Towards Consensus Gene Ages

Benjamin J. Liebeskind[1,2,*], Claire D. McWhite[1], and Edward M. Marcotte[1]

[1]Department of Molecular Biosciences, Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin

[2]Center for Computational Biology and Bioinformatics, University of Texas at Austin

Corresponding author: E-mail: bliebeskind@austin.utexas.edu.

Accepted: May 1, 2016

Data deposition: This project has been deposited at Zenodo under the accession http://dx.doi.org/10.5281/zenodo.51708.

Abstract

Correctly estimating the age of a gene or gene family is important for a variety of fields, including molecular evolution, comparative genomics, and phylogenetics, and increasingly for systems biology and disease genetics. However, most studies use only a point estimate of a gene's age, neglecting the substantial uncertainty involved in this estimation. Here, we characterize this uncertainty by investigating the effect of algorithm choice on gene-age inference and calculate consensus gene ages with attendant error distributions for a variety of model eukaryotes. We use 13 orthology inference algorithms to create gene-age datasets and then characterize the error around each age-call on a per-gene and per-algorithm basis. Systematic error was found to be a large factor in estimating gene age, suggesting that simple consensus algorithms are not enough to give a reliable point estimate. We also found that different sources of error can affect downstream analyses, such as gene ontology enrichment. Our consensus gene-age datasets, with associated error terms, are made fully available at so that researchers can propagate this uncertainty through their analyses (geneages.org).

Key words: ortholog, phylostratigraphy, LECA, LUCA.

Introduction

From their inception, whole genome datasets have been used to infer the evolutionary history of gene families (Mushegian and Koonin 1996). The age of a gene family, its provenance, and its evolutionary history, such as loss and duplication events, can inform us about its function (Capra et al. 2013). For instance, gene age has been found to correlate with disease association (Domazet-Lošo and Tautz 2008; Maxwell et al. 2014), evolutionary rate (Wolf et al. 2009), and the number of associated protein-interaction partners (Kim and Marcotte 2008), and a gene's phylogenetic distribution can be used to infer aspects of its function (Pellegrini et al. 1999). Gene ages can also be used to estimate the gene content of ancient organisms, such as the last universal common ancestor (LUCA, Mushegian and Koonin 1996), or the last eukaryotic common ancestor (LECA, Thiergart et al. 2012; Koumandou et al. 2013). Accordingly, an analysis of gene family ages on a genomic scale can inform the phylogenetic history of important phenotypes, such as eyes or the nervous system (Rivera et al. 2010; Liebeskind et al. 2016). In more recent years, gene age has been used to annotate systems biology datasets (Conaco et al. 2012; Alié et al. 2015; Wan et al. 2015), with the promise of elucidating the evolutionary history of core cellular machinery.

Such studies rely first and foremost upon the correct identification of homologs and/or orthologs. These two relationships form the basis of the gene-age determination in nearly all studies, with orthology being the more common criterion (Gabaldón 2008; Maxwell et al. 2014). Orthology is a pairwise relationship between two genes that occurs when their most recent common ancestor (MRCA) lies at a speciation event in a phylogenetic gene tree. This is in contrast to paralogs, whose MRCA lies at a gene duplication event (nodes on gene trees represent either speciation or gene duplication events, barring horizontal gene transfer) (Fitch 1970, 2000). Orthologs tend to display higher functional conservation than paralogs (although perhaps only weakly, Chen and Zhang 2012, – see Gabaldón and Koonin 2013, for a review), hence their use as a basis of cross-species comparison. Typically, studies of gene age will consider an orthologous group to be all the descendent lineages of the deepest speciation node, or the divergence between the two most distant homologs, if that is

the criterion being used, as in "phylostratigraphy" (Domazet-Lošo and Tautz 2008). Then, the age of the gene group is defined as the MRCA of the species found in that group.

Inferring a gene family's age thus relies on the accuracy of orthology assignment, but inferring correct orthologs is notoriously difficult, with no one of the more than 30 algorithms outperforming all others (Sonnhammer et al. 2014; Altenhoff et al. 2016). In particular, algorithms differ strongly in the tradeoff between recall and precision (Altenhoff et al. 2016). And many, perhaps all, algorithms may underestimate the size of orthologous groups due to inherent sensitivity limits in homology searches, thereby underestimating gene age (Moyers and Zhang 2015, 2016). Yet most studies on gene age rely on only one kind of algorithm, either using a pre-existing method or establishing an *ad hoc* protocol, most of which resemble one of the pre-existing algorithms (Maxwell et al. 2014). Although methods for probabilistic orthology assignment do exist (Ullah et al. 2015), available methods are not currently scalable to large genomic datasets using protein sequences, and at any rate still rely on a preliminary clustering step to infer gene families. Consensus algorithms also exist, some of which seem to substantially improve performance on established benchmarks (Pereira et al. 2014; Maher and Hernandez 2015). However, these methods still give only a point estimate. Another approach is to propagate the uncertainty that necessarily arises in orthology inference through subsequent analyses. However, it is unclear what the relevant sources of uncertainty are in orthology inference, and most consensus algorithms do not keep track of the different sources of error.

To remedy this situation, we characterized the error structure of gene-age estimation using 13 popular orthology inference algorithms. In doing so, we identify common types of errors and, after correcting these, present consensus gene-age calls for several model organisms (table 1). We provide these gene-age estimates along with a detailed analysis and annotation of the uncertainty associated with each age call so that this uncertainty can be propagated through future analyses, as we show for functional term enrichment. The consensus gene ages we calculate can be used for annotating genomic datasets in a variety of fields, and the analysis of error will help to prioritize genes for manual annotation and aspects of orthology inference for future study.

Methods

Data Collection

In order to fairly consider the range of orthology algorithms, we took advantage of the reference datasets managed by the Quest for Orthologs (QFO) consortium. QFO researchers have established community standards and benchmarks for orthology inference and have made their benchmarking results publicly available (Sonnhammer et al. 2014; Altenhoff et al. 2016). Fifteen algorithms have submitted their orthology estimates on 66 reference proteomes (http://www.ebi.ac.uk/reference_proteomes, last accessed 19 May 2016) to QFO's benchmarking tool (http://orthology.benchmarkservice.org/cgi-bin/gateway.pl, last accessed 19 May 2016) (Altenhoff et al. 2016). Importantly, these algorithms are widely used and capture a variety of methods commonly used in the literature to infer orthology and gene age (Vilella et al. 2009; Huerta-Cepas et al. 2008; Pryszcz et al. 2011; DeLuca et al. 2012; Mi et al. 2013; Altenhoff et al. 2015; Huerta-Cepas et al. 2016; Sonnhammer and Östlund 2015; Linard et al. 2015). It is, therefore, expected that nearly every study of gene age, regardless of the method used, will closely resemble the results of at least one of the algorithms we explore here. We omitted two of these because they either did not have full taxon coverage (RBH), or their results were so different from all the others that it dominated the variance in all downstream analyses (OMA_GETHOGS). Pairwise orthology calls for the 13 remaining algorithms were converted into tables for each gene, which were then used for subsequent analyses. The reference species tree was downloaded from SwissTree (Boeckmann et al. 2015) on 2015 Jun 15 (ftp://ftp.lausanne.isb-sib.ch/databases/SwissTree/speciestree.nhx, last accessed 19 May 2016) and was pruned to match the taxa in the Quest for Orthologs reference proteomes (http://www.ebi.ac.uk/reference_proteomes, last accessed 19 May 2016). The results below are with reference to the human proteome, but the same methods were applied to a variety of model organism proteomes (table 1).

Custom programs were written to perform the analyses below, and these are publicly available, as are iPython notebooks used for plotting. These, and the datasets supporting the conclusions in this article, are available at geneages.org, and the associated GitHub repository (https://github.com/marcottelab/Gene-Ages, last accessed 19 May 2016) with the following commit id: fee8d009d4d5ee24c3ae3cb0763439c48d4705e6. Scripts relied heavily on the python packages dendropy (Sukumaran and Holder 2010), BioPython (Cock et al. 2009), and pandas (McKinney 2013).

Protein Age Calls

We inferred the ages for each gene by mapping the species in that gene's ortholog group onto a reference species tree from SwissTree, which was derived from a consensus of trees found in the literature (Boeckmann et al. 2015). The age of a protein is calculated on the species tree by finding the MRCA node of the taxa that have orthologs of that protein. This node is the "node age." We use the node age to calculate a simple statistic that captures the uncertainty around the age-call, called the "node-error." This is the average number of branches (patristic distance, with equal weights of 1 for all branches) between the age calls any two algorithms. We also used the average node-error between pairs of algorithms as input for a heuristic search in PAUP (Swofford 2003) to cluster algorithms by similarity (see below).

Table 1
Species for which final consensus tables were constructed. Tables are available at https://github.com/marcottelab/Gene-Ages

Common name	Uniprot ID	False-negative analysis
Anopheles gambiae (Mosquito)	ANOGA	No
Bos taurus (Cattle)	BOVIN	No
Branchiostoma floridae (Lancelet)	BRAFL	No
Caenorhabditis elegans (Worm)	CAEEL	Yes
Candida albicans	CANAL	Yes
Canis lupus familiaris (Dog)	CANFA	No
Gallus gallus (Chicken)	CHICK	Yes
Ciona intestinalis (Tunicate)	CIOIN	No
Cryptococcus neoformans	CRYNJ	No
Danio rerio (Zebrafish)	DANRE	Yes
Drosophila melanogaster (Fly)	DROME	Yes
Homo sapiens (Human)	HUMAN	Yes
Ixodes scapularis (Tick)	IXOSC	No
Macaca mulatta (Rhesus macaque)	MACMU	No
Monosiga brevicollis (Choanoflagellate)	MONBE	No
Monodelphis domestica (Opossum)	MONDO	No
Mus musculus (Mouse)	MOUSE	Yes
Nematostella vectensis (Sea anemone)	NEMVE	No
Neurospora crassa	NEUCR	No
Ornithorhynchus anatinus (Platypus)	ORNAN	No
Pan troglodytes (Chimp)	PANTR	No
Phaeosphaeria nodorum	PHANO	No
Rattus rattus (Rat)	RAT	Yes
Saccaromyces cerevisiae (Budding yeast)	YEAST	Yes
Schistosoma mansoni (Blood fluke)	SCHMA	No
Schizosaccharomyces pombe (Fission yeast)	SCHPO	Yes
Sclerotinia sclerotiorum	SCLS1	No
Takifugu rubripes (Pufferfish)	TAKRU	No
Ustilago maydis (Corn smut/Huitlacoche	USTMA	No
Xenopus tropicalis (Frog)	XENTR	No
Yarrowia lipolytica	YARLI	No

To simplify the comparison of algorithm performance on human genes, we broke the reference species tree into eight age categories – the "binned age" (fig. 1). These categories form nested clades, with the exception of the category "Euk + Bacteria." This non-phylogenetic category captures the substantial number of eukaryotic genes that were horizontally transferred from bacteria after eukaryotes diverged from the rest of archaea (Méheust et al. 2015; Pittis and Gabaldón 2016), and is defined as genes present in eukaryotes and bacteria but not archaea. These "binned ages" conform to the interior labels given by SwissTree.

Filtering False Positives and Negatives

Before calculating a consensus, we flag algorithms that may have committed false-positive or false-negative errors on a per-gene basis. These algorithms are then removed from consideration of that gene's age. False positives are orthology calls that are substantially more distant than orthology calls by other algorithms, and have the effect of driving age deeper in the tree. These are found as follows. For each algorithm and each protein: (1) the node age is calculated, (2) the number of taxa in the species tree descended from this node is found, and (3) the number of taxa containing orthologs of the focal protein is subtracted from the number of descendant taxa. This number is the number of taxa without the orthogroup that are descended from an ancestor that putatively had the orthogroup, and is, therefore, proportional, but not identical, to the number of inferred losses of the orthogroup. For each algorithm and each protein, if this number is two standard deviations above the pooled algorithm mean for the focal protein, that algorithm's age call is considered a false positive and is thrown out.

False negatives are cases where an algorithm fails to make an orthology call, driving the inferred age to shallower nodes in the species tree. We identify one possible cause of this, which we call "over-splitting." This is when a group of co-orthologs is not correctly recognized by an algorithm and only one or a few of its members are found as orthologs to a more distant species, while the others are split off into their own orthogroups. The members that are split off would then be called at an incorrectly young age. To identify these errors, we used PhylomeDB's (Huerta-Cepas et al. 2008) orthogroups as a standard. For each protein and each algorithm (except for PhylomeDB), if the focal algorithm called a younger age than PhylomeDB and a co-ortholog of the focal protein could be found where the focal algorithm called the same node age as PhylomeDB did on the focal protein, then this algorithm was considered to be over-splitting the focal protein, and was not considered in this protein's age call. This error calculation was not performed on proteins where PhylomeDB was flagged as a false positive.

We also investigated the correlation between several intrinsic properties of proteins and their node error statistic. These included protein length, number of domains (as annotated by HMMER, Eddy 1998), and evolutionary rate. Rate was calculated for each protein by performing pairwise Needleman–Wunsch alignments for all one-to-one orthologs (annotated by PhylomeDB) between human and mouse, and human and yeast.

Consensus Ages

We generated consensus binned ages after removing algorithms flagged with false positives and negatives as described above. The number of algorithms favoring each binned age is counted and then normalized by the number of contributing algorithms to give a distribution over age calls. For subsequent analyses, we used the mode of this distribution as the consensus age.

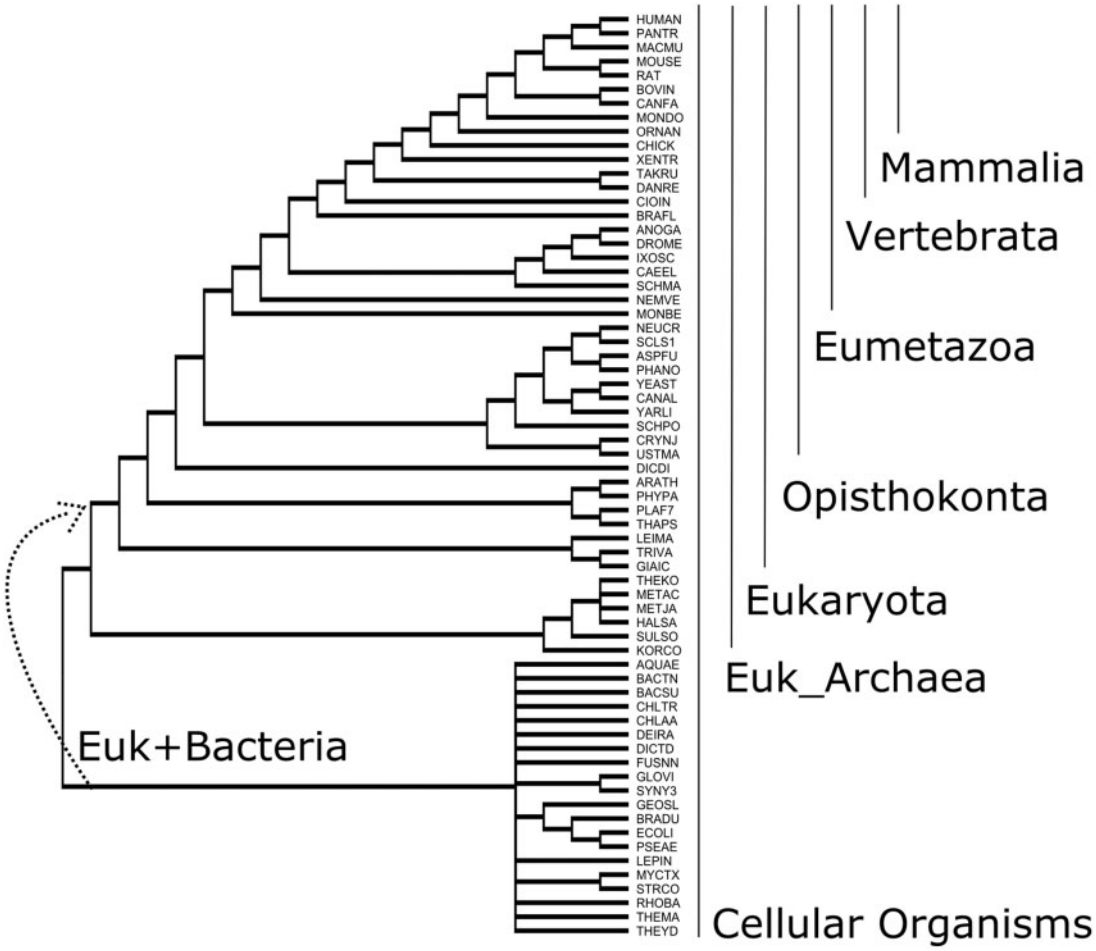


Fig. 1.—The reference species tree and age categories used for gene-age inference. This tree is based on SwissTree (Boeckmann et al. 2015) and reflects a consensus of recent large-scale phylogenies. Tip names are Uniprot species identifiers.

Results

The Effect of Algorithm Choice on the Distribution of Human Gene Ages

We investigated the effect of algorithm choice on gene-age estimation by assigning each human gene to the "binned" age category in which the MRCA of the species in its orthogroup fall. We then calculated the distributions over the different age categories for the human proteome inferred by each algorithm. The algorithms fell into two distinct groups with respect to the distribution of age classes (fig. 2). Hierarchically clustering the algorithms by the average number of branches between their per-gene age calls recapitulated this grouping, and we define the two groups based on the midpoint root of this tree. One group tended to find that most orthogroups could be traced to the MRCA of vertebrates, whereas the other group found a much older mode age dating back to LECA. We call these two groups, the "young" and the "old" group, respectively, although, of course, there are many more subtle and interesting distinctions between the algorithms.

Orthology inference algorithms are typically classed into graph-based and tree-based methods (Sonnhammer et al. 2014). However, we found that even though tree-based methods tended to fall in the "old" group, this was not universally the case, or were all graph-based methods found in the "young" group. The use of species tree information was not a determining factor either (fig. 2). The bimodal nature of the age calls, either "young" or "old", is, therefore, not simply a reflection of the graph/tree distinction, although it is clearly correlated. What is the source of this bimodality? One obvious answer is systematic error in the "young" group algorithms, the "old" group, or both. Systematic error in the young group would be equivalent to false negatives, that is, missing orthology assignments, whereas systematic error in the old group is equivalent to false positives, or spurious orthology assignments. This would have the effect of pushing

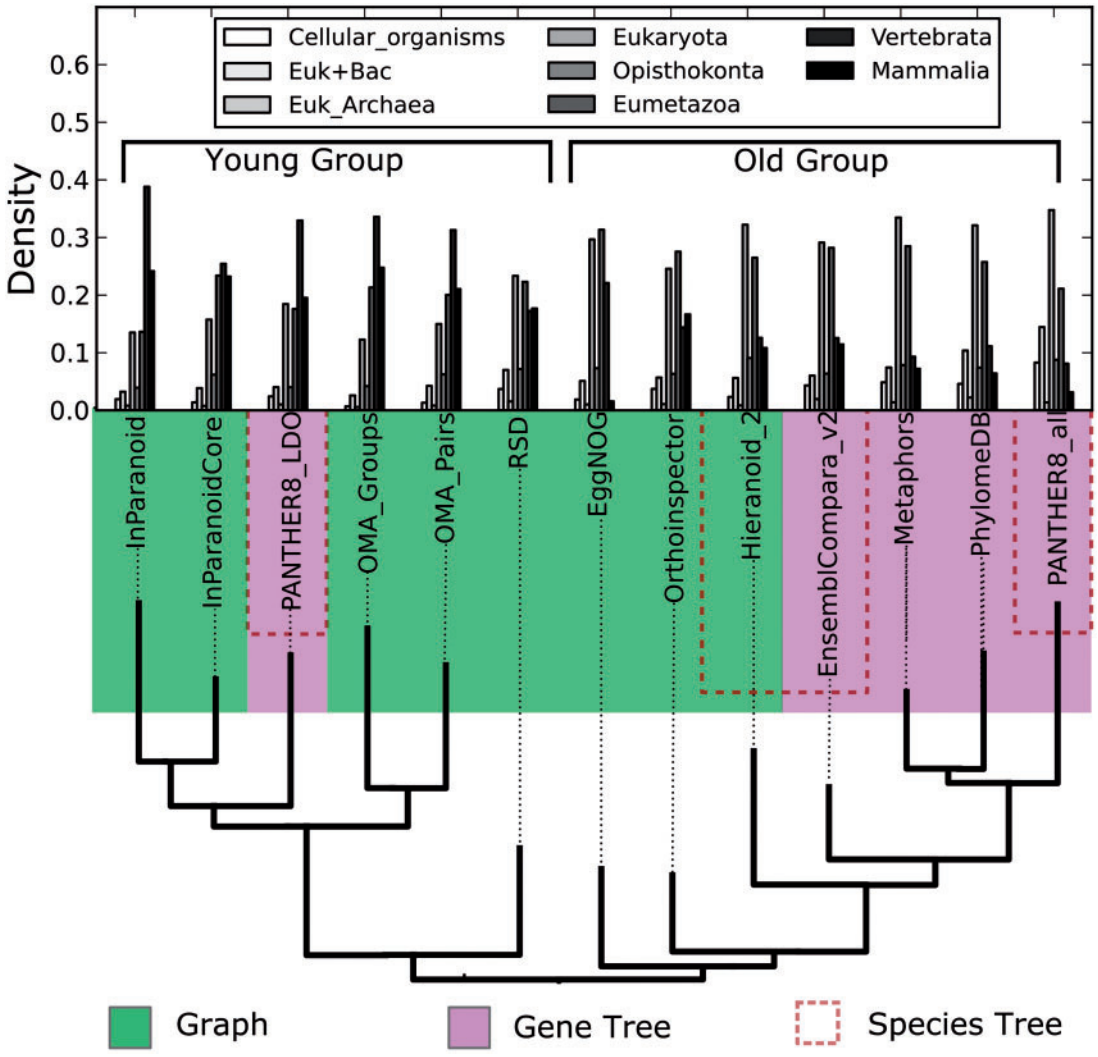


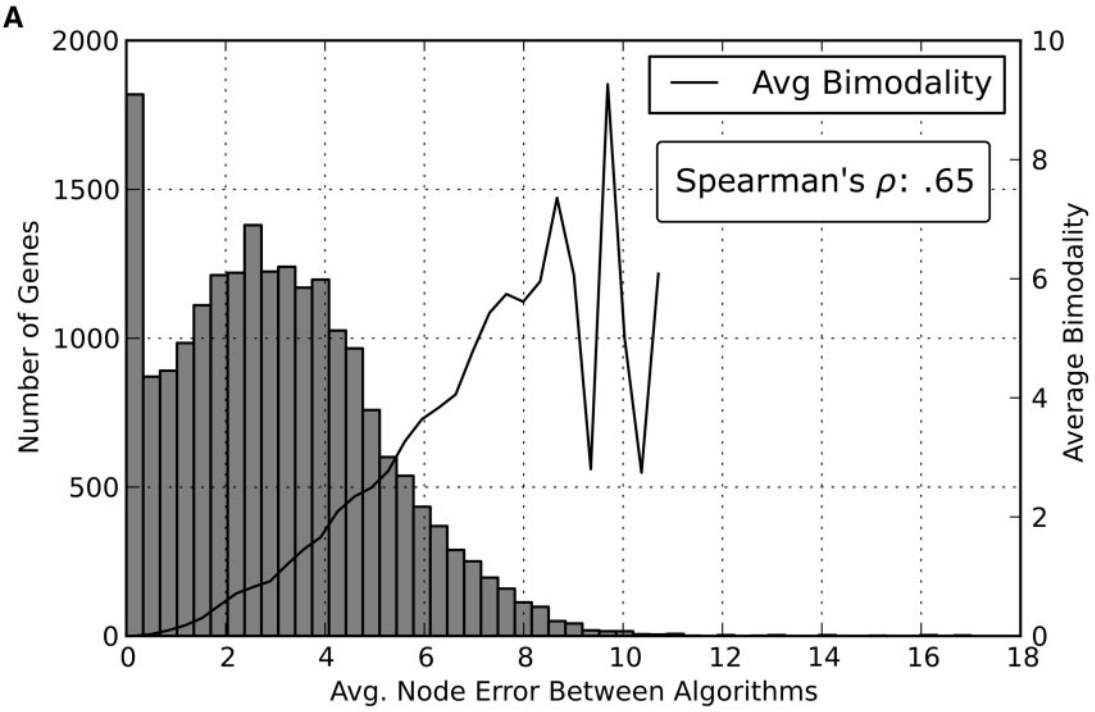
Fig. 2.—Distribution of age categories in the human proteome inferred by 13 different orthology inference algorithms. Algorithms were clustered according to the average pairwise distance between their age-calls, counted in units of braches (patristic distance with equal weights of 1 for all branches). The distance tree is rooted at the midpoint. Algorithms are colored by the methods they use to infer orthology. They either use a graph-based or a gene tree-based strategy, either with, or without, the use of a species tree (dotted outline).

the age of the group away from or towards the root of the tree, respectively.

Identifying Systematic Error

We first investigated whether the bimodality of age-calls played out on the single gene level or whether the two groups apparent in figure 2 were due to the effects of averaging across genes, with error being randomly distributed among proteins. To do so, we calculated a simple statistic that captured how bimodal a protein's age calls were between the two groups of algorithms ("old" and "young"). This statistic, which we call bimodality, is the difference between node-error within the two groups and between them, with more highly bimodal proteins having more variation between groups. Over 80% of proteins had some degree of bimodality corresponding to these two age groups, or none, as is expected given the hierarchical clustering in figure 1. The remaining genes were anti-correlated with the "old"/"young" groupings. Furthermore, the degree of bimodality between the "young" and "old" algorithm groups correlates well with the amount of error associated with each protein (Spearman's ρ: 0.69) (fig. 3). That is, proteins with a large amount of error tend to be more bimodal. The bimodality between algorithms is, therefore, a systematic phenomenon and a major source of error in these datasets. Unfortunately, in the case of highly polarized genes, we cannot know *a priori* whether the "old" or "young" age is the correct one. It is,

A



B

	"Young"			"Old"			
	InParanoid	OMA_Groups	PANTHER_LDO	EGGNOG	Hieranoid_2	Metaphors	**Bimodality**
Q5H910	Mammalia	Mammalia	Mammalia	Eukaryota	Eukaryota	Eukaryota	**14**
Q7Z5J4	Vertebrata	Vertebrata	Eumetazoa	Vertebrata	Eumetazoa	Eumetazoa	**0.3**

Fig. 3.—Error statistics. (A) The distribution of average node error, a measure of disagreement among the algorithms for a given gene, is given, along with a plot of the average bimodaliy in each bin. Genes with more error tend to be more bimodal between "old" and "young" algorithms. (B) Example of a strongly bimodal and weakly bimodal gene with a few representative algorithms. The ages are given as categories for clarity, but the bimodality statistic is calculated according to the number of branches between node age-calls (see the Methods section).

therefore, important to propagate this uncertainty through further analyses, and the bimodality statistic is included with our consensus age estimates.

We also investigated whether aspects of the individual proteins contributed to systematic error. For instance, it may be difficult to infer correct evolutionary relationships for small proteins, or those with many domains. At least one orthology inference algorithm uses this idea to "correct" for protein length (Emms and Kelly 2015). However, we found that protein length has a weak positive correlation with age-call error, and that the number of domains also correlates weakly (Spearman's ρ: < 0.2 in both cases). Several simulation studies have found that increased evolutionary rate can lead to false-negatives in phylostratigraphic studies (Moyers and Zhang 2015, 2016). We found that similarity of one-to-one orthologs and node error have a negative correlation, as predicted by these studies, but weakly for human–mouse orthologs (Spearman's $\rho = -0.13$) and only slightly stronger for human–yeast orthologs (-0.36). We did find, however, that proteins in the youngest age class, "Mammalia," tended to have a higher rate of evolution than older age-classes. We cannot, therefore, rule out the possibility that some of the "youngest" proteins may be due to the limited sensitivity of the homology searches (see also supplemental iPython notebook "errorCorrelation_plotting" (Supplementary Material online) at https://github.com/marcottelab/Gene-Ages/tree/master/Notebooks).

Systematic False Negatives

What are the causes of systematic false negatives and can we identify them without *a priori* knowledge of the true orthogroup? One clue comes from the different age-category distributions between PANTHER8_all and PANTHER8_LDO (Mi et al. 2013). These two sets of orthology calls are based on the same set of gene trees, but differ in their definition of

orthology. "LDO" stands for "least diverged ortholog," and only considers the least diverged among a set of co-orthologs to be the true ortholog of an outgroup. This can be contrasted to the traditional phylogenetic definition of orthology where all co-orthologs are equally orthologous to the outgroup (fig. 4; Fitch 2000). Although it may be useful to split co-orthologous groups, as the LDO definition does, in cases where orthology is being used for, for example, gene function annotation, it is inappropriate for defining the age of a gene or gene family because the age must be in reference to the topology of the phylogenetic tree. The fact that PANTHER8_LDO's age category distribution resembled that of several graph-based methods, and the fact that it clustered with them based on its per-gene age calls (fig. 2), suggests that these methods may be splitting up co-orthologous groups as well.

There is no gold standard set of co-orthologs in this dataset, so we used the database PhylomeDB as a reference for identifying co-ortholog over-splitting. PhylomeDB was chosen because it infers gene trees under Maximum-Likelihood with well-characterized models, and is, therefore, similar to how most researchers infer orthologs when analyzing one gene family at a time. We nevertheless recognize the limitation inherent in choosing a single, imperfect set as a reference. PhylomeDB summary files for 10 species in PhylomeDB's model species collection (PhyC2) that overlapped with species in our tree (table 1) were downloaded, and we determined groups of co-orthologs that were then used for the analysis. Briefly, for protein (A), if an algorithm called a younger age (Y) and PhylomeDB an older age (O), and if in the co-orthologs of (A) we could find a protein (B) which that algorithm called at age (O), then (B) was identified as the LDO, age (O) was assumed to be the true age, and that algorithm was determined to be over-splitting the co-ortholog group (fig. 4). This was not carried out for proteins on which PhylomeDB's age call was determined to be a false positive (see below). We note that this method for identifying co-ortholog over-splitting is not ideal, because it relies on a single, imperfect algorithm (PhylomeDB). It is conservative, however, because algorithms will only be trimmed if they give a member of the co-orthologs the *exact* same age on the species tree as that called by PhylomeDB on the focal gene. More thorough analyses of whether graph-based methods are consistently missing co-orthologs will be necessary in the future.

Identifying False Positives

If genes of distant organisms are incorrectly inferred to be part of an orthology group, it will drive the age of the orthogroup towards the root of the tree. Recent HGT events are a biological source of such errors, but some algorithmic error is expected to play a role as well. Such problems are perhaps more likely to occur in tree-based algorithms, where slight rearrangements that do not strongly affect the likelihood of the tree can have an outsized effect on the inference of gene gains and losses (Hahn 2007). In such cases, the large number of taxa that fall between the true in-group taxa and the false positive out-group taxa will be inferred to have lost the orthogroup. We used this criterion on a per-gene basis to identify algorithms that were likely to have false positives and genes that were likely to be the result of recent HGT events. Algorithms that had an outsized number of taxa missing from an orthogroup were considered false positives and removed from downstream analysis of that orthogroup's age. After trimming these outliers, genes that were in the 95th percentile of inferred losses were flagged as being potential recent HGT events (i.e., horizontally transferred long after LECA). These potential HGT genes are an interesting set in themselves: 66% are from the Euk + Bacteria category, they are hugely enriched for metabolic genes (gProfiler P value = $9.08e^{-116}$), and several are associated with human diseases.

We found that, as expected, algorithms in the "old" group tended to commit more false positive errors, and algorithms in the "young" group committed more false negative errors (fig. 5). Because PhylomeDB was used as a basis for identifying false negatives, its false negative rate could not be quantified.

Consensus

These analyses suggested a way to more robustly estimate consensus gene ages and to calculate a posterior distribution over the estimate. We used the methods described above to identify algorithms that may have committed false positive or false negative errors and then removed these algorithms from the consideration on a per-protein basis. After doing so, we generated consensus tables based on the remaining algorithms for the human proteome and for a number of other model eukaryotes (table 1), and we make these tables available. Because our tree is best sampled within the opisthokonts (fungi, animals, and closely related protists), we restricted our analyses to this lineage. These tables contain a consensus age category for each protein based on the mode age call of non-trimmed algorithms. Older genes were found to be involved in key components of cell biology. Genes in the Euk + Bac group were found to be highly enriched for mitochondrial function, and genes that date back to the Euk_Archaea node were enriched for translational machinery, as has been shown previously (Thiergart et al. 2012; Koumandou et al. 2013). Many of these older genes are also associated with hereditary diseases that represent a deficiency in a cell function associated with that evolutionary epoch. For instance, the cytoskeletal system and cilium date to LECA (Koumandou et al. 2013), and genes in this age category are enriched for diseases affecting the cilium, such as primary ciliary dyskinesia and Bardet–Biedl syndrome (fig. 6).

These enrichment terms are derived from the point estimates of consensus ages, but we also provide other data that can be used to propagate uncertainty to downstream

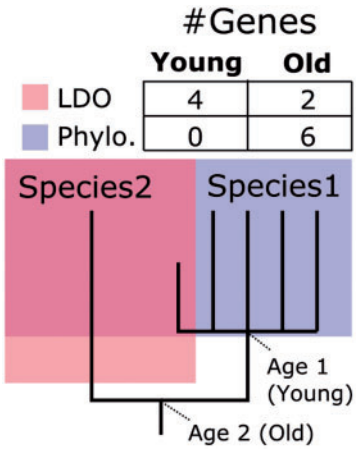


Fig. 4.—Determination of false negatives due to co-ortholog over-splitting. This tree compares the ages given by least derived orthology (LDO) and traditional, phylogenetic orthology (Phylo.). Given a group of co-orthologs in Species 1, LDO will give only the co-ortholog with the shortest distance to an outgroup (gene in Species 2) the status of ortholog to this outgroup (red box). All others are put in separate orthogroups. Hence, LDO produces more genes that are mapped (incorrectly) to a younger age (Y), whereas traditional, phylogenetic orthology (blue box) includes all co-orthologs to the orthogroup, thereby mapping more genes to the older age (O).

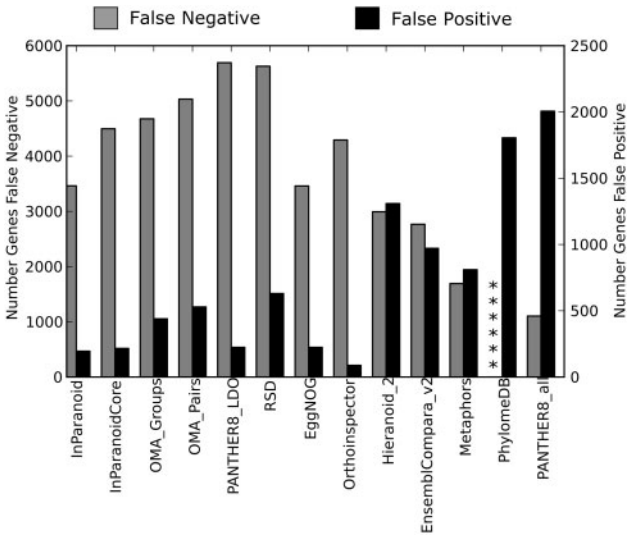


Fig. 5.—Errors committed by different algorithms. False positives and negative are defined in the text. PhylomeDB was used as a standard for false negative, so its false negative count could not be determined.

analyses. For each gene, the distribution over age-calls from the non-trimmed algorithms is given, as well as the number of contributing algorithms and the entropy of the age call distribution. About 87% of human proteins had at least five algorithms contributing after trimming, and 59% had at least 10 out of a total of 13 original algorithms. In addition, the tables contain information on whether the protein was flagged as being a potential horizontal gene transfer event. Finally, we include the node error and bimodality statistics, both of which are measures of uncertainty that reference the reference species tree.

We note that in several cases, we have made *ad hoc* decisions during the building of the consensus. For instance, algorithms were flagged as false positives if the number of taxa inferred to have lost the orthogroup was two standard deviations above the mean of all algorithms. These decisions were informed by the underlying distributions of values. Nevertheless, we supply the source data files, scripts used for these analyses, as well as interactive iPython notebooks, and we invite researches to explore and change parameters if they desire (https://github.com/marcottelab/Gene-Ages).

Error Propagation

How can our error annotations be used in downstream analyses? Here we give an example of a simple stability analysis for gene ontology enrichment that uses these error terms. It has previously been shown that eukaryotic genes vertically acquired from archaea are enriched for translation and RNA processing, whereas genes acquired horizontally from bacteria at the root of eukaryotes are enriched for metabolic processes (fig. 6; Thiergart et al. 2012; Koumandou et al. 2013). This conclusion relies on functional term enrichment, but what is the effect of different sources of error on these sorts of enrichment analyses? To investigate the robustness of this conclusion to different sources of error, we used the program g:Profiler (Reimand et al. 2007) to perform functional enrichment analysis on the two age classes "Euk_Archaea" and "Euk + Bacteria" after filtering the datasets at varying levels of stringency (fig. 7*A*). We found that removing genes that were flagged as a possible late HGT event had a strong effect on the average *P* values of functional annotation terms in the Euk + Bacteria age class but not the Euk_Archaea class (fig. 7*B*). This may be due to these genes being more commonly lost or to many bacterial genes being more recent HGT events (and hence being filtered out). The latter possibility would mean that many genes in this age category could be misidentified as being present in LECA, so these genes are good candidates for manual curation. Notably, filtering on different error terms can increase or decrease the significance of different terms, and, depending on the filtering strategy, the significance ranking of terms can be switched (fig. 7*C* and *D*). Analyses that rely on smaller test-sets of genes are likely to be much more strongly affected that these proteome-wide searches.

Discussion

Most studies of gene age use a single point estimate arising from one of a variety of methods. Given our analysis of some the most popular orthology inference algorithms, we find that

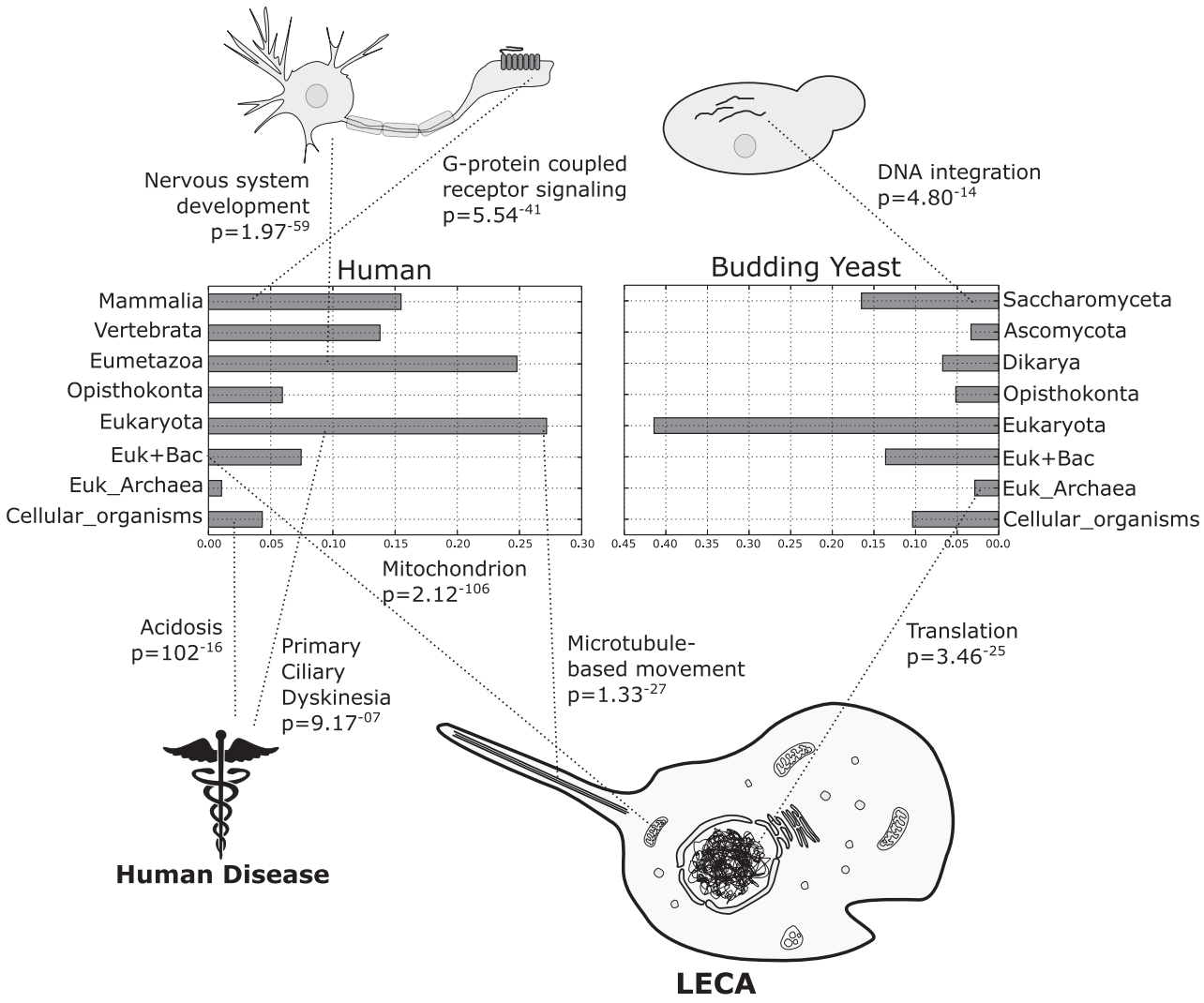


Fig. 6.—Enrichment of gene ontology terms and human disease terms (OMIM) in the different consensus age classes for human and budding yeast (*Saccharomyces cerevisiae*). The distribution of age classes are shown for each species. Older genes tend to be enriched for core cellular machinery and heritable diseases. Newer genes are associated with lineage-specific function, such as nervous system development and olfaction (*via* G-protein coupled receptors) in mammals, and DNA integration in yeast. *P* values are derived from g:Profiler (Reimand et al. 2007).

point estimates of gene age will be wrong for (at least) thousands of genes in a human-sized proteome (fig. 4). More troubling is the fact that algorithms appear to fall into two classes, each of which presumably has a systematic bias towards either false positives ("old group") or false negatives ("young group"). This systematic bias happens on a per-gene basis, meaning that simple voting methods will not be able to resolve conflicts. Even with the ideal sampling of algorithms, which we approximate here by exploring a wide diversity of popular algorithms, the effective voting population will still drop to two on highly polarized genes.

Many areas of computational biology have faced a similar problem, namely, the need to keep track of error in several components of a workflow, and to correctly propagate this error through the whole analysis (Guang et al. 2016). One illustrative example is multiple sequence alignment and phylogenetic inference. The former is a necessary precursor to the latter, and each involves estimation error. Methods have been developed to infer the posterior distributions of both steps simultaneously (Suchard and Redelings 2006), which is computationally intractable for all but the smallest datasets, or to perform each step iteratively in a maximum likelihood framework (Liu et al. 2009). We argue that, eventually, such steps will have to be taken with orthology inference and gene-age estimation. Using a point estimate at each step in the analysis makes the assumption that each inference step has no uncertainty associated with it, which we can clearly reject in the case of gene-age estimation.

A

	Entropy	HGT flag	#Algorithms Contributing	Bimodality
D1	No Trimming	No Trimming	No Trimming	No Trimming
D2	**Keep < 1**	No Trimming	No Trimming	No Trimming
D3	No Trimming	**Remove**	No Trimming	No Trimming
D4	No Trimming	No Trimming	**Keep > 3**	No Trimming
D5	No Trimming	No Trimming	No Trimming	**Keep < 5**
D6	**Keep < 1**	**Remove**	**Keep > 3**	**Keep < 5**

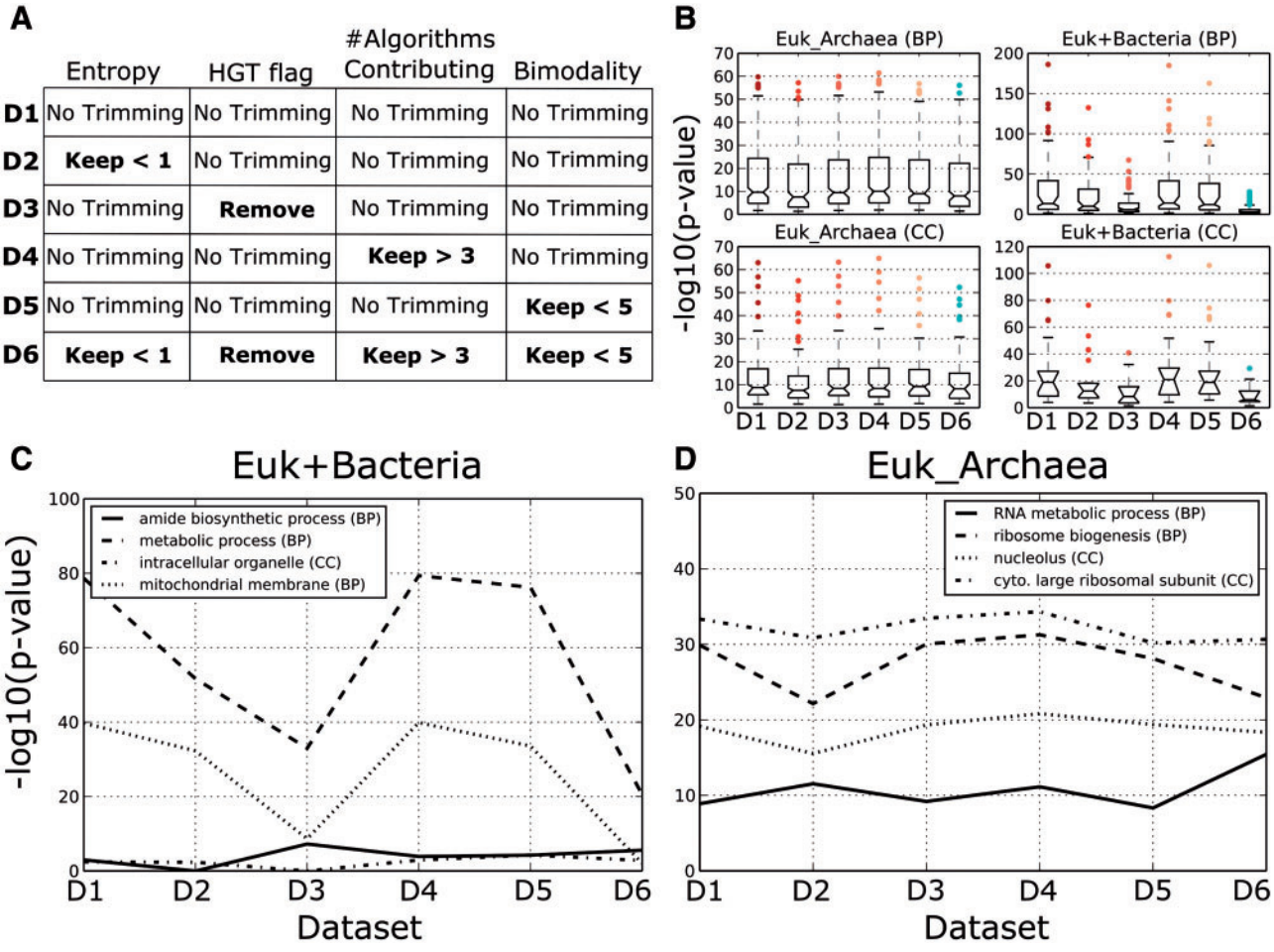


Fig. 7.—Effect of filtering on functional term enrichment analysis. (A) Datasets 1–6 were trimmed based on four sources of error: entropy of age-calls, whether the genes were flagged as potential horizontal gene transfer (HGT) events, the number of algorithms contributing to the final age call (after filtering algorithms, as described in the Methods and Results section), and the polarization of each gene. Dataset 6 was filtered on all four criteria (B) Negative $\log_{10}$ *P* values for the five datasets for two age categories (Euk_Archaea, and Euk + Bacteria) and two gene ontology terms (biological process, and cellular compartment). (C and D) Negative $\log_{10}$ *P* values across datasets for eight functional terms in the two age categories. These terms show a variety of ways that significance can be affected by filtering.

Some methods for probabilistic orthology inference do exist (Ullah et al. 2015). These use gene tree models with free parameters for gene duplication, loss, and sometimes HGT, which then contribute to the likelihood along with the multiple sequence alignment. However, these methods are in their infancy, and are not usually scalable to large datasets or widely used. In the meantime, it is important to have an understanding of common sources of error in gene-age estimation. We provide that information along with consensus age calls for a variety of model organisms so that researchers can incorporate error propagation into their analyses in a way that is appropriate to their question of interest.

Several error terms are likely to be important for a broad range of analyses. The first and most straightforward is the entropy of the age-call estimate after filtering false positives and negatives. This statistic gives a quick idea of how certain an age-call is, with higher entropies being less certain. It is defined with reference to our age categories, so if researchers need to use other age categories, they must use the node age of the gene, which we also provide. HGT events are also likely to affect some datasets, especially when genes originating in Bacteria are involved (fig. 6). A large number of eukaryotic genes are likely transfers from Bacteria (Thiergart et al. 2012), but these may have been transferred at any point on the phylogeny. We define one age category, Euk + Bacteria, to describe all genes transferred before LECA, with later transfers hopefully being caught by our flag. If researchers are primarily interested in HGT, we suggest a much fuller analysis, as our simple method is likely to miss many HGT events. Finally, the bimodality of the age-call between "young" and "old" algorithm types is a key statistic. The systematic biases in the different algorithm types mean that many datasets will be

radically different and difficult to compare, and it may account for some of the differences between studies of ancient gene repertoires that used either graph or tree-based methods. Genes that are highly polarized are good candidates for manual curation, because it is unlikely that any *ad hoc* algorithm will differ substantially enough from those we sampled here to be decisive.

Our analysis also identifies several areas for further study. In particular, our analysis of systematic false-negatives suggests that the way in which orthology inference algorithms handle co-orthology is a major source of difference among them. We relied on co-ortholog groups from PhylomeDB because no gold-standard sets are available, and PhylomeDB's method of Maximum-Likelihood gene tree inference without the use of reference species tree is the most similar to how researchers infer phylogenies on single gene families. However, a fuller analysis of co-ortholog oversplitting is clearly wanting, and we also provide final datasets without the false-negative filter that relies on PhylomeDB.

Although we have characterized only two components of a typical computational biology workflow, orthology inference, and gene-age estimation, it would be ideal to characterize error distributions for all the steps in an analysis, which has not been done with gene age data to our knowledge (but see Thompson et al. (2014) for an interesting example on gene-expression data, and Guang et al. (2016) for a general review). The datasets we provide here will hopefully help guide future research efforts aimed at a more formal, probabilistic way to handle error in gene-age estimation, perhaps even in the context of an entire workflow. Until such methods are available, we advocate using our error annotations or a similar analysis in any study incorporating gene-age data.

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors would especially like to acknowledge the Quest for Orthologs consortium and those who contributed their algorithms to the benchmarking tool for making their data freely available. B.J.L. was funded by NIH fellowship 1F32GM112504-01A1. E.M.M. acknowledges funding from the NIH, NSF, CPRIT, ARO (61789-MA-MUR), and Welch Foundation (F1515).

Literature Cited

Alié A, et al. 2015. The ancestral gene repertoire of animal stem cells. Proc Natl Acad Sci. 112:E7093–E7100.

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

Altenhoff AM, et al. 2016. Standardized benchmarking in the quest for orthologs. Nat. Methods 13(5):425–430.

Boeckmann B, et al. 2015. Quest for orthologs entails quest for tree of life: in search of the gene stream. Genome Biol Evol. 7:1988–1999.

Capra JA, Stolzer M, Durand D, Pollard KS. 2013. How old is my gene? Trends Genet. 29:659–668.

Chen X, Zhang J. 2012. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. PLoS Comput Biol. 8:e1002784.

Cock PJA, et al. 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423.

Conaco C, et al. 2012. Functionalization of a protosynaptic gene expression network. Proc Natl Acad Sci. 109:10612–10618.

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

Domazet-Lošo T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol. 25:2699–2707.

Eddy SR. 1998. Profile hidden Markov models. Bioinform Oxf Engl. 14:755–763.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:1–14.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Biol. 19:99–113.

Fitch WM. 2000. Homology a personal view on some of the problems. Trends Genet TIG 16:227–231.

Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? Genome Biol. 9:235.

Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 14:360–366.

Guang A, Zapata F, Howison M, Lawrence CE, Dunn CW. 2016. An integrated perspective on phylogenetic workflows. Trends Ecol Evol. 31:116–126.

Hahn MW. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol. 8:R141.

Huerta-Cepas J, Bueno A, Dopazo J, Gabaldón T. 2008. PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Res. 36:D491–D496.

Huerta-Cepas J, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44(D1):D286–D293.

Kim WK, Marcotte EM. 2008. Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence. PLoS Comput Biol. 4:e1000232.

Koumandou VL, et al. 2013. Molecular paleontology and complexity in the last eukaryotic common ancestor. Crit Rev Biochem Mol Biol. 48:373–396.

Liebeskind BJ, Hillis DM, Zakon HH, Hofmann HA. 2016. Complex homology and the evolution of nervous systems. Trends Ecol Evol. 31:127–135.

Linard B, et al. 2015. OrthoInspector 2.0: software and database updates. Bioinform Oxf Engl. 31:447–448.

Liu K, Raghavan S, Nelesen S, Linder CR, Warnow T. 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 324:1561–1564.

Maher MC, Hernandez RD. 2015. Rock, paper, scissors: harnessing complementarity in ortholog detection methods improves comparative genomic inference. G3 Genes Genomes Genet. 5:629–638.

Maxwell EK, et al. 2014. Evolutionary profiling reveals the heterogeneous origins of classes of human disease genes: implications for modeling disease genetics in animals. BMC Evol Biol. 14:212.

McKinney W. 2013. Python for data analysis. Beijing: O'Reilly
Méheust R, Lopez P, Bapteste E. 2015. Metabolic bacterial genes and the construction of high-level composite lineages of life. Trends Ecol Evol. 30:127–129.
Mi H, Muruganujan A, Thomas PD. 2013. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41:D377–D386.
Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol Evol. 32:258–267.
Moyers BA, Zhang J. 2016. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Mol Biol Evol. 33(5):1245–1256.
Mushegian AR, Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci. 93:10268–10273.
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci. 96:4285–4288.
Pereira C, Denise A, Lespinet O. 2014. A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics 15:S16.
Pittis AA, Gabaldón T. 2016. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature 531:101–104.
Pryszcz LP, Huerta-Cepas J, Gabaldón T. 2011. MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res. 39:e32–e32.
Reimand J, Kull M, Peterson H, Hansen J, Vilo J. 2007. g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 35:W193–W200.
Rivera AS, et al. 2010. Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach. BMC Evol Biol. 10:123.
Sonnhammer ELL, et al. 2014. Big data and other challenges in the quest for orthologs. Bioinformatics 30:2993–2998.
Sonnhammer ELL, Östlund G. 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res. 43:D234–D239.
Suchard MA, Redelings BD. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinforma Oxf Engl. 22:2047–2048.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Swofford DL. 2003. Phylogenetic analysis using parsimony (*and other methods). Sunderland, MA: Sinauer Associates.
Thiergart T, Landan G, Schenk M, Dagan T, Martin WF. 2012. An evolutionary network of genes present in the eukaryote common ancestor polls genomes on eukaryotic and mitochondrial origin. Genome Biol Evol. 4:466–485.
Thompson A, Vo D, Comfort C, Zakon HH. 2014. Expression evolution facilitated the convergent neofunctionalization of a sodium channel gene. Mol Biol Evol. 31:1941–1955.
Ullah I, Sjöstrand J, Andersson P, Sennblad B, Lagergren J. 2015. Integrating sequence evolution into probabilistic orthology analysis. Syst Biol. 64:969–982.
Vilella AJ, et al. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327–335.
Wan C, et al. 2015. Panorama of ancient metazoan macromolecular complexes. Nature 525:339–344.
Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proc Natl Acad Sci. 106:7273–7280.

Associate editor: Eugene Koonin