

Rock, Paper, Scissors: Harnessing Complementarity in Ortholog Detection Methods Improves Comparative Genomic Inference

M. Cyrus Maher^{*,1} and Ryan D. Hernandez^{†,‡,§,2}

^{*}Department of Epidemiology and Biostatistics, University of California, San Francisco, and [†]Department of Bioengineering and Therapeutic Sciences, [‡]Institute for Human Genetics, and [§]Institute for Quantitative Biosciences (QB3), University of California, San Francisco, San Francisco, California 94158

ABSTRACT Ortholog detection (OD) is a lynchpin of most statistical methods in comparative genomics. This task involves accurately identifying genes across species that descend from a common ancestral sequence. OD methods comprise a wide variety of approaches, each with their own benefits and costs under a variety of evolutionary and practical scenarios. In this article, we examine the proteomes of ten mammals by using four methodologically distinct, rigorously filtered OD methods. In head-to-head comparisons, we find that these algorithms significantly outperform one another for 38–45% of the genes analyzed. We leverage this high complementarity through the development MOSAIC, or **M**ultiple **O**rthologous **S**equences **A**nalysis and **I**ntegration by **C**luster optimization, the first tool for integrating methodologically diverse OD methods. Relative to the four methods examined, MOSAIC more than *quintuples* the number of alignments for which all species are present while simultaneously maintaining or improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality. Further, this improvement in alignment quality yields more confidently aligned sites and higher levels of overall conservation, while simultaneously detecting of up to 180% more positively selected sites. We close by highlighting a MOSAIC-specific positively selected sites near the active site of TPSAB1, an enzyme linked to asthma, heart disease, and irritable bowel disease. MOSAIC alignments, source code, and full documentation are available at <http://pythonhosted.org/bio-MOSAIC>.

KEYWORDS

multiple
sequence
alignment
ortholog
detection
comparative
genomics
positive selection
open source
software

Orthologs are genes that derive from a common ancestral gene but that have diverged from one another through speciation. This route 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 particular gene from its alignment with orthologous sequences. To enable this analysis, orthologs must first be inferred, making ortholog detection (OD) an

indispensable first step in a variety of phylogenetic inference tasks (Ciccarelli *et al.* 2006; Yandell and Ence 2012).

In general, existing OD methods can be classified as tree-based, graph-based, or a hybrid of the two (Altenhoff and Dessimoz 2012). Tree-based methods may use reconciliation techniques between gene and species trees or may rely on the gene tree alone. Graph-based methods can use a variety of metrics to quantify similarity between sequences, including Basic Local Alignment Search Tool (BLAST) scores or sequence identity. Information about the conserved gene neighborhood also may 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 a graph's existing connections (Kuzniar *et al.* 2008).

Unfortunately, the few benchmarking studies that have sampled broadly from this methodologic diversity have provided equivocal results. Although there are general patterns in relative effectiveness between methods, performance is highly context-dependent and does not always favor more sophisticated approaches (Hulsen *et al.* 2006; Chen *et al.* 2007; Altenhoff and Dessimoz 2009a). This is discouraging

Copyright © 2015 Maher and Hernandez

doi: 10.1534/g3.115.017095

Manuscript received October 29, 2014; accepted for publication February 6, 2015; published Early Online February 23, 2015.

This is an open-access article distributed under the terms of the Creative Commons Attribution Unported License (<http://creativecommons.org/licenses/by/3.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supporting information is available online at <http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.115.017095/-/DC1>

¹Present address: 1700 4th Street, San Francisco, CA 94158.

²Corresponding author: UCSF, Box 2530, 1700 4th Street, San Francisco, CA 94158. Phone: 415-514-9813. E-mail: ryan.hernandez@ucsf.edu

from the point of view of identifying a single best OD method, but it also suggests a new and relatively facile avenue for methodologic improvement. By harnessing differences between OD methods, a wide variety of algorithms may play complementary roles within a cooperative inference framework.

We begin our analysis with a comprehensive comparison of four popular and methodologically distinct OD methods: (1) MultiParanoid, a reciprocal-BLAST plus Markov clustering method (Alexeyenko *et al.* 2006); (2) TBA, a synteny-based aligner used to produce University of California Santa Cruz's MultiZ alignments (Blanchette *et al.* 2004); (3) six-frame translated BLAT, a fast, approximately-scored protein query approach that does not rely on predicted proteomes (Kent 2002); and (4) OMA, a well-established tree-graph hybrid method (Altenhoff *et al.* 2011). Applying these methods to OD in a range of primates and closely related mammals, we demonstrate that methodological performance varies widely by species and appears to depend critically on genome quality.

Next, we characterize the striking performance gains yielded by combining these methods. This is demonstrated using sequence identity, phylogenetic tree concordance, and hidden Markov model-based functional agreement. This improvement in alignment quality translates to higher estimated levels of overall conservation. At the same time, we detect up to 180% more positively selected sites. We close by highlighting a novel positively selected site (PSS) near the active site of TPSAB1, an enzyme linked to asthma and irritable bowel disease.

The implementation of this novel approach for the integration of diverse ortholog detection methods is presented as the software tool, MOSAIC, or Multiple Orthologous Sequence Analysis and Integration by Cluster optimization. MOSAIC is implemented as a well-documented 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 (version GRCh37.p9), we sought to retrieve orthologs for chimp, gorilla, orangutan, rhesus macaque, marmoset, bushbaby, cat, cow, and horse using four methodologically distinct methods: (1) MultiParanoid (Alexeyenko *et al.* 2006); (2) TBA (Blanchette *et al.* 2004); (3) six-frame translated BLAT (Kent 2002); and (4) OMA (Altenhoff *et al.* 2011; July 2013 release). We used MultiParanoid over OrthoMCL (Li *et al.* 2003) because the latter produced no errors or output after careful execution of the thirteen-stage analysis protocol. For all methods, genomic data were retrieved for the genome builds listed in Table 1.

■ **Table 1 Genome builds**

Genome	Version	Release
Chimp	panTro4	Feb-11
Gorilla	gorGor3.1	May-10
Orangutan	ponAbe2	Jul-11
Rhesus macaque	rheMac3	Oct-10
Marmoset	calJac3	Mar-09
Bushbaby	otoGar3	May-11
Cat	felCat5	Sep-11
Cow	bosTau7	Oct-11
Horse	equCab2	Sep-07

For MultiParanoid (Alexeyenko *et al.* 2006), an all-vs.-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.

For BLAT (Kent 2002), genomes for each species of interest were downloaded from the NCBI Entrez Genome database (McEntyre and Ostell 2002). Queries were conducted using the following command structure:

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

In the case of MultiZ (Blanchette *et al.* 2004), CCDS orthologs were downloaded directly from the UCSC genome browser (Kent *et al.* 2002). For OMA (Altenhoff *et al.* 2011), ortholog predictions were downloaded from omabrowser.org (December 2012 release). For genes with more than one CCDS, orthologs were mapped to each analyzed transcript. Finally, ortholog predictions from metaPhOrs (Pryszcz *et al.* 2011) were retrieved from release v201009 (June 2012).

Filtering putatively non-orthologous sequences

All ortholog detection methods produce false positives. For example, this can result when a gene deletion on one lineage means that no true ortholog exists in a given species. Typically, these issues are dealt with through rigorous filtering of input alignments. The intuition is that by applying a stringent sequence similarity filter, we can remove the vast majority of evolutionarily unrelated genes. We use this filtering approach to ensure that only credible, putatively orthologous sequences are included in the analysis. Because of heterogeneity in genome quality, similarity cutoffs were chosen heuristically, 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. Specifically, we first chose a cutoff based on the species-specific levels of percent identity to human. We then updated these numbers based on spot checks of borderline alignment cases. 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. For applications where consistency across methods is not important, these cutoffs could be chosen using downstream quality metrics such as those presented in Figure 4. Note that such an approach would still require the user to specify a tradeoff between the quality and number of orthologs.

MOSAIC: OD INTEGRATION AS CLUSTER OPTIMIZATION

MOSAIC provides a highly flexible, graph-based framework for integrating diverse OD methods. For a given reference sequence, proposal orthologs are conceptualized as nodes in a graph, connected with edges weighted according to the pairwise similarity between sequences (Figure 1). The task of OD integration is then to choose proposal orthologs from each species such that a chosen measure of intra-cluster similarity is optimized.

MOSAIC optimizes (weighted) pairwise similarities

To begin, MOSAIC calculates pairwise similarities between all orthologs from different species. Percent identity- and BLAST-based

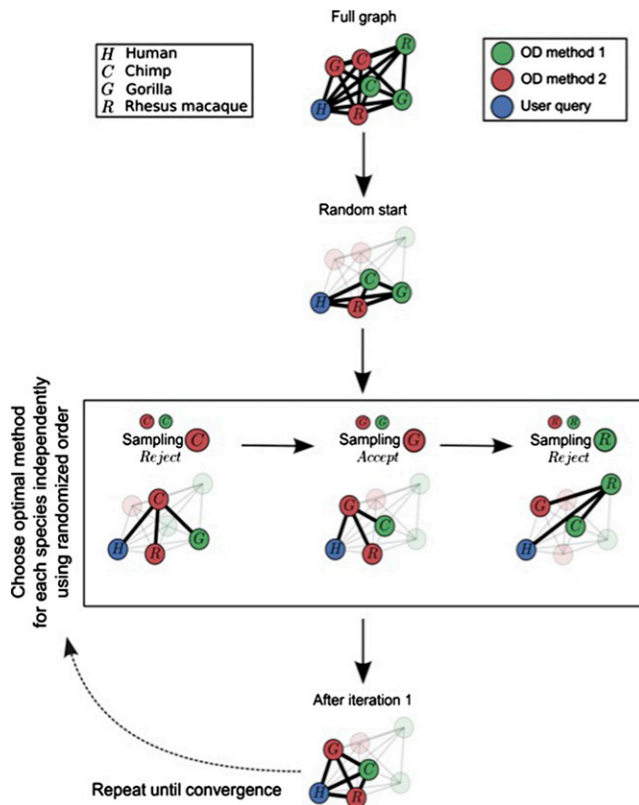


Figure 1 A schematic of the sequence selection algorithm. Steps: (1) Construct graph; (2) Choose the sequence from a random OD method for each species; (3) Iterate through species. For each species, pick the orthologs with highest similarity to the current best choices for all other species; (4) Return current best choices if no changes are made after iterating through all species; (5) To find global optimum, repeat steps 1-4 with random sampling paths.

similarity metrics are provided by default, but user-defined similarity metrics are also accepted. These similarity scores define edge weights, which are used to construct a graph such as the one presented at the top of Figure 1. Once this full graph is constructed, it is highly recommended that it be quality filtered using user-specified similarity cutoffs. This step is necessary to minimize the effect of gene loss, duplication, etc. Once the graph is cleaned, MOSAIC then chooses at most one proposal ortholog from each species so that the overall pairwise similarity between accepted sequences is optimized.

To accommodate user priorities, pairwise similarities can be weighted such that sequences from different species contribute unequally to the total similarity score. For uniform weights, this is equivalent to maximizing the average pairwise similarity. In the case where only similarity to a reference sequence is of interest, this reduces to simply accepting the ortholog for each species that is most similar to the reference.

Optimization is carried out using cyclic coordinate descent

For m OD methods and s species, there are up to m^s possible integrated alignments. In the case analyzed in this paper, $m = 4$ and $s = 10$. This translates to over a million possible integrated alignments for each of the ~25,000 reference sequences considered. It is clear to see from this example that an exhaustive optimization becomes quickly infeasible. Therefore, MOSAIC chooses optimal clusters using cyclic

coordinate descent (CCD), an efficient non-derivative optimization algorithm (Bertsekas 1999).

In Figure 1, we illustrate the way CCD functions in the context of MOSAIC. After the full graph that includes all orthologous sequences is built, random orthologs from each species are chosen as the current best. MOSAIC then loops through the species of interest in a random order. For each species, MOSAIC chooses the sequence that optimizes cluster tightness, given the current best sequences for all other species. This process is repeated until no further improvements can be made to cluster tightness. Finally, because CCD is prone to finding local rather than global optima, this entire process is repeated multiple times with random starting points and sampling paths.

Scoring and optimization procedures for this study

For the alignments presented here, we consider a protein set with relatively low levels of evolutionary divergence. We chose percent identity as our metric for sequence divergence. Note that several other popular scoring functions are implemented in MOSAIC. For more distantly related species, the application of scoring matrices (Dayhoff *et al.* 1978; Henikoff 1992) or Hidden Markov Models (Ebersberger *et al.* 2009) may be preferable. To reduce computational costs related to pairwise alignment, we considered only similarities between orthologs and the human target sequence. The optimization procedure was then equivalent to choosing, for each species, the ortholog among all methods that is most similar to the human sequence. This approach corresponds to the arguments `edgefunc='perID'`, `optrule='pairwise'` when calling the Mosaic constructor in `mosaic.py` (see: <http://pythonhosted.org/bio-MOSAIC/Module.html>).

Example: measuring similarity

Percent identity was 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), because there are two 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
```

A note on gene loss, duplication, and divergent evolution

For any query protein, there is a risk that a related gene in another genome has been deleted, or has not maintained the same function and so provides inapplicable evolutionary information. In the case of deletions, it is unlikely that a non-homologous gene would be suitably similar to be classified as an ortholog. Divergence in function would be expected to sharply increase sequence divergence. In many cases such functionally distinct can be removed by suitably stringent sequence similarity filters.

Another pitfall in ortholog detection is gene duplication within a particular lineage. This results in so-called in-paralogs, which may inject additional bias if, compared to the query protein, the most functionally similar of the set is not the most similar at the sequence level. Although this is possible, it is the exception and not the rule under reasonable models of evolution. Indeed, experimental data from several model systems has demonstrated that there is an extremely high correlation between functional conservation and sequence conservation (Mashiyama *et al.* 2014; Zhao *et al.* 2014). Taking the single most similar sequence from a paralogous group is therefore a rational and

effective (albeit imperfect) approach to this problem. For this reason, MOSAIC does not exclude putatively orthologous sequences that have paralogs in the source genomes. It rather picks the paralogous sequence that is most likely to share the same function as the other putatively orthologous sequences under examination. We will show that this decision allows us to capture more putative orthologs while simultaneously improving ortholog quality by all commonly used metrics.

In summary, MOSAIC is adapted to producing multiple sequence alignments (MSAs) that are functionally informative at the site-level. For other applications, researchers may wish to infer genomic events such as gene loss, duplication, horizontal gene transfer, and/or incomplete lineage sorting (e.g., Capra *et al.* 2013). This involves jointly examining functionally diverged paralogous groups alongside their corresponding orthologs. This task generally requires a combination of tools such as MultiParanoid (to infer paralogs; Remm *et al.* 2001), RaxML (to build gene and species trees; Stamatakis 2014), and Notung (to reconcile gene trees with species trees and infer evolutionary events; Stolzer *et al.* 2012). For applications such as this, MOSAIC alignments can still be leveraged to help ensure the inclusion of relevant sequences. Likewise, reconstructed evolutionary histories can be used to flag, among tens of thousands of automatically generated MOSAIC alignments, those exceptional cases that could benefit most from manual inspection.

Multiple sequence alignment

Retrieved sequences were jointly aligned to query proteins using MSAProbs (Liu *et al.* 2010), a multithreaded aligner with better performance benchmarks than many top aligners, including ClustalW, MAFFT, MUSCLE, ProbCons, and Probalign (Liu *et al.* 2010). Importantly, MSAProbs has the further advantage of providing, for each column of an alignment, dependable estimates of the confidence of the alignment at the site.

Quality assessment

One approach to evaluating the performance of ortholog detection methods is to restrict analysis to validation sets, which usually consist of small, curated gene groups from unicellular model organisms (e.g., Salichos and Rokas 2011). Known true positive and true negative relationships allow researchers to calculate power and sensitivity, and even provide the possibility of applying supervised learning techniques such as support vector machines or random forest classifiers. Although such results are assuredly internally valid, it is unclear whether these results generalize beyond this small and highly biased subset of genes. As a simple example, “true” orthologous relationships often are restricted to cases where synteny is also maintained. Filtering out nonsyntenic orthologous sequences will thus significantly bias performance metrics toward OD methods that use syntenic information. Furthermore, this approach to quality assessment does not allow researchers to evaluate performance on their own arbitrary dataset. It is for these reasons that we decided to use the quality metrics described below. Strictly speaking, the details necessary to truly establish orthology are buried deep in evolutionary time. We therefore believe that ortholog detection is perhaps better viewed as an unsupervised learning problem that is amenable to graph-based cluster optimization.

Sequence identity

MOSAIC optimizes pairwise sequence similarity. In this example, sequence identity is used as the similarity measure, and pairwise similarities are weighted such that only concordance with the human reference sequence is considered. To achieve greater separation

between metrics used for optimization and assessment, comparisons of sequence identity were performed in the context of the full MSAs. We believe this choice is sensible because it is the quality of the MSA that is of primary importance to many downstream phylogenetic inference tasks. In addition, this approach allows us to indirectly incorporate information about intra-cluster similarity. As an MSA incorporates increasingly divergent sequences, performance relative to pairwise alignments is expected to progressively degrade.

Tree concordance

For each MSA, gene trees were built using RAXML (Stamatakis and Alachiotis 2010). An unweighted Robinson-Foulds (RF) distance (Robinson and Foulds 1981) was then calculated between each gene tree and the known species tree using the python module dendropy (Sukumaran and Holder 2010). 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 (Yu *et al.* 2011). To summarize the genome-wide distribution of normalized RF distances, we took the area under the curve of the cumulative distribution function. This was limited to values less than 0.4, since beyond this value there is little difference between the observed curves (see Supporting Information, Figure S5). This metric is superior to, e.g., calculating the proportion of genes below a given threshold because it up-weights smaller RF distances as opposed to, in effect, using non-zero uniform weights below the cutoff value.

Functional concordance

Profile HMMs were downloaded from the PfamA protein families database (Punta *et al.* 2012). Each sequence was then annotated using the top scoring functional class retrieved by querying that sequence against the database of all PfamA protein family HMMs. This search was conducted using HMMER3 (Eddy 2011). 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. It is important to note that not all PfamA HMMs are functionally validated. In cases where experimental validation is unavailable, these HMMs provide a family-specific scoring function that nevertheless yields information not contained in naïve sequence identity measures.

Evolutionary analysis

Gene-level conservation: Alignments were analyzed using phylogenetic analysis by maximum likelihood (PAML) (Yang 2007). For each alignment three models were fit: (1) a neutral model where the ratio of the non-synonymous to synonymous substitution rates (dN/dS) is fixed at one, (2) a conservation model where dN/dS is less than or equal to one, and (3) a positive selection model where some fraction of the sequence is fit under the conservation model, while another dN/dS parameter is estimated freely for the remainder of the sequence. Because evolutionary models are not in general nested, we performed model selection via the popular Akaike information criterion, a method that penalizes a model's fit by its number of included variables (Akaike 1973) and is asymptotically equivalent to maximizing the model's predictive performance on unseen data (Stone 1977).

Despite rigorous model selection procedures, in rare cases PAML may estimate very high levels of selection over a tiny proportion of

a given sequence (even a single site), leading to greatly inflated average levels of dN/dS. To reduce the influence of outlying estimates of selection, all dN/dS values greater than 3 were excluded for the analysis. For all methods, this corresponded to less than 0.05% of all sequences.

Site-level positive selection: The program sitewise likelihood ratio (SLR) (Massingham and Goldman 2005) was used to estimate the number of positively selected sites in each sequence. To eliminate false positives caused by poorly aligned sites, we filtered out all sites estimated by MSAProbs to be aligned to less than 95% confidence. All included positively selected sites estimated at 95% confidence or greater by SLR were included in the subsequence comparison.

To assess agreement in PSS, we calculated the degree of overlap between PSS from all pairs of methods. This was calculated as the size of the genome-wide intersection between sites divided by the union of said sites.

Mapping positively selected sites onto three-dimensional structures

We leveraged UniProt mapping files (<http://www.uniprot.org/docs/pdbtosp>; accessed 9/30/14) to determine which proteins had a relevant structure in the Protein Data Bank (PDB; Berman 2000). We then aligned sequences between PDB structures and candidate genes to determine the degree of coverage and to obtain a mapping between residues. We found 2003 genes for which there was a structure with greater than 70% coverage. Of these, 787 had PSS results from all five ortholog detection methods from at least one species. Reasons for missing data comprise more than 4 missing species in source MSAs and lack of convergence in the PSS calculation. Within this set of 787 genes, 76 proteins had PSS from MOSAIC that were not found with any of the component methods. From this point, the example of TPSAB1 was quickly identified by manual inspection. We then downloaded PDB structure 2ZEC to visualize the location of positively selected sites. To validate sequences used in the analysis, we blasted each ortholog against the human SwissProt database. TPSAB1 was the most similar human protein in each case, confirming that we had retrieved best bidirectional hits. Annotations for each protein were also consistent with alpha/beta tryptases activity (Table S1).

RESULTS AND DISCUSSION

Ortholog detection methods frequently outperform one another

We begin with a comprehensive comparison of four popular, methodologically distinct OD methods. In Figure 2, we show the head-to-head performances of these different methods for a range of primates and closely related mammals. Performance is assessed using alignments between all human consensus coding sequences (Pruitt *et al.* 2009) and their corresponding orthologs from each method. For each possible ortholog (defined by human target sequence and species of origin), we examine whether sequence identity to human is at least five percentage points greater for one method *vs.* another. We otherwise consider the two methods to be tied. By this metric, one method significantly outperforms another 38–45% of the time. Importantly, no method uniformly outperforms all others, underlining the complementarity of the chosen algorithms. For each method, distributions of percent identity and relative performance by species are presented in Figure S1 and Figure S2,

Combining multiple sequence alignments with MOSAIC: It is well-known in theory (Wolpert and Macready 1997) and in practice (Van

Der Laan and Gruber 2010) 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 (Van Der Laan *et al.* 2007; Rokach 2009; Chandrasekaran and Jordan 2013). 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 (Kuncheva and Whitaker 2003).

Having observed the complementarity between OD methods, we sought to develop a structure for the automatic integration of methodologically distinct OD methods such as those described above. We term this framework MOSAIC, or Multiple Orthologous Sequence Analysis and Integration by Cluster optimization. MOSAIC allows for the flexible integration of diverse OD methods through the application of standard or user-defined metrics of sequence similarity and ortholog cluster quality. By the use of the specified similarity metrics, clusters of proposed orthologs are built. These orthologs are then adopted or rejected to optimize cluster completeness and quality (*e.g.*, similarity to a reference sequence or average pairwise similarity).

Having presented a schematic of the algorithm itself in Figure 1, we provide in Figure 3 a view of example inputs and output MSAs. These are illustrations of real alignments for carbonic anhydrase 12, an enzyme critical to a number of biological functions, including the formation of bone, saliva, and gastric acid (Pruitt *et al.* 2014). MSA columns that are aligned to below 95% confidence are displayed in red and masked from the analysis. Orthologs that were not returned for a given species are denoted with a horizontal black bar. Those that were filtered using pre-integration sequence identity cutoffs (see the section *Materials and Methods*) are indicated with gray bars. Sequence identity is measured based on pairwise realignment to the human

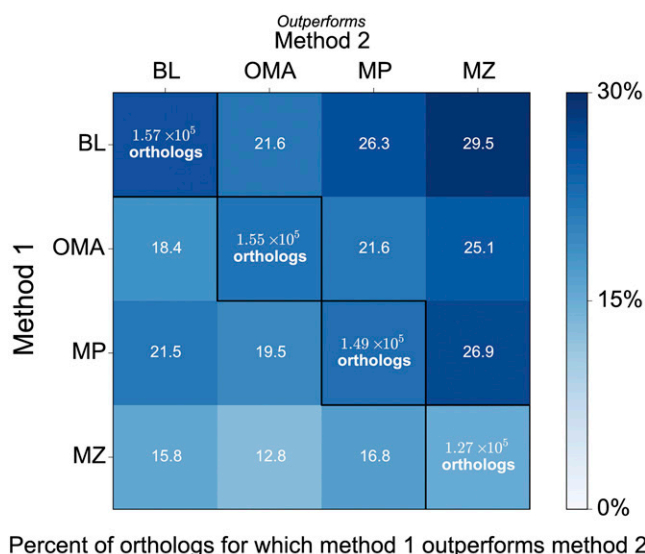


Figure 2 Comparison of sequence identity levels between methods. Heat map of the percent of orthologs for which MultiParanoid (MP), OMA (OMA), BLAT (BL), and MultiZ (MZ) outperform one another. Performance is based on percent identity of each method's orthologs to the human sequence. One method is considered 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 orthologs for which a given method outperforms all the others.

sequence. Note that, just as when employing a single ortholog detection method, this filtering step is critical to guaranteeing alignment quality.

Method integration increases the number of included sequences

The gains afforded by MOSAIC vary by species and increase with the number of methods that are included (Figure 4A). When all four component methods are included, MOSAIC more than quintuples the number of alignments for which all species are present (Figure 4B). We observe in Figure 4A 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. See Figure S3 for further details). In the sections that follow, we will demonstrate that MOSAIC captures these additional sequences while simultaneously improving functional-, phylogenetic-, and sequence identity-based measures of ortholog quality.

MOSAIC improves sequence identity

MOSAIC achieves massive gains in the number of retrieved orthologs while slightly improving average levels of sequence identity. Although MOSAIC directly optimizes sequence identity, this result is non-circular for two reasons. First, average levels of sequence identity could be reduced by preferentially adding sequences from the lower end of the sequence identity distribution. This result would be consistent with a scenario in which most methods correctly inferred that a gene

was deleted on a particular lineage. Second, MOSAIC optimizes sequence identity measured from pairwise global alignments. In the validation phase, we calculated this metric in the context of the full MSA. That is, we do not realign to the human sequence in a pairwise fashion as we do in the optimization phase. Rather, we measure sequence identity based on the alignment specified within the full MSA. We believe this choice is sensible because it is the quality of the MSA that is of primary importance to many downstream phylogenetic inference tasks. In addition, this approach allows us to indirectly incorporate information about intra-cluster similarity. As an MSA incorporates increasingly divergent sequences, performance relative to pairwise alignments is expected to progressively degrade.

MOSAIC improves functional concordance

We employed profile HMMs from the Protein Families Database A (PfamA) (Punta *et al.* 2012) and HMMER3 (Eddy 2011) to ascertain putative functional concordance between proposed orthologs and the human consensus coding sequences of interest. PfamA builds HMMs via curated alignments of small numbers of representative members from each protein family. It is important to note that not all PfamA HMMs are functionally validated. In cases where experimental validation is unavailable, these HMMs provide a family-specific scoring function that yields information not contained in naïve sequence identity measures.

With HMMER3, we queried protein sequences against all PfamA protein family profiles, annotating each protein according to its top

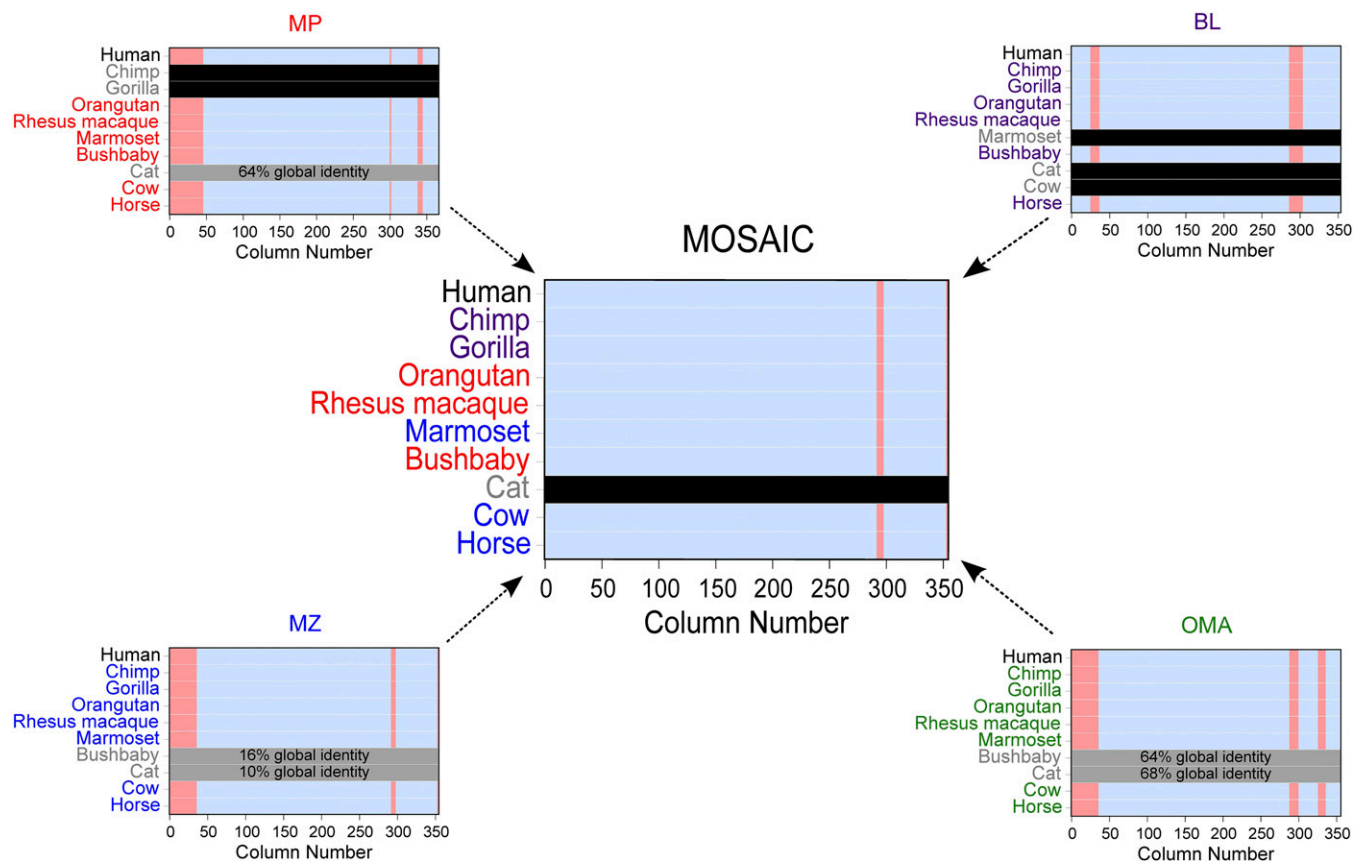


Figure 3 Illustration of integration process for carbonic anhydrase 12. MSA columns that are aligned to below 95% confidence are displayed in red. Orthologs that were not returned for a given species are denoted with a horizontal black bar. Those that were filtered using pre-integration sequence identity cutoffs are indicated with gray bars with the global percent identity from pairwise alignment to human included. Species name label colors denote the species of origin for orthologs in the MOSAIC alignment.

protein family hit. This allowed for an ascertainment of functional concordance that is more comprehensive than relying on gene-by-gene annotation across species, while retaining many of the advantages of manual curation where it exists. This assessment reveals that, for the set of orthologous sequences proposed by all methods, MOSAIC provides levels of functional concordance that are slightly better than the best performing component method (Figure 5). Gains are particularly large for gorilla, bushbaby, and cat orthologs (Figure S4).

MOSAIC improves phylogenetic concordance

Phylogenetic concordance was ascertained by calculating the normalized, unweighted Robinson-Foulds distance (Robinson and Foulds 1981) between gene trees and the established species tree (Altenhoff and Dessimoz 2009b). 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 Robinson-Foulds 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 and incomplete lineage sorting can lead to true discordance between the species tree and the phylogenetic history of a particular gene (Maddison and Knowles 2006). At greater levels of divergence, loss of signal and homoplasy may similarly confound the analysis in some cases. 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 5 presents a comparison of genome-wide phylogenetic concordance (see the section *Materials and Methods* for details on this metric). MultiZ performs the best of any individual method, likely due to its utilization of syntenic information. Surprisingly, OMA, the OD method that incorporates phylogenetic tree information, exhibits the worst performance according to this tree-based metric. MOSAIC, on the other hand, provides significant performance gain over all component methods, including a 59% increase in phylogenetic concordance compared to OMA.

Increased ortholog quality leads to more conservation and positively selected sites

Having demonstrated an increase in ortholog quality using tree-, function-, and similarity-based measures of quality, we next sought to assess the influence of increased alignment quality on estimated levels of selection. To assess gene-level conservation, we applied PAML (Yang 2007) with automated likelihood-based model selection. To ascertain site-level positive selection, we used SLR, a method shown

to have a greater power and a lower false positive rate than PAML's popular Bayes Empirical Bayes method (Massingham and Goldman 2005).

Because varying numbers of sequences can sway evolutionary estimates in unpredictable ways due to, *e.g.*, inhomogeneous levels of selection across organisms, we assessed the performance of MOSAIC relative to each method by matching the species present in each alignment. We refer to this approach as MOSAIC_{matched}. In the case of both PAML and SLR, synonymous substitution rates in coding DNA are used as a background against which to test for changes in rates of non-synonymous substitution. We compared our performance to that of metaPhOrs, an OD integration method that works on tree-based methods only. Since the metaPhOrs database provides only protein sequences for its alignments, no evolutionary comparison with this method was possible given the available data. However, we demonstrate in Figure S6 that MOSAIC outperforms metaPhOrs according to the metrics presented in Figure 3, despite integrating nearly half the number of OD methods in this example. An analysis with a matching number of OD methods was not possible because metaPhOrs is available only as a pre-calculated database.

In Figure 6B, we see that MOSAIC leads to greater gene-level conservation (lower dN/dS) compared with every method except Blat, for which the difference was not statistically significant. Full distributions of dN/dS for each method are presented in Figure S7. Despite greater levels of conservation, MOSAIC was able to detect ~30–180% more positively selected sites than any of its component methods. This was not due to an increase in the inferred rate of positive selection. Rather, most of this increase in power was attributable to the fact that more sites were aligned to high confidence and therefore included in the analysis. This step of filtering for alignment quality is important because site-wise estimates of positive selection are highly sensitive to short poorly aligned regions (Jordan and Goldman 2012).

To investigate the quality of the positively selected sites detected by MOSAIC, we assessed concordance with and between component methods. For a pair of method, we measure overlap by dividing the total size of the intersection between positively selected sites by the total size of the union. These results are shown in Figure 6B. We observe that the minimum overlap between MOSAIC and a component method (MOSAIC/Blat) is still better than the best overlap between component methods (MultiZ/OMA). Averaging over comparisons, we find the improvement in concordance with *vs.* between component methods is statistically significant beyond computational precision ($P < 1e-16$).

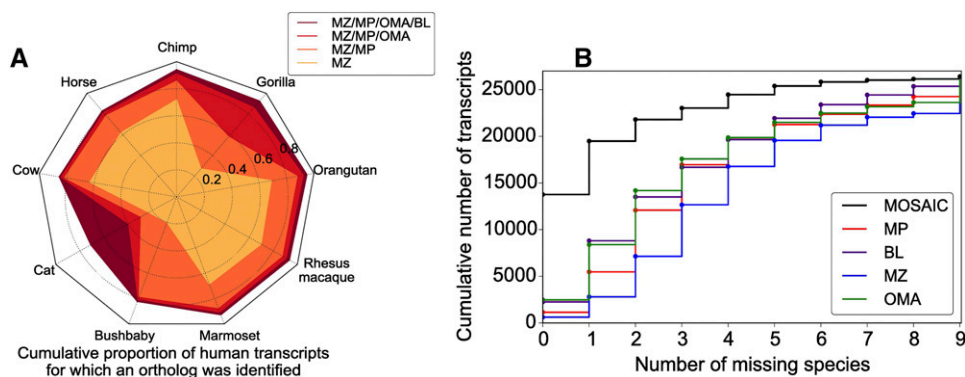


Figure 4 OD power and the effect of pooling methods (A) The cumulative proportion of human transcripts for which an ortholog was detected, stratified by species. Envelopes illustrate results from pooling an increasing number of methods. (B) The cumulative number of human transcripts as a function of the maximum number of missing species allowed.

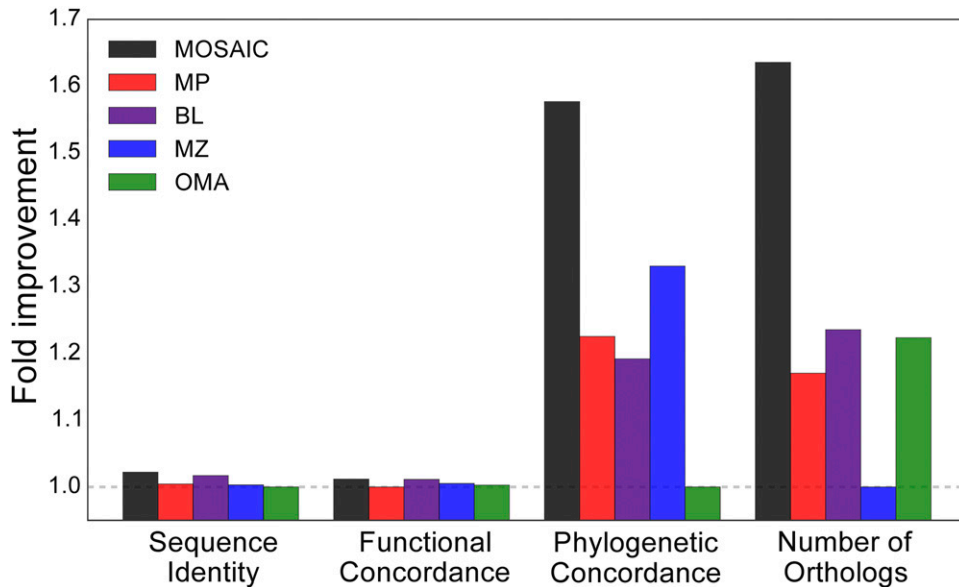


Figure 5 MOSAIC improves alignment quality. We show the fold improvement of each method over the worst performing method in four categories: sequence identity, functional concordance, phylogenetic concordance, and number of orthologs detected.

Understanding MOSAIC's improvements in performance

As we have mentioned, ensemble-based inference frameworks have been shown to operate effectively in a wide variety of statistical contexts. For the methods examined here, MOSAIC may provide increased MSA quality for a variety of reasons, not all of which are related to the performance of the component algorithms. For example, the ability to specify a custom scoring function for pairwise similarity gives MOSAIC an advantage over component methods. For example, BLAST-based scoring such as that used by MultiParanoid may in some cases reward substitutions over than sequence identity. This is the result of building substitution matrices on libraries of sequences that are much more highly diverged than those within mammalian proteomes.

The quality of a genome assembly also plays an important role. For methods like MultiZ, improperly assembled genomic segments may provide misleading information about orthologous relationships. Similarly, naïve methods such as BLAT will miss proteins whose exons are spread across unassembled genomic segments. Other methods such as MultiParanoid and OMA may be robust

to this effect since they draw from more sophisticated proteome predictions. However, the stringent and sometimes arbitrary filtering required for proteome prediction cause these methods to miss sequences that can be found using BLAT's six-frame genomic translation.

Better alignments may yield new insights into human evolution

We next sought to examine the biological significance of some of the positively selected sites identified uniquely by MOSAIC. This led us to Tryptase Alpha/Beta 1 (TPSAB1), a tetrameric serine protease that has been implicated in the pathogenesis of asthma (Taira *et al.* 2002; Cui *et al.* 2014), heart disease (Bot *et al.* 2014), inflammatory bowel disease (Hamilton *et al.* 2011), and other disorders with allergic and/or inflammatory components (Sommerhoff and Schaschke 2007). Shown in Figure 7 is the three-dimensional structure of a TPSAB1 tetramer with inhibitor (white) bound (Costanzo *et al.* 2008). In orange, distal to the active site, is the positively selected residue detected by component methods and by MOSAIC. Note that positive selection at this location is active only outside of the great apes, with a fixed lysine

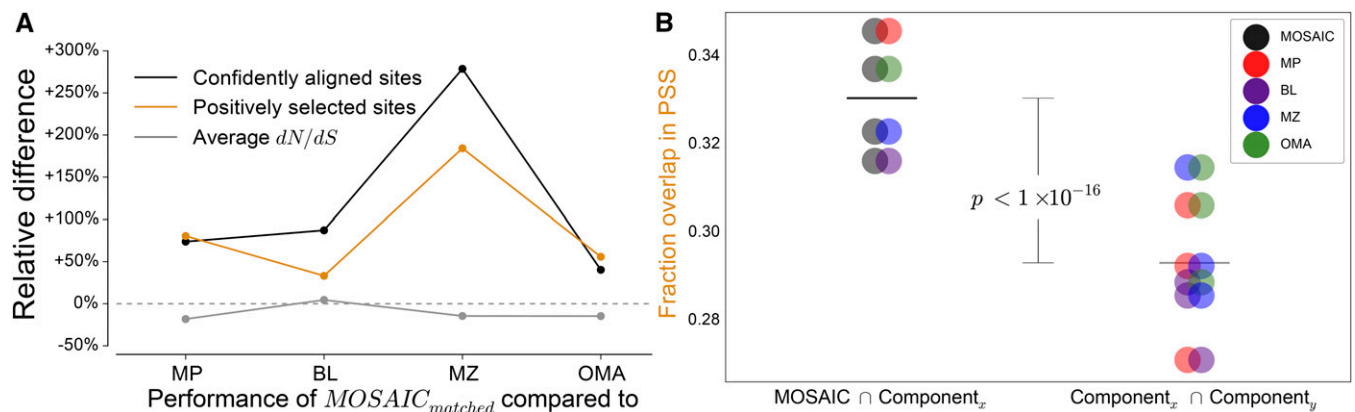


Figure 6 A comparison of evolutionary estimates. (A) The relative difference of $MOSAIC_{matched}$ vs. each component method for: (1) the number of positively selected sites, (2) the number of confidently aligned sites, and for reference, (3) the average level of conservation across all alignments. (B) The agreement between positively selected sites (1) between MOSAIC and component methods, and (2) among component methods. Fractional overlap values are plotted as Venn diagrams to illustrate the two methods being compared.

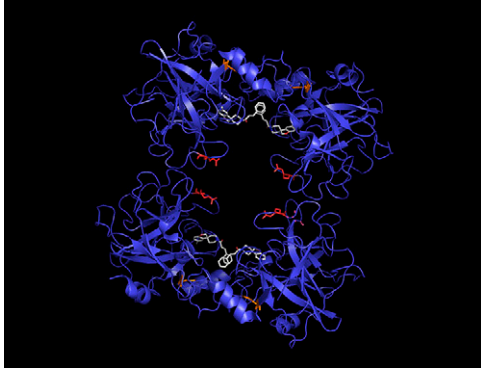


Figure 7 Example: a MOSAIC-specific PSS in Trypsin Alpha/Beta 1 (TPSAB1). The tetrameric TPSAB1 structure is shown with positively selected sites highlighted. The site detected by component methods and by MOSAIC is colored orange, whereas the MOSAIC-specific PSS is featured in red. A bound inhibitor (white) pinpoints the active site of the enzyme.

observed in human, chimp, gorilla, and orangutan (Figure S9 and Figure S10).

In red, directly within the proteolytic pore, is the site identified by MOSAIC as positively selected. This residue is a positively charged arginine in humans. This would be expected to modify the electrostatics of ligand binding. In chimp, we instead observe a kink-inducing proline. We might anticipate this change to have a large steric effect, possibly allowing the inward-facing unstructured loop to act as a more rigid lid closing over top of the substrate, or as a modifier of subunit contacts. Importantly, these changes occurred repeatedly in mammals. Proline is observed at this position in rhesus macaque and marmoset. Arginine, on the other hand, is present in gorilla and horse (Figure S9 and Figure S10). In orangutan, we observe a histidine: another positively charged amino acid.

Throughout this examination, we must be cognizant that tryptases evolved rapidly during primate evolution (Trivedi *et al.* 2007). The expansion of this gene family can itself be viewed as an example of positive selection. However, the presence of several paralogs creates the risk of inappropriately aligning pseudo-orthologous sequences that have evolved to serve divergent functions. Given the challenges, this case study provides an excellent opportunity to compare the high-throughput performance of MOSAIC to that of manually curated alignments.

As a first step, we showed that each proposal ortholog was a best bidirectional hit to TPSAB1 (Table S1). Next, we compared our sequences to those retrieved manually by Trivedi *et al.* (2007). Although we notice a few minor discrepancies between the two sets of alignments (see Figure S9 vs. Figure S11, reproduced from Trivedi *et al.* 2007), these differences do not alter our conclusion of human-relevant positive selection at the highlighted site in the proteolytic core of TPSAB1. Illustrations of component alignments from each method are shown in Figure S8.

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. Further, given the same number of species, MOSAIC alignments include more columns aligned with high confidence. This translates to higher levels of estimated conservation, and simulta-

neously, a greatly increased number of positively selected sites detected. Moreover, MOSAIC's positively selected sites agree better with those from component methods than component results do with each other. This suggests that not only does MOSAIC detect more positively selected sites—these sites are more reproducible and are detected due to an increase in alignment quality. Finally, we illustrated the significance of this increase in power by highlighting a positively selected site near the active site of the tryptase TPSAB1. Given the role of this enzyme in asthma and other allergic and inflammatory disorders, we feel that this case study is worthy of experimental follow-up.

In summary, MOSAIC provides the unique flexibility to incorporate any OD method that may be available now or in the future. It can therefore capture the entire swath of methodologic diversity, thereby improving OD performance, and allowing researchers to take advantage of methodological gains in a variety of areas of OD research. In addition, it provides the flexibility to adapt scoring and optimization procedures to the set of species under study. In future work, it will be interesting to ascertain how optimal procedures vary between species sets that have differing mean levels of divergence and markedly different patterns of evolution. For example, mammals and prokaryotes will likely have distinct optimal parameter values within MOSAIC. This tool is available 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>.

ACKNOWLEDGMENTS

We thank Raul Torres, Lawrence Uricchio, Nicolas Strauli, and Zachary Szpiech for their feedback regarding the manuscript. We would also like to thank Nathan Pearson for inspiring the paper title. This work was partially supported by the National Institutes of Health (grant numbers HG007644, P60MD006902, UL1RR024131, 1R21HG007233, 1R21CA178706, and 1R01HL117004) and an Alfred P. Sloan Foundation Research Fellowship to R.D.H. M.C.M. was supported by the Epidemiology and Translational Science program at the University of California, San Francisco, a National Institutes of Health F31 Predoctoral Fellowship (grant number 1 F31 CA180609-01), and a University of California, San Francisco Lloyd M. Kozloff Fellowship.

LITERATURE CITED

- Akaike, H., 1973 Information theory and an extension of the maximum likelihood principle, pp. 267–281 in 2nd International Symposium on Information Theory, edited by F. Czaki, and B. N. Petrov. Akademiai Kiado, Budapest.
- Alexeyenko, A., I. Tamas, G. Liu, and E. L. L. Sonnhammer, 2006 Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 22: e9–e15.
- Altenhoff, A. M., and C. Dessimoz, 2009a Phylogenetic and functional assessment of orthologs inference projects and methods. *PLOS Comput. Biol.* 5: e1000262.
- Altenhoff, A. M., and C. Dessimoz, 2012 Inferring orthology, pp. 259–279 in *Evolutionary Genomics. Vol. 855. Methods in Molecular Biology*, edited by M. Anisimova. Humana Press, Totowa, NJ.
- Altenhoff, A. M., A. Schneider, G. H. Gonnet, and C. Dessimoz, 2011 OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39: D289–D294.
- Berman, H. M., 2000 The Protein Data Bank. *Nucleic Acids Res.* 28: 235–242.
- Bertsekas, D., 1999 *Nonlinear Programming*. Ed. 2. Athena Scientific, Nashua, NH.
- Blanchette, M., W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit *et al.*, 2004 Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14: 708–715.

- Bot, I., G.-P. Shi, and P. T. Kovanen, 2015 Mast cells as effectors in atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* 35: 265–271.
- Capra, J. A., M. Stolzer, D. Durand, and K. S. Pollard, 2013 How old is my gene? *Trends Genet.* 29: 659–668.
- Chandrasekaran, V., and M. I. Jordan, 2013 Computational and statistical tradeoffs via convex relaxation. *Proc. Natl. Acad. Sci. USA* 110: E1181–E1190.
- Chen, F., A. J. Mackey, J. K. Vermunt, and D. S. Roos, 2007 Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2: e383.
- Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel *et al.*, 2006 Toward automatic reconstruction of a highly resolved tree of life. *Science* 311: 1283–1287.
- Costanzo, M. J., S. C. Yabut, H.-C. Zhang *et al.*, 2008 Potent, nonpeptide inhibitors of human mast cell tryptase. Synthesis and biological evaluation of novel spirocyclic piperidine amide derivatives. *Bioorg. Med. Chem. Lett.* 18: 2114–2121.
- Cui, Y., J. S. Dahlin, R. Feinstein, L. G. Bankova, W. Xing *et al.*, 2014 Mouse mast cell protease-6 and MHC are involved in the development of experimental asthma. *J. Immunol.* 193: 4783–4789.
- Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt, 1978 A model of evolutionary change in proteins, pp. 345–358 in *Atlas of Protein Sequence and Structure*, edited by M. O. Dayhoff. National Biomedical Research Foundation, Washington, DC.
- Ebersberger, I., S. Strauss, and A. von Haeseler, 2009 HaMSTR: profile hidden Markov model based search for orthologs in ESTs. *BMC Evol. Biol.* 9: 157.
- Eddy, S. R., 2011 Accelerated profile HMM searches. *PLoS Comput. Biol.* 7: e1002195.
- Hamilton, M. J., M. J. Sinnamon, G. D. Lyng *et al.*, 2011 Essential role for mast cell tryptase in acute experimental colitis. *Proc. Natl. Acad. Sci. USA* 108: 290–295.
- Henikoff, S., 1992 Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89: 10915–10919.
- Hulsen, T., M. A. Huynen, J. de Vlieg, and P. M. A. Groenen, 2006 Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.* 7: R31.
- Jordan, G., and N. Goldman, 2012 The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* 29: 1125–1139.
- Kent, W. J., 2002 BLAT—the BLAST-like alignment tool. *Genome Res.* 12: 656–664.
- Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The Human Genome Browser at UCSC. *Genome Res.* 12: 996–1006.
- Kuncheva, L. I., and C. J. Whitaker, 2003 Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* 51: 181–207.
- Kuzniar, A., R. C. van Ham, S. Pongor, and J. A. Leunissen, 2008 The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* 24: 539–551.
- Van der Laan, M. J., and S. Gruber, 2010 Collaborative double robust targeted maximum likelihood estimation. *Int. J. Biostat.* 6: 17.
- Van der Laan, M. J., E. C. Polley, and A. E. Hubbard, 2007 Super learner. *Stat. Appl. Genet. Mol. Biol.* 6: Article 25.
- Li, L., C. J. Stoeckert, and D. S. Roos, 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13: 2178–2189.
- Liu, Y., B. Schmidt, and D. L. Maskell, 2010 MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* 26: 1958–1964.
- Maddison, W. P., and L. L. Knowles, 2006 Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* 55: 21–30.
- Mashiyama, S. T., M. M. Malabanan, E. Akiva *et al.*, 2014 Large-scale determination of sequence, structure, and function relationships in cytosolic glutathione transferases across the biosphere. *PLoS Biol.* 12: e1001843.
- Massingham, T., and N. Goldman, 2005 Detecting amino acid sites under positive selection and purifying selection. *Genetics* 169: 1753–1762.
- J. McEntyre, and J. Ostell (Editors), 2002 The NCBI Handbook. National Center for Biotechnology Information, Bethesda, MD.
- Pruitt, K. D., J. Harrow, R. A. Harte, C. Wallin, M. Diekhans *et al.*, 2009 The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* 19: 1316–1323.
- Pruitt, K. D., G. R. Brown, S. M. Hiatt, F. Thibaud-Nissen, A. Astashyn *et al.*, 2014 RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* 42: D756–D763.
- Pryszcz, L. P., J. Huerta-Cepas, and T. Gabaldón, 2011 MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.* 39: e32.
- Punta, M., P. C. Coghill, R. Y. Eberhardt, J. Mistry, J. B. Tate *et al.*, 2012 The Pfam protein families database. *Nucleic Acids Res.* 40: D290–D301.
- Remm, M., C. E. Storm, and E. L. Sonnhammer, 2001 Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314: 1041–1052.
- Robinson, D. F., and L. R. Foulds, 1981 Comparison of phylogenetic trees. *Math. Biosci.* 53: 131–147.
- Rokach, L., 2009 Ensemble-based classifiers. *Artif. Intell. Rev.* 33: 1–39.
- Salichos, L., and A. Rokas, 2011 Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One* 6: e18755.
- Sommerhoff, C. P., and N. Schaschke, 2007 Mast cell tryptase beta as a target in allergic inflammation: an evolving story. *Curr. Pharm. Des.* 13: 313–332.
- Stamatakis, A., and N. Alachiotis, 2010 Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics* 26: i132–i139.
- Stamatakis, A., 2014 RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.
- Stolzer, M., H. Lai, M. Xu, D. Sathaye, B. Vernet *et al.*, 2012 Inferring duplications, losses, transfers and incomplete lineage sorting with non-binary species trees. *Bioinformatics* 28: i409–i415.
- Stone, M., 1977 An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Stat. Soc.* 39: 44–47.
- Sukumaran, J., and M. T. Holder, 2010 DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 26: 1569–1571.
- Taira, M., J. Tamaoki, M. Kondo, K. Kawatani, and A. Nagai, 2002 Serum B12 tryptase level as a marker of allergic airway inflammation in asthma. *J. Asthma* 39: 315–322.
- Trivedi, N. N., Q. Tong, K. Raman, V. J. Bhagwandin, and G. H. Caughey, 2007 Mast cell alpha and beta tryptases changed rapidly during primate speciation and evolved from gamma-like transmembrane peptidases in ancestral vertebrates. *J. Immunol.* 179: 6072–6079.
- Wolpert, D. H., and W. G. Macready, 1997 No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1: 67–82.
- Yandell, M., and D. Ence, 2012 A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13: 329–342.
- Yang, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586–1591.
- Yu, C., N. Zavaljevski, V. Desai, and J. Reifman, 2011 QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Res.* 39: e88.
- Zhao, S., A. Sakai, X. Zhang *et al.*, 2014 Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife* 3.

Communicating editor: J. C. Fay