated according to the percent of sites in the human protein sequence that are identical to it (indels and substitutions are penalized equivalently per base, though affine scoring could be accommodated). The best-scoring ortholog among all methods is then chosen for each species.

Combining methods increases the number of included sequences

To assess the efficacy of MOSAIC, we first examined the total number of species included in alignments to human CCDS sequences. For MOSAIC and each OD method, we observe the number of alignments to human CCDS as a function of the maximum number of missing species allowed. Strikingly, the integration of methods more than quintuples the number of alignments for which all species are present (Figure 2a). As expected, the gains afforded by MOSAIC are species-specific and increase as a function of the number of methods that are included (Figure 2b). Using MultiZ as a baseline, we observe once again that the largest improvements are seen for gorilla, bushbaby, and cat. Importantly, orthologs for each of these three species are rescued by different methods (OMA for gorilla, MultiParanoid for bushbaby, and BLAT for cat), further demonstrating the power of integrating diverse OD methods.

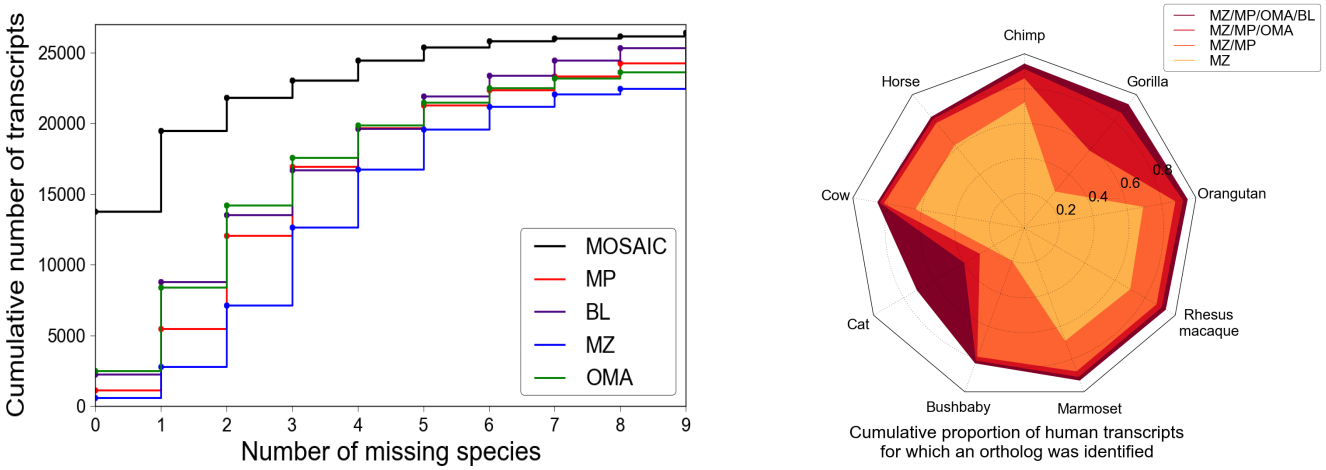

Figure 2. OD power and the effect of pooling methods *A.) The cumulative number of human transcripts as a function of the maximum number of missing species allowed. B.) The cumulative proportion of human transcripts for which an ortholog was detected, stratified by species. Envelopes illustrate results from pooling an increasing number of methods.*

The addition of new sequences does not sacrifice average levels of sequence identity

In our current examples, MOSAIC optimizes sequence identity to human, limiting the utility of this metric for assessing performance. Indeed, for a given putative ortholog, MOSAIC is guaranteed to improve or maintain percent identity compared to its constituent methods. Counterintuitively, this provides no assurance that MOSAIC will provide gains in *average* levels of percent identity. For example, average levels of percent identity could decrease if MOSAIC ensures the inclusion of a greater number of species by pulling in poorly scoring sequences that were initially filtered out by the majority of component methods. However, we see that this is not the case. Indeed, for each species MOSAIC retrieves a much larger number of sequences than any method alone, while maintaining levels of percent identity comparable to those of the best performing method (Figure 3).

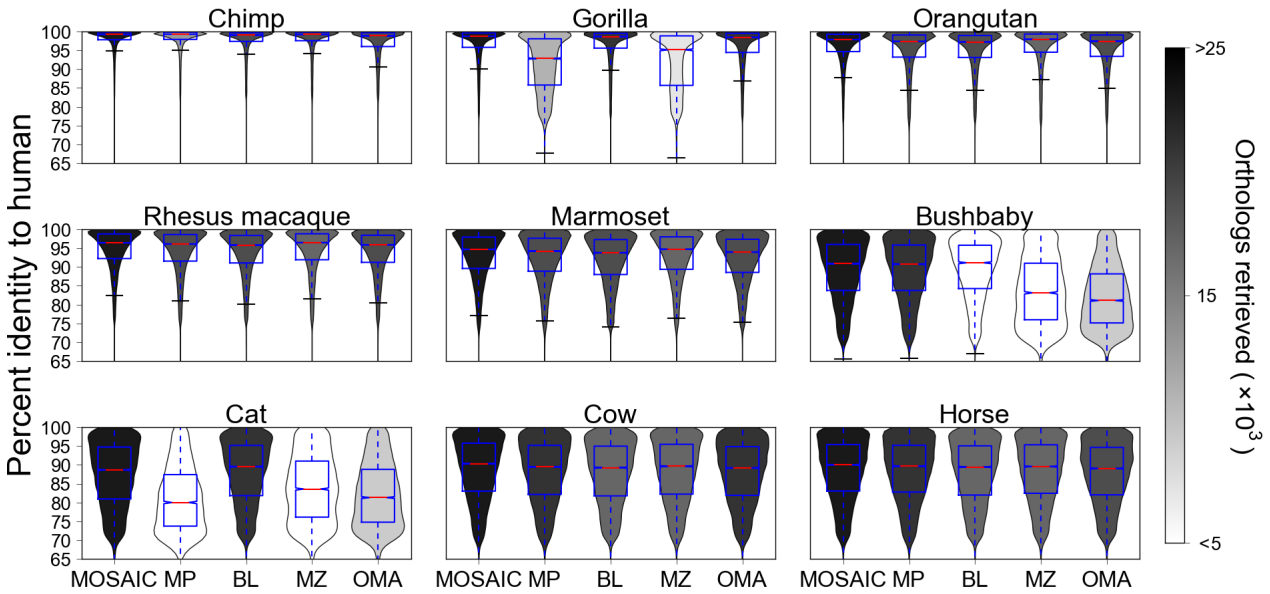

Figure 3. The effect of method integration on sequence identity. *A comparison of the overall distributions of percent identity to human for MOSAIC and its component methods. As in Figure 1B, smoothed distributions underlying the boxplots are shaded according to the number of human transcripts for which an ortholog was proposed. White denotes 5000 sequences or less. Darker shades signify increasingly larger numbers of detected orthologs.*

Integrating methods leads to higher levels of phylogenetic and functional concordance

To further examine the effect of MOSAIC on alignment quality, we compared phylogenetic and functional concordance across methods. Phylogenetic concordance was ascertained by calculating the normalized, unweighted Robinson-Foulds (RF) distance [23] between gene trees and the established species tree [24]. This metric is equal to the sum of the number of splits in one tree that are not present in the other, scaled by the total number of splits present across the two trees. Accordingly, larger RF distances correspond to worse agreement between gene and species trees. On a gene-by-gene basis, this metric should be interpreted with caution, since post-speciation admixture can lead to true discordance between the species tree and the phylogenetic history of a particular gene [25]. However, at

the level of the genome, higher concordance between gene trees and the known speciation process strongly suggests a relative improvement in OD.

Figure 4a presents the cumulative proportion of alignments included as a function of the maximum allowable RF distance. Multiz is seen to perform the best of any individual method, likely due to its utilization of syntenic information. Surprisingly, the tree-based OD method, OMA, is seen to be the worst performing method according to this tree-based metric. Combining all methods using MOSAIC leads to a strong enrichment of highly concordant gene trees, while providing performance that is competitive to MultiZ at more permissive RF distance cutoffs.

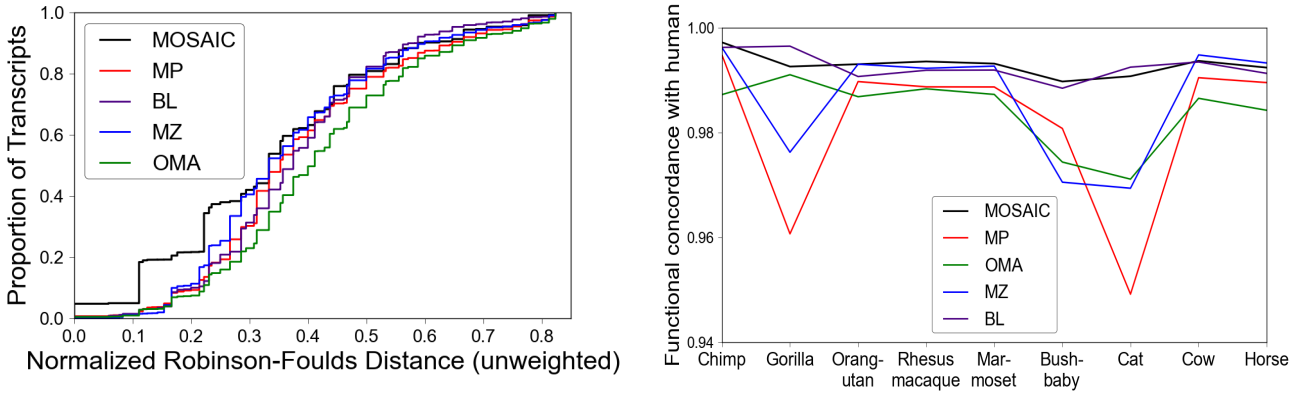

Figure 4. The effect of method integration on tree-based and functional concordance. *A.) The cumulative proportion of human transcripts as a function of the maximum allowable Robinson-Foulds distance between the gene tree and the species tree. B.) The rate of concordance between functional annotations for proposal orthologs and human transcripts.*

In addition, we used profile HMMs from the Protein Families Database A (PfamA) [26] and HMMER3 [27] to ascertain functional concordance between proposed orthologs and the human CCDS of interest. PfamA builds HMMs via curated alignments of small numbers of representative members from each protein family. Using HMMER3, we queried protein

sequences against all PfamA protein family profiles, annotating each protein according to its top protein family hit. This allowed for an ascertainment of functional concordance that is vastly more comprehensive than relying on gene-by-gene annotation across species, while retaining many of the advantages of manual curation. This assessment reveals that, for the set of orthologous sequences proposed by all methods, MOSAIC provides levels of functional concordance that are comparable to the best performing method and considerably better than most methods for gorilla, bushbaby, and cat (Figure 4b).

MOSAIC significantly outperforms metaPhOrs

We have shown that MOSAIC provides a large increase in the number of detected orthologs relative to its component methods, while simultaneously maintaining or improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality. Next, we sought to compare this method of OD integration to the only alternative of which we are aware: metaPhOrs [12]. Using an approach based on tree overlap, metaPhOrs integrates ortholog predictions using phylogenetic trees from seven databases: PhylomeDB, Ensembl, TreeFam, EggNOG, OrthoMCL, COG, and Fungal Orthogroups.

We compared MOSAIC and metaPhOrs based on the number of retrieved orthologs, average differences in sequence identity, and comparative levels of functional and phylogenetic concordance. We observe that MOSAIC provides large increases in the number of retrieved orthologs, while providing slight improvements in sequence identity for those cases where proposal orthologs are available from both methods (Figure 5). For the cases where MOSAIC predicted an ortholog but metaPhOrs did not, we examined the level of sequence identity in these sequences compared to the species-specific average returned by

metaPhOrs. We find that these additional sequences display levels of sequence identity comparable to those provided by metaPhOrs. Finally, we observe that MOSAIC yields a slight increase in functional concordance, as well as a 40% increase in tree concordance, measured as the area under the curve below an RF distance of 0.5. A 0.5 threshold was chosen because there is little differentiation between methods after this point.

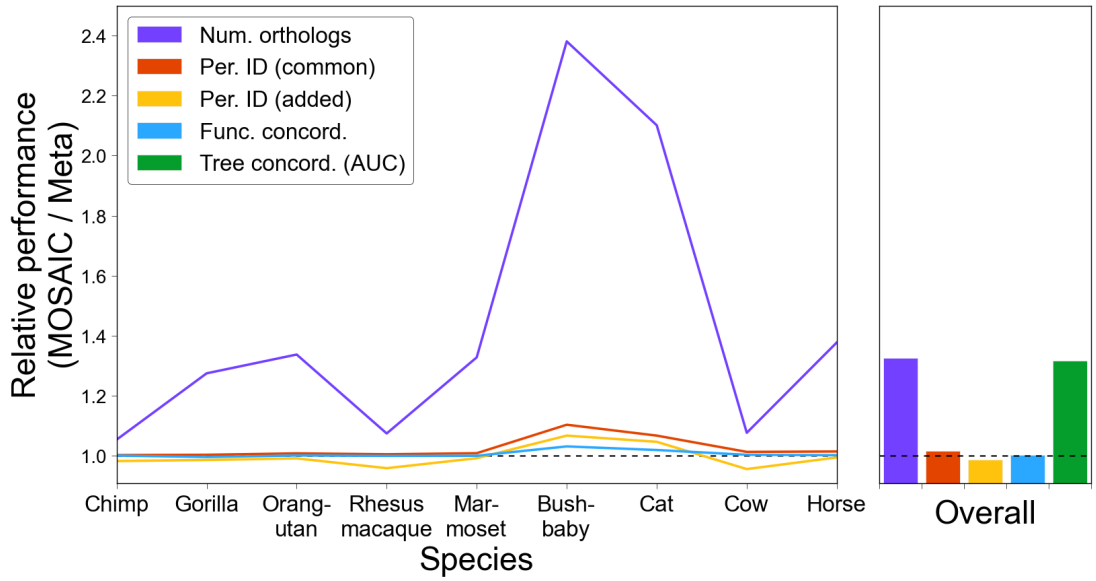

Figure 5. A comparison between MOSAIC and metaPhOrs. *The relative performance between MOSAIC and metaPhOrs according to five metrics: 1.) the number of orthologs detected (purple); 2.) the percent identity to human for orthologs present in both (red); 3.) the percent identity to human for orthologs unique to MOSAIC compared to metaPhOrs species-specific average (yellow); 4.) rate of functional concordance between proposal orthologs and human transcripts (blue); and 5.) concordance between gene and species trees, as measured by a normalized, unweighted Robinson-Foulds distance (green).*

Conclusions:

In this paper we have introduced a novel algorithm, MOSAIC, which is capable of integrating an arbitrary number of methodologically diverse ortholog detection methods. We have demonstrated that MOSAIC provides large increases in power relative to its component

methods, while simultaneously maintaining or improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality. Furthermore, despite including several fewer methods, MOSAIC demonstrates significant improvements over, metaPhOrs, the only other OD integration method of which we are aware.

In summary, MOSAIC provides the unique flexibility to incorporate any OD method, thereby increasing methodological diversity, improving OD performance, and allowing researchers to take advantage of methodological gains in a variety of areas of OD research. MOSAIC is a python package that can be installed using `easy_install bio-mosaic` from the command-line. MOSAIC alignments, source code, and full documentation are available at http://pythonhosted.org/bio-MOSAIC.

Materials and methods:

Retrieval of orthologs

For each human consensus coding sequence, we sought to retrieve orthologs for chimp, gorilla, orangutan, rhesus macaque, marmoset, bushbaby, cat, cow, and horse. In the case of MultiZ [9], CCDS orthologs were downloaded directly from the UCSC genome browser [29]. For OMA [11], ortholog predictions were downloaded from omabrowser.org. For genes with more than one CCDS, orthologs were mapped to each analyzed transcript.

For BLAT [10], genomes for each species of interest were downloaded from the NCBI Entrez Genome database [30]. Queries were conducted using the following command structure:

blat -q=prot -t=dnax -minIdentity=70 –extendThroughN [genome file] [query file] [output file]

For MultiParanoid [8], an all-versus-all blast search was run using the following command structure:

blastp -db $blastdatabase -query [query file] -out [output file] -evalue .01 -num_threads [number of threads] -outfmt 6 -db_soft_mask 21 -word_size 3 -use_sw_tback

From this output, ortholog predictions were produced using the standard MultiParanoid protocol.

To remove possibly spurious orthologs, proposals from each method were then filtered according to a species-specific sequence identity cutoff, as described below.

Filtering and integration of orthologs

For each proposed ortholog for a given CCDS, the CCDS and the orthologous sequence under consideration were globally realigned using the program stretcher from the EMBOSS toolkit [31]. Percent identity was then calculated as the percent of sites in the human sequence that were identical in the orthologous sequence. For example, the hypothetical sequence below would be scored as 71% identical (5/7), since there are 2 mismatches between the seven sites present in the human sequence and the character to which those sites are aligned in the chimp sequence (sites where the human sequence has been deleted or the outgroup has an insertion are ignored):

Human A W V A - T F D

Chimp - W V R Y T F D

All orthologs with percent identity below a critical threshold were removed from all subsequent analyses. This cutoff was chosen considering the known level of genome-wide divergence between human and the species of interest, as well as the overall distributions of percent identity between putative orthologs in the two species. These cutoffs were as follows: chimp: 82%, gorilla: 77%, orangutan: 75%, rhesus macaque: 73%. A cutoff of 70% was employed for marmoset, bushbaby, cat, cow, and horse.

After filtering, the orthologs with the highest percent identity from each species were accepted into the integrated orthologous cluster. These sequences were then aligned using MSAprobs [32].

Quality assessment

Sequence identity

Sequences aligned pairwise often display higher agreement than is observed in the comparison of the same two sequences within the context of a multiple sequence alignment (MSA). Since it is the quality of the MSA that is of primary importance to many downstream phylogenetic inference tasks, we appraised sequence identity with respect to the MSA rather than the pairwise alignment. This allows us to indirectly incorporate information about intra-cluster similarity, since MSAs between divergent sequences are expected to exhibit a larger degradation in performance relative to pairwise alignments.

Tree concordance

For each MSA, gene trees were built using RAxML [33]. An unweighted Robinson-Foulds (RF) distance [23] was then calculated between each gene tree and the known species tree using the python module dendropy [34]. Briefly, the unweighted RF distance counts the number of operations required to transform one tree into the other. This quantity is equal to the total number of splits that are present in one tree but not the other. To normalize for variations in tree size, we then divided this distance by the sum of the total number of splits in the gene and species trees [35].

Functional concordance

Profile HMMs were downloaded from the PfamA protein families database [26]. Each sequence was then annotated using the top scoring function retrieved by querying that sequence against the database of all PfamA protein family HMMs. This search was conducted using HMMER3 [27]. Functional concordance was then measured as a binary quantity,

corresponding to whether or not a putative orthologous sequence had the same inferred function as its cognate human sequence.

All data were plotted using the python module matplotlib [36].

List of abbreviations:

Ortholog detection (OD), multiple sequence alignment (MSA), hidden markov model (HMM), Robinson-Foulds (RF), MultiParanoid (MP), MultiZ (MZ), BLAT (BL), OMA (OM)

Competing interests:

The authors have no competing interests to declare.

Authors' contributions:

MCM and RDH conceived the study. MCM wrote the software and gathered and analyzed the data. MCM and RDH drafted the paper. Both authors read and approved the final manuscript.

Acknowledgements:

The authors would like to thank Raul Torres, Lawrence Urrichio, and Zachary Szpiech for their feedback regarding this manuscript. This work was partially supported by the National Institutes of Health (grants P60MD006902, UL1RR024131, 1R21HG007233, 1R21CA178706, and 1R01HL117004 to R.D.H.). M.C.M. was supported by the Epidemiology and Translational Science program at UCSF, an NIH F31 Predoctoral Fellowship, and a Lloyd M. Kozloff Fellowship.

References:

1. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science (New York, NY)* 2006, **311**:1283–7.
2. Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation.** *Nature reviews Genetics* 2012, **13**:329–42.
3. Orthology I, Altenhoff AM, Dessimoz C: **Evolutionary Genomics**. 2012, **855**.
4. Kuzniar A, van Ham RCHJ, Pongor S, Leunissen J a M: **The quest for orthologs: finding the corresponding gene across genomes.** *Trends in genetics : TIG* 2008, **24**:539–51.
5. Chen F, Mackey AJ, Vermunt JK, Roos DS: **Assessing performance of orthology detection strategies applied to eukaryotic genomes.** *PloS one* 2007, **2**:e383.
6. Hulsen T, Huynen MA, de Vlieg J, Groenen PMA: **Benchmarking ortholog identification methods using functional genomics data.** *Genome biology* 2006, **7**:R31.
7. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS computational biology* 2009, **5**:e1000262.
8. Alexeyenko A, Tamas I, Liu G, Sonnhammer ELL: **Automatic clustering of orthologs and inparalogs shared by multiple proteomes.** *Bioinformatics* 2006, **22**:e9–15.
9. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF a, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome research* 2004, **14**:708–15.
10. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome research* 2002, **12**:656–64.
11. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C: **OMA 2011: orthology inference among 1000 complete genomes.** *Nucleic acids research* 2011, **39**(Database issue):D289–94.
12. Pryszcz LP, Huerta-Cepas J, Gabaldón T: **MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score.** *Nucleic acids research* 2011, **39**:e32.
13. **No Free Lunch Theorems For Optimization - Evolutionary Computation, IEEE Transactions on - 78.pdf** [http://ti.arc.nasa.gov/m/profile/dhw/papers/78.pdf]
14. Van der Laan MJ, Gruber S: **Collaborative double robust targeted maximum likelihood estimation.** *The international journal of biostatistics* 2010, **6**:Article 17.
15. Van der Laan MJ, Polley EC, Hubbard AE: **Super learner.** *Statistical applications in genetics and molecular biology* 2007, **6**:Article25.
16. Chandrasekaran V, Jordan MI: **Computational and statistical tradeoffs via convex relaxation.** *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**:E1181–90.
17. Rokach L: **Ensemble-based classifiers**. *Artificial Intelligence Review* 2009, **33**:1–39.
18. Kuncheva LI, Whitaker CJ: **Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy**. *Machine Learning* 2003, **51**:181–207.
19. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner M-M, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, et al.: **The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.** *Genome research* 2009, **19**:1316–23.
20. Henikoff S: **Amino Acid Substitution Matrices from Protein Blocks**. *Proceedings of the National Academy of Sciences* 1992, **89**:10915–10919.

21. Dayhoff MO, Schwartz RM, Orcutt BC: **A model of evolutionary change in proteins**. In *Atlas of protein sequence and structure*. Edited by Dayhoff MO. Nature Biomedical Research; 1978:345–358.
22. Ebersberger I, Strauss S, von Haeseler A: **HaMStR: profile hidden markov model based search for orthologs in ESTs.** *BMC evolutionary biology* 2009, **9**:157.
23. Robinson DF, Foulds LR: **Comparison of phylogenetic trees**. *Mathematical Biosciences* 1981, **53**:131–147.
24. Altenhoff AM, Dessimoz C: **Phylogenetic and functional assessment of orthologs inference projects and methods.** *PLoS computational biology* 2009, **5**:e1000262.
25. Maddison WP, Knowles LL: **Inferring phylogeny despite incomplete lineage sorting.** *Systematic biology* 2006, **55**:21–30.
26. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database.** *Nucleic acids research* 2012, **40**(Database issue):D290–301.
27. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS computational biology* 2011, **7**:e1002195.
28. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics (Oxford, England)* 2009, **25**:1422–3.
29. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D: **The Human Genome Browser at UCSC**. *Genome Research* 2002, **12**:996–1006.
30. McEntyre .J, Ostell J (Eds): *The NCBI Handbook*. Bethesda, MD: National Center for Biotechnology Information; 2002.
31. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends in genetics : TIG* 2000, **16**:276–7.
32. Liu Y, Schmidt B, Maskell DL: **MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities.** *Bioinformatics (Oxford, England)* 2010, **26**:1958–64.
33. Stamatakis A, Alachiotis N: **Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data.** *Bioinformatics (Oxford, England)* 2010, **26**:i132–9.
34. Sukumaran J, Holder MT: **DendroPy: a Python library for phylogenetic computing.** *Bioinformatics (Oxford, England)* 2010, **26**:1569–71.
35. Yu C, Zavaljevski N, Desai V, Reifman J: **QuartetS: a fast and accurate algorithm for large-scale orthology detection.** *Nucleic acids research* 2011, **39**:e88.
36. Hunter JD: **Matplotlib: A 2D graphics environment**. *Computing in Science & Engineering* 2007, **9**:90–95.