Toward Reducing Phylostratigraphic Errors and Biases

Bryan A. Moyers[1] and Jianzhi Zhang[2,*]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama

[2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan

*Corresponding author: E-mail: jianzhi@umich.edu.

Accepted: July 28, 2018

Abstract

Phylostratigraphy is a method for estimating gene age, usually applied to large numbers of genes in order to detect nonrandom age-distributions of gene properties that could shed light on mechanisms of gene origination and evolution. However, phylostratigraphy underestimates gene age with a nonnegligible probability. The underestimation is severer for genes with certain properties, creating spurious age distributions of these properties and those correlated with these properties. Here we explore three strategies to reduce phylostratigraphic error/bias. First, we test several alternative homology detection methods (PSIBLAST, HMMER, PHMMER, OMA, and GLAM2Scan) in phylostratigraphy, but fail to find any that noticeably outperforms the commonly used BLASTP. Second, using machine learning, we look for predictors of error-prone genes to exclude from phylostratigraphy, but cannot identify reliable predictors. Finally, we remove from phylostratigraphic analysis genes exhibiting errors in simulation, which by definition minimizes error/bias if the simulation is sufficiently realistic. Using this last approach, we show that some previously reported phylostratigraphic trends (e.g., younger proteins tend to evolve more rapidly and be shorter) disappear or even reverse, reconfirming the necessity of controlling phylostratigraphic error/bias. Taken together, our analyses demonstrate that phylostratigraphic errors/biases are refractory to several potential solutions but can be controlled at least partially by the exclusion of error-prone genes identified via realistic simulations. These results are expected to stimulate the judicious use of error-aware phylostratigraphy and reevaluation of previous phylostratigraphic findings.

Key words: BLASTP, gene age, HMMER, OMA, PHMMER, PSIBLAST.

Introduction

Phylostratigraphy is a method for estimating the evolutionary age of a gene. It uses homology detection programs, typically the BLAST (Basic Local Alignment Search Tool) suite of algorithms, to identify homologs of a query gene in a target database, most often a subset of the NCBI nonredundant database that is sometimes combined with additional sequences (Domazet-Lošo and Tautz 2003; Domazet-Lošo et al. 2007; Neme and Tautz 2013; Domazet-Lošo et al. 2017). The age of the query gene is the time since the most recent common ancestor between the query and its most distant homolog detected. When phylostratigraphy is applied to a large set of genes, one can analyze phylostratigraphic trends by correlating various gene properties with the estimated gene age; these trends have been used to infer mechanisms of gene emergence and evolution (Albà and Castresana 2005; Domazet-Lošo and Tautz 2008; Prat et al. 2009; Wolf et al. 2009; Cai and Petrov 2010; Domazet-Lošo and Tautz 2010; Carvunis et al. 2012; Hemmrich et al. 2012; Sestak et al. 2013).

However, because homology detection programs use sequence similarity to approximate homology, errors are inevitable. A false negative error causing underestimation of gene age occurs when a distant homolog is missed due to the undetectably low sequence similarity between the homolog and query. Extensive computer simulations showed that such phylostratigraphic errors are nonnegligible, occurring in at least 5–14% of genes (Elhaik et al. 2006; Albà and Castresana 2007; Moyers and Zhang 2015, 2016, 2017). More worrisome than the precise amount of error is the fact that error is nonrandom; higher error rates are associated with certain gene properties such as higher evolutionary rates and shorter sequence lengths (Moyers and Zhang 2015). So long as nonnegligible errors are nonrandom, observed phylostratigraphic trends may be attributable in whole or in part to phylostratigraphic error (Moyers and Zhang 2015, 2016, 2017). Even gene properties having no apparent direct impact on phylostratigraphic error can be influenced due to their association with gene properties that affect the performance

of phylostratigraphy (Moyers and Zhang 2016). Although synteny could help in gene age estimation (McLysaght and Hurst 2016), its utility is usually limited due to the decay of synteny in long-term evolution. Consequently, phylostratigraphy remains the most widely used method.

There is therefore a pressing need to reduce the amount and effects of phylostratigraphic error. At least three broadly applicable approaches have the potential to reduce phylostratigraphic error. In this work, we explore these three approaches. First, there is an abundance of tools for homology detection, some of which may outperform the existing pipeline. Although BLASTP (Altschul et al. 1990) is used most commonly in phylostratigraphy, there are a number of other tools including, for example, PSIBLAST (Position-Specific Information BLAST) (Altschul et al. 1997), PHMMER and HMMER (Finn et al. 2011), the MEME (Multiple Em for Motif Elicitation) suite of algorithms (Bailey et al. 2009), PSIPRED (PSI-blast based secondary structure PREDiction) (Buchan et al. 2013), and HHSEARCH (Söding 2005). Additionally, each of these tools has several parameters to tune the performance of the program which may produce more accurate results. We apply a set of these programs to simulated sequences and identify an ideal set of parameters for each. In addition to measuring the error rates, it is also important to determine whether or not these programs have the same biases as BLASTP; so we assess the correlation between homology detection error and various sequence features to determine if any of the methods is unbiased. Second, one may assess, *a priori*, a gene's propensity for homology detection error based on its sequence and evolutionary features. We investigate multiple machine learning methods to determine if there is a sufficiently sensitive and precise method to identify error-prone genes. Third, one could evaluate the error-prone status of a gene through realistic simulation, and then remove error-prone genes from phylostratigraphic analysis. Although previous researchers have claimed to do so (Domazet-Lošo et al. 2017), we demonstrated recently that their method for control is insufficient (Moyers and Zhang 2017). Error-aware phylostratigraphic analysis should be applied to only those genes shown not to produce error in simulation rather than all genes except those shown to produce error, because a sizable fraction of genes is not amenable to realistic simulation. We use this error-aware methodology to reexamine some previously reported phylostratigraphic trends.

In addition to false negative errors, phylostratigraphy is also subject to false positive errors, which cause overestimation of gene age when a nonhomolog is detected as a homolog due to sequence similarity caused by convergent sequence evolution or chance. However, because false positives owing to chance sequence similarity are avoidable by using stringent *E*-value cutoffs and those owing to rampant convergent sequence evolution (Li et al. 2010) are rarely known, false positive errors are generally believed negligible in phylostratigraphy. Nevertheless, this belief requires validation, which we now provide in the context of BLASTP and other homology detection tools. Following convention, we use phylostratigraphic error to refer to false negative error unless mentioned.

Materials and Methods

Sequence Acquisition

We acquired from OrthoMaM (Ranwez et al. 2007) 4,942 human sequences with one-to-one orthologs in 14 mammalian species diverged ~90 mya (Hedges et al. 2006). The species were *Homo sapiens*, *Pan troglodytes, Gorilla gorilla, Pongo pygmaeus, Nomascus leucogenys, Macaca mulatta, Callithrix jacchus, Tarsius syrichta, Otolemur garnettii, Microcebus murinus, Rattus norvegicus, Mus musculus, Dipodomys ordii*, and *Cavia porcellus*. We also collected all human sequences available in OrthoMaM, regardless of conservation level. Separately, we acquired a full database of human protein sequences from Ensembl, current as of September 20, 2016, available at http://ftp.ensembl.org/pub/current_fasta/homo_sapiens/pep/, last accessed August 6, 2018.

Estimating Evolutionary Rates

From the orthologs of 14 mammalian species, we employed TreePuzzle (Schmidt et al. 2002) to infer evolutionary rate information including average evolutionary rate and among-site rate heterogeneity patterns of each of the 4,942 human proteins, using the JTT–f matrix (Jones et al. 1992) with a discrete gamma model with 16 rate heterogeneity categories. Most proteins have one or more long series of conserved sites, a feature that most homology detection programs rely upon.

Parameters in Computer Simulation

We simulated three sets of proteins. In set I, we simulated the evolution of 4,942 human proteins, using the evolutionary rate and rate heterogeneity parameters estimated above. To start the simulation, we shuffled the amino acid residues within each human protein (without shuffling site-specific substitution rates) to destroy any remaining paralogy among proteins, ensuring a set of truly unrelated sequences.

In set II, we randomly picked 4,942 proteins from all human proteins downloaded from Ensemble and used their lengths in the simulation of 4,942 proteins. For each protein, its amino acid sequence is randomly constructed by assigning a random amino acid to each site, with the probability of a given amino acid appearing being equal to the frequency of that amino acid across all proteins in set I. We assigned evolutionary rate and rate heterogeneity information using a sampling method described previously (Moyers and Zhang 2016). Briefly, we computed the absolute evolutionary rate at each site of each of the 4,942 proteins from set I. We then

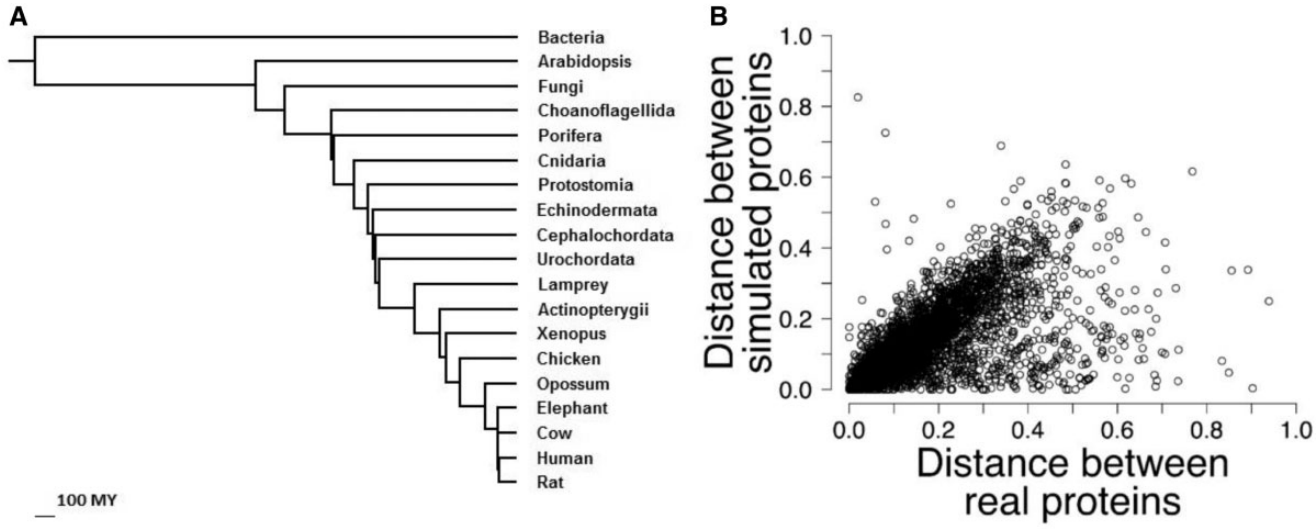

Fig. 1.—Simulation for the assessment of phylostratigraphic error. (*A*) The phylogenetic tree along which simulation of protein sequence evolution was performed. Branch lengths follow TimeTree estimates of divergence times from the human. (*B*) Comparison of maximum-likelihood genetic distance determined by TreePuzzle between human and rat real and simulated proteins (Pearson's $r = 0.61$, $P = 2.2 \times 10^{-316}$, slope = 0.85). Each circle represents one pair of orthologous proteins. Because the slope is < 1, the simulation apparently under-evolved the sequences, making our estimate of the false negative rate of phylostratigraphy conservative.

concatenated the 4,942 evolutionary rate strings into a large ring structure. Then, for each of the 4,942 proteins of set II, we sampled a continuous string of sites equal to the length of the protein in question as their site-specific evolutionary rates, requiring that the sampled string does not have all sites with the same rate.

In set III, we sought to investigate more extreme models of evolution. Because all rates were sampled from 4,942 proteins with full conservation across 14 mammals examined, it is highly likely that these evolutionary rates are not representative of those of all human proteins. We therefore created a set of proteins with higher evolutionary rates by using the exact same methodology as described for our set II, but multiplying all site-specific rates by a factor of five.

Simulation of Sequence Evolution

We simulated sequence evolution along the tree in fig. 1*A*; the branch lengths are based on average divergence times of the relevant species listed in TimeTree (Hedges et al. 2006). ROSE (Stoye et al. 1998) was used for the simulation, allowing the evolutionary rate for each site to be set by the user. We determined insertion and deletion thresholds based upon observed indel counts in our initial alignments of 4,942 mammalian sequences, according to the methodology described in Moyers and Zhang (2016). For each protein, we simulated evolution using a JTT–f matrix with observed amino acid frequencies from the alignment. Details of the simulation were previously described in Moyers and Zhang (2015, 2016, 2017).

Overview of the Homology Detection Programs Used

We performed phylostratigraphy using six programs: BLASTP, PSIBLAST, PHMMER, HMMER, OMA, and GLAM2Scan. We here provide a brief introduction to each of them, and refer readers to the corresponding original papers for deeper descriptions.

BLAST (Altschul et al. 1990, 1997) is a heuristic algorithm for homolog detection that relies on both overall sequence similarity between a query and a database entry and multiple high-scoring matches. BLAST begins its homolog search by taking "words" of a user-defined length from the query sequence and searching for high-scoring matches to these words among the entries in the database. All database entries containing a user-defined (default = 3) number of high-scoring matches with individual words are further investigated by extending the alignment and using a dynamic programing algorithm to score the alignment. If the score drops below a threshold, the comparison is stopped. Once the full score is determined, the algorithm compares the realized score with a distribution of scores based on the expected maximum score obtained from a search using a randomized query. If the realized score is sufficiently far on the right tail of this extreme value distribution, it is classified as a hit. BLASTP is one of the BLAST programs for which queries and database entries are both protein sequences.

PSIBLAST (Altschul et al. 1997) is a modification of the BLAST algorithm in which a set of homologs is used to construct a position-specific scoring matrix (PSSM). This PSSM is then used as the query to a database to detect further homologs, operating under the same fundamental process that

BLAST uses. The additional homologs can then be incorporated into the PSSM for further runs, if the user desires. The logic of this method is that by accounting for sites with greater variation, the program can detect more distant homologs. The potential danger is that by accounting for variant sites, one might include a hit which is not a true homolog into the PSSM. This has the risk of inflating the false positive rate.

PHMMER (Finn et al. 2011) is typically used as a sequence similarity search tool that generates homologs that can then be used as inputs to the HMMER algorithm described below. Using a substitution matrix to determine the score of an alignment, PHMMER searches a target database for matches to a query. The manual describes the algorithm as "BLASTP-like". Based on the query sequence offered, PHMMER creates a hidden Markov model (HMM) which uses a pre-defined substitution matrix to parameterize the model. This HMM is then used as a query for searching the database.

HMMER (Finn et al. 2011) is an iterative, profile-based algorithm that searches a target database using an HMM query. The algorithm compares query sequences to target sequences to produce an *E*-value, which is the log-odds score for the full alignment between the target and query. Like PSIBLAST, this method can then be used to incorporate new sequences into the HMM and the algorithm can be run again with a new query.

OMA (Train et al. 2017) is a method for orthology inference based on sequence similarity. The program begins by performing an all-against-all Smith-Waterman alignment of the provided proteins. Then, for each protein it determines mutually best-scoring proteins. It then performs clustering to create ortholog groups of the proteins provided.

We chose to test one additional program, GLAM2Scan. GLAM2Scan is part of the MEME suite of algorithms (Bailey et al. 2009) and was not designed as a tool for homolog detection. Instead, its purpose is to identify sequences in a target database which most closely match a user-defined motif. This is useful for identifying particular signal sequences or other commonly occurring amino acid strings. It offers a potential benefit in terms of homology detection in that it focuses only on well-conserved strings of amino acids. Because it does not directly incorporate more variant sites into the alignment, we reasoned that this method may be worth investigating as a potential tool in phylostratigraphy. The algorithm itself finds among a target database a user-defined number of alignments between a motif and target sequences. It further reports the number of exact matches to the motif. Users can trim the reported alignments based on the total similarity to the motif of interest.

Phylostratigraphy of Simulated Sequences Using Various Programs

Phylostratigraphy was conducted using several programs, including BLASTP, PSIBLAST, PHMMER, HMMER, OMA, and GLAM2Scan. In all cases, the simulated human sequences were used as the query, whereas all other simulated extant sequences in the tree of fig. 1*A* were combined into a single target database.

For BLASTP, in addition to using the default parameters, we also performed phylostratigraphic runs wherein we varied independently several parameters including Gap Extension and Gap Opening (using all possible combinations allowed by the program, as outlined in the BLASTP Manual), Composition based statistics (setting to 0 and to 1), Threshold (testing values of 8 through 15), window size (testing 0), and word size (testing 2 through 6). In total, 30 phylostratigraphic runs were performed for BLASTP for each of the three protein sets. For all runs we set the *E*-value to 100, which allowed us to progressively restrict *E*-value from 100 to 1E-10 for each run and observe the results.

For PHMMER, in addition to using default parameters, we also performed phylostratigraphic runs wherein we modified three parameters. We tested values of gap extension penalties from 0.0 to 0.9 in steps of 0.1. For each extension penalty, we varied gap opening penalty from 0.0 to 0.4 in steps of 0.1. We also varied the matrix used by PHMMER, testing all matrices allowed by the program. In total, we performed 60 phylostratigraphic runs using PHMMER for each of the three protein sets. The same *E*-values as in the BLASTP analysis were used.

For each of PSIBLAST and HMMER, we ran the initial BLASTP and PHMMER searches using the optimal parameters as determined from each of BLASTP and PHMMER. Using these starting points, we tested default parameters for each of PSIBLAST and HMMER using from 1 to 5 iterations of the programs. In total, we performed 5 phylostratigraphic runs for each of these programs for each of the three protein sets. The same *E*-values as in the BLASTP analysis were used.

For OMA (Train et al. 2017), we used the default parameters, supplying a species tree to guide the clustering of sequences. OMA's low speed prohibited us from examining other parameter settings. Based on the overall performance of OMA at its default parameters, which are presumably near the optimum, OMA is unlikely to outperform other methods appreciably.

For GLAM2Scan, we first used default BLASTP settings to identify homologs of a gene in the target database. Once such sequences were identified, we used the MEME algorithm (Bailey et al. 2009) to identify motifs in the alignment of hits. We chose the top motif and used GLAM2Scan to find matches to the motif in the target database, returning 36 hits which ensured that at least some false positives would arise in each scan. From there, for each protein we determined the age of a protein based on the hits that remained when we required that at least 10% of amino acid alignments were identical, 20% were identical, and so on until requiring 100% of amino acids were identical. We reasoned that requiring more identical hits would, to a point, exclude false positive hits in the database, and would with further restriction begin

to exclude true positive hits as well. In total, we performed 10 phylostratigraphic runs using GLAM2Scan for each of the three protein sets.

Identification of Optimal Parameters

In order to identify the optimal parameters under a particular simulation and homology detection program, we first identified the lowest false positive rate achieved among the set of parameters and *E*-value cutoffs. Because false positive rates were not found to be time-dependent, we used the presence of a false positive hit in bacteria as a measure for the false positive rate for a given parameter set at a given *E*-value cutoff. For each program, once a group of parameter sets and *E*-value cutoffs were identified as having the lowest achieved false positive rate, we compared among all of these to identify the set that had the lowest false negative rate, as defined by ability to correctly identify the bacterial homolog. Whichever parameter set had the lowest degree of false negatives was selected as the optimal parameter set.

Real Phylostratigraphy

We performed real phylostratigraphic analysis using the following three protein sets. First, we performed phylostratigraphy using the 4,942 human proteins acquired from OrthoMaM that have one-to-one orthologs in 13 other mammals examined (Ranwez et al. 2007). Second, we performed phylostratigraphy of the protein sequences of 4,942 randomly chosen human genes available from OrthoMaM. Finally, we performed phylostratigraphy using 4,942 randomly chosen proteins from the Ensembl collection of human proteins. For all phylostratigraphic runs, we used the BLASTP algorithm with default parameters and an *E*-value of 0.001. The target database was the NCBI nonredundant protein database. We identified the species represented by each hit, and determined which phylostratum each gene fell into based on species lists taken from NCBI which represented the following clades: Primate, Euarchontoglires, Boreoeutheria, Eutheria, Mammalia, Amniota, Tetrapoda, Craniata, Vertebrata, Chordata, Protostomia, Cnidaria, Eumetazoa, Eukaryota, and Bacteria.

Statistical Analyses

All statistical analyses were performed using R version 3.2.3. For the creation of support vector machines (SVMs), we used the R packages "MASS" and "e1071" (Venables and Ripley 2002). For the creation of random forests, we used the R package "randomForest" (Liaw and Wiener 2002). Hypergeometric tests were performed using the methodology previously described in Rivals et al. (2007).

Results

General Procedure for Assessing False Negative and False Positive Rates of Phylostratigraphy

To identify all false negative and false positive errors in evaluating the performance of homology detection tools, one must have a set of genes for which all true homologous relationships are known. Therefore, computer simulation is the only reliable approach. To this end, we generate a random protein sequence at the common ancestor of bacteria and eukaryotes, and simulate its evolution along the tree in fig. 1*A* until a protein ortholog is generated for each extant species in the tree. We then estimate the age of each simulated human protein by phylostratigraphy. If the human protein is found younger than its true age in the simulation, a false negative error is recorded. If the human protein hits a nonortholog, a false positive error is recorded.

To represent different classes of proteins and different kinds of challenges that a homology detection tool might face, we simulated the evolution of three sets of proteins (see Materials and Methods). Briefly, in set I, we simulated the evolution of all 4,942 human proteins that have detectable one-to-one orthologs in each of 13 examined placental mammals that last shared a common ancestor ~90 mya, using parameters estimated from orthologous protein sequence alignments (see Materials and Methods). Obviously, these proteins do not unbiasedly represent all human proteins, because of the exclusion of proteins lacking detectable orthologs in any of the 13 mammals. In set II, we simulated the evolution of 4,942 artificially constructed proteins based on the length distribution of all human proteins and the evolutionary rate distribution of the 4,942 proteins used in set I (see Materials and Methods). Because short proteins are more likely than long proteins to lack detectable homologs, proteins in set II are expected to be on average shorter than those of set I. In addition, because rapidly evolving proteins tend to miss homologs in BLASTP searches (Moyers and Zhang 2015), the 4,942 proteins in set I are expected to evolve slower than randomly chosen human proteins. Hence, in set III, we simulated protein evolution as in set II except that the evolutionary rate of each protein is five times that in set II. The three protein sets varied in two important parameters influencing the performance of phylostratigraphy: protein length and evolutionary rate (supplementary figs. S1 and S2, Supplementary Material online). Protein set I and the other two sets also have small differences in the maximum length of the most conserved block of amino acids (supplementary fig. S3, Supplementary Material online). We confirmed that our simulation was overall conservative because our simulated sequences in set I have lower divergences than those of the corresponding real sequences (fig. 1*B*). We note that protein sets II and III have two qualities often suggested to belong to "young genes": they are shorter, and they are faster-evolving. Because these protein sets include either a representation of

shorter proteins or a combination of shorter and faster evolving proteins, they offer some insight into what error trends we would expect to see if we could simulate the evolution of apparently young genes.

Identifying the Optimal Parameter Set

We wish to determine if there is a better approach to phylostratigraphy than using the BLASTP algorithm with default parameters. If another method has a much lower false negative error rate than BLASTP without having a substantially higher false positive error rate, it should be further considered for use in phylostratigraphy. Additionally, homology detection tools have a wide array of parameters that can be altered, and alteration of these parameters might improve performance. For a fair assessment of each of these tools, we should compare each of their best performances with one another.

We assessed six different tools for homology detection, which we briefly describe here. For a more thorough overview, see Materials and Methods. Tools were selected based on the following three criteria: (i) applicability to our methodology, (ii) relevance to phylostratigraphy, and (iii) usage in the biological community. We were restricted in tool selection by the qualities that our simulation of evolution respects. For instance, tools such as HHSearch (Söding 2005) and PSIPRED (Buchan et al. 2013) could not be tested because they require comparison of query and target sequences to established databases for which protein structural information is available but our sequences have no meaningful relation to those databases. The eggNOG method (Jensen et al. 2008) could not be included because it presumes an enormous set of clustered sequences from over 2,000 genomes, to which our simulated data bears no relation. There are many tools that might address questions of gene similarity possibly due to homology, but they are distinct from the phylostratigraphic approach of classifying genes on the basis of detectable homology. Such methods include, for example, the one in which genes are clustered based on functional similarity (Yi et al. 2007). Finally, tool selection was prioritized based on their relevance to the biological community. Although other tools may exist and prove useful for this problem, the selected tools are well-known, well-characterized, and have seen wide usage in relevant communities.

The BLASTP algorithm (Altschul et al. 1990, 1997), described as the "workhorse of phylostratigraphy" (Domazet-Lošo et al. 2017), uses a heuristic algorithm to extend small "words" of potential homology to create a sufficiently high-scoring alignment to mark two sequences as homologs. An expansion on the BLAST algorithm, PSIBLAST (Altschul et al. 1997) employs iterative BLAST searches using a PSSM as the query, where the matrix is updated at each iteration to incorporate new putative homologs. This additional iteration has been argued to find more distantly related homologs than does BLASTP, but may pose problems of falsely identifying nonhomologs as homologs (i.e., false positives). PHMMER and HMMER (Finn et al. 2011) are algorithms that use an HMM approach to identifying homologs, differing largely in the query input—PHMMER takes a single sequence as input, while HMMER takes multiple homologous sequences as input from which an HMM is built. Our primary interest was in HMMER's use of an HMM and its application of an overall probability of alignment, as opposed to the extreme-value-distribution-based assignment of *E*-values to alignments found in the BLAST suite of algorithms. PHMMER was included because it is a prerequisite to the use of HMMER, and it was instructive to compare the performance of PHMMER against that of BLAST given that both are based on a single sequence but have different scoring methods. OMA (Train et al. 2017) offers an example of an exhaustive protein comparison via Smith-Waterman alignment followed by clustering of sequences based on similarity profiles. It is described as an orthology-based program, but this orthology is based wholly on sequence similarity and clustering of sequences. Finally, GLAM2Scan is part of the MEME suite of algorithms (Bailey et al. 2009) and is designed to find instances of a given motif within a database rather than to produce large alignments between two sequences. Because this tool de-emphasizes surrounding sequences, we reasoned that it might be able to detect homologs with small, highly conserved segments among more rapidly evolving sequences.

We separately applied the six homology detection programs (BLASTP, PSIBLAST, PHMMER, HMMER, OMA, and GLAM2Scan) to our three simulated protein sets. For each simulation set and each program, we determined the optimal set of parameters that minimizes first false positives and then false negatives (supplementary figs. S4–S8, Supplementary Material online). To identify the optimal parameter set, we first determined the minimum false positive rate of all parameter sets at each *E*-value when considering false hits in bacteria. For the set of runs which reached this minimum false positive rate, we chose the combination of parameter set and *E*-value cutoff that produced the lowest false negative rate. This allowed us to identify the optimal parameter set for each tool under each simulation. It should be noted that some tools had different optimal parameter sets under different contexts (supplementary figs. S4–S8, Supplementary Material online), implying that in a real phylostratigraphic experiment there may be no single optimal parameter set that reduces homology detection error uniformly.

No Tool Substantially Outperforms BLASTP

On the basis of their ability to detect homologs in each phylostratum, we compared the five tools under their respective optimal parameters as well as BLASTP and OMA under default parameters (fig. 2, supplementary fig. S9, Supplementary Material online). We note first that GLAM2Scan appears to be unsuited for this kind of analysis due to high false positives

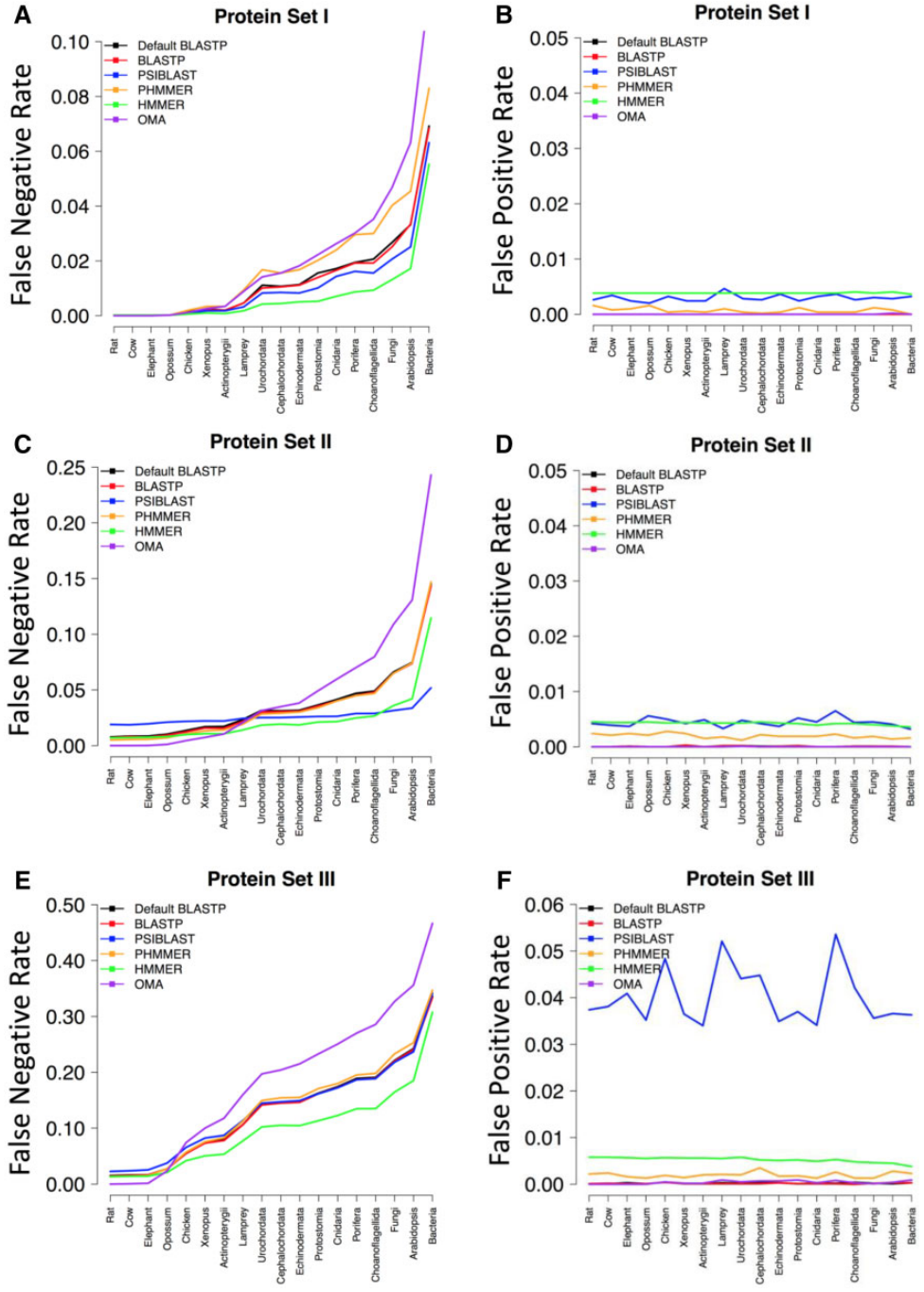

Fig. 2.—False negative and false positive error rates in phylostratigraphy using BLASTP default parameters and the optimal parameters of five programs. False negative (*A*) and positive (*B*) rates for protein set I. False negative (*C*) and positive (*D*) rates for protein set II. False negative (*E*) and positive (*F*) rates for protein set III. GLAM2Scan is not shown, but can be seen in relation to other programs in supplementary fig. S9, Supplementary Material online.

and false negatives (supplementary fig. S9, Supplementary Material online), which is not surprising given that it was never intended for this purpose; we therefore do not discuss its results further. Similarly, while OMA outperforms all other tools in terms of false negative rate for closely related species, the false negative rate becomes much higher than other programs for distantly related species under all evolutionary models. It is therefore not discussed further. Among the other programs, we note that the dynamics of performance depend at least partially on the qualities of the protein set under consideration, as is clear when comparing a given tool's performance between simulations. We also note that, generally

speaking, HMMER and PSIBLAST tend to outperform BLASTP in terms of false negative rate, whereas BLASTP tends to have the lowest false positive rate. However, these differences tend to be generally marginal, suggesting that default BLASTP may be the preferred workhorse under most conditions.

Unsurprisingly, a more realistic protein size distribution increases false negative errors substantially (fig. 2*C*), as does increased evolutionary rates (fig. 2*E*), but these increases in error also slightly change the dynamics of comparative tool performance. Although BLASTP outperforms PHMMER for protein set I, this difference virtually disappears under protein set II and III (fig. 2*A* vs. 2*C* and 2*E*). It is also interesting to note that under some conditions the ideal program in terms of false negative rate changes depending on the divergence time under consideration (fig. 2*C*), with PSIBLAST outperforming other tools in detection of homologs in bacteria, Arabidopsis, and fungi but being outperformed in other phylostrata. In closely related species, PSIBLAST has substantially higher false negative error rates than does BLASTP, PHMMER, or HMMER. Generally, HMMER outperforms other tools, but this becomes clearer the higher the error rates become (fig. 2*A* vs. 2*E*). Finally, we find that false positive rate, while it differs among programs, is generally negligible (<1%), except when using PSIBLAST in the case of fast-evolving proteins (fig. 2*F*).

In terms of the absolute error rate, we note that the false negative error rate falls between 5% and 10% for our protein set I, depending on the program used (fig. 2*A*). However, the false negative error approaches 25% for protein set II and 50% for set III (fig. 2*E*). These error rates refer to the percentage of genes missing a homolog in bacteria despite that each gene has a bacterial homolog simulated. Error rises with increasing divergence time for all programs, from ~0.02% when searching for homologs in rat to 2.5% in fungi and ~3.3% in Arabidopsis when BLASTP is used on protein set I. In protein set II, these numbers are 0.7%, 6.5%, and 7.4%; in protein set III, they are 1.5%, 22.1%, and 24.3%. The amount of error in BLASTP is well correlated with the amount of time which has passed since the most recent common ancestor of these species with human (Spearman's rho = 0.991, 0.998, 1.00 for protein sets I–III, respectively).

False Negative Errors Are Biased in All Tools

The above results show that using other homology detection tools does not appreciably reduce the error when compared with the standard practice of using BLASTP under default parameters. Nevertheless, there is a separate question of whether the errors are random or biased. We therefore determined the correlation of estimated gene age with sequence length, evolutionary rate, and the maximum length of conserved block for each of BLASTP, PSIBLAST, PHMMER, and HMMER in each of our three protein sets. We find that, in almost all cases, homology detection error creates spurious correlations between the estimated gene age and the three gene properties examined (table 1), showing that no program is without bias because all of these trends are, by definition, due exclusively to error. Furthermore, the spurious correlations created by different programs have the same direction although the magnitude of the correlation varies (table 1). Because any reduction in error is generally joined with a substantially increased computational load, we see no reason to use alternative homology detection programs. Hence, BLASTP is as reasonable a choice for phylostratigraphy as other tools are.

Table 1

Spurious Correlations (Spearman's ρ) between Estimated Gene Age and Biological Features

	BLASTP	PSIBLAST	PHMMER	HMMER
Protein set I				
Protein length	0.14**	0.16**	0.03	0.11**
Evolutionary rate	−0.37***	−0.36***	−0.02	−0.34***
Block length[a]	0.35***	0.35***	0.03*	0.32***
Protein set II				
Protein length	0.31***	0.30***	0.28***	0.27***
Evolutionary rate	−022***	−0.08**	−0.22***	−0.22***
Block length[a]	0.37***	0.28***	0.37***	0.33***
Protein set III				
Protein length	0.32***	0.36***	0.29***	0.28***
Evolutionary rate	−0.12**	−0.12**	−0.13**	−0.13**
Block length[a]	0.41***	0.44***	0.42***	0.38***

[a]Length of the longest block of conserved residues.
*$P < 0.05$; **$P < 1 \times 10^{-10}$; ***$P < 1 \times 10^{-100}$.

Predictive Models of Propensity for Error

Another possible way to remove the effects of homology detection error in phylostratigraphy is the application of a model that identifies *a priori* those genes likely to be subject to homology detection error and exclude them from phylostratigraphic analysis. We reasoned that if we could construct a model that is able to identify 90% or more of error-prone genes without removing a substantial proportion of nonerror-prone genes, this would be an effective model. Here, error-prone genes are defined as those whose estimated ages differ from the true ages simulated. We therefore used the BLASTP-estimated gene ages of the three sets of simulated proteins to construct SVM and random forest models. We used 10-fold cross-validation to determine the average sensitivity, specificity, and precision of each model, where sensitivity is the number of correctly identified error-prone genes divided by the total number of error-prone genes, specificity is the number of correctly ignored nonerror-prone genes divided by the total number of nonerror-prone genes, and precision is the number of correctly identified error-prone genes divided by the total number of genes identified as being error-prone. We tried all combinations of the three parameters (protein length, evolutionary rate, and maximum length of conserved block) as predictors

Table 2
Performances of the Best-Performing Machine Learning Models in Identifying Error-Prone Genes for Each Protein Set by SVM and Random Forest Methods

	SVM			Random Forest		
	Protein Set I	**Protein Set II**	**Protein Set III**	**Protein Set I**	**Protein Set II**	**Protein Set III**
Best-performing model[a]	Error ~ L+E+B	Error ~ L+E+B	Error ~ L*E*B	Error ~ L+E+B	Error ~ L+E+B	Error ~ B
Sensitivity	0.504	0.253	0.512	0.711	0.629	0.633
Specificity	0.987	0.984	0.863	0.967	0.900	0.730
Precision	0.768	0.718	0.653	0.519	0.360	0.336

[a]L = protein length; E = evolutionary rate; B = maximum length of conserved block.

Table 3
Spearman's Correlation between Gene Age and Gene Properties in Real Phylostratigraphy

Gene Properties Correlated	**4,942 Proteins Randomly Chosen from Ensembl**	**4,942 Proteins Randomly Chosen from OrthoMaM**	**4,942 Proteins Used in Simulation**	**4,619 Nonerror-Prone Proteins**
Protein length and age	0.16**	−0.010	−0.09**	−0.12**
Evolutionary rate and age	NA	−0.18**	−0.05*	0.002

*$P < 0.05$; **$P < 1 \times 10^{-10}$; ***$P < 1 \times 10^{-100}$.
N/A, not applicable because the evolutionary rate cannot be estimated for some genes due to the lack of detectable orthologs.

and used error measured by a missed bacterial homolog as a response variable. We then determined which set of predictors produced the model with the greatest sensitivity.

We found that none of the models were sufficiently sensitive, though random forests performed better than SVM models (table 2). We also created models wherein the response variable was whether or not a homolog was found in fungi. This, however, did not change the results (supplementary table S1, Supplementary Material online). Hence, we consider machine learning unsuccessful in predicting error-prone genes.

Error-Aware Phylostratigraphy

Having investigated several methods to remove the effects of error and found none, we are left with only one method: removal of error-prone genes identified by simulation of gene evolution. Here, error-prone genes are those exhibiting any amount of homology detection error in a realistic simulation of evolution. In the context of simulation, removing error-prone genes results in an error-free data set, because all remaining genes have correctly estimated ages. Because not all genes are amenable to realistic simulation, we emphasize that, in applying this method in phylostratigraphy, one should restrict to genes that can be realistically simulated and are not found error-prone (Moyers and Zhang 2017). Here we investigate if using this method alters any previously observed phylostratigraphic trend. To that end, we performed a real phylostratigraphic analysis of the 4,942 human genes that have one-to-one orthologs in each of 13 other mammals examined, against the NCBI nonredundant protein database. We then removed 323 genes found by default BLASTP in our simulation of protein set I to be subject to homology detection error. The remaining 4,619 error-free genes still have a substantial variation in estimated age (supplementary fig. S10, Supplementary Material online).

We examined the correlation between estimated gene age and 2 traits (protein length and evolutionary rate) in the 4,619 genes. Surprisingly, we observed a significantly negative correlation between gene age and protein length (table 3), which is opposite to what was previously reported from phylostratigraphy that did not remove error-prone genes (Carvunis et al. 2012). We confirmed that indeed a significant positive correlation is observed in a random set of 4,942 genes from the Ensembl human genome (table 3). This correlation disappears for a random set of 4,942 human genes from OrthoMaM, becomes significantly negative for the 4,942 human genes with orthologs in all of the 13 other mammals examined, and is even more negative upon the removal of error-prone genes (table 3). Previous phylostratigraphic studies also reported a significant negative correlation between gene age and protein evolutionary rate (Albà and Castresana 2007), which is also present in our 4,942 genes (table 3). But, when only the error-free genes are considered, the correlation becomes positive, albeit not statistically significantly (table 3). These findings suggest that the previous phylostratigraphic findings were artifacts. In the past, phylostratigraphic findings of similar absolute strength were used to derive models of gene maturation (Carvunis et al. 2012; Abrusán 2013). Our discovery that these trends disappear or reverse to a similar absolute strength in error-aware phylostratigraphy casts doubts on these models.

Discussion

We have reconfirmed here that false negative error is prevalent in phylostratigraphy and demonstrated that the severity of this problem is greater than what was previously shown via conservative simulations. Although the 4,942 human genes with one-to-one orthologs in 13 other mammals (protein set I) have only a moderate degree of error (6.5% of genes missed their bacterial homolog in our simulation), this error rate is an underestimate because these genes are necessarily less prone to error given that their homologs are identified from 13 other mammals. Even changing length distributions to be more realistic (protein set II; supplementary fig. S1, Supplementary Material online) without substantially changing evolutionary rate or rate heterogeneity properties (supplementary figs. S2 and S3, Supplementary Material online) produces greatly increased error. We observed that 14.3% of genes could not find a bacterial homolog in our protein set II. When higher evolutionary rates are introduced (protein set III), we find that this error rate can be substantially increased, with 33.4% missing a bacterial homolog. Note that there are other substitution patterns such as those described by the covarion model that would further increase the error rate (Moyers and Zhang 2015). For two reasons, we have chosen not to include such a model in this study. First, the precise degree of covariation is not clear, because current estimates are based on extremely conserved proteins (Wang et al. 2009). Thus, estimates of covarion rates suffer from a similar problem as other phylostratigraphic findings: those proteins with high rates of covariation are theoretically unlikely to have enough information for estimating covarion rates. Therefore, inclusion of the covarion model here would have questionable applicability. Second, the general findings of including a covarion model in our simulation are predictable: it would increase the amount of false negative error, as has been observed in BLASTP (Moyers and Zhang 2015). Fundamentally, a covarion model breaks regions of conservation among homologs by allowing highly conserved sites to become less well conserved in a subset of branches. All methods relevant to phylostratigraphy rely on sequence similarity, so it is expected that introduction of a covarion model would increase false negative error for all programs.

It is true that missing a bacterial homolog does not guarantee that a human gene will also miss a homolog in a more closely related species, but this kind of error still does occur. We observed 0.02%, 0.74%, and 1.5% error in detecting rat homologs in gene sets I, II, and III, respectively. Because genes that appear to be restricted to closely related species tend to be shorter and fast-evolving, these genes are expected to be more subject to error, and some fraction of them appear young specifically because of this kind of error. This is in concordance with the results of Domazet-Loso et al. (2017), who reevaluated the species-specific status of 15 ORFs (as assigned by phylostratigraphy) and found that one third of them were falsely classified as species-specific. It is also shown by innovative new tools recently introduced (Martín-Durán et al. 2017).

We found that other homology-detection tools do not noticeably outperform the standard BLASTP in terms of producing false negative errors. Any concerns surrounding false positive error are not well-supported by our results. Although we find that profile-based homology detection programs (PSIBLAST and HMMER) generally have a higher degree of false positive error than does BLASTP, this error is small (<1% of genes) except for small, fast-evolving genes using PSIBLAST (fig. 2*F*). Moreover, we find that false positive error is not time-dependent whereas false negative error is. False positive error is therefore less likely to introduce spurious trends with gene evolution when it does occur. It may be argued that proteins from distinct families sometimes bear regions of homology to one another. Because our simulation does not retain precise amino acid sequence of existing genes and does not include such patterns, one might argue that it is therefore underestimating the false positive rate. We reject this argument on the grounds that such shared sequence identity is the result of true homology—that is, derivation from the same ancestral sequence. This can occur through nonhomologous recombination or the same general model of divergence that produces false negative error, with sequence identity retained in a subset of proteins. Both of these explanations were found for example in a study of archaeal restriction endonucleases (Sukackaite et al. 2007). Inclusion of such a model in our simulations would overestimate rates of false positive error.

We have here demonstrated that error-aware phylostratigraphy is not merely a conservative approach to phylostratigraphy, but can provide novel biological insights. For instance, the previously reported negative correlation between gene age and evolutionary rate that ignited the initial suspicion about phylostratigraphy (Elhaik et al. 2006) disappears when only nonerror-prone genes are analyzed (table 3). More strikingly, the positive correlation between protein length and gene age even becomes reversed when only nonerror-prone genes are considered (table 3). Although it may appear that these trends—both the original trends and those seen only in error-aware phylostratigraphy—are relatively weak, much has been made of them. Trends of similar strength have been used to suggest that young genes evolve quickly and are shorter in several studies. One study (Abrusán 2013) presented associations corresponding to similar correlation strength between gene age and properties as diverse as number of transcription factors regulating a gene and the percentage of protein sequence present in alpha helices to argue for a model of gene integration into network with age. The finding here that some of these trends disappear weakens evidence for these claims. Our finding that some of these trends change direction suggests alternative models of gene

maturation with evidence of comparable strength found in previous studies.

Future work should investigate alternative methods for identifying error-prone genes. Our simulation set is inappropriate for use with certain homology detection methods. This is partially due to the fact that our simulation does not take into account such properties as protein structure, and partially because some tools are applicable to nucleotide sequences and we cannot accurately simulate the more complicated evolution of protein-coding nucleotide sequences. If a new simulation set can be performed that captures such features as structural evolution and other sequence evolution constraints, the phylostratigraphic effects of homology detection error may be reduced. The error-prone status of genes might be further probed by using a larger number of genes for simulation with their native, as opposed to sampled or simulated, properties. There is an inherent problem here, because simulation requires inference of evolutionary parameters and inference of evolutionary parameters requires detectable homologs. Thus, there is a set of genes which, by definition, cannot be simulated according to their native parameters—those which have no detectable homologs. Additionally, for those genes with few detectable homologs (or when using fewer homologs to infer evolutionary parameters), issues of stochasticity become greater and simulations are more likely to be inaccurate. Therefore, error-aware phylostratigraphy may have a necessary limitation in which sequences it can evaluate. We hope that prior phylostratigraphic findings will be reevaluated in this context, and that future work will account for phylostratigraphic error in inferring evolutionary mechanisms.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Chuan Xu, Zhengting Zou, and three anonymous reviewers for valuable comments. This work was supported in part by the U.S. National Institutes of Health research grant R01GM120093 to J.Z.

Literature Cited

Abrusán G. 2013. Integration of new genes into cellular networks, and their structural maturation. Genetics 195(4):1407–1417.
Albà MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. Mol Biol Evol. 22(3):598–606.
Albà MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. BMC Evol Biol. 7(1):53–58.
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.
Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.
Bailey TL, et al. 2009. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 37(Web Server):W202–W208.
Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT. 2013. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res. 41(W1):W349–W357.
Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol. 2:393–409.
Carvunis A-R, et al. 2012. Proto-genes and de novo gene birth. Nature 487(7407):370–374.
Domazet-Lošo T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23(11):533–533.
Domazet-Lošo T, et al. 2017. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Mol Biol Evol. 34(4):843–856.
Domazet-Lošo T, Tautz D. 2003. An evolutionary analysis of orphan genes in Drosophila. Genome Res. 13(10):2213–2219.
Domazet-Lošo T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol. 25(12):2699–2707.
Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature 468(7325):815–818.
Elhaik E, Sabath N, Graur D. 2006. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. Mol Biol Evol. 23(1):1–3.
Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39(suppl):W29–W37.
Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22(23):2971–2972.
Hemmrich G, et al. 2012. Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity. Mol Biol Evol. 29(11):3267–3280.
Jensen LJ, et al. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 36(Database):D250–D254.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8(3):275–282.
Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene Prestin unites echolocating bats and whales. Curr Biol. 20(2):R55–R56.
Liaw A, Wiener M. 2002. Classification and Regression by randomForest. R News. 2:18–22.
Martín-Durán JM, Ryan JF, Vellutini BC, Pang K, Hejnol A. 2017. Increased taxon sampling reveals thousands of hidden orthologs in flatworms. Genome Res. 27(7):1263–1272.
McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 17(9):567–578.
Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol Biol Evol. 32(1):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.
Moyers BA, Zhang J. 2017. Further simulations and analyses demonstrate open problems of phylostratigraphy. Genome Biol Evol. 9(6):1519–1527.
Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics 14(1):117–113.
Prat Y, Fromer M, Linial N, Linial M. 2009. Codon usage is associated with the evolutionary age of genes in metazoan genomes. BMC Evol Biol. 9:285.

Ranwez V, et al. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol Biol. 7:241.
Rivals I, Personnaz L, Taing L, Potier MC. 2007. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23(4):401–407.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18(3):502–504.
Sestak MS, Božičević V, Bakarić R, Dunjko V, Domazet-Lošo T. 2013. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. Front Zool. 10(1):18.
Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7):951–960.
Stoye J, Evers D, Meyer F. 1998. Rose: generating sequence families. Bioinformatics 14(2):157–163.
Sukackaite R, et al. 2007. Restriction endonuclease BpuJI specific for the 5′-CCCGT sequence is related to the archaeal Holliday junction resolvase family. Nucleic Acids Res. 35(7):2377–2389.
Train CM, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C. 2017. Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. Bioinformatics 33(14):i75–i82.
Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. Verlag, New York: Springer.
Wang HC1, Susko E, Roger AJ. 2009. PROCOV: maximum likelihood estimation of protein phylogeny under covarion models and site-specific covarion pattern analysis. BMC Evol Biol. 9(1):225.
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 USA. 106(18):7273–7280.
Yi G, Sze SH, Thon MR. 2007. Identifying clusters of functionally related genes in genomes. Bioinformatics 23(9):1053–1060.

Associate editor: Bill Martin