On Gene Age, Gene Origins, and Evolutionary Trends

By

Bryan Anthony Moyers

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2017

Doctoral Committee:

Professor Jianzhi Zhang, Chair
Professor Patricia Wittkopp
Associate Professor Jun Li
Assistant Professor Stephen Smith
Professor Edward Ionides

Dedication

To my wife, Carmen, who always makes sure I enjoy life as I study it.

Acknowledgements

I would like to thank Jianzhi (George) Zhang, for his constant guidance, mentorship, and support. Whatever my goals or interest, George has always offered me resources and respect. He has consistently helped me to remain focused on the biology rather than the computation, a decisive quality in my career success. I would not be half the scientist I am without such a skilled and careful mentor.

Drs. Margit Burmeister, Jun Li, and Michael Boehnke deserve special thanks, as they offered me extensive mentorship while I considered career options in and outside of science. Without their support, I would likely have followed a much different path.

I would also like to thank the members of the Zhang lab, past and present, who have always offered insightful discussion. In particular, Jianrong Yang, Chuan Li, Wei-chin Ho, Zhenting Zou, Jinrui Xu, Chungoo Park, Brian Metzger, and Nagajuran Vijay, all of whom have offered extensive advice to me on multiple occasions.

Special thanks go to my classmates and cohort in the Bioinformatics PhD and Masters programs, who have helped me both in and out of academic life. In particular, Ari Allyn-Feuer, Shweta Ramdas, Brandon Govindarajoo, Raymond Cavalcante, Brittany Nelson, Tony Chen, Laura Seaman, Shiya Song, Yindra Puentes, Patricia Fly, and Alexandr Kalinin.

Extracurricular pursuits often reinvigorated my passion for science, and I owe thanks to the many people that encouraged me and collaborated with me in that realm. Thanks go out to Ada Hagan, Alisha John, Kevin Boehnke, So-Hae (Irene) Park, Shweta Ramdas, and Dr. Scott Barolo for their hard work and dedication to creating MiSciWriters and to improving my writing ability (despite my best efforts to resist them). Drs. Katherine Moynihan and Mutsumi Yoshida guided me through my time in the Technology Transfer Fellows program, in which I saw the diverse results of applied and basic research on the market. Drs. John Godfrey, Lori Isom, Joseph Kolars, and Senait Fisseha were all pivotal in the establishment of the EM-PACE program, along with far more students than I can mention. Particular thanks go out to Eden Dulka, who helped me keep my spirits and sanity in the cradle of life.

Last, but far from least, I thank my family; my loving wife, Carmen, who has fought with me through hardship, uncertainty, and all the adventures of life; my parents, Tony and Regina, who have always challenged me to struggle for truth, however difficult, and have loved me, no matter where I land; and my brother, Matthew, who always reminds me to not get too lost in a single pursuit and remember that life has so much joy to offer. Without the light they bring to my life, the appeal of pursuing scientific truth would be infinitely diminished.

Table of Contents

List of Tables

List of Figures

List Appendices

Abstract

Novel genes are a contributor to species diversity and specialization. Determining when, how, and in which lineages novel genes formed is a major challenge in evolutionary biology. A key step in this process is identifying novel genes. Phylostratigraphy is a method developed to identify novel sequences. This method relies on the detection of homologs, existing sequences in different species which derive from a common ancestral sequence. This method uses homology detection programs, such as the BLAST suite of algorithms, to identify genes that are specific to a lineage and infer from there when this sequence arose. When done for large numbers of sequences, they can be grouped by age and trends with gene age can be identified. This methodology assumes that homology detection error—the failure of a homology detection program to accurately detect homologs—is negligible. I show that this is a faulty assumption. I demonstrate that homology detection error is more common than previously believed, and that it is non-random. Homology detection error is biased in a way that may produce spurious biological trends. I demonstrate that this kind of error has major influence on theories of gene emergence. I further develop a methodology which addresses and mitigates the effects of error on phylostratigraphy, and use this method to approach phylostratigraphic problems and produce novel biological insight. In total, this thesis demonstrates a major problem in phylostratigraphic methodology, produces a new methodology which addresses these limitations, and applies this methodology to investigate problems of gene age, the mechanisms by which genes emerge, and trends in evolution.

Chapter 1

Introduction to the Field

"In principle, the recovery of homology only requires a source of information with two properties: sufficiently numerous and sufficiently independent items to preclude, on grounds of mathematical probability alone, any independent origin in two separate lineages."
- Stephen Jay Gould, 1986

Introduction

We see among the myriad species on Earth both incredible diversity of structure and cases of curious similarities. We naturally wonder how each of these arises. Based on modern understandings of biology, we can be sure that the genetic material of organisms helps to determine their development, structure, and function. When comparing the genetic material between species, we can similarly see remarkable similarities and differences—both small changes at single points in the genome as well as enormous structural differences between the genomes of organisms. Of particular interest is the fact that when comparing the genes in any two organisms, some genes appear in very similar forms in both species, whereas other genes are found in only one of the two species. Given that common ancestry of all life, one must wonder how these differences in gene number arise. This thesis focuses on methodologies for determining evolutionary trends of gene origin as well as the contributions of various mechanisms to novel gene formation.

Homology

Homology is perhaps one of the oldest concepts in western biology. Aristotle used homologies as a key factor in classifying animals in his *History of Animals* (Aristotle, 1984). In this context, the term generally referred to parts or behaviors of animals which were analogous in form or function. This concept has influenced biology ever since, most famously in the classification system of Linnaeus in his *Systema Naturae* (Linnaeus, 1735), which went on to influence all of European biology. Such homologies were typically explained by the perfect formation of the universe for humans in particular and all animals in general (Paley, 1802). It was Darwin's classic text *On the Origin of Species by Natural Selection* (Darwin, 1859) which provided an alternative explanation: these homologies exist because they were inherited from a common ancestor.

Mendel unwittingly lent credence to this idea with his studies on inheritance. His experiments showed that variation within the population could be inherited with predictable patterns. But it would be a long time before Friedrich Meischer's discovery of nuclein (Dahm, 2005), now known as Deoxyribonucleic Acid (DNA), and its establishment as the hereditary factor (Avery, Macleod, & McCarty, 1944). Once this was done, it was possible to start identifying more precisely the segments of DNA which corresponded to certain heritable traits (Rubin & Lewis, 2000). Developmental biology, and indeed even Mendel's experiments, showed that the relationship was not a simple one-to-one relationship between genes and homologous traits. Nonetheless, it was clear through key experiments that some genes clearly played a similar role in forming homologous structures between species, as when a mouse gene spurred the formation

of an eye in *Drosophila melanogaster* (Gehring, 2002). It became only natural to integrate evolution with the molecular revolution and begin to ask about homologies between genes.

It is worth here identifying three distinct types of homologies. The first is historical homology, which is the relationship between structures such that if you followed two organisms back along their ancestral paths, tracking the analogous feature in parent and offspring , the features in each species would correspond to the same feature in their common ancestor. As an example, one could identify the wing of a pigeon and the wing of a cardinal as historically homologous. However, the wing of a pigeon and the wing of a bat would not have historical homology, as the wing formed independently in mammals. An important feature of this historical homology is that it is inferred, as the tape of evolution cannot be rewound and played for us to observe the historical relationships. Instead, we must infer historical homologies through observation and proxy homologies. The first of these was structural similarity.

Structural similarity is, briefly, the kind of homology used from Aristotle to Darwin. By examining the structure, anatomy, and sometimes function of the parts of various organisms as well as the structure of the organisms as a whole, biologists classify structural similarities as historically homologous. In order to establish historical homology from happenstance structural similarity, various methods which were based on parsimony were introduced. This structural similarity was insufficient to establish the homologous nature of genes, as their structure was difficult to identify, and it was not clear if a similar function between genes in different organisms corresponded to similar structure. It was therefore necessary to introduce a new method for homology: sequence homology.

Once DNA had been identified as the hereditary molecule, the ability to determine its sequence became crucial. An important step toward this goal was the work of Rosalind Franklin, Maurice Wilkins, Francis Crick and James Watson (Watson & Crick, 1953), which enabled Frederick Sanger and colleagues to introduce a method for rapidly sequencing segments of DNA (Sanger, Nicklen, & Coulson, 1977). This opened up the possibility for researchers to compare gene sequences and determine if segments of DNA which performed similar functions in different organisms also looked similar. Like the form of structural similarity previously described, if two sequences in different organisms had a similar enough sequence, they were called homologous, and it was assumed that they bore a historical homology. Since then, the efficiency, speed, and power of sequencing techniques has astronomically improved (Heather & Chain, 2016). This has necessitated the automation of the detection of sequence homology.

Even before the massive amount of sequence information existed, it was necessary to automate the process of comparing gene sequences. The length of these sequences and the known molecular processes of nucleotide substitution, insertions, deletions, and inversions, made it difficult to compare sequences with the human eye. The Smith-Waterman algorithm was developed to explore the possible alignments between two sequences and give a quantifiable result about their homologous status (T. F. Smith & Waterman, 1981). Generally, the logic goes thus: If two sequences shared a common ancestor, then they have had some period of time to acquire differences between one another. Sequence space is enormous—a gene with only 30 nucleotides has a potential 4^30 sequences. It therefore seems reasonable that if two sequences are very close in sequence space, this is due to a historically homologous relationship with one

another. Conversely, it was extremely unlikely that two gene sequences which were not historically homologous would wander close to each other in sequence space (Figure 1-1, red and green points). The Smith-Waterman algorithm allowed researchers to define a score which quantified how close together two sequences must be in order to be considered historical homologs (Figure 1-1, expanding circles).

The major limitation of the Smith-Waterman algorithm was its exhaustive comparison of all possible alignments between two sequences, many of which were not worth considering, practically. This feature of the algorithm meant that it worked in O(n*m) time, where m and n refer to the length of the two sequences being compared. While this was tractable for comparing small numbers of sequences, it was intractable for the increasingly large amount of sequence data that was being generated (Benson et al., 2013). This required the development of heuristic algorithms. Many such methods of sequence comparison have been developed (S F Altschul et al., 1997; Finn, Clements, & Eddy, 2011; Grundy & Grundy, 1998; H. Li & Homer, 2010), but they all follow the same general principles as the Smith-Waterman algorithm: if two sequences are close enough in some description of sequence space, they are homologs and as such share a common ancestry.

There is, however, a common problem with all of these algorithms: they cannot tell you anything about the actual historical homology relationship between two sequences, only a measure of how similar two sequences are. If a researcher is too restrictive with the score cutoff or if evolution is proceeding rapidly, historical homologs may not be identified as sequence homologs (Figure 1-

1**)**. If a researcher is too loose with the score cutoff then non-historical homologs may be counted as historical homologs (Figure 1-1).

The evaluation of the accuracy of these tools has typically been done by comparison of their results to well-curated "Gold Standard" databases such as the Structural Classification of Proteins (SCOP) (Conte et al., 2000). These databases are constructed and curated based on expert study of protein function, their three-dimensional structures, and other characteristics. The underlying logic of their construction is that by comparing many aspects of a gene—including its sequence properties, function, genomic location, and structure—in relation to the property of other proteins, researchers can construct relationships between genes for which we can be highly confident of their accuracy. The sequence homology tools developed and the recommendation of score cutoffs are designed with the goal of reconstructing these gold standards as accurately as possible.

An interesting question regarding origins can arise when comparing the gene sets of two species. When using one of these tools to identify homologs between the species pair, one often finds that a substantial number of genes will have a historical homolog in the other species, but each species will have some number of genes which does not bear homology with any gene in the other species. Just as one might ask where the homologous wings of al birds come from, one might also ask where homologous genes come from.

Gene formation

When a gene appears in one species but not in some other species, it implies that since the divergence of the most recent common ancestor of the two, either one species has lost a gene, one has gained a gene through gene duplication, or a new genic sequence has somehow formed from non-genic sequence. There are several possibilities of where a new gene may have come from. I briefly give here a description of the major mechanisms.

Because homology detection programs generally rely upon sequence identity, genetic divergence is one of the first possibilities to explain a lack of homologs between species. After splitting of the population, the same genetic sequence will undergo independent sequence changes in each lineage. Over sufficient time, this may cause a sufficient lack of sequence identity for two true historical homologs to no longer be identified as such. This is thought to be a relatively rare occurrence in the lack of a change in gene function, as evidenced by the work of Alba and Castrasanna 2007 (Albà & Castresana, 2007). This is because gene function often relies on conserved stretches of the protein to maintain the appropriate structure. These conserved sequences are a major part of the correct function of homology detection programs. However, this phenomenon does occur at least sometimes, and if species have undergone a change in functional constraint which affects a given gene, then two historical homologs may diverge wholly independently and lose detectable homology.

Related to this possibility, a new sequence may be produced by duplication of an existing gene sequence (M Lynch & Conery, 2000). This is a well-documented method for the formation of new genes, and was proposed as a major drive of evolutionary innovation by Susumu Ohno (Ohno, 1970). However, gene duplicates are at least sometimes identifiable as homologs on the

basis of either sequence homology (Henikoff et al., 1997) or syntenic (gene order) analysis (Byrne & Wolfe, 2005). In fact, if a sequence has historical homology with another sequence in the same species, these are a special kind of historical homolog called a paralog. However, it is possible that a duplication event will create a situation in which the two paralogs individually change their functional constraints or one of the two has a change in its functional constraint (Jianzhi Zhang, 2013). There are two specific models in which this kind of change is expected to have an effect on sequence identity between paralogs.

The first model is subfunctionalization. If a protein is performing more than one function, it is sometimes possible that its sequence is constrained in such a way that the gene cannot specialize more closely to any of its multiple functions. We can take the example of an enzyme that displays promiscuity in its target (Figure 1-2). If a single gene is responsible for breaking down multiple metabolites, it may not be particularly good at breaking down any one metabolite. When a duplication occurs, the two daughter genes are free to specialize more closely to one metabolite, or one set of metabolites. This allows finer-tuning, and may drive a loss of some conserved sites between the two proteins. This can drive the loss of detectable homology.

The second model is neofunctionalization. After the duplication of a sequence, one of the two copies is free to accrue mutations. This is because a loss or change of some essential function in one copy is compensated for by the existence of the second copy. Frequently, this will allow one of the two daughter genes to accrue mutations, lose function, and undergo pseudogenization. However, it sometimes happens that the mutations acquired by one of the copies grants a novel function (M Lynch & Conery, 2000). The gene can then specialize for this new function, which

may select for very different conserved regions (Figure 1-3. This can therefore drive loss of detectable homology.

It is possible that one may compare the genes of several species and discover a pattern that suggests that a gene has potentially been lost many times independently (Figure 1-4). While it is possible that such a pattern may have occurred, it is also possible that a gene has been donated from one species to one or more other species in a process known as Horizontal Gene Transfer (HGT) (Soucy, Huang, & Gogarten, 2015). This has been well-documented in bacteria (Martínez, 2008; Ochman, Lawrence, & Groisman, 2000; Pál, Papp, & Lercher, 2005) but there is evidence functional proteins have also been passed to eukaryotes from other kingdoms of life, despite initial skepticism (Hotopp et al., 2006; Keeling & Palmer, 2008; Salzberg, White, Peterson, & Eisen, 2001). However, due to both a lack of confirmed cases in many species and lack of plausible biological mechanisms for this being a common occurrence outside of bacteria, it is generally thought that this is rare.

Another possible mechanism is the *de novo* formation of a new gene in a given species. This can occur in a number of methods, as outlined by McLysaght and Hurst in their thorough review on the topic (Mclysaght & Hurst, 2016). It is thought that *de novo* gene birth can occur in several ways. The recruitment of an open reading frame along with the relevant transcriptional and translational signals can produce a novel gene (Knowles & McLysaght, 2009). Alternatively, the fusion of a gene fragment due to transposition or duplication of a genomic segment can induce the creation of a novel gene sequence (Mclysaght & Hurst, 2016; Song, Wachi, Doi, Ishino, &

Mutsuhashi, 1987). The formation of a novel sequence due to frame shift mutations, after gene duplication for instance, has also been suggested as a source for novel sequence formation (Vandenbussche, Theissen, Van de Peer, & Gerats, 2003). These mechanisms have generally been disregarded as realistic sources for functional new genes due to the random nature of the resulting sequences. This sentiment is epitomized by Francois Jacob's 1977 quote, "The probability that a functional gene would form by random association of amino acids is practically zero," (Jacob, 1977). And, even granting that this is possible, there are many outstanding questions which are being approached by researchers: How are regulatory signals recruited to new locations, whether or not there is an open reading frame (A. Carvunis et al., 2015; Eichenlaub & Ettwiller, 2011)? How is the leap made from a transcribed gene to a translated gene (Banfai et al., 2012; Ingolia et al., 2014)? How frequently do these events happen (A.-R. Carvunis et al., 2012; Neme & Tautz, 2013)?

These questions of homology and gene origin have spawned a renewed interest in general evolutionary patterns of how genes emerge, under what conditions, and how their properties changed over time. This has led to the desire to identify the formation of *de novo* genes. However, distinguishing *de novo* genes from genes formed by other mechanisms is challenging for at least two major reasons. First, lack of homologous sequences could be merely due to sequence divergence. Second, an apparently new open reading frame may not be transcribed or translated, or perform any function at all. It has been therefore suggested that three criteria are necessary to confidently identify a recent *de novo* gene (Knowles & McLysaght, 2009). One must find 1) a gene which is both transcribed and translated 2) which has an orthologous sequence in related species which does not code for a protein and 3) the ancestral sequence is

also noncoding. If you cannot find the appropriate genomic region in closely-related species, this suggests that some other event has occurred, such as a horizontal gene transfer, a translocation, a gene fusion, or some form of duplication. If you find the appropriate sequence in nearby species and find that a protein is produced, then the gene has not been recently born *de novo*.

Once such a gene has been identified, its properties can be studied, and we can make attempts to reconstruct its history. However, this method is extremely limited, and can only examine case-studies. It tells us relatively little about the broad strokes and patterns of gene formation over evolutionary time. For that, a high-throughput method is required.

Phylostratigraphy and *de novo* Gene Birth

In this earnest effort to identify patterns of gene formation, an interesting idea was introduced in genomics: the idea of phylostratigraphy. In a sense, the concept was a molecular twist on an old idea. It was already recognized that one could approximately date the emergence of biological structures by identifying all species with that structure and determining approximately when the most recent common ancestor of that species lived. For instance, the bilateral body plan can be dated to the Ediacaran period some 600 million years ago (Peterson, Cotton, Gehling, & Pisani, 2008), the emergence of the clade Bilatera. Phylostratigraphy was a method introduced by Diethard Tautz and colleagues in 2007 (Domazet-Lošo, Brajkovic, & Tautz, 2007) with similar logic. If one can identify all species which have the homolog for a given gene, then that gene emerged at approximately the date of the most recent common ancestor for all of those species. This method had two major advantages. First, it allowed gene emergence studies to see further

back in time by comparing distantly-diverged species rather than only closely-related species. Second, it could be done in a high-throughput manner, instead of on a gene-by-gene basis.

The method is this: identify a query species, and create a database of all of its genes. Create a target database consisting of all of the known sequences of many species, whose evolutionary relationship to your query species is known. Then, for every query protein, perform a homology search to identify which species do or do not have homologs of that gene. Once all homologs are identified, the researcher can determine the approximate age of a protein by identifying the Last Common Ancestor of all species with a homolog. The age of a gene is thus defined in this method by its detectable homologs.

This method also allowed the study of trends with gene age. Typically, after performing this homology detection for all proteins, researchers will attempt to identify some relevant association of gene age and a biological property. Many such analyses have been performed, and associations have been found between gene age and length (A.-R. Carvunis et al., 2012; Wolf, Novichkov, Karev, Koonin, & Lipman, 2009), evolutionary rate (Albà & Castresana, 2005), tissue expression in various developmental times and species (Domazet-Lošo & Tautz, 2010), the emergence of multicellularity (Hemmrich et al., 2012), the formation of the head and neck sensory systems (M S Sestak, Bozicevic, Bakaric, Dunjko, & Domazet-Loso, 2013), and several other properties (Abrusán, 2013; J. J. Cai & Petrov, 2010; A.-R. Carvunis et al., 2012; Domazet-Lošo & Tautz, 2008; Prat, Fromer, Linial, & Linial, 2009). Most importantly for discussion of novel gene emergence, it has been used to suggest that *de novo* gene birth occurs frequently (A.-R. Carvunis et al., 2012; Neme & Tautz, 2013). It is a powerful method that has

presumably opened new avenues of evolutionary thinking, and is influencing the development of evolutionary theory.

Outstanding problems with phylostratigraphy

Phylostratigraphy poses several concerns, both practical and theoretical. First, its definition for the age of a gene and the meaning of homology differ significantly from traditional meanings. The method of phylostratigraphy uses an operational definition of homology based on homology detection algorithms such as BLASTP or PHMMER. These tools have been used in the past to confirm historical homology. However, phyostratigraphy turns this on its head and says that if two genes are not detectable as homologs through these programs, then they are not historical homologs (Figure 1-1). These leads to substantial questions regarding the definition of a gene's age, and what that means in biological terms.

In phylostratigraphy, if a pair of historically homologous genes undergo sufficient evolution, they will be considered entirely new genes due to the limitations of our tools to detect them. This process is sped up in the case of duplication mechanisms (Pegueroles, Laurie, & Alba, 2013). As it currently stands, knowing the relative contributions of divergence, duplication-divergence, *de novo* gene birth, and other mechanisms to such novel sequences is not known. It is therefore unclear what this definition of "gene age" tells us biologically. It also means that the method will sometimes rank genes as much younger than their actual historical time of formation. These theoretical concerns translate to several more concrete technical concerns regarding the efficacy of homology detection programs to reproduce historical homology. Two

kinds of error are important for the efficacy of homology detection programs: false positives and false negatives.

False negatives occur when two proteins which are historical homologs are not detected as homologs by a homology detection program (Figure 1-1). This can occur for a number of reasons, as outlined previously. The degree to which the BLAST algorithm is subject to this kind of error in phylostratigraphy was first investigated by Elhaik and Graur in 2006 (Elhaik, Sabath, & Graur, 2006), who simulated the evolution of the DNA sequence of many genes and then performed homology detection using BLASTN to see if the program could recapitulate the known, simulated historical homology. These researchers noted that, depending on the evolutionary rate of a gene, the BLASTN algorithm might make an extremely high number of false negative errors. This assertion was challenged in 2007 when Alba and Castressana performed a more refined simulation (Albà & Castresana, 2007). They noted that Elhaik and Graur used nucleotide sequences in their simulation even though it was known that homology detection using protein sequences was a more sensitive method. They also noted that Elhaik and Graur, when simulating evolution, allowed all sites to evolve at the same rate. This was an inaccurate feature of the simulation, as it is known that some sites are highly conserved (Masatoshi Nei & Kumar, 2000). The BLAST suite of algorithms relies on these highly-conserved sites to detect homologs (see Chapter 2). Alba and Castresanna therefore performed a more accurate simulation using protein sequences and respecting rate heterogeneity among sites. They found that BLAST error was minimal. However, a major problem for their simulation was that they only inferred evolutionary rates from proteins which had detectable homology out to 450 MYA. If homology detection programs do not always recapitulate historically homologous

relationships, this selection method would choose proteins that tend to evolve very slowly, and which have long blocks of conserved sites which allow detection out to such great distances. This subset of proteins may not be representative of the evolutionary trends of all proteins. Additionally, since they used only 19 genes from which to infer such information, small sample size may have skewed their results.

False positive errors would occur when two genes which do not share historical homology are falsely called as homologs by a homology detection program (Figure 1-1). Because of the vast size of sequence space and the mechanics of sequence evolution, it is expected very few proteins which start from random points in sequence space will wander close enough toward each other to be called as false positives. Nonetheless, it is conceivable that two sequences may strike upon a similar function which happens to select for the same sequence despite a lack of historical homology. While convergent evolution between homologous proteins (Christin, Weinreich, & Besnard, 2010; J Zhang & Kumar, 1997) and non-homologous proteins (Chen, DeVries, & Cheng, 1997) have been noted, it seems unlikely that the full sequence of two non-homologous proteins would converge to largely the same sequence. We can imagine, for example, that if two separate organisms have a duplication occur in a pair of non-homologous enzymes, these enzymes may be recruited to break down the same nutrient available to both species. It could happen, then that in order to break down this nutrient their active site must converge upon a highly similar sequence. So it may be possible that this convergent molecular evolution occurs, though it is unlikely.

In the assessment of phylostratigraphy, we expect that false negatives will play a role in observed associations of traits with age. There are three relevant evolutionary parameters which may reduce the ability of BLAST or similar algorithms (see chapters 2 and 4) to detect historical homology: evolutionary rate, sequence length, and the prevalence of conserved blocks of sites. If a sequence is very short, then even a small number of substitutions may cause a large portion of an alignment with its homologs to be mismatched. Because homology detection programs wish to exclude potential false positive errors and shorter proteins have a smaller sequence space, shorter proteins are less likely to be called as historical homologs. If a protein evolves very quickly, again mismatches in the alignment between true historical homologs will break down more quickly. This will make it more likely that BLAST will not be confident in their homologous relationship, and will thus exclude them as homologs. Finally, the prevalence of conserved blocks of sites plays an important role in the BLAST suite of algorithms (see chapter 2). BLAST relies upon highly-conserved blocks of sites across homologs to establish an initial match. If a given protein does not have many or any blocks which are highly conserved, BLAST will never even consider two of these historically homologous sequences as potential homologs.

These features, by themselves, pose major problems for theory developed via phylostratigraphy. Phylostratigraphic studies have claimed that proteins become longer as they age (A.-R. Carvunis et al., 2012; Wolf et al., 2009), and that their evolutionary rate slows as they age (Albà & Castresana, 2005). These two traits may be explained equally well by homology detection error from the above theoretical considerations. It is therefore important to assess the contribution of homology detection error to these trends. However, this opens the possibility that less obvious phylostratigraphic trends are similarly due to homology detection error. For instance, it is known

that genes which are highly expressed tend to evolve more slowly (Jianzhi Zhang & Yang, 2015). It is therefore possible that trends which show older genes are more highly expressed (A.-R. Carvunis et al., 2012) are due to homology detection error. Similarly, if a mutant gene is associated with a genetic disorder, it is likely that that gene evolves more slowly, as a fast evolutionary rate would produce a greater prevalence of the disorder and therefore be selected against. It is therefore possible that a finding that disease genes tend to be older (Domazet-Lošo & Tautz, 2008) may be due to homology detection error.

Thesis Overview

The above considerations suggest that a critical evaluation of phylostratigraphy in light of homology detection error was necessary. I therefore set out to investigate the contributions of error to phylostratigraphy and theory developed using this method. I develop a framework for evaluating the contribution of homology detection error to phylostratigraphy under various evolutionary contexts. I then apply this framework to various problems in phylostratigraphy, and identify problems with theory as developed by current phylostratigraphic methods. I then develop an error-aware phylostratigraphic framework and use it to identify new trends which are robust to homology detection error. Using this framework, I make initial estimates of the contributions of various gene birth mechanisms to novel gene sequence formation.

Chapter 2 of this thesis, quantifies the amount of error that occurred when using a more representative set of genes than was present in Alba and Castresanna (Albà & Castresana, 2007). I demonstrate that when genes which are less highly conserved are used to perform a simulation, error rates can be non-negligible. I also demonstrate that more realistic modes of evolution are

likely to increase the degree of false negative error. I demonstrate that several previously-reported relationships are also present in simulation alone, where all proteins are equally old. These associations are therefore likely due to homology detection error. Finally, I demonstrate that not all such trends are attributable to homology detection error, and it is therefore important to assess the potential contribution of homology detection error to any phylostratigraphic finding in order to be sure of its reality.

In Chapter 3, I apply these findings to a particular report using phylostratigraphy, the claim that *de novo* gene birth is extremely common. In 2012, Carvunis *et al* claimed that *de novo* gene birth was extremely common (A.-R. Carvunis et al., 2012), and that it in fact contributed more to the formation of new genes than did duplication. If this were true, this would be extremely surprising. Their hypothesis was termed the "proto-gene" hypothesis and was expressed in terms of a model wherein a non-coding sequence became a coding sequence through the formation of an open reading frame. It was expected, due to the frequency of stop codons in the genome, that these novel ORFs would at first be short. It was also expected that they would be fast-evolving and lowly expressed due to an initial lack of function. Authors also asserted that if genes survived, they would increase in length, slow their evolutionary rate, and become more highly expressed, though no clear reason for these expectations were given. Nonetheless, authors used phylostratigraphy to date the age of genes in yeast, and searched for correlations of these properties with age. They found these properties, and others which they claimed supported their model. As previously stated, many of these properties are expected to be associated with gene age on the basis of false negative homology detection errors alone. I therefore set out to assess the contribution of error to these findings by simulating the evolution of the yeast genes in

question. I found that homology detection error alone could fully explain the strength of observed trends. I further pointed out that the qualitative nature of their trends were not fully predicted by nor consistent with their proto-gene model. I concluded that it is currently not clear that *de novo* gene birth is a more common contributor to novel gene formation than is duplication. Supplementary data for this chapter can be found in Appendix A.

In late 2016, my work came under attack by established phylostratigraphy researchers who argued that error was not a major contributor to evolutionary trends identified by phylostratigraphy (Domazet-Lošo et al., 2016). In Chapter 4, I offer a response to the criticisms of our work, re-analyzing their data to account for error. I demonstrate that error contributes significantly and disproportionately to phylostratigraphic trends, and that phylostratigraphy cannot be done in the absence of corrections for homology detection error. Supplementary data for this chapter can be found in Appendix B.

Having established major problems with phylostratigraphic method, I next sought to improve upon the state of the field by searching for a more accurate and biologically meaningful phylostratigraphic method. Additionally, several criticisms of my research were published along with outstanding problems to explore for phylostratigraphy, including the contributions of false positive error. These concerns are the focus of Chapter 5 of this thesis. There are four methods to potentially eliminate the effects of homology detection error: 1) Choose a more accurate homology detection tool or method to reduce the incidence of false negatives and thus their impact on observed trends. 2) Develop a model to identify error-prone genes *a priori* based on their properties to remove them from analysis, thus reducing error and its impact on observed

trends. 3) Restrict phylostratigraphy to closely-related clades, where error may be much lower due to the small amount of divergence time that has occurred, allowing genetic distance to accumulate. 4) When performing phylostratigraphy, restrict to only those genes which can have their error rate determined via simulation, and then remove any error-prone genes from analysis to be confident of observe trends. I performed new simulations using both real genes and genes with simulated sequences and properties to better match the full range of genetic property space. I applied several homology detection algorithms, machine learning models, and evolutionary contexts to the data to identify which of the proposed methods of reducing error were effective. I found that the fourth was the only method in which error and its effects could be largely eliminated. I then performed real phylostratigraphy on the same set of human genes which we had simulated, after removing any error-prone genes. I show that this improved phylostratigraphic method produced trends in complete opposition to previously-published findings. In addition, the data presented in this chapter serve to refute several criticisms and apparent limitations of my previous simulations. I therefore established a more accurate and biologically meaningful phylostratigraphy, as well as established a framework in which error-aware phylostratigraphy must be used in drawing biological conclusions. Supplementary data for this chapter can be found in Appendix C.

In chapter 6, I sought to apply this new framework in a relevant context. I chose to further investigate the relative contributions of divergence, duplication, and *de novo* gene birth to novel sequences. As stated, there are thought to be three major sources for novel sequences—i.e., those sequences which do not have detectable homologs beyond some particular clade. They can arise due to sequence divergence, rapid sequence divergence following a gene duplication, or the

appearance of a *de novo* gene. However, phylostratigraphic analysis alone cannot determine their relative contributions. In my previous studies, I have essentially investigated the contribution of the first of these three mechanisms to the formation of new genes by simulating the evolution of genes and then determining their apparent ages. I assessed the contribution of duplications to novel sequence formation by similarly simulating many models of duplication followed by punctuated and continued models of rapid or modified evolution. These simulations of duplication showed how sequences might be retained and the mechanics of novel sequence formation under various models of duplication. It is arguable that we could assess the contributions of *de novo* gene birth similarly, but the mechanics of *de novo* gene birth are not entirely known. While there is conjecture that they start short and fast-evolving then generally become longer and slow their evolution, the precise mechanics are unclear. Further, even these conjectures cannot be trusted, as they come from non-error-aware phylostratigraphic contexts, and are therefore influenced by homology detection error. I therefore approached the problem indirectly by comparing the number of novel sequences actually observed in phylostratigraphy to the number of novel sequences derived via reasonable simulation of gene evolution with periodic duplications in the genome. I found that even under a relatively extreme model, divergence and duplication-divergence could only account for approximately half of the observed novel sequences. This implied that *de novo* gene birth has been common throughout evolution. Supplementary data for this chapter can be found in Appendix D.

Chapters 2 through 6 demonstrate that homology detection error is a significant confounding factor in phylostratigraphic analyses and creates an error-aware context in which phylostratigraphy can be performed. Chapter 7 assesses remaining outstanding problems in the

field, some of which were uncovered by my analyses. I conclude with several suggested future directions.

References

Abrusán, G. (2013). Integration of new genes into cellular networks, and their structural maturation. *Genetics*, *195*(4), 1407–17. http://doi.org/10.1534/genetics.113.152256

Albà, M. M., & Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular Biology and Evolution*, *22*(3), 598–606. http://doi.org/10.1093/molbev/msi045

Albà, M. M., & Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology*, *7*(53). http://doi.org/10.1186/1471-2148-7-53

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract

Aristotle. (1984). *Complete Works of Aristotle Vol 1*. (J. Barnes, Ed.). Princeton University Press.

Avery, O. T., Macleod, C. M., & McCarty, M. (1944). STUDIES ON THE CHEMICAL INDUCING NATURE TYPES OF THE SUBSTANCE TRANSFORMATION. *Journal of Experimental Medicine*, *79*(2), 137–158.

Banfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W., … Lipovich, L. (2012). Long

noncoding RNAs are rarely translated in two human cell lines. *Genome Research*, *22*, 1646–1657. http://doi.org/10.1101/gr.134767.111.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(November 2012), 36–42. http://doi.org/10.1093/nar/gks1195

Byrne, K. P., & Wolfe, K. H. (2005). The Yeast Gene Order Browser : Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, *15*, 1456–1461. http://doi.org/10.1101/gr.3672305.

Cai, J. J., & Petrov, D. A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biology and Evolution*, *2*, 393–409. http://doi.org/10.1093/gbe/evq019

Carvunis, A., Wang, T., Skola, D., Yu, A., Chen, J., Kreisberg, J. F., & Ideker, T. (2015). Evidence for a common evolutionary rate in metazoan transcriptional networks. *eLife*, *4*, 1–22. http://doi.org/10.7554/eLife.11615

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*, 370–374. http://doi.org/10.1038/nature11184

Chen, L., DeVries, A. L., & Cheng, C.-H. (1997). Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proceedings of the National Academy of Sciences*, *94*(April), 3817–3822.

Christin, P., Weinreich, D. M., & Besnard, G. (2010). Causes and evolutionary significance of genetic convergence. *Trends in Genetics*, *26*(9), 400–405. http://doi.org/10.1016/j.tig.2010.06.005

Conte, L. Lo, Ailey, B., Hubbard, T. J. P., Brenner, S. E., Murzin, A. G., & Chothia, C. (2000). SCOP : a Structural Classification of Proteins database. *Nucleic Acids Research\, 28*(1), 257–259.

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology*, *278*(2), 274–288. http://doi.org/10.1016/j.ydbio.2004.11.028

Darwin, C. (1859). *The Origin of Species*. Wordsworth Editions.

Domazet-Lošo, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics : TIG*, *23*(11), 531–3. http://doi.org/10.1016/j.tig.2007.07.007

Domazet-Lošo, T., & Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular Biology and Evolution*, *25*(12), 2699–707. http://doi.org/10.1093/molbev/msn214

Domazet-Lošo, T., & Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, *468*, 815–8. http://doi.org/10.1038/nature09632

Eichenlaub, M. P., & Ettwiller, L. (2011). De Novo Genesis of Enhancers in Vertebrates. *PLoS Biology*, *9*(11), 1–11. http://doi.org/10.1371/journal.pbio.1001188

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. *Molecular Biology and Evolution*, *23*(1), 1–3. http://doi.org/10.1093/molbev/msj006

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(SUPPL. 2), 29–37.

http://doi.org/10.1093/nar/gkr367

Gehring, W. J. (2002). The genetic control of eye development and its implications for the evolution of the various eye-types. *International Journal of Developmental Biology*, *73*, 65–73.

Grundy, W. N., & Grundy, W. N. (1998). Family-based Homology Detection via Pairwise Sequences Comparison. *Proccedings of the Second Annual International Conference on Computertational Molecular Biology*, 94–100. Retrieved from file:///H|/projekte/literatur/modelling/files/Grundy_1998.pdf

Heather, J. M., & Chain, B. (2016). Genomics The sequence of sequencers : The history of sequencing DNA. *Genomics*, *107*(1), 1–8. http://doi.org/10.1016/j.ygeno.2015.11.003

Hemmrich, G., Khalturin, K., Boehm, A.-M., Puchert, M., Anton-Erxleben, F., Wittlieb, J., … Bosch, T. C. G. (2012). Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity. *Molecular Biology and Evolution*, *29*(11), 3267–80. http://doi.org/10.1093/molbev/mss134

Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K., & Hood, L. (1997). Gene Families : The Taxonomy of Protein Paralogs and Chimeras. *Science*, *278*(5338), 609–614.

Hotopp, J. C. D., Clark, M. E., Oliveira, D. C. S. G., Foster, J. M., Fischer, P., Torres, M. C. M., … Werren, J. H. (2006). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, *327*, 1753–1756.

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., … Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports*, *8*, 1365–1379. http://doi.org/10.1016/j.celrep.2014.07.045

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*, 1161–1166. http://doi.org/10.1126/science.860134

Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, *9*, 605–618. http://doi.org/10.1038/nrg2386

Knowles, D. G., & McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, 1–9. http://doi.org/10.1101/gr.095026.109

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, *11*(5), 473–483. http://doi.org/10.1093/bib/bbq015

Linnaeus, C. (1735). *Systema Naturae* (First). Hes & De Graff Pub B V.

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11073452

Martínez, J. L. (2008). Antibiotics and Antibiotic Resistance. *Science*, *321*(July), 365–368.

Mclysaght, A., & Hurst, L. D. (2016). Open questions in the study of de novo genes: what, how and why. *Nature Reviews Genetics*, *17*(9), 567–578. http://doi.org/10.1038/nrg.2016.78

Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, *14*(117). http://doi.org/10.1186/1471-2164-14-117

Ochman, H., Lawrence, J. G., & Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, *405*, 299–305.

Ohno, S. (1970). *Evolution by gene duplication*. Berlin: Springer-Verlag.

Pál, C., Papp, B., & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, *37*(12), 1372–1375. http://doi.org/10.1038/ng1686

Paley, W. (1802). *Natural Theology*. (M. D. Eddy & D. Knight, Eds.) (First). Oxford University Press.

Pegueroles, C., Laurie, S., & Alba, M. M. (2013). Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology and Evolution*, *30*(8), 1830–1842. http://doi.org/10.1093/molbev/mst083

Peterson, K. J., Cotton, J. A., Gehling, J. G., & Pisani, D. (2008). The Ediacaran emergence of bilaterians : congruence between the genetic and the geological fossil records. *Philosophical Transactions of the Royal Society*, *363*, 1435–1443. http://doi.org/10.1098/rstb.2007.2233

Prat, Y., Fromer, M., Linial, N., & Linial, M. (2009). Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC Evolutionary Biology*, *9*, 285. http://doi.org/10.1186/1471-2148-9-285

Rubin, G. M., & Lewis, E. B. (2000). A Brief History of Drosophila's Contributions to Genome Research. *Science*, *287*(March), 2216–2218.

Salzberg, S. L., White, O., Peterson, J., & Eisen, J. (2001). Microbial Genes in the Human Genome : Lateral Transfer or Gene Loss ? *Science*, *292*(June), 1903–1906.

Sanger, F., Nicklen, S., & Coulson, R. (1977). DNA sequencing with chain-terminating. *Proceedings of the National Academy of Sciences*, *74*(12), 5463–5467.

Sestak, M. S., Bozicevic, V., Bakaric, R., Dunjko, V., & Domazet-Loso, T. (2013). Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Front Zool*, *10*(1), 18. http://doi.org/10.1186/1742-9994-10-18

Smith, T. F., & Waterman, M. (1981). Comparison of Biosequences. *Advances in Applied Mathematics, 2,* 482–489.

Song, M. D., Wachi, M., Doi, M., Ishino, F., & Mutsuhashi, M. (1987). Evolution of an inducible penicillin-target protein in methicillin-resistant Staphylococcus aureus by gene fusion. *Federation of European Biochemical Sciences, 221*(I), 167–171.

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics, 16*(8), 472–482. http://doi.org/10.1038/nrg3962

Vandenbussche, M., Theissen, G., Van de Peer, Y., & Gerats, T. (2003). Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. *Nucleic Acids Research, 31*(15), 4401–4409. http://doi.org/10.1093/nar/gkg642

Watson, J., & Crick, F. (1953). Molecular Structure fo Nucleic Acids. *Nature, 171,* 737–738.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V, & Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America, 106*(18), 7273–80. http://doi.org/10.1073/pnas.0901808106

Zhang, J. (2013). Gene duplication. In J. Losos (Ed.), *The Princeton Guide to Evolution* (pp. 397–405). Princeton, New Jersey.

Zhang, J., & Kumar, S. (1997). Detection of convergent and parallel evolution at the amino acid sequence level. *Molecular Biology and Evolution, 14*(5), 527–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9159930

Zhang, J., & Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics, 16*(7), 409–420. http://doi.org/10.1038/nrg3950

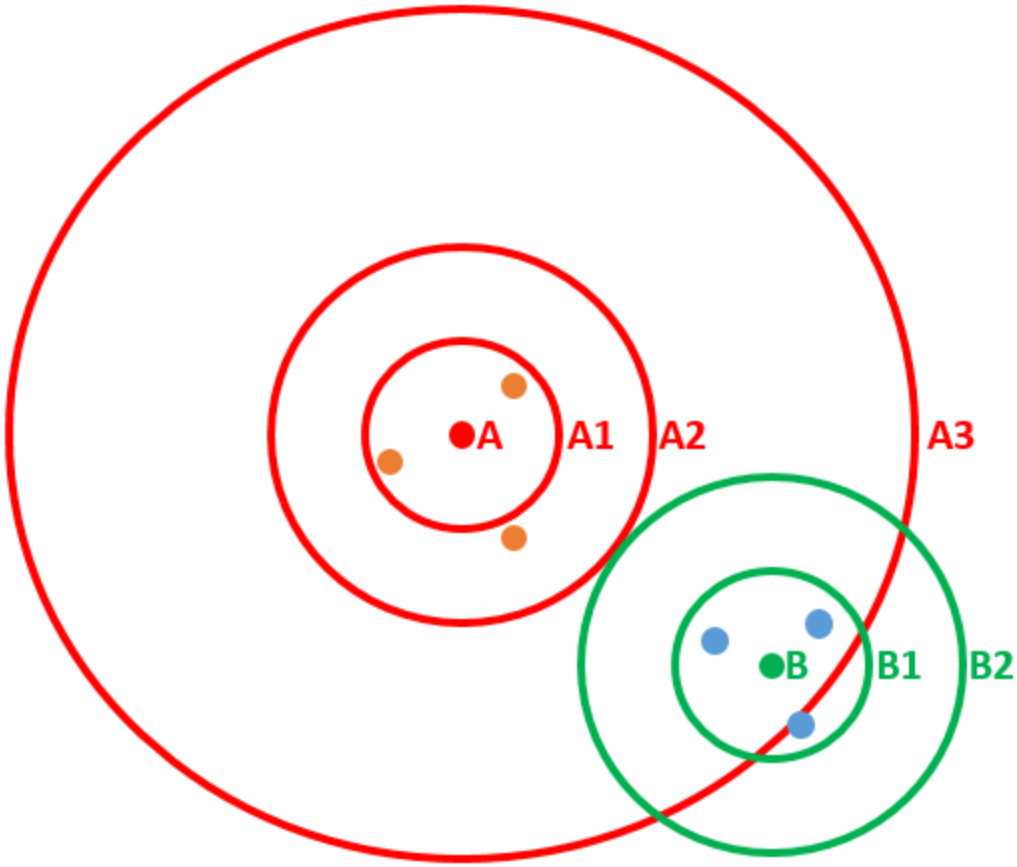

Figure 1- 1 Two-dimensional rendering of sequence space

Consider two proteins query proteins which do not share historical homology, A (red dot) and B (green dot). Consider also sets of proteins in other species which do share historical homology with either protein A (orange dots) or B (blue dots). We infer historical homology based on sequence homology. Sequence homology is granted based on how close to the query protein a given target protein is in sequence space (represented by expanding circles A1-3 and B1-2). If a sequence similarity criterion is too rigid (A1), it will fail to detect some true historical homologs resulting in a false negative error. However, if the criteria are too lax (A3) one will falsely identify some genes as having sequence (and therefore historical) homology when no historical homology is shared, resulting in a false positive error. Additionally, sometimes extending criteria identifies no further homologous proteins (B1 versus B2).

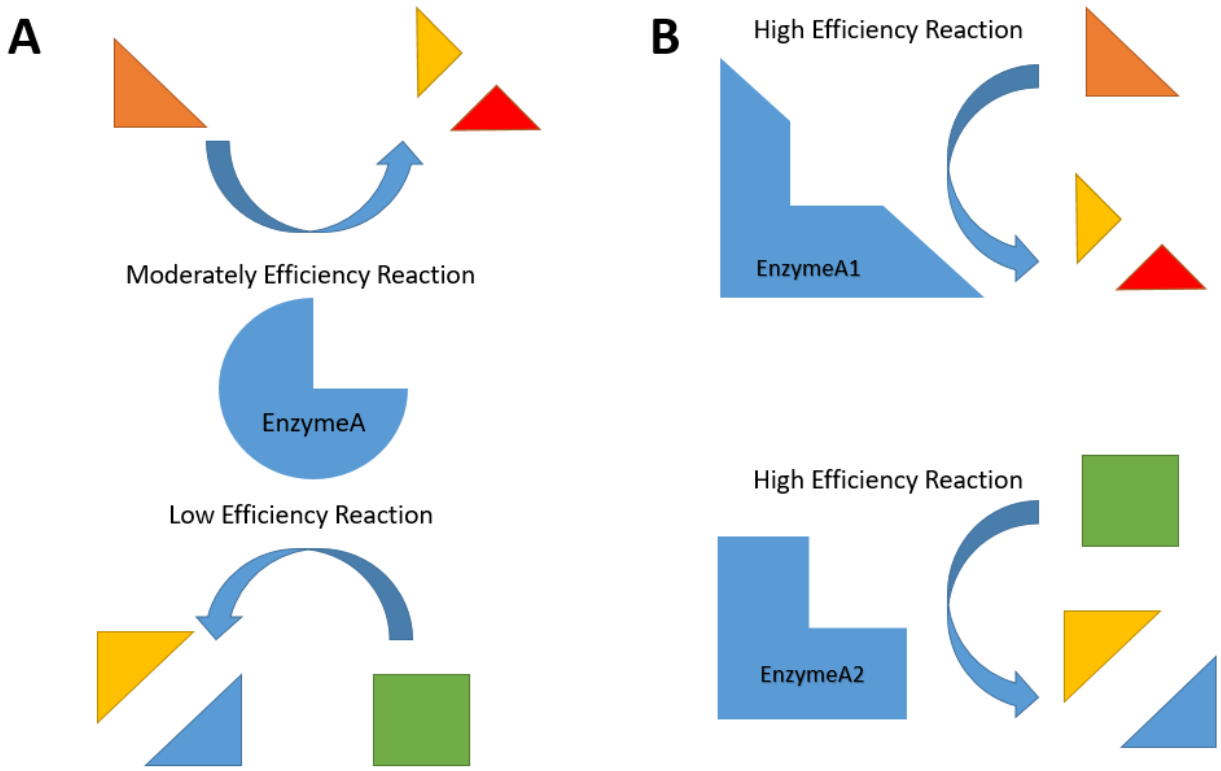

Figure 1- 2 Schematic of Subfunctionalization

(A) Prior to duplication, an enzyme may perform multiple reactions at lower than peak efficiency. Because its active site is restricted by both functions, it cannot specialize for either. (B) After duplication, each copy is able to specialize for one particular reaction without destroying the organism's ability to perform either reaction.

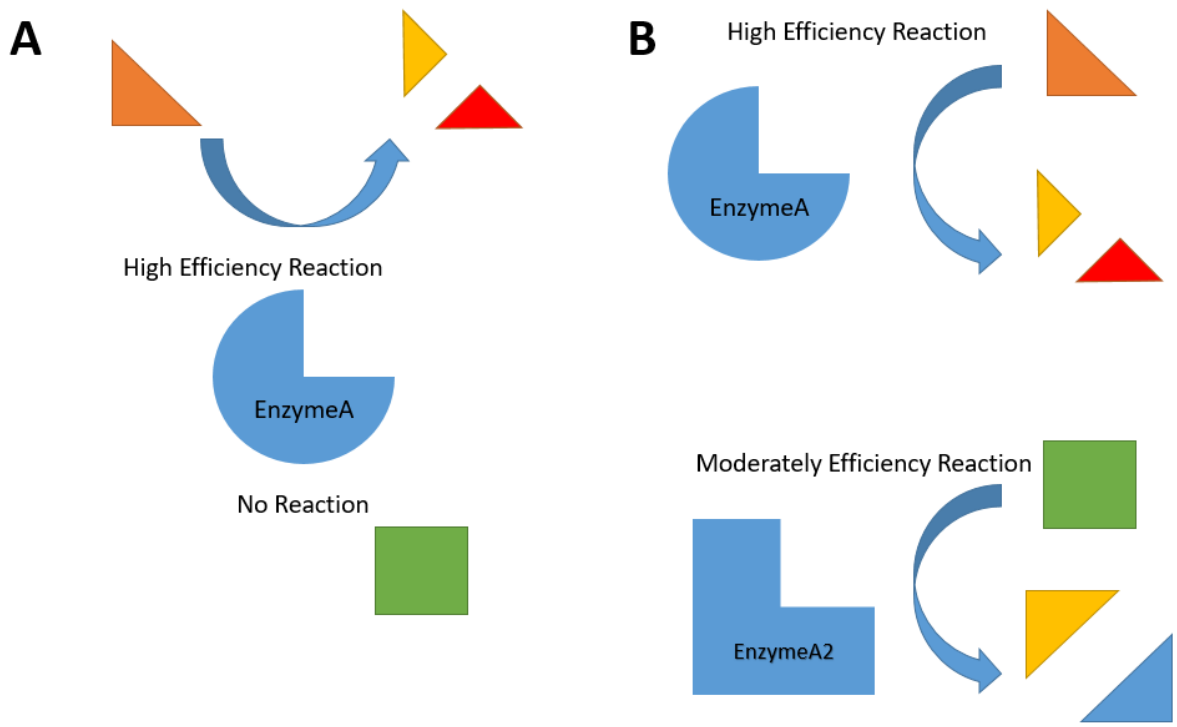

Figure 1- 3 Schematic of Neofunctionalization

(A) Prior to duplication, an enzyme may be restricted to performing one reaction or a specific set of reactions, unable to break down a metabolite present in the environment. (B) After duplication, one of the two copies lacks selective pressure, and may by mutations to its active site gain the ability to bind to and break down a novel metabolite.

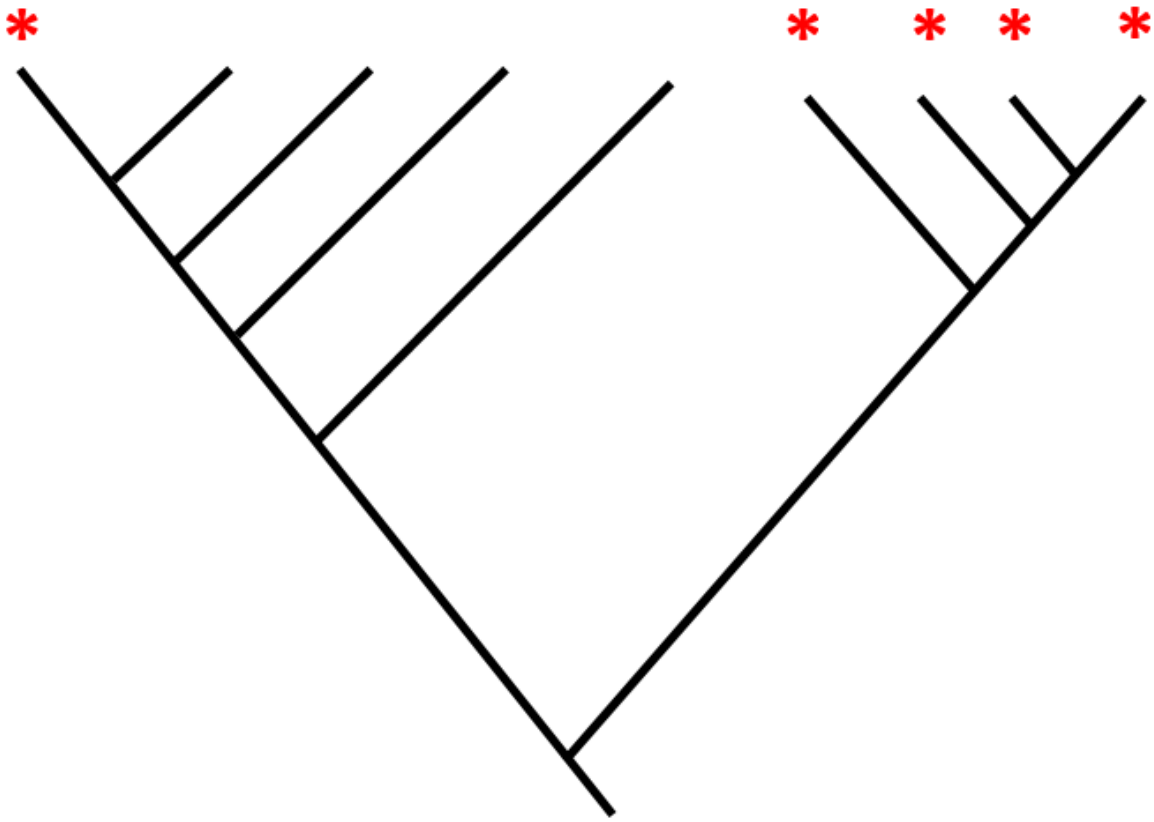

Figure 1- 4 Potential horizontal gene transfer event

Consider a scenario in which a given gene has strong support for having homologs in the five species marked with an asterisk. There are at least two potential explanations for the pattern of homolog conservation. First, the gene was present in the base of the tree. This would require the gene to have been lost independently in four lineages. Second, the gene could have undergone a horizontal gene transfer event from one of the species on the right to the single homolog-bearing species on the left. In some cases, the parsimonious explanation is a horizontal gene transfer

Chapter 2

Phylostratigraphic bias creates spurious patterns of genome evolution

Published as: Moyers, B.A. and Zhang, J (2015) Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.*, 32:258-267.

Abstract

Phylostratigraphy is a method for dating the evolutionary emergence of a gene or gene family by identifying its homologs across the tree of life, typically by using BLAST searches. Applying this method to all genes in a species, or genomic phylostratigraphy, allows investigation of genome-wide patterns in new gene origination at different evolutionary times and thus has been extensively used. However, gene age estimation depends on the challenging task of detecting distant homologs via sequence similarity, which is expected to have differential accuracies for different genes. Here we evaluate the accuracy of phylostratigraphy by realistic computer simulation with parameters estimated from genomic data, and investigate the impact of its error on findings of genome evolution. We show that (1) phylostratigraphy substantially underestimates gene age for a considerable fraction of genes, (2) the error is especially serious when the protein evolves rapidly, is short, and/or its most conserved block of sites is small, (3) these errors create spurious non-uniform distributions of various gene properties among age groups, many of which cannot be predicted a priori. Given the high likelihood that conclusions about gene age are faulty, we advocate the use of realistic simulation to determine if observations

from phylostratigraphy are explainable, at least qualitatively, by a null model of biased measurement, and in all cases, critical evaluation of results.

Introduction

The term phylostratigraphy was first introduced in 2007 to refer to a method of dating the emergence of genes and gene families (Domazet-Lošo et al., 2007). The method actually predates the term, and has been used to approach a large number of questions. For example, phylostratigraphic analyses showed that, compared to relatively old genes, relatively young genes evolve faster (Albà & Castresana, 2005), have lower expressions (J. J. Cai & Petrov, 2010; Wolf et al., 2009), encode shorter proteins (Wolf et al., 2009), are subject to weaker purifying selection and stronger positive selection (J. J. Cai & Petrov, 2010), are less likely to be associated with human disease (Domazet-Lošo & Tautz, 2008), are less frequently expressed during the phylotypic stage in animal embryonic development (Domazet-Lošo & Tautz, 2010), and have different synonymous codon usage (Prat et al., 2009). The method has also been applied to investigate the modes of gene origination (A.-R. Carvunis et al., 2012), the life cycle of genes (Abrusán, 2013), and the evolution of developmental structures and cell types in a variety of taxa (Hemmrich et al., 2012; Martin Sebastijan Sestak, Božičević, Bakarić, Dunjko, & Domazet-Lošo, 2013).

Each phylostratigraphic study has a focal species. The age of a gene from the focal species is defined by the time since the divergence between the focal species and its most distantly related taxon in which a homolog of the gene is found. This exercise requires a method for homolog detection, for which the most common tool by far is Basic Local Alignment Search Tool

(BLAST) (Stephen F Altschul, Gish, Miller, Myers, & Lipman, 1990) and its derivatives (blast.ncbi.nlm.nih.gov/Blast.cgi). We present below a highly simplified overview of the BLAST algorithm for reference (Madden & Morgulis, 2009). BLAST 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 programming algorithm to score the alignment. Missing a true homolog may result in gene age underestimation (if the most distant true homolog is missed) or a false conclusion that a particular lineage has lost a gene (if a homolog is not found in a species but found in a more distant species). Therefore, conclusions based on phylostratigraphic analysis critically rely on the correct identification of homologs by BLAST.

Importantly, BLAST error may vary nonrandomly among genes and create biased results. For instance, because detection of homologs is affected by sequence similarity and because sequence similarity is lost faster for rapidly evolving genes than for slowly evolving genes, the former are expected to have a higher BLAST error rate than the latter, which would create a spurious pattern of faster evolution of younger genes. This possibility was investigated by Elhaik *et al.* (Elhaik et al., 2006) using computer simulation. Specifically, they simulated DNA sequence evolution along an evolutionary tree and used BLAST to search for homologs that were

generated in the simulation. False negative error rates as high as 100% were observed, with quickly-evolving genes having larger errors and hence looking younger.

Elhaik *et al.*'s study, however, was criticized for two reasons (Albà & Castresana, 2007). First, they simulated nucleotide sequence evolution, but amino acid sequences allow for more sensitive detection of distant homologs and are preferred in phylostratigraphy. Second, all sites in a sequence had the same evolutionary rate in the simulation, a major deviation from the general observation in real gene and protein sequences that the evolutionary rate varies among sites, often referred to as "among-site rate heterogeneity" (Jianzhi Zhang & Gu, 1998). The rate heterogeneity is important in homolog detection, because BLAST relies on highly-conserved "words" among homologs. Even very short conserved sequences (e.g., three letters) can greatly enhance BLAST's performance. Because of these two major weaknesses, Elhaik *et al.*'s results were considered unreliable and a new simulation was conducted by Albà and Castresana (2007). These authors estimated the among-site rate heterogeneity of 14 proteins and simulated protein sequence evolution either with or without rate heterogeneity. They reported that gene age was underestimated by BLAST, but the fraction of genes affected is small when the sequences were simulated with rate heterogeneity. They concluded that BLAST error is not an important element in phylostratigraphic analysis. While Albà and Castresana's simulation is more realistic, it also has serious drawbacks. First, their simulation was based on only 14 real genes, which may not be representative. Second and more importantly, the rate heterogeneity patterns were derived from the multiple sequence alignments of either seven vertebrates with a ~450 MY-old common ancestor or nine bilaterians with a ~980 MY-old common ancestor. Thus, their study actually excluded those rapidly evolving genes whose vertebrate or bilaterian homologs are

missed by BLAST. In other words, they studied a biased sample of relatively slowly evolving genes, which would lead to an underestimation of BLAST error.

Because of the widespread use of phylostratigraphy, understanding how BLAST error affects the reliability of phylostratigraphy will have important implications for a diverse array of evolutionary studies. Given the limitations of the previous researches on the subject, we undertake a genome-scale investigation. We simulate the evolution of protein sequences using parameters estimated from the alignments of 6695 orthologous genes found in 12 *Drosophila* species. These species share a most recent common ancestor ~62 MYA (Tamura, Subramanian, & Kumar, 2004), allowing for the study of both slowly-evolving genes and faster-evolving genes than were represented in Albà and Castresana (2007). We simulate evolution across a wide range of divergence times and hence can gauge gene age estimation error with a greater precision than previous studies. We report that BLAST error is abundant and may be responsible for many patterns of genome evolution previously identified in phylostratigraphic studies.

Methods

Simulation of protein sequence evolution

We acquired 6698 protein alignments among the 12 *Drosophila* species from FlyBase (ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments/all_species.guide_tree.longest.translation.tar.gz). The 12 species are *D. simulans, D. sechellia, D. melanogaster, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis,* *D. virilis,* and *D. grimshawi*. We also acquired 5217 protein alignments among 12 mammalian species from Orthomam (Ranwez et al., 2007). The mammalian species were chosen such that

there were 12 species and we retained at least 5000 proteins which had a full alignment. This resulted in selecting species that diverged as much as 92 MYA. The species included were rhesus macaque (*Macaca mulatta*), treeshrew (*Tupaia belangeri*), orangutan (*Pongo pygmaeus*), galago (*Otolemur garnettii*), rat (*Rattus norvegicus*), squirrel (*Ictidomys tridecemlineatus*), marmoset (*Callithrix jacchus*), guinea pig (*Cavia porcellus*), rabbit (*Oryctolagus cuniculus*), gibbon (*Nomascus Leucogenys*), human (*Homo sapiens*), and mouse (*Mus musculus*).

We estimated among-site rate heterogeneity, amino acid frequency, and *D. melanogaster*–*D.grimshawi* or human–mouse genetic distance (i.e., number of substitutions per site) for each protein using TreePuzzle (Schmidt, Strimmer, Vingron, & von Haeseler, 2002). We used the JTT-f matrix (Jones, Taylor, & Thornton, 1992) with the observed amino acid frequencies in the protein and a discrete gamma model with 16 rate categories for parameter estimation. Three alignments were excluded from the *Drosophila* data due to one or more species having only gaps or ambiguous characters for the entire alignment.

We used three evolutionary guide trees. The first tree (Figure 2-1A) was constructed according to the divergence times estimated in TimeTree (Hedges, Dudley, & Kumar, 2006). For each species, we used the mean estimate of divergence time from *D. melanogaster*, with the following exceptions. Nematode and sponge average divergence times were swapped, because they had very wide margins on their estimates and the average divergence times would misplace them compared to the known phylogeny. INT1 and INT2 were entirely fictional, providing a smoother range of divergence times for a more informative analysis. The second guide tree (Figure 2-3A) was constructed according to the divergence times provided by Domazet-Lošo and

Tautz (2007). The third guide tree (Figure 2-4) was constructed using TimeTree divergence time estimates for a phylogeny provided by Domazet-Lošo and Tautz (2008).

Once the above information was acquired, we simulated sequence evolution using ROSE (Stoye, Evers, & Meyer, 1998), which allows the evolutionary rate for each site to be specified by the user. Additionally, following Albà and Castresana (2007), we set an insertion and deletion (indel) threshold to 0.0001. For each branch in the simulation, the expected number of insertion attempts and the expected number of deletion attempts both equal the expected number of amino acid substitutions for that branch times 0.0001. A random location along the protein is chosen to place an indel. If the amino acid substitution rate at the random location is greater than the average substitution rate for the protein, the indel occurs; otherwise, the indel does not occur. A proposed indel length between 1 and 14 amino acids is decided based on a predetermined probability function. In our simulation, the probability was set at 0.1 for any length between 1 and 6 amino acids and 0.05 for any length between 7 and 14 amino acids. In the case of a deletion, only those sites with amino acid substitution rates higher than the average for the protein will be deleted, with the occurrence of a site with a lower-than-average rate truncating the deletion. In the case of an insertion, all new sites are set to have amino acid substitution rates equal to the average substitution rate of the protein. For each protein, we simulated its evolution using a JTT-f matrix with observed amino acid frequencies from the alignment. We calculated the mean evolutionary rate of a protein by the number of substitutions per site per MY between *D. melanogaster* and *D. grimshawi* or between human and mouse. Based on TimeTree, the former pair of species diverged 62 MYA and the latter 92 MYA. The sequence provided as the

start sequence for evolution was the *D. melanogaster* sequence or human sequence. The simulation of sequence evolution was performed 10 times for each protein.

Covarion model of sequence evolution

Under the covarion model, we simulated sequence evolution in 50 MY chunks. After each 50 MY iteration, we selected a subset of sites accounting for y = 0%, 1%, 2%, or 5% of the protein length, and shuffled their evolutionary rates. We then continued evolution along that lineage for another 50 MY and repeated until the entire lineage had been evolved. In cases where we were required to evolve for $x < 50$ MY, (xy/50)% of sites were shuffled in their evolutionary rates. We also ran simulations in which we excluded the most conserved one or two rate categories from being shuffled. In these constrained covarion models, at each 50 MY iteration, we selected 0%, 1%, 2%, or 5% of sites such that no sites from the most conserved one or two rate categories were selected but the appropriate percentage of the full protein length was selected and shuffled. Evolution was continued according to this pattern until the entire lineage had been evolved.

Detection of homologs using BLASTP

We downloaded BLASTP (version 2.2.28+) from NCBI. For each run, we took the simulation-generated fruit fly (or human) database consisting of 6695 (or 5217) protein sequences and performed BLASTP searches against the simulation-generated sequence database from each of the other 11 species for that run. We used an E-value cutoff of 1E-3 unless otherwise mentioned. Results of true homologs found were stored. We then dated each gene to the common ancestor

of the query species and all taxa in which true positive hits were found. This represented the "age" of the protein for that run.

Analysis of BLASTP results: rate of new gene origination

We divided the average number of new gene originations in a tree branch over 10 simulations by the evolutionary time represented by the branch. This is not identical to the method used by Domazet-Lošo and Tautz (2007), who corrected for paralogs. But, because our study did not involve gene duplication, we did not perform this correction.

Analysis of BLASTP results: human disease genes

We downloaded the MORBIDMap (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005), and restricted the data to only those genes marked with "[3]" (mutation was positioned by mapping the wild-type gene and the mutation is associated with the disorder). We then determined which genes in each age group were disease genes and plotted the percentage of such genes against phylostratum. We further used Spearman's rank correlation to determine if there was a significant correlation between the inferred age of a gene and its status as disease gene.

Results

Characterizing gene age estimation errors

We acquired from FlyBase (St Pierre, Ponting, Stefancsik, & McQuilton, 2014) 6695 orthologous protein alignments from 12 *Drosophila* species that diverged ~62 MYA (Tamura et al. 2004). For each protein, we used TreePuzzle (Schmidt et al., 2002) to classify all sites into 16

equal-sized rate bins according to a discrete gamma model of among-site rate heterogeneity and estimated the relative rates of the 16 bins. We also inferred the mean absolute evolutionary rate across all sites of a protein by dividing the number of substitutions per site in the protein between *D. melanogaster* and *D. grimshawi* by 2×62 MY (Tamura et al., 2004). Using all of these parameters, we simulated the evolution of 6695 proteins using ROSE (Stoye et al., 1998) along a tree with 11 taxa, representing species from fruit fly to bacteria (Figure 2-1A). The divergence times among these taxa were assumed to equal what TimeTree (Hedges et al., 2006) estimated (see Materials and Methods). Using the extant sequences generated from the simulation, we constructed protein databases and used BLASTP, a derivative of BLAST for searching protein homologs, to detect orthologs of the simulated fruit fly queries in the other 10 extant taxa. Unless necessary for distinction, BLASTP is simply referred to as BLAST in this chapter. Because in the simulation all genes originated in the common ancestor of eukaryotes and bacteria, any inferred gene age other than that was considered an estimation error. Following Albà and Castressana (2007), we repeated this simulation 10 times to examine the stochasticity of the obtained results. Unless otherwise noted, the averages from the 10 simulations were presented.

BLAST searches require specifying an E-value cutoff to guard against false positives. Because it was suggested that the E-value cutoff of 1E-3 be used in phylostratigraphy (Domazet-Lošo & Tautz, 2003), we used this cutoff in our simulation unless otherwise noted. We found from our simulation that in 13.85% of cases a homolog was not detected in the most distant taxa (Table 2-1). This indicates that age estimation error is a relatively common phenomenon. We also found

that in 2.77% of cases no homolog was found in any taxon (Table 2-1), indicating that age underestimation can be extreme.

To examine the frequency of gene age underestimation under different E-value cutoffs, we tried cutoffs from 1E-1 to 1E-10. Because we are examining false negative errors, the error rate should increase as the E-value cutoff becomes smaller. This is indeed the case, although the variation in error rate under different cutoffs is relatively small (Table 2-1).

It might be justifiably argued that in real phylostratigraphy there can be numerous potential orthologs that correspond to a particular divergence time (e.g., many bacteria rather than one), which may improve age estimation. In order to examine the error rate under this scenario, we performed an additional database search using the simulated bacterial protein as the query and the simulated proteins for all other taxa as the database, providing 10 representatives of the "most distant homolog". We found that in 12.03% of cases, no homologs were found (under the E-value cutoff of 1E-3). Thus, the use of multiple species for a given divergence has virtually no impact on the error rate.

While it is expected that more distant homologs are more difficult to detect, the exact relationship between divergence time and mean detectability for a group of genes has not been examined. Using the simulated data, we plotted the fraction of fruit fly genes whose homologs are not detected in a taxon as a function of the time since the separation between that taxon and fruit fly (Figure 2-1B). Although the probability of missing a homolog by BLAST clearly increases with the divergence time, the relationship is decidedly non-linear ($F = 333.5$, $P = 7.1\times10^{-7}$, Ramsey RESET test, (Ramsey, 1969)). Rather, it can be approximated by a log-linear

curve (Figure 2-1B), with a faster increase in error rates for shorter divergence times and a slower increase for longer divergence times.

Properties of genes that influence its age underestimation

We sought to determine which properties of a gene influence its age underestimation by BLAST. Due to the way the BLAST algorithm works, two likely candidates are the rate of protein sequence evolution and the length of the protein. Indeed, we found highly significant correlations between the inferred gene age and both rate (Spearman's $\rho = -0.57$, $P < 2.2\times10^{-308}$; Figure 2-2A) and protein length ($\rho = 0.19$, $P < 1.1\times10^{-53}$; Figure 2-2B). Both of these associations have been noted before in real phylostratigraphic studies (J. J. Cai & Petrov, 2010; Wolf et al., 2009), but are replicated by our simulation where all genes are equally old. Hence, the trends previously observed in phylostratigraphic analyses may be entirely due to BLAST errors. We further reasoned that, because of the requirement for high-scoring matches of "words" in BLAST searches, longer stretches of conserved blocks would result in fewer BLAST errors. Indeed, we find the error rate to increase quickly as the maximum length of the stretch of the most conserved category of sites decreases, especially when the mean evolutionary rate is high (Figure 2-2C).

To examine if the above three protein characteristics (mean evolutionary rate, protein length, and maximum length of the stretch of the most conserved category of sites) have independent contributions to gene age underestimation, we conducted a partial correlation between each of these characteristics and the inferred gene age, after controlling the other two characteristics. Significant partial correlations were found for evolutionary rate ($\rho = -0.32$, $P < 1.3\times10^{-171}$),

protein length ($\rho = 0.11$, $P < 5.5\times10^{-19}$), and maximum length of the stretch of the most conserved category of sites ($\rho = 0.21$, $P = 4.2\times10^{-68}$), demonstrating that these factors have independent influences on gene age underestimation.

The above simulation assumed that a site has a constant evolutionary rate throughout the tree, which may not be true in reality because of potential evolutionary alterations in the functional constraint of the site due to either protein functional changes (Jianzhi Zhang, 2006) or epistasis (Breen, Kemena, Vlasov, Notredame, & Kondrashov, 2012). To examine the level of gene age underestimation under this scenario, we simulated a covarion model of sequence evolution (Fitch, 1971; Penny, McComish, Charleston, & Hendy, 2001) along the tree in Fig. 1A. To implement this model, at certain evolutionary times, we randomly picked a subset of sites and shuffled their rate categories. This was done for a total of 1%, 2%, or 5% of sites every 50 MY of evolution. As a negative control, 0% of sites were shuffled in rate categories. We then attempted to detect the bacterial homologs of fruit fly proteins. We found that the covarion evolution substantially increases the BLAST error rate. When 5% of sites are shuffled in their evolutionary rates per 50 MY, more than 67% of bacterial homologs could not be detected, compared to 14% when no site is shuffled (Tabel 2-2). Even a tiny amount of covarion evolution (1% per 50 MY) increases the probability of gene age underestimation by more than a factor of 0.25 (Table 2-2). Considering that functionally most critical residues in a protein may be largely immune to covarion evolution, we conducted an additional simulation shuffling 0%, 1%, 2%, or 5% of sites every 50 MY, but excluding the sites belonging to the lowest one or two rate categories from being picked for rate shuffling. Our result showed only a small increase in age estimation error by these constrained covarion models, compared with no rate shuffling (Table 2-

2). The reality is probably somewhere between the full convarion model and the constrained covarion models, although the fraction of sites subject to covarion evolution and the frequency of rate changes are currently unknown.

Gene age underestimation generates spurious patterns of genome evolution

Because phylostratigraphy by homology detection underestimates gene age and because the probability and extent of the underestimation vary among genes, it is possible for phylostratigraphic errors to create spurious patterns of genome evolution. As demonstrated in our simulation (Figure 2-2), that young genes tend to evolve rapidly (Albà & Castresana, 2005) and encode short proteins (Wolf et al., 2009) is explainable by gene age estimation error. While one can predict *a priori*, based on how BLAST works, that these correlations are likely artifacts, whether many other phylostraigraphy-based discoveries are genuine or artifactual cannot be easily predicted. Below we two additional phylostraigraphy-based discoveries and examined whether they could have resulted from gene age underestimations.

We first examined two genomic patterns reported in Domazet-Lošo and Tautz (2007), a paper of special importance to the phylostratigraphy field because the term phylostratigraphy was coined in this paper. Using *D. melanogaster* as the focal species, these authors reported a peak in the number of new gene originations per MY in the common ancestor of bilatera. Because these authors used a phylogeny that is different from the one used in our main simulation, we conducted another simulation using their tree (Figure 2-3A).

While all genes were simulated to have originated in the common ancestor of all cellular life, 17% were inferred by phylostratigraphy to have originated more recently. More disturbingly, the inferred number of new gene originations per MY is not uniform throughout evolution (X^2 = 46.38, P =5.1×10^{-7}, chi-squared test), creating an intriguing pattern of rapid new gene origination at certain evolutionary times and slow new gene origination at other times (Figure 2-3B). Nevertheless, we did not observe in our simulation the peak of gene origination in the common ancestor of bilateria as reported by Domazet-Lošo and Tautz (2007). Inaccuracies in tree topology and divergence times may account for the disparity between our simulation result and what was discovered by Domazet-Lošo and Tautz, given that the divergence times surrounding the ancestral node of the common ancestor of bilateria are relatively short (Figure 2-3A).

All of the above simulations and analyses used *D. melanogaster* as the focal species. It would be important to examine if our findings apply to other species. To this end, we used simulation to examine a result from Domazet-Lošo and Tautz (2008). These authors reported that disease genes tend to be older, and found a remarkable dearth of disease genes in the youngest group of genes. We conducted a simulation according to the species relationships considered in their paper and constructed this tree using divergence time estimates from TimeTree (Figure 2-4). Using human as the focal species, we acquired orthologous proteins from Orthomam (Ranwez et al., 2007) using taxa diverged as much as 92 MY from human. We inferred evolutionary rate and rate heterogeneity using TreePuzzle, evolved sequences using ROSE, and detected homologs using BLASTP. From the simulated data, we observed a positive correlation between the inferred age of a gene and its probability of being a disease gene (Spearman's $\rho = 0.623$, P = 0.004; Figure 2-4). Because the true ages of all genes are the same in our simulation, our finding

demonstrates that Domazet-Lošo and Tautz's finding was at least partly an artifact of gene age estimation error.

Discussion

Homology detection programs make a major common assumption. If two sequences are similar enough on some measure, they are homologs—they share a common ancestry. The researcher has freedom in deciding where the similarity cutoff should be. This does not imply the converse assumption—that is, if sequences are not similar then they do not share a common ancestry. However, in phylostratigraphy this second assumption is made, because genes are grouped and analyzed based on their detected homologs. It is thus critical to understand the amount of type-II error (i.e., false negatives) in homology detection used for phylostratigraphic analyses.

We have systematically quantified the bias and effects of false negative errors of BLAST homolog detection on gene age estimation. Under our model of sequence evolution, BLAST results in common errors in gene age underestimation, some of which are extreme. For four reasons, our results are likely to be conservative. First, our simulation used parameters estimated from proteins that can be detected from all 12 *Drosophila* genomes. There are proteins that cannot be detected from all 12 *Drosophila* genomes (Palmieri, Kosiol, & Schlötterer, 2014). Apart from the true gene loss or new gene origination, some of them may actually exist in all 12 genomes but are undetectable due to the limited power of homology detection. Not including such genes in our simulation reduces the apparent error rate of BLAST. Second, we estimated protein evolutionary rate per MY by comparing two *Drosophila* species and assumed that this rate applies to other organisms including fungi and bacteria. Because mutation rate tends to be

constant per cell division (Michael Lynch, 2010) and the average (germline) cell cycle tends to be shorter in smaller organisms, mutation rate per year is expected to be much higher in smaller organisms such as bacteria than in *Drosophila*. In other words, we underestimated the amount of BLAST error for a protein by assuming a constant evolutionary rate per MY across the tree of life. Third, our main simulation assumed that the evolutionary rate of a site relative to the average of all sites in a protein is a constant. When this assumption is violated, BLAST error tends to increase, as shown in our simulation of the covarion evolution. Fourth, our simulation parameters were estimated from one-to-one orthologous proteins and the simulation considered neither gene duplication nor gene loss. In reality, gene duplication is quite common in genome evolution (K. Wolfe, 2004; Jianzhi Zhang, 2003) and it often results in a change in evolutionary rate associated with post-duplication changes in gene function (Pegueroles et al., 2013; J Zhang, Rosenberg, & Nei, 1998). This rate change will likely increase the BLAST error rate. Gene loss can further compromise gene age estimation if a gene loss occurs to the most distant taxa where the homolog would otherwise be detected. Taken together, it is most likely that the actual frequency of gene age estimation error by BLAST is greater than what is shown in this study.

There also exists the possibility of overestimation of gene age, especially in the context of horizontal gene transfer. Imagine a gene that originated recently in bacteria but was horizontally transferred to some eukaryotes. Phylostraigraphy could mistakenly date the gene to the common ancestor of eukaryotes and bacteria. In future research, it would be important to explore the impacts of increasingly accurate and complex models of sequence and genomic evolution mentioned above on gene age estimation.

By itself, the high error rate should encourage skepticism toward the statement that any gene is of a particular age. We find, however, that this error is associated with the mean evolutionary rate of the protein, protein length, and the maximum length of the most conserved stretch of sites. Thus, one may be able to temper this skepticism by further analyses (e.g., by controlling the confounding factors). However, additional research will be needed to determine if these qualities can be parsed away from the effects of true gene age.

We demonstrated in some cases that the gene age estimation error can result in statistically highly significant and biologically intriguing findings without any true biological meaning or, at the very least, with misinterpreted biological meaning. Some of these spurious patterns may be predicted *a priori* given our understanding of how BLAST works and the correlates of factors that most seriously impact the performance of BLAST. For instance, given that fast protein sequence evolution leads to gene age underestimation and that lowly expressed genes tend to evolve rapidly (Pal, Papp, & Hurst, 2001), one could predict that phylostratigraphic bias would create a positive correlation between gene expression level and age. Thus, the report that young genes tend to be lowly expressed (Wolf et al., 2009) may be entirely artifactual. Because gene expression level is correlated with codon usage bias, phylostratigraphic bias would also lead to the observation that genes with different ages have different codon usage (Prat et al., 2009). Similarly, because the evolutionary rate of a protein is negatively correlated with the strength of purifying selection and positively correlated with the strength of positive selection acting on the protein, the discovery that, compared to old genes, young genes are subject to weaker purifying selection and stronger positive selection (J. J. Cai & Petrov, 2010) can be artifactual. However, not all patterns created by phylostratigraphic bias can be predicted *a priori*, such as the apparent

statistically-significant peak we observed in gene fixation, or the apparent ancient origin of disease-associated genes. It is therefore crucial to consider phylostratigraphic error as the first possible cause of any nonrandom pattern observed in phylostratigraphic studies. Further, many phylostratigraphic studies did not start with clear hypotheses, but attempted to explain whatever patterns that were observed in such studies. The danger of offering post hoc explanations has been eloquently discussed in the context of gene ontology analysis (Pavlidis, Jensen, Stephan, & Stamatakis, 2012) and applies to phylostratigraphy.

Nevertheless, we do not imply that all phylostratigraphic results are artifacts. In fact, most of our simulations do not exactly recapitulate empirical findings, although one cannot exclude the possibility that the disparity is due to the use of inaccurate parameters (e.g., divergence times between taxa) and/or simplified models (e.g., constant evolutionary rate for a site) in the simulations. Some of the disparities are so large that it is highly probable that true biological signals exist. For instance, the age distribution of *D. melanogaster* genes in real phylostratigraphic analysis shows a peak for very young genes, but the corresponding distribution based on the simulated data does not have this peak (Figure 2-5). Because it is improbable for BLAST to miss the honeybee homolog of a *Drosophila* gene if the homolog truly exists, the most likely cause of the disparity is an unusually high rate of new gene origination in *Drosophila* after its separation from the honeybee. Furthermore, because the BLAST error rate increases with (real) gene age (Figure 2-1B), the overall error will be smaller than what is shown here if a large fraction of genes in a genome are younger than what was assumed in our simulation. But, due to the BLAST error, it is difficult to know the true gene age and hence difficult to assess the likelihood of this scenario.

In order to analyze the effects of phylostratigraphic error on any particular data set, one must assess the probability that a given gene has been subject to BLAST error. This is most easily determined by a simulation of protein evolution, but simulation has its own limitations. For instance, it requires at least the knowledge of the protein's rate of evolution and rate heterogeneity, typically inferred from the multiple sequence alignment of homologs. But this begs the question, as the purpose of BLAST is to identify these homologs. One could attempt to estimate rate heterogeneity of genes by using homologs detectable by BLAST, but this may produce biased estimates. Furthermore, due to the limited understanding of the evolutionary models of individual proteins, investigators tend to assume relatively simple models, which can result in biased parameter estimation and unreliable simulations (J Zhang, 1999). Additionally, in the case of true orphan genes, these homologs do not even exist in principle, independent of our ability to find them. More studies are needed to design methods that differentiate true biological signals from artifacts in phylostratigraphic analysis.

We must also note that we studied only false negative errors in homolog search. In real phylostratigraphic analysis, the only indicator for gene age classification is how far out a hit is found. This method does not and cannot differentiate between the hit of a true homolog and a false one. In our analysis we were not able to assess the degree of false positive errors. This is because the starting point for our protein evolution included a number of paralogous proteins, for which we would expect to find BLAST hits. We did not bypass this problem by using random sequences, because these sequences might not represent real functional constraints and cannot

represent convergent sequence evolution that may happen in nature (J Zhang & Kumar, 1997). We see this as an open problem in future research.

Our analysis focused on BLAST, because this is the method that has been used in the vast majority of phylostratigraphic studies. Future studies should explore whether other methods such as HMMer (Finn et al., 2011) and PSI-BLAST (S F Altschul et al., 1997) perform better than BLAST for gene age estimation.

References

Abrusán, G. (2013). Integration of new genes into cellular networks, and their structural maturation. *Genetics*, *195*(4), 1407–17. http://doi.org/10.1534/genetics.113.152256

Albà, M. M., & Castresana, J. (2005). Inverse relationship between evolutionary rate and age of mammalian genes. *Molecular Biology and Evolution*, *22*(3), 598–606. http://doi.org/10.1093/molbev/msi045

Albà, M. M., & Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology*, *7*(53). http://doi.org/10.1186/1471-2148-7-53

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Jounal of Molecular Biology*, *215*(3), 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rend

ertype=abstract

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., & Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature*, *490*, 535–8. http://doi.org/10.1038/nature11510

Cai, J. J., & Petrov, D. A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biology and Evolution*, *2*, 393–409. http://doi.org/10.1093/gbe/evq019

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*, 370–374. http://doi.org/10.1038/nature11184

Domazet-Lošo, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics : TIG*, *23*(11), 531–3. http://doi.org/10.1016/j.tig.2007.07.007

Domazet-Lošo, T., & Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome Research*, *13*, 2213–9. http://doi.org/10.1101/gr.1311003

Domazet-Lošo, T., & Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular Biology and Evolution*, *25*(12), 2699–707. http://doi.org/10.1093/molbev/msn214

Domazet-Lošo, T., & Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, *468*, 815–8. http://doi.org/10.1038/nature09632

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. *Molecular Biology and Evolution*, *23*(1), 1–3. http://doi.org/10.1093/molbev/msj006

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(SUPPL. 2), 29–37. http://doi.org/10.1093/nar/gkr367

Fitch, W. M. (1971). Rate of change of concomitantly variable codons. *Journal of Molecular Evolution*, *1*, 84–96. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4377447

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, *33*, D514-7. http://doi.org/10.1093/nar/gki033

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, *22*(23), 2971–2. http://doi.org/10.1093/bioinformatics/btl505

Hemmrich, G., Khalturin, K., Boehm, A.-M., Puchert, M., Anton-Erxleben, F., Wittlieb, J., … Bosch, T. C. G. (2012). Molecular signatures of the three stem cell lineages in hydra and the emergence of stem cell function at the base of multicellularity. *Molecular Biology and Evolution*, *29*(11), 3267–80. http://doi.org/10.1093/molbev/mss134

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275–282. http://doi.org/10.1093/bioinformatics/8.3.275

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics*, *26*(8), 345–52. http://doi.org/10.1016/j.tig.2010.05.003

Madden, T., & Morgulis, A. (2009). BLAST Command Line Applications User Manual, *1997*,

1–42.

Pal, C., Papp, B., & Hurst, L. D. (2001). Highly Expressed Genes in Yeast Evolve Slowly. *Genetics*, *158*, 927–931.

Palmieri, N., Kosiol, C., & Schlötterer, C. (2014). The life cycle of Drosophila orphan genes. *eLife*, *3*, e01311. http://doi.org/10.7554/eLife.01311

Pavlidis, P., Jensen, J. D., Stephan, W., & Stamatakis, A. (2012). A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, *29*(10), 3237–48. http://doi.org/10.1093/molbev/mss136

Pegueroles, C., Laurie, S., & Alba, M. M. (2013). Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology and Evolution*, *30*(8), 1830–1842. http://doi.org/10.1093/molbev/mst083

Penny, D., McComish, B. J., Charleston, M. A., & Hendy, M. D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution*, *53*(6), 711–23. http://doi.org/10.1007/s002390010258

Prat, Y., Fromer, M., Linial, N., & Linial, M. (2009). Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC Evolutionary Biology*, *9*, 285. http://doi.org/10.1186/1471-2148-9-285

Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least-squares Regression Analysis. *Journal of the Royal Statistical Society*, *31*(2), 350–371.

Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.-K., & Douzery, E. J. (2007). OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evolutionary Biology*, *7*(1), 241. http://doi.org/10.1186/1471-2148-7-241

Schmidt, H. A., Strimmer, K., Vingron, M., & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, *18*(3), 502–504. http://doi.org/10.1093/bioinformatics/18.3.502

Sestak, M. S., Božičević, V., Bakarić, R., Dunjko, V., & Domazet-Lošo, T. (2013). Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Frontiers in Zoology*, *10*(1), 18. http://doi.org/10.1186/1742-9994-10-18

St Pierre, S. E., Ponting, L., Stefancsik, R., & McQuilton, P. (2014). FlyBase 102--advanced approaches to interrogating FlyBase. *Nucleic Acids Research*, *42*(Database issue), D780-8. http://doi.org/10.1093/nar/gkt1092

Stoye, J., Evers, D., & Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics*, *14*(2), 157–163.

Tamura, K., Subramanian, S., & Kumar, S. (2004). Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. *Molecular Biology and Evolution*, *21*(1), 36–44. http://doi.org/10.1093/molbev/msg236

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V, & Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(18), 7273–80. http://doi.org/10.1073/pnas.0901808106

Wolfe, K. (2004). Evolutionary genomics: yeasts accelerate beyond BLAST. *Current Biology*, *14*(10), R392-4. http://doi.org/10.1016/j.cub.2004.05.015

Zhang, J. (1999). Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Molecular Biology and Evolution*, *16*(6), 868–75. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10368963

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*(6), 292–298. http://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, J. (2006). Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nature Genetics*, *38*(7), 819–23. http://doi.org/10.1038/ng1812

Zhang, J., & Gu, X. (1998). Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics*, *149*(3), 1615–25.

Zhang, J., & Kumar, S. (1997). Detection of convergent and parallel evolution at the amino acid sequence level. *Molecular Biology and Evolution*, *14*(5), 527–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9159930

Zhang, J., Rosenberg, H. F., & Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(7), 3708–13.

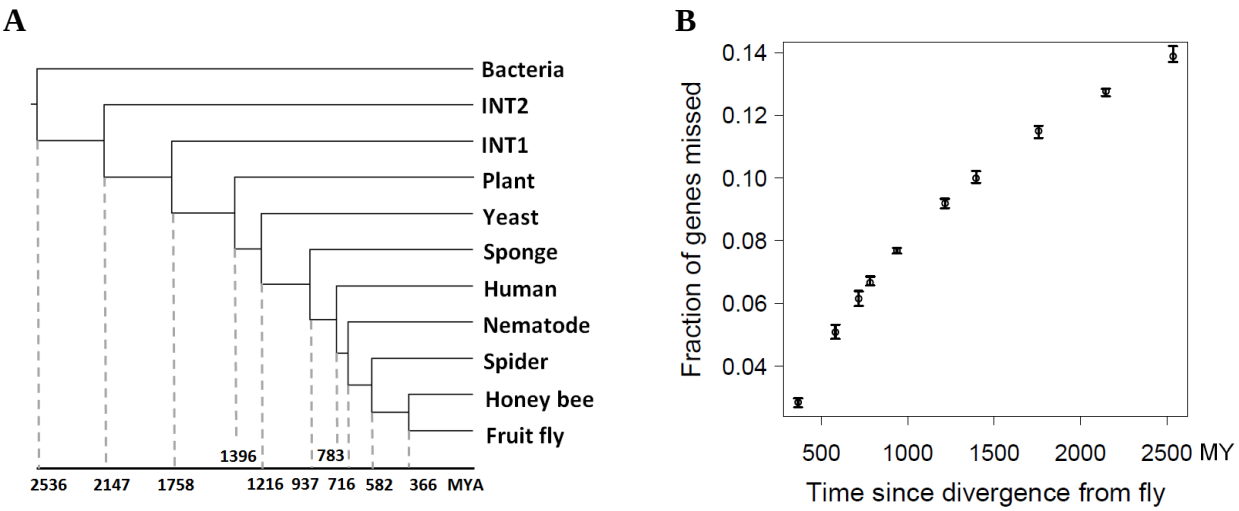

Figure 2- 1 BLAST error rates at different divergence times

(A) Phylogeny showing the relationship of simulated sequences in this study. Organism names are for reference only. Branch lengths are proportional to divergence times, the sources of which are detailed in Materials and Methods. INT1 and INT2 are not true taxa, but are equally spaced between plant and bacterial divergence to allow a smoother range of distances. (B) Fraction of proteins from a taxon that are missed by BLAST increases nonlinearly with the time since the divergence between the taxon and the query taxon (fruit fly). This function is most likely log-linear (deltaAIC=-23.87 compared to the linear model). Shown are the averages from 10 simulations, with the error bars depicting the range from the 10 simulations.

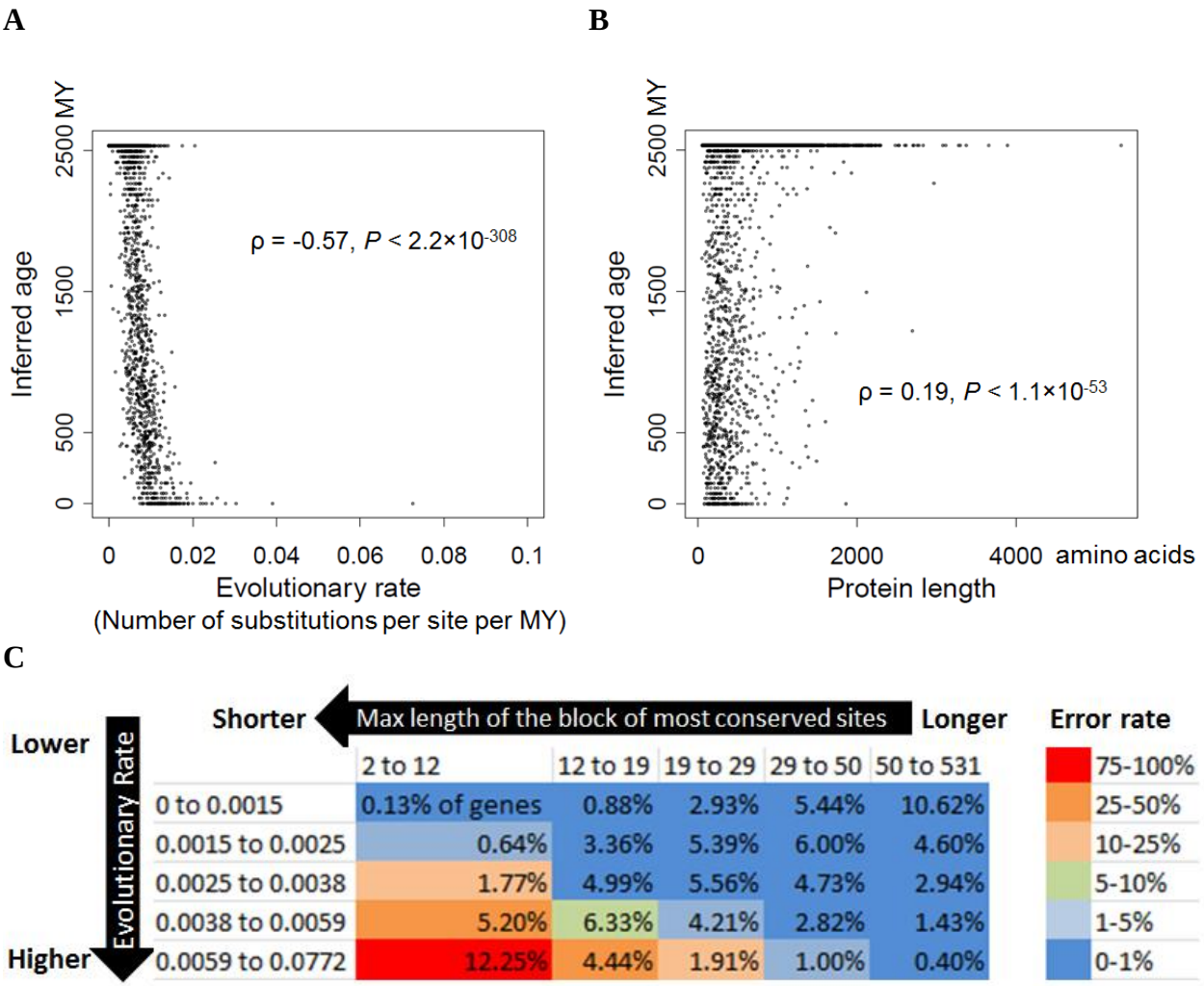

Figure 2- 2 Gene properties influencing BLAST error

Gene age inference by BLAST is influenced by (A) protein evolutionary rate, (B) protein length, and (C) the maximum length of the block of the most conserved sites in the protein. Presented are the average results from 10 simulations. In (A) and (B), each circle represents one fruit fly protein, whose age equals the average inferred age over 10 simulations. In (C), each row and each column represents an equal number of genes. The number in each bin corresponds to the mean number of genes from 10 simulations that fall into the bin. The color of each bin represents the average error rate in that bin, with the color scheme shown on the right of the figure. Error was considered when a gene was inferred to have originated after the separation between bacteria and eukaryotes. Max length is in the unit of amino acid. As shown in the main text by partial correlations, each of the three factors has a significant contribution to BLAST error even when the other two are controlled.

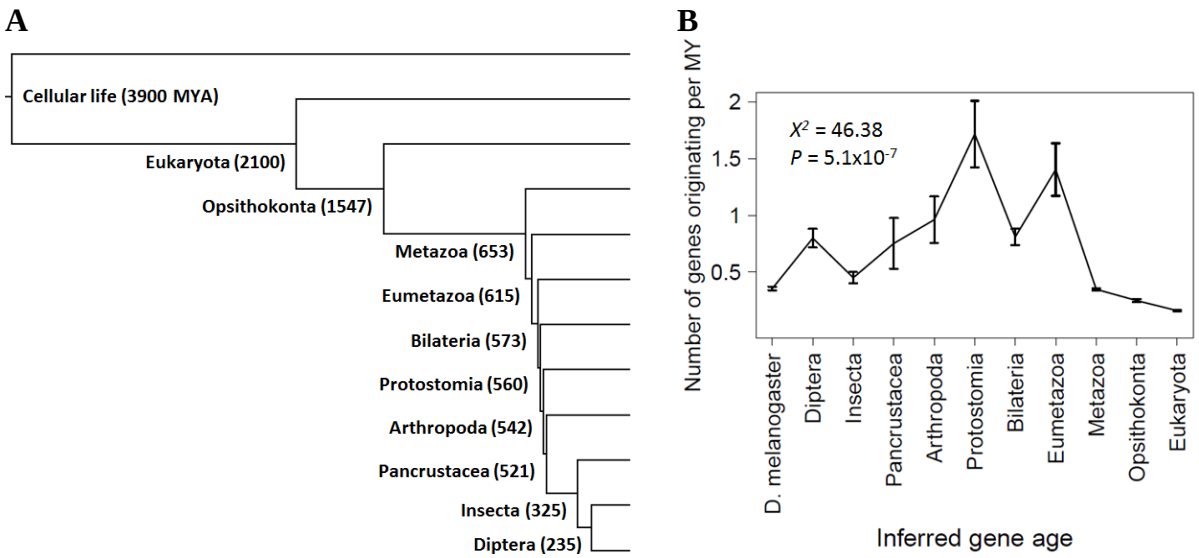

Figure 2- 3 BLAST error mimics phylostratigraphic findings in Drosophila

Shown are results from analysis of simulated data, in which all proteins originated in the common ancestor of cellular life. (A) Phylogeny along which protein evolution is simulated. Both the tree topology and node ages (shown in parentheses) are from Domazet-Loso and Tautz (2007). (B) The inferred number of new gene originations per MY determined by dividing the number of genes inferred to have originated in a tree branch by the time represented by the branch, averaged over 10 simulations. Error bars represent standard deviations. The null hypothesis of equal numbers of gene originations per MY across all strata was examined by a chi-squared test

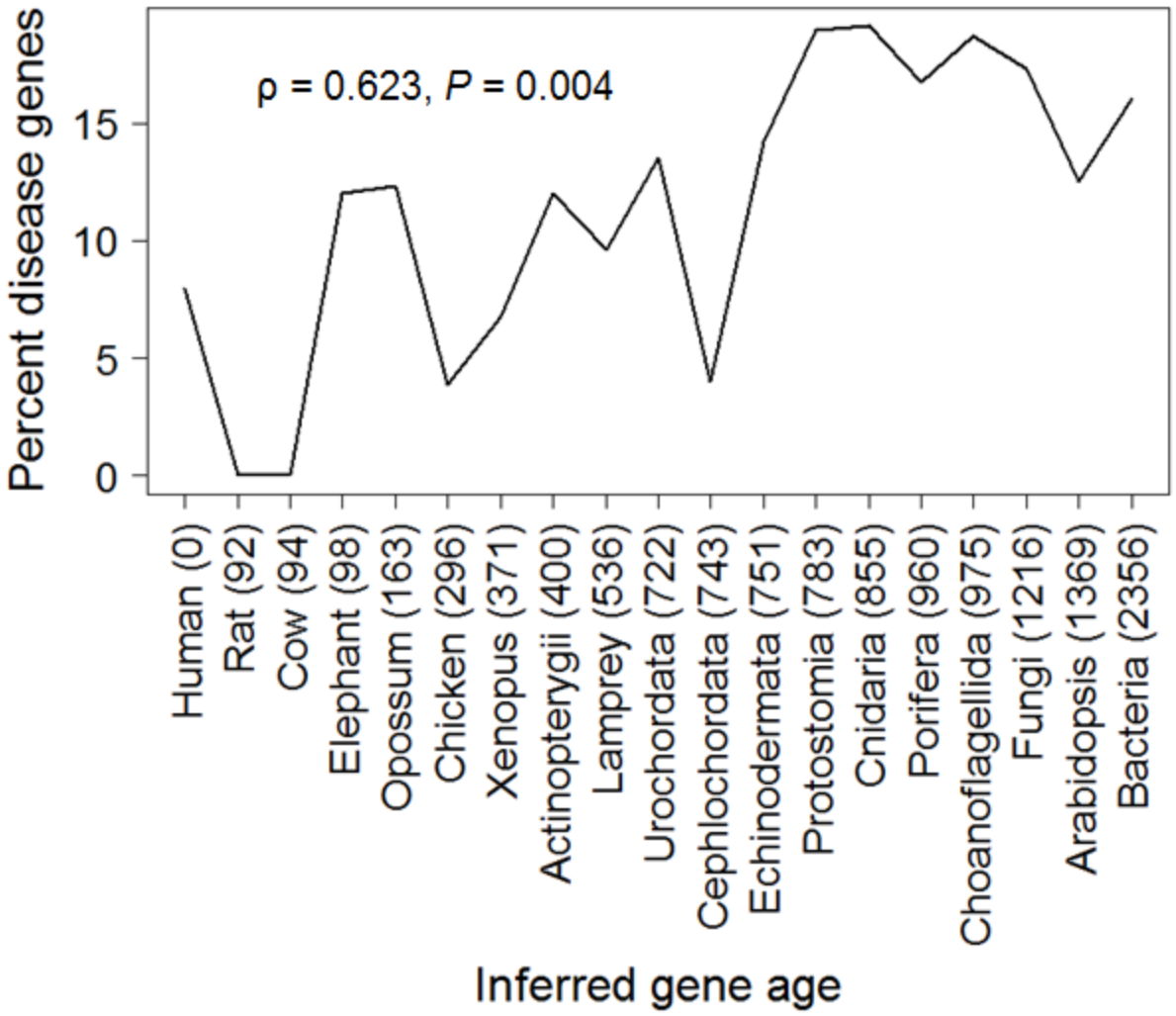

Figure 2- 4 BLAST error mimics phylostratigraphic findings in Human

BLAST error mimics the finding in human genomic phylostratigraphy that old genes are more likely than young genes to be disease genes. Shown are results from analysis of simulated data, in which all proteins originated in the common ancestor of eukaryotes and bacteria. The time (in MY) since divergence between each taxon and human is from TimeTree and is shown in parentheses.

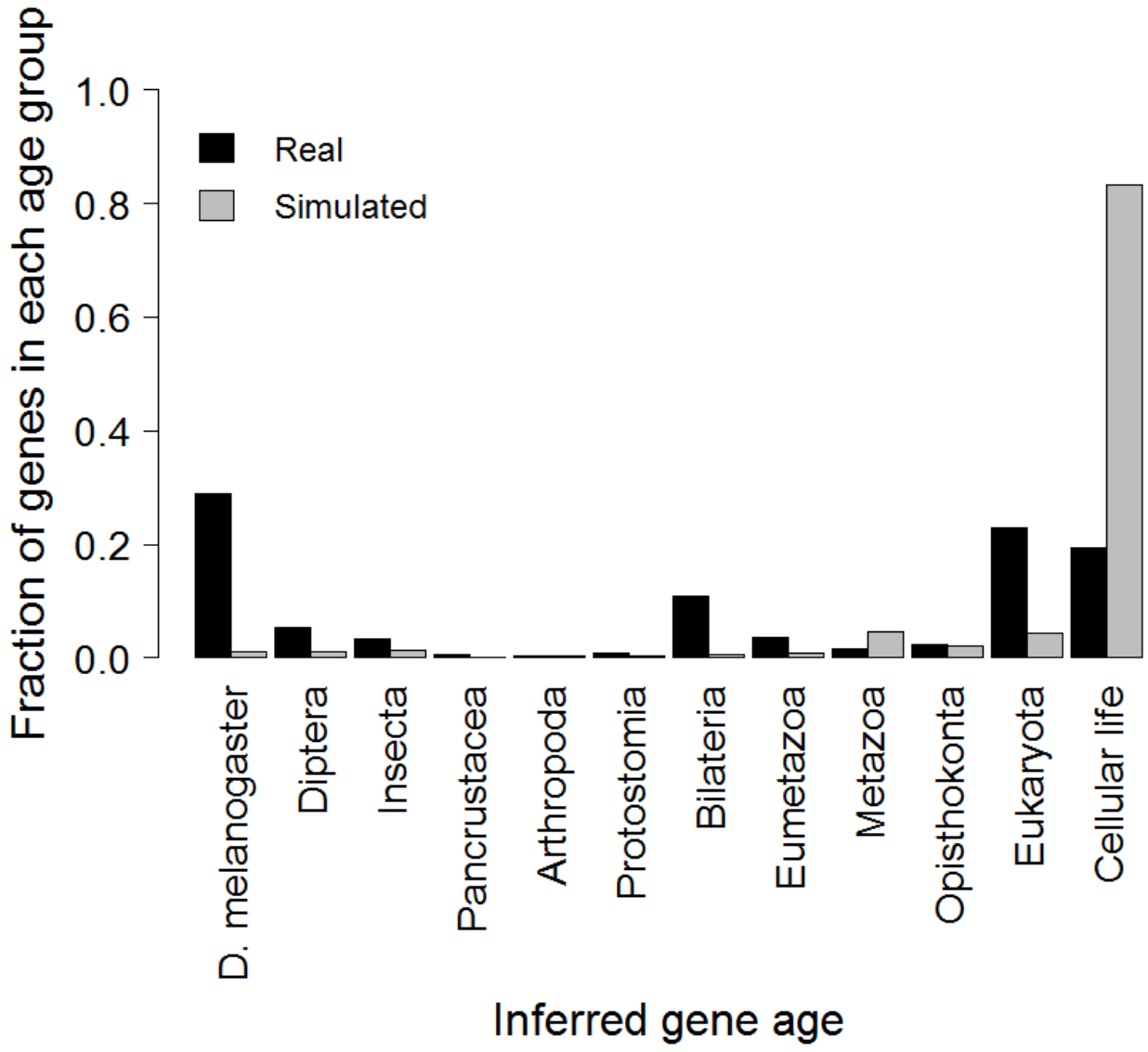

Figure 2- 5 Not all phylostratigraphic signals are due to error

Phylostratigraphy produces signals beyond what BLAST error can account for. Black bars represent the percentage of fruit fly genes inferred to be in each phylostratum based on the real phylostratigraphic analysis of Domazet-Lošo and Tautz (2007). Grey bars represent the percentage of fruit fly genes inferred to be in each phylostratum in our simulated phylostratigraphic analysis. The simulation is the same as in Fig. 3.

Table 2- 1 False negative error rates of BLASTP at various E-value Cutoffs

E-value cutoff	1E-1	1E-2	1E-3	1E-4	1E-5	1E-6	1E-7	1E-8	1E-9	1E-10
Fly homolog not found in bacteria	12.78%* (0.11%)**	13.33% (0.11%)	13.85% (0.14%)	14.32% (0.18%)	14.78% (0.18%)	15.22% (0.15%)	15.67% (0.17%)	16.10% (0.17%)	16.53% (0.13%)	16.96% (0.10%)
Fly homolog not found in any taxon	2.18% (0.09%)	2.48% (0.10%)	2.77% (0.11%)	3.05% (0.14%)	3.32% (0.13%)	3.58% (0.13%)	3.85% (0.10%)	4.11% (0.12%)	4.34% (0.10%)	4.62% (0.11%)
Bacterial homolog not found in any taxon	10.85% (0.12%)	11.47% (0.09%)	12.03% (0.12%)	12.46% (0.11%)	12.88% (0.08%)	13.28% (0.09%)	13.73% (0.12%)	14.12% (0.10%)	14.51% (0.11%)	14.87% (0.12%)

*The top value in each row represents the mean percentage over ten runs.

**The bottom value in each row represents the standard deviation of the percentage over ten runs.

*Table 2- 2 BLASTP error rates under covation evolution**

Rates shuffled per 50 MY	All rate categories shuffled	Lowest rate category cannot be shuffled	Lowest two rate categories cannot be shuffled
0% of sites	14.05%	14.05%	14.05%
1% of sites	17.81%	14.97%	14.51%
2% of sites	32.97%	15.24%	15.23%
5% of sites	67.08%	16.60%	16.52%

*Presented are the fraction of proteins in which the bacterial homolog of a fly protein is not found.

Chapter 3

Evaluating phylostratigraphic evidence for widespread *de novo* gene birth in genome evolution

Published as: Moyers, B.A. and Zhang, J. (2016) Evaluating phylostratigraphic evidence for widespread *de novo* gene birth in evolution. *Mol. Biol. Evol.* 33:1245-1256.

Abstract

The source of genetic novelty is an area of wide interest and intense investigation. Although gene duplication is conventionally thought to dominate the production of new genes, this view was recently challenged by a proposal of widespread *de novo* gene origination in eukaryotic evolution. Specifically, distributions of various gene properties such as coding sequence length, expression level, codon usage, and probability of being subject to purifying selection among groups of genes with different estimated ages were reported to support a model in which new protein-coding proto-genes arise from noncoding DNA and gradually integrate into cellular networks. Here we show that the genomic patterns asserted to support widespread *de novo* gene origination are largely attributable to biases in gene age estimation by phylostratigraphy, because such patterns are also observed in phylostratigraphic analysis of simulated genes bearing identical ages. Furthermore, there is no evidence of purifying selection on very young *de novo* genes previously claimed to show such signals. Together, these findings are consistent with the prevailing view that *de novo* gene birth is a relatively minor contributor to new genes in genome

evolution. They also illustrate the danger of using phylostratigraphy in the study of new gene origination without considering its inherent bias.

Introduction

Different species tend to have different numbers of genes. The human genome, for instance, has somewhere between 19,000 and 25,000 protein-coding genes (Ezkurdia et al., 2014; Hattori, 2005). By contrast, there are approximately 13,000 protein-coding genes in the genome of the fruit fly *Drosophila melanogaster* (Misra et al., 2002). There is some amount of overlap between these two gene sets, but there are also genes unique to each of the two organisms. The question of how these differences in gene number and content arise has been an area of interest and investigation for decades (Kaessmann, Vinckenbosch, & Long, 2009; Long, Betrán, Thornton, & Wang, 2003; M Nei, 1969; Ohno, 1970; K. H. Wolfe, 2001; Jianzhi Zhang, 2003, 2013). In general, these differences are attributable to differential gene gains and losses in different evolutionary lineages. In terms of gene gains, three distinct mechanisms are known: horizontal gene transfer, gene (and genome) duplication, and *de novo* gene birth. While the first two mechanisms and their contributions to organismal adaptation have been abundantly documented (Koonin, Makarova, & Aravind, 2001; Pál et al., 2005; Qian & Zhang, 2014; Jianzhi Zhang, 2013) , the arising of genes from non-genic material via *de novo* gene birth (Tautz & Domazet-Lošo, 2011) was thought nigh-impossible for a long time (Jacob, 1977). Although the last decade has seen the discovery of *de novo* gene birth in several species (Begun, Lindfors, Kern, & Jones, 2007; J. Cai, Zhao, Jiang, & Wang, 2008; Heinen, Staubach, Häming, & Tautz, 2009; Knowles & McLysaght, 2009; Levine, Jones, Kern, Lindfors, & Begun, 2006; C.-Y. Li et al., 2010; Wu, Irwin, & Zhang, 2011; Xiao et al., 2009; Yang & Huang, 2011), the number of reported cases remains small. Because horizontal gene transfer merely transfers genes between

species, gene duplication is commonly regarded as the dominant source of new genes while *de novo* gene birth is thought to have a minimal contribution.

The above view was recently challenged by Carvunis and colleagues, who claimed that *de novo* gene birth is common in evolution and is a larger source of new genes than gene duplication (A.-R. Carvunis et al., 2012). Specifically, they proposed that non-genic sequences are spuriously transcribed and translated, and the protein products may by chance possess biological functions, which could be selected for, resulting in a gradual enhancement of the protein function in evolution. They named the open reading frames (ORFs) that are transcribed and translated but have not fully established their functions as proto-genes. They asserted that their model predicts a number of trends as proto-genes gradually age, including, for example, increases in ORF length, expression level, codon usage bias, and probability of being under purifying selection. The ideal test of their hypothesis would be to conduct laboratory evolution experiments and watch in real time how a non-genic sequence turns into a functional protein-coding gene. But because such evolutionary events are expected to be rare and the evolutionary processes slow, the authors took an indirect approach by comparing various properties among different age groups of proto-genes and genes from the genome of the budding yeast *Saccharomyces cerevisiae*, where gene ages were estimated using phylostratigraphy (Domazet-Lošo et al., 2007). In phylostratigraphy, the age of a gene from a focal species is defined by the time since the divergence between the focal species and its most distantly related taxon in which a homolog of the gene is found by a commonly used homology detection tool such as BLAST. Carvunis et al. reported that multiple trends predicted by their model were observed. The same claim was made in a similar study of vertebrates (Neme & Tautz, 2013). Carvunis et al. further noted that 143

proto-genes originated in *S. cerevisiae* since its divergence from its sister species *S. paradoxus* and 19 of them are under purifying selection in *S. cerevisiae*. By contrast, they noted that no more than five genes were estimated to have been generated by gene duplication in the same period of time. These results led Carvunis et al. to conclude that *de novo* gene birth is widespread and is a bigger source of new genes than is gene duplication. A subsequent study based on a similar analysis of age distributions of gene properties suggested that proto-genes are gradually integrated into cellular networks by for instance gradual gains of protein interactions and genetic interactions (Abrusán, 2013).

While nothing is wrong with the theoretical model of *de novo* gene birth, whether the reported genomic patterns signify *de novo* gene birth and subsequent evolution is questionable for two reasons. First, some of the asserted predictions from the *de novo* gene birth model do not seem to be definitive. For example, it is unclear why the ORF of a gene should continually increase in length with time. Although it is easy to imagine scenarios where length increases are beneficial, one can also come up with situations where length reductions are advantageous. Because of the frequency of stop codons in the genome, it is likely that a *de novo* gene will be short and will increase in its early lifespan as a proto-gene. But it is not clear that this trend would be monotonic or prolonged for hundreds of millions of years. Once a function is established, why would increasing rather than decreasing its length tend to enhance or refine its function? Even if increasing the ORF length is beneficial to the functional refinement of a proto-gene, why should the length continue to rise even long after the proto-gene has become a well-established gene (e.g., when the gene is over 500 million years old), as was observed by Carvunis and colleagues? Second, phylostratigraphy tends to underestimate gene age and the probability and amount of

underestimation differ among genes (Moyers and Zhang 2015). For example, the probability of age underestimation decreases with the increase of ORF length, which could in principle explain Carvunis et al.'s observation of a gradual increase in ORF length with the estimated gene age. In this work, we show that the age distributions of various gene properties supporting widespread *de novo* gene birth are in fact largely attributable to age estimation errors created by phylostratigraphy. As such, there is no valid evidence to date for a larger contribution of *de novo* gene birth than gene duplication to new gene origination.

Methods

Yeast genes

For simulation of sequence evolution, we acquired 5261 orthologous sequence alignments in protein format from the *sensu stricto* group of yeast species from http://www.saccharomycessensustricto.org/current//aligns/coding_allfiles.fasta.tgz (Scannell et al., 2011). Except for two alignments, all contain five orthologous sequences from five *sensu stricto* yeast species. The simulation of the 5259 genes that have alignments of five sequences used parameters estimated from the alignments. The simulation of other genes in *S. cerevisiae* used parameters estimated from a set of *sensu stricto* restricted genes.

To identify *sensu stricto* restricted genes, we acquired protein databases of four yeast species outside of the *sensu stricto* group. These species were *S. castellii* and *S. kluyveri*, downloaded from the Saccharomyces Genome Database at http://www.yeastgenome.org/download-data/sequence (Cherry et al., 2012), as well as *K. thermotolerens* and *Z. rouxii*, acquired from

the Genolevures Consortium (Souciet, Dujon, & Gaillardin, 2009). Using the alignments acquired from Scannell et al. (2011), we created five databases, one for each of the *sensu stricto* species. We then performed a BLASTP (E-value = 0.01, in following with Carvunis *et al.*) search using each of these individually as a query, and the target being an aggregate of the *S. castellii*, *S. kluyveri*, *K. thermotolerens*, and *Z. rouxii* proteins. We identified proteins for which none of the five *sensu stricto* yeast homologs found a hit in the target database, amounting to 148 genes. These 148 genes exist in all five *sensu stricto* yeasts but are not found in the four outgroup species. While homology detection error may explain the apparent restriction of these genes to the *sensu stricto* group, this is not a problem for our simulation, because it is exactly our goal to identify patterns of genes that appear to be *sensu stricto* restricted, whether or not they are in reality.

Main simulation of evolution

The evolutionary tree including the relative branch lengths used in simulation was from a previous study of yeast genes (Wapinski, Pfeffer, Friedman, & Regev, 2007). For each of the 5259 proteins with alignments of five sequences, we used TreePuzzle (Schmidt et al., 2002) to classify all sites into 16 equal-sized rate bins according to a discrete gamma model of among-site rate heterogeneity and estimated the relative rates of the 16 bins. We also inferred the mean evolutionary rate across all sites of the protein between *S. cerevisiae* and *S. bayanus*; all branch lengths for the protein concerned were then estimated using the relative tree branches aforementioned. Using all of these parameters, we simulated the evolution of these proteins using ROSE (Stoye et al., 1998), which allows the evolutionary rate for each site to be specified by the user, along the tree in Figure 3-1A. ROSE evolves sequences through amino acid

substitutions and insertions and deletions (indels). For each branch of the tree, ROSE first performs the amino acid substitution function, and then performs the indel function. If the branch is an internal branch in the tree, it then copies the resulting amino acid sequence to the base of each of the two branches after the split.

We used the JTT-f model in the ROSE simulation of protein sequence evolution, where "f" refers to the amino acid compositions of the protein concerned (Masatoshi Nei & Kumar, 2000). Each site along the protein has a particular relative rate. The relative rate for a site is multiplied by the length of the branch to obtain the expected amount of evolution along the branch at the site. ROSE makes substitutions based on this expected amount of evolution and the substitution matrix supplied. This is repeated for all sites along the amino acid sequence.

For indels, there are two parameters that determine indel formation in ROSE, the indel threshold and the indel function. The indel threshold measures how frequently indels occur and was determined in the following manner. Taking the alignments of the yeast *sensu stricto* orthologs acquired from Scannell et al. (2011) and using a custom script, we determined the minimum number of indels necessary to produce the observed gapped alignments. From this information, we determined the number of indels per amino acid, averaged over all proteins. This indel threshold was then applied to all proteins in simulation. The indel function is a vector that sums to 1 and gives, at each vector site *i*, the probability of an indel of size *i*, given that an indel is occurring. For the indel function, we took the observed frequencies of indel sizes from 1 amino acid to 30 amino acids long (accounting for > 99% of all observed indels), and adjusted these frequencies to sum to 1. Sequence simulation was performed once for each protein.

Simulation of other proteins

Sequences were acquired as described above, but we could not determine evolutionary rate or rate heterogeneity for proteins lacking an alignment or the two proteins from Scannell et al. (2011) that do not have alignments of all five orthologous sequences. We used parameters estimated from the group of *sensu stricto* limited genes to simulate these proteins. To do this, we took each protein in this group and multiplied the relative rates of all sites by the average evolutionary rate for the protein. This gave us an absolute evolutionary rate for each site. We then concatenated these numerical vectors into a single vector from which we could sample rates for each protein (Figure A-2). We specifically sampled the inferred absolute substitution rates of a contiguous set of sites. From there, we performed a simulation of evolution as described above. This simulation likely rendered our estimate of phylostratigraphic error rate conservative, because on average *sensu stricto* limited genes are expected to evolve more slowly than the 619 genes which do not have homologs in all *sensu stricto* species, as fast evolution is a reason for an apparently young gene age (Moyers and Zhang 2015). Note that smORF sequences were not simulated. Instead, they were universally assigned to age group 0, as in Carvunis et al. (2012).

Protein phylostratigraphy

To perform protein phylostratigraphy, we used BLASTP with a permissive e-value of 0.01, following the methods of Carvunis et al. (2012). We used the simulated sequences corresponding to *S. cerevisiae* as the query, and each other species as an independent database.

We ran BLASTP searches for each simulated species independently rather than as a single aggregate database to increase sensitivity of homology detection.

Carvunis *et al*. conducted BLASTP, TBLASTX, and TBLASTN searches; the latter two searches require the use of DNA sequences. We chose not to simulate the evolution of protein-coding DNA sequences because realistic simulation of codon sequence evolution is difficult and because protein-based homology searches are generally much more sensitive than DNA-based homology searches.

NCBI homology searches

We acquired from Saccharomyces Genome Database (SGD) the DNA and protein sequences of Carvunis et al.'s 16 genes of age group 1 that were purported to be under purifying selection. We used the NCBI BLAST tool to perform BLASTN, TBLASTN, and TBLASTX searches against the full non-redundant database of all species. We restricted results to a permissive e-value of 0.01, and only considered hits that had at least 40% query coverage.

Testing purifying selection in 16 young genes

We downloaded the reference sequence for each of the 16 young genes in question from the SGD, and noted exactly which nucleotides were not overlapped by another annotated open reading frame. We then acquired single nucleotide polymorphisms (SNPs) for all chromosomes in all strains, available at ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/latest/cere_matches.tgz. We extracted the SNPs of 38 strains present in both the SGRP data and the phylogeny in Liti et

al. (Liti et al., 2009). We extracted only those SNPs for which quality score was 55 or greater, following Carvunis et al. (2012). We modified the reference sequence for each strain, producing FASTA files containing each strain's sequence. We removed all sections of the sequence which were overlapped with another ORF. In order to retain full codons, we removed any codon which had even partial overlap with another ORF. We then aligned these sequences using MUSCLE (Edgar, 2004). We performed Fisher's exact test using the observed numbers of synonymous and nonsynonymous SNPs and the potential numbers of synonymous and nonsynonymous sites estimated assuming 70% of random mutations are nonsynonymous (Jianzhi Zhang, Kumar, & Nei, 1997). In no case was the result significantly different from the neutral expectation.

The 38 strains used are as follows: DBVPG6040, NCYC361, S288c, W303, 378604X, YJM789, YS2, YS4, YS9, 273614N, YIIc17_E5, RM11_1A, YJM975, YJM978, YJM981, DBVPG1853, 322134S, BC187, DBVPG6765, DBVPG1788, L-1374, L-1528, DBVPG1106, DBVPG137, SK1, DBVPG6044, NCYC110, Y55, UWOPS87_2421, UWOPS83_787_3, UWOPS03_461_4, UWOPS05_227_2, UWOPS05_217_3, K11, Y12, Y9, YPS606, and YPS128.

Other datasets

We were provided with various gene properties from Carvunis et al. via email communication. We downloaded datasets used by Abrusan (2013) from the supplementary data of that paper. The definitions and measurements of all of these properties were detailed in the respective publications (Carvunis et al. 2012; Abrusan 2013).

Results

Phylostratigraphy of simulated genes

To examine whether gene age estimation error caused by phylostratigraphy could create spurious age distributions of gene properties resembling Carvunis et al.'s observations, we conducted a computer simulation of the evolution of all *S. cerevisiae* protein sequences along the tree shown in Figure 3-1A using protein-specific parameters for site-specific rates and overall evolutionary rate. All *S. cerevisiae* protein sequences were simulated to have orthologs in all of the species shown in the tree (Figure 3-1A). That is, they all have the same age of 10, and there is no *de novo* gene origination in our simulation. We then applied phylostratigraphy to estimate the ages of the *S. cerevisiae* proteins by BLASTing them against the simulated sequences in all other species. These ages are referred to as estimated ages of simulated proteins (Figure 3-1B). We subsequently computed age distributions of various properties of *S. cerevisiae* proteins using the above estimated ages (Figure 3-2 and 3-3). Note that we used the properties provided by Carvunis et al. for each *S. cerevisiae* protein in these distributions; the only difference is the estimated gene age. In other words, we ask what would be the observed age distributions of gene properties if all *S. cerevisiae* genes have the same true age with no *de novo* gene birth. If the age distributions we observed resemble what Carvunis et al. observed, their observations cannot be used to support the *de novo* gene birth hypothesis because these observations are expected even in the absence of *de novo* gene birth.

To derive protein-specific parameters for simulation, we acquired 5261 published orthologous protein sequence alignments from five *sensu stricto* yeast species (*S. cerevisiae, S. paradox, S.*

mikatae, *S. kudriavzevii*, and *S. bayanus*) (Scannell et al., 2011). For each of these proteins, we estimated the mean substitution rate per amino acid site and the substitution rate at each site relative to the mean rate of the protein (see Materials and Methods). These parameters were used in the simulation of the evolution of the protein (see Materials and Methods). For 619 *S. cerevisiae* proteins that do not have homologs in all five *sensu stricto* yeast species, we simulated their evolution in a conservative manner by sampling rate heterogeneity patterns and mean evolutionary rates from *sensu stricto* restricted proteins (see Materials and Methods). In all, we simulated the evolution of all 5878 proteins present in the Carvunis *et al*. dataset. The genetic distance of simulated orthologous proteins matches well that of real proteins (Figure A-1).

Because the true ages are 10 for all genes in the simulation (Figure 3-1A), any observed age distribution in which not all genes are in age group 10 is spurious. We found that, for 11.4% of simulated proteins, a homolog could not be found in the most distant species considered (*Schizosaccharomyces pombe*) (Figure 3-1B), which was estimated to diverge from *S. cerevisiae* approximately 788 million years (MY) ago (Heckman et al., 2001; Hedges et al., 2006). The error rate of 11.4% is likely an underestimate, because a portion of our genes were evolved in a conservative manner (see Materials and Methods) and because we assumed that each site has a fixed substitution rate throughout its evolution, which is known to result in an underestimation of the error rate (Moyers and Zhang 2015). Of the 669 simulated proteins whose ages were underestimated by phylostratigraphy, 185 had estimated ages of 1-4 (Figure 3-1B). These genes would therefore be considered "candidate proto-genes" under Carvunis *et al.*'s definition, although they originated hundreds of millions of years ago in our simulation. Most strikingly, phylostratigraphy determined that two of these genes are *S. cerevisiae*-specific, despite that they

originated in the common ancestor of *S. cerevisiae* and *S. pombe*. Nevertheless, the number of genes with estimated age 1-9 is greater in the actual data than in the simulated data (Figure 3-1B). While this disparity may indicate the presence of some *de novo* genes, it may also be due to the fact that our simulation is conservative. That is, evolutionary processes that are not simulated here, such as gene duplication followed by rapid divergence and changes in the evolutionary rate of a site during evolution, could be responsible for this disparity.

Age distribution of six gene properties with statistical support

We next compared the age distributions between the real genes and simulated genes for each gene property used by Carvunis et al. as evidence for their model of widespread *de novo* gene birth. If the age distributions for a gene property are similar between the real genes and simulated genes, the age distribution observed by Carvunis et al. for the real genes can be explained by phylostratigraphy errors and hence cannot be used to support their model.

We first examined the six trends for which statistical support was previously provided (A.-R. Carvunis et al., 2012). These trends are significant increases in ORF length (Figure 3-2A), mRNA abundance (Figure 3-2B), proportion of genes in proximity of transcription factor binding sites (Figure 3-2C), proportion of genes under significant purifying selection (Figure 3-2D), proportion of genes with optimal AUG context (Figure 3-2E), and codon adaptation index (Figure 3-2F) with gene age estimated through phylostratigraphy. Here, proportion of genes under significant purifying selection was determined by testing the action of purifying selection on each gene based on sequence polymorphisms among eight *S. cerevisiae* strains. All gene properties are defined as in Carvunis et al. (2012) and the property data were acquired from the

authors. We found that, while qualitative appearances differed between the real and simulated data in these age distributions (Figure 3-2), statistical trends, quantified by Kendall's τ as in Carvunis et al. (2012), were almost identical between the two (Table 3-1). Kendall's τ was used following Carvunis *et al*. Using Spearman's ρ did not alter our results. Both effect size (i.e., correlation coefficient) and significance level were reasonably well matched. This implies that the observed statistical trends of various gene properties with regard to gene age can be largely explained by gene age estimation errors.

Carvunis et al. included in their analysis ~108,000 so-called small ORFs (smORFs) that were arbitrarily assigned the age of 0. These *S. cerevisiae* smORFs are not annotated genes, are at least 30-nucleotide long, and are free from overlap with annotated features on the same strand. The similarity in the above six trends between real and simulated data holds whether or not these smORFs were included in our analysis (Table 3-1).

Some of the *S. cerevisiae* genes analyzed are paralogous to one another, but our simulation and subsequent phylostratigraphy treated them as unrelated genes, rendering our result from the simulated data not directly comparable with that from the real data. To solve this problem, we performed an all-against-all BLASTP search of the original *S. cerevisiae* proteins and recorded paralogous relationships. From this information, we used the oldest age among each gene family as the age of all genes in that family. This modification of phylostratigraphically estimated gene age on our simulated data did not change our results on the genomic trends studied above (Table 3-1).

Age distributions of four gene properties without statistical support

Carvunis et al. (2012) also reported four additional trends without providing statistical support, including changes in amino acid usage, hydropathicity, proportion of transmembrane regions, and proportion of disordered regions with estimated gene age. For the majority of these, the simulated data do not qualitatively match the real data (Figure 3-3A, B, and C). A notable exception is the patterns found in amino acid usage, where simulated data matches real data quite closely (Figure 3-3D). Note, however, no explicit explanation was provided by Carvunis et al. why these observed trends are expected from the *de novo* gene birth model (see Discussion). As such, we do not see these trends as evidence for or against the *de novo* gene birth model.

Age distributions of gene properties reflecting genetic integrations

Subsequent to Carvunis et al.'s study, Abrusán used Carvunis et al.'s data in conjunction with the data in Wapinski (2007) to examine the phylostratigraphically-based age distributions of a number of additional gene properties that he proposed to reflect gradual genetic integrations of *de novo* genes into cellular networks or maturation of protein structures (Abrusán, 2013). These included many factors that seemed to be reasonable proxies for the integration of a gene into the gene network, such as genetic coregulation, number of protein-protein interactions, number of genetic interactions, number of feed-forward loops regulating a gene, number of transcription factors regulating a gene, and epistatic effects However, there were also a number of factors with questionable relationships to a gene's integration, such as percent of a gene which was made up of alpha-helices or beta-sheets and the propensity of a protein to aggregate. Interestingly, all significant trends he found in real genes are also significant in simulated genes,

except for the case of alpha helices (Table 3-2). We note that, in several but not all cases, effect sizes are comparable as well (Table 3-2). Even in those cases where the effect size appears quite different between real data and simulated data, the differences do not necessarily support the *de novo* gene birth model, because the differences may be attributable to new genes created via gene duplication in the real data (He & Zhang, 2005). Furthermore, it is unclear whether several of the trends observed (e.g., decrease in percent in beta sheets) indicate structure maturation of *de novo* genes. These appear to be *post hoc* explanations rather than *a priori* predictions of the *de novo* gene birth model (see Discussion).

Number of young genes under purifying selection

Carvunis et al. (2012) noted that they observed 19 genes that are both *S. cerevisiae*-specific and under within-species purifying selection. Based on their new analyses (Carvunis, personal communication), this number now drops to 16. The abundance of these genes was suggested by Carvunis et al. to be evidence of high rates of *de novo* (functional) gene birth in comparison to gene duplication (A.-R. Carvunis et al., 2012; Gao & Innan, 2004).

However, we noticed that 15 of the 16 genes are each overlapped with another gene on the opposite strand and the overlapping regions constitute between 73% and 93% of each of these 15 genes (Table 3-3). The remaining gene, *YOL166C*, has no overlap with any annotated gene in *S. cerevisiae*. When searching for homologs in other fungal species, Carvunis et al. removed sections of query genes which overlapped. We searched for homologs using the full sequences of these query genes and discovered that many of them are present in other species (Table 3-3). All hits occurred in true ORFs in the target sequence, which were at least 80 amino acids long

and were frequently annotated and known to be transcribed. If these 15 genes are *S. cerevisiae*-specific, they are not expected to have long ORFs (≥ 80 codons) in other species even when the opposite strand has an overlapping gene. Thus, we conclude that these 15 genes are not *S. cerevisae*-specific and that Carvunis et al.'s results were erroneous because of their use of short query sequences that rendered BLAST powerless. The gene of most interest is *YOL166C*, because it is not overlapped by any other gene and has no hit in any other sequenced species. There are two major questions to be addressed about this gene. First, is there a homologous sequence in *S. paradoxus*, the species known to be the closest to *S. cerevisiae,* such that one can identify the source of *YOL166C*? Second, is there direct evidence for translation of this gene? To approach the first question, we looked for the *S. paradoxus* genomic region aligned to *S. cerevisiae* chromosome 15, base pairs 1 to 2078, a region encompassing *YOL166C*. No such alignment exists in this region, according to the Saccharomyces Genome Resequencing Project (SGRP) Genome Browser. We further checked for the homologs of *YOL166C*'s neighboring genes *TEL15L* and *YOL165C*. *TEL15L* found a significant hit in the *S. paradoxus* retrotransposons Ty5-10p and Ty5-5p, but *YOL165C* had no hit in *S. paradoxus*. *YOL165C* and *YOL166C* are in the subtelomeric region of chromosome 15 in *S. cerevisiae*. These regions are generally quite unstable (Brown, Murray, & Verstrepen, 2010), so it is not surprising that an orthologous region could not be found. Additionally, when BLASTed against the *S. cerevisiae* genome, *YOL166C* only finds itself as a hit.

To approach the second question, we searched for direct evidence of translation of *YOL166C*. Carvunis et al. did not find evidence of the translation of this gene under either rich or starved conditions based on yeast ribosome profiling data (Ingolia, Ghaemmaghami, Newman, &

Weissman, 2009). Several papers report changes in the transcript concentration of *YOL166C* under different conditions (Fisk et al., 2006), but there is no evidence that *YOL166C* is expressed at the protein level. Based on these analyses, *YOL166C* does not meet the strict definition of a *de novo* gene (see Discussion). However, it also does not appear to be an instance of gene duplication. This leaves open the possibility that this is an example of a *de novo* gene birth.

A major question remains about whether or not these 16 genes are under selective constraint. Carvunis et al. estimated the nonsynonymous to synonymous substitution rate ratio on a phylogeny of eight *S. cerevisiae* strains and found this ratio to be significantly lower than 1, an indication of the action of purifying selection. However, their method is commonly used for testing selection in gene sequences collected from different species and is inappropriate for testing selection in sequences from the same species, because, for intra-specific data, different regions of the genome can have different phylogenies due to recombination. Additionally, because the majority of the sequence was overlapped by another gene, inferring selective constraint can be confounded (Wei & Zhang, 2015). So, in the cases of these genes, only their non-overlapped portions should be used to infer selection. To increase the accuracy and power of selection detection, we used 38 *S. cerevisiae* strains in the Saccharomyces Genome Resequencing Project (Cherry et al., 2012) and counted the number of synonymous and nonsynonymous polymorphisms in the region of a gene that is non-overlapping with other genes (Table 3-3). Using Fisher's exact test, we then examined whether the ratio between the observed number of nonsynonymous polymorphisms to that of synonymous polymorphisms is significantly different from the corresponding ratio under neutrality, which was calculated from the potential numbers of nonsynonymous and synonymous sites in the same region (Zhang et al.

1998). In none of the 16 genes could the null hypothesis of neutrality be rejected in favor of the action of purifying selection or positive selection. This is probably unsurprising, because no evidence was found for their translation by Carvunis et al. and these genes probably bear no protein function. As a comparison, the same selection test was conducted for 100 randomly picked genes classified to age group 10 by Carvunis et al., and 86 of them were found to be under significant purifying selection. However, these genes are among the longest and most conserved genes in the set, and it can be assumed that power for extremely short genes or gene fragments would be very low.

Discussion

The origin of new protein-coding genes from non-coding sequences is a fascinating hypothesis that has been supported by the discoveries of dozens of cases of *de novo* gene birth in human, *Drosophila*, yeast, and other species (J. Cai et al., 2008; Clark et al., 2007; Heinen et al., 2009; Knowles & McLysaght, 2009; Levine et al., 2006; C.-Y. Li et al., 2010; Wu et al., 2011; Xiao et al., 2009; Yang & Huang, 2011). Previous studies established a set of criteria for identifying *de novo* gene birth: (1) the candidate *de novo* protein-coding gene is transcribed and translated, (2) its homologous sequence can be found in the syntenic region in related species but the sequence has no protein-coding capacity, and (3) the sequence is ancestrally non-coding (Knowles & McLysaght, 2009). One should add the fourth criterion of action of natural selection for a *de novo* gene to be considered functional. Satisfying all these criteria would prove *de novo* gene birth beyond reasonable doubt.

However, not all of the above criteria were used and satisfied in Carvunis et al.'s study. Instead, Carvunis et al. relied on estimating gene age by phylostratigraphy and using age distributions of various gene properties to test widespread *de novo* gene birth. For their approach to work, gene age estimation must be reliable and *de novo* gene birth must be widespread. Unfortunately, phylostratigraphy is known to be biased (Elhaik et al. 2006; Moyers and Zhang 2015). Thus, only those trends that are predicted by the *de novo* gene birth model but cannot be produced by phylostratigraphic bias may be used to support the model. But, we found that essentially every trend reported by Carvunis et al. (2012) and Abrusán (2013) are explainable at least to some extent by phylostratigraphic bias. One might argue that the age distributions observed from the actual data are not exactly the same as those observed from the simulated data, providing evidence for the *de novo* gene birth hypothesis. This argument is flawed for two reasons. First, a realistic simulation requires many parameters. Because not all parameters are known, we conducted conservative simulations. For example, the substitution rate of a site is unlikely to be constant in evolution (Fitch, 1971; Penny et al., 2001; Zou & Zhang, 2015) and this inconstancy increases phylostratigraphic error (Moyers and Zhang 2015). But because of the lack of information on the extent of this rate variation over time, we assumed no such variation in our simulation, rendering the phylostratigraphic error underestimated and our results conservative. Furthermore, the parameters chosen in simulating genes that are not found in all five *sensu stricto* yeast species also made the results conservative. Thus, the fact that the observed trends in real data are not exactly the same as in the simulated data does not necessarily indicate the existence of biological signals. Second, even if a biological signal truly exists, it does not necessarily support the *de novo* gene birth hypothesis. For instance, in Figure 3-2B, one can see a grey peak at age 7, indicating that genes of age 7 have unusually high expressions. This feature

in the real data is not present in the simulated data, so might mean a true biological signal. Nevertheless, this signal is not predicted by the *de novo* gene birth model and thus cannot be used to support the model.

A common pitfall of phylostratigraphy-based studies is to report whatever nonrandom trends observed and then provide *post hoc* explanations, as if all nonrandom trends have biological meanings. The problem of these kinds of explanations has been pointed out in other contexts (Pavlidis et al., 2012). Carvunis et al.'s and Abrusán's studies also fall into this trap. Many of the trends they reported are not predicted *a priori* from the *de novo* gene birth model. These trends include ORF length in Figure 3-2, all four properties in Figure 3-3, genetic co-regulation, % alpha helices, and % beta sheets in Table 3-2. As mentioned, there is no particular reason why the refinement of the biological function of an ORF has to occur by increasing the ORF length rather than decreasing the length. Similarly, there is no prediction that as proto-genes age and mature, the mean hydropathicity should decrease, trans-membrane fraction of the protein should decrease, disordered fraction should increase, and certain amino acid frequencies should increase or decrease. In fact, the authors offer no explanation of why these trends are expected under the *de novo* gene birth model. Even for the trends that may be predicted by the *de novo* gene birth model, one cannot explain why some of them continue even for genes with age 10 (e.g., expression level and codon adaptation index), as if the maturation of *de novo* genes takes more than 500 MY. Phylostratigraphic error remains the simplest and best explanation of the observed trends, whether or not they are predicted from the *de novo* gene birth model.

One might ask why phylostratigraphic error could result in seemingly nonrandom age distributions of so many gene properties. Based on the property of BLAST search, we previously predicted and demonstrated that gene age underestimation in phylostratigraphy is more severe when the protein under investigation is shorter or evolves faster (Moyers and Zhang 2015). Thus, the increase in ORF length with age observed in the simulated data (Figure 3-2A) is a known bias of phylostratigraphy. Lower protein evolutionary rates are caused by stronger purifying selection, so it is unsurprising that phylostratigraphic error causes a positive correlation between gene age and proportion of genes under purifying selection (Figure 3-2D). Because protein evolutionary rate is strongly negatively correlated with its mRNA expression level (Jianzhi Zhang & Yang, 2015), mRNA expression level must also impact phylostratigraphic error, as seen in our simulated data (Table 3-1). Hence, a positive correlation between gene age and expression level (Figure 3-2B) reflects an expected bias of phylostratigraphy. Phylostratigraphic error is also expected to create a positive correlation between gene age and codon adaptation index (CAI) (Figure 3-2F), because CAI is positively correlated with gene expression level (Sharp & Li, 1987). Because the expression level of a gene is positively correlated with the probability that the gene is in proximity of TF binding sites (Wong et al., 2015) ($\tau = 0.094$ in our data, $p < 1E\text{-}300$), phylostratigraphic error also causes a positive correlation between gene age and proportion in proximity of TF binding sites (Figure 3-2C). It was reported (Miyasaka, Kanai, Tanaka, Akiyama, & Hirano, 2002) and confirmed here that the expression level of a gene is positively correlated with the probability that the gene has an optimal AUG context ($\tau = 0.057$, $p < 1E\text{-}300$), potentially explaining why a positive correlation between gene age and proportion in optimal AUG context is created by phylostratigraphic error (Figure 3-2E). Amino acid usage is known to be correlated with gene expression level (Akashi

& Gojobori, 2002), potentially explaining the observed trends in Figure 3-3D. In fact, we found that all gene properties examined by Carvunis et al. are significantly correlated with one or more of the three factors that impact phylostratigraphic bias: ORF length, evolutionary rate, and expression level (Table 3-4; Table A-1).

The contribution of *de novo* gene birth compared with gene duplication to the origin of new (functional) genes is an important subject of evolutionary genomics. Carvunis et al. suggested that there have been 16 *de novo* births of functional genes in *S. cerevisiae* since its split from *S. paradoxus*. They compared this to a suggested five genes formed by duplication in the same time period (Gao & Innan, 2004), though this duplicate gene number has since been challenged (Casola, Conant, & Hahn, 2012). If correct, Carvunis et al.'s comparison would contradict the paradigm that duplication is the primary source of new genes. We found that 15 of the 16 genes claimed by Carvunis et al. to be *S. cerevisiae*-specific and under selection have homologous ORFs in at least one other species and that none of the 16 bear significant signals of natural selection or have evidence for translation. To our knowledge, there are only two verified instances of functional *de novo* gene births in *S. cerevisiae* (J. Cai et al., 2008; D. Li et al., 2010), whereas approximately 144 functional duplications occurred in that time based on the inference from gene family expansions since the common ancestor of *sensu stricto* yeasts (Hahn, Bie, Stajich, Nguyen, & Cristianini, 2005). While these estimates may not be precise, gene duplication appears to surpass *de novo* gene birth by two orders of magnitude in terms of contribution to the number of new functional genes. Of course, apart from this rate difference, the two mechanisms of new gene origination may supply different kinds of genetic materials. Gene duplication confers a functional gene structure to the daughter gene, whereas *de novo* gene

birth provides something closer to a blank slate, a near-random form and function that may or may not be useful. It is possible that *de novo* gene births offer a greater degree of novelty, even if they contribute less frequently to the genome.

The investigation of *de novo* gene birth mechanisms brings up the question of what is meant by a (functional) gene. There is no shortage of answers to this question (Demerec, 1933; Gerstein et al., 2007). Clearly, in the *de novo* gene birth model discussed here, what is meant is a functional, protein-coding gene. It is thus important to prove the functionality of a gene by demonstrating that it is under purifying or positive selection. Given the widespread trancription of intergenic sequences in eukaryotes (Johnson, Edwards, Shoemaker, & Schadt, 2005) and widespread translation of non-coding RNAs (at least based on ribosome profiling data) (Ingolia et al., 2014), it is probably not rare for a random non-coding sequence to be spuriously transcribed and translated. For example, over 100 human pseudogenes were reported to be translated, but the vast majority of them are not under purifying selection at the protein level (Xu and Zhang 2015). If one starts to call all such sequences as *de novo* genes, *de novo* gene birth rate is expected to be high, even if only a tiny fraction of them are functional. The real question is the birth rate of *de novo* genes that have selected functions. It is thus imperative to require the fourth criterion (natural selection) in identifying *de novo* genes. Nonetheless, we recognize that statistical tests of natural selection may be powerless for species-specific genes because only intraspecific polymorphism data may be used and because newly created *de novo* genes may be short. Thus, it appears that a more productive approach to estimating the rate of *de novo* gene birth is to identify *de novo* genes that arose in the common ansctor of a few closely related species such as that of *S. cerevisiae* and *S. paradoxus* rather than in *S. cerevisiae*. While Carvunis et al. and this

study focused on protein-coding genes, non-coding RNAs may also play important biological functions. It is possible that the larger part of genetic novelty in evolution is in the aspect of non-coding RNA genes. When searching for *de novo* genes in the future, it may be beneficial to expand the scope of "gene" to include this group.

In conclusion, it is clear that *de novo* gene birth plays some role in the formation of new genes in yeast, given previously identified cases. However, compared with gene duplication, the relative contribution of *de novo* gene birth to new genes is minor. Moving forward, evidence for *de novo* gene birth will need to be evaluated gene by gene based on the criteria mentioned rather than in aggregate, because current genomic studies for these trends are insufficient and confounded by phylostratigraphic error.

References

Abrusán, G. (2013). Integration of new genes into cellular networks, and their structural maturation. *Genetics*, *195*(4), 1407–17. http://doi.org/10.1534/genetics.113.152256

Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(6), 3695–3700. http://doi.org/10.1073/pnas.062526999

Begun, D. J., Lindfors, H. A., Kern, A. D., & Jones, C. D. (2007). Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. *Genetics*, *176*(2), 1131–1137. http://doi.org/10.1534/genetics.106.069245

Brown, C. A., Murray, A. W., & Verstrepen, K. J. (2010). Rapid expansion and functional

divergence of subtelomeric gene families in yeasts. *Current Biology*, *20*(10), 895–903. http://doi.org/10.1016/j.cub.2010.04.027

Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008). De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics*, *179*, 487–496. http://doi.org/10.1534/genetics.107.084491

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*, 370–374. http://doi.org/10.1038/nature11184

Casola, C., Conant, G. C., & Hahn, M. W. (2012). Very low rate of gene conversion in the yeast genome. *Molecular Biology and Evolution*, *29*(12), 3817–3826. http://doi.org/10.1093/molbev/mss192

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., … Wong, E. D. (2012). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, *40*, 700–705. http://doi.org/10.1093/nar/gkr1029

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., … MacCallum, I. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, *450*, 203–18. http://doi.org/10.1038/nature06341

Demerec, M. (1933). What is a gene? *The Journal of Heredity*, *24*, 368–378.

Domazet-Lošo, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics : TIG*, *23*(11), 531–3. http://doi.org/10.1016/j.tig.2007.07.007

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. http://doi.org/10.1093/nar/gkh340

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. *Molecular Biology and Evolution, 23*(1), 1–3. http://doi.org/10.1093/molbev/msj006

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., … Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Human Molecular Genetics*, 1–13. http://doi.org/10.1093/hmg/ddu309

Fisk, D. G., Ball, C. A., Dolinski, K., Engel, S. R., Hong, E. L., Issel-Tarver, L., … Cherry, J. M. (2006). Saccharomyces cerevisiae S288C genome annotation: a working hypothesis. *Yeast, 23*, 857–865. http://doi.org/10.1002/yea

Fitch, W. M. (1971). Rate of change of concomitantly variable codons. *Journal of Molecular Evolution, 1*, 84–96. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4377447

Gao, L.-Z., & Innan, H. (2004). Very low gene duplication rate in the yeast genome. *Science, 306*, 1367–1370. http://doi.org/10.1126/science.1102033

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., … Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research, 17*, 669–681. http://doi.org/10.1101/gr.6339607

Hahn, M. W., Bie, T. De, Stajich, J. E., Nguyen, C., & Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research, 15*, 1153–1160. http://doi.org/10.1101/gr.3567505

Hattori, M. (2005). Finishing the euchromatic sequence of the human genome. *Nature, 50*(2), 162–168. http://doi.org/10.1038/nature03001

He, X., & Zhang, J. (2005). Gene complexity and gene duplicability. *Current Biology*, *15*(11), 1016–1021. http://doi.org/10.1016/j.cub.2005.04.035

Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L., & Hedges, S. B. (2001). Molecular evidence for the early colonization of land by fungi and plants. *Science*, *293*, 1129–1133. http://doi.org/10.1126/science.1061457

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, *22*(23), 2971–2. http://doi.org/10.1093/bioinformatics/btl505

Heinen, T. J., Staubach, F., Häming, D., & Tautz, D. (2009). Emergence of a new gene from an intergenic region. *Current Biology*, *19*(18), 1527–1531. http://doi.org/10.1016/j.cub.2009.07.049

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., … Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Reports*, *8*, 1365–1379. http://doi.org/10.1016/j.celrep.2014.07.045

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, *324*, 218–223. http://doi.org/10.1126/science.1168978

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*, 1161–1166. http://doi.org/10.1126/science.860134

Johnson, J. M., Edwards, S., Shoemaker, D., & Schadt, E. E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics*, *21*(2), 93–102. http://doi.org/10.1016/j.tig.2004.12.009

Kaessmann, H., Vinckenbosch, N., & Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics*, *10*(1), 19–31. http://doi.org/10.1038/nrg2487

Knowles, D. G., & McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, 1–9. http://doi.org/10.1101/gr.095026.109

Koonin, E. V, Makarova, K. S., & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology*, *55*, 709–742.

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., & Begun, D. J. (2006). Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 9935–9. http://doi.org/10.1073/pnas.0509809103

Li, C.-Y., Zhang, Y., Wang, Z., Zhang, Y., Cao, C., Zhang, P.-W., … Wei, L. (2010). A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Computational Biology*, *6*(3). http://doi.org/10.1371/journal.pcbi.1000734

Li, D., Dong, Y., Jiang, Y., Jiang, H., Cai, J., & Wang, W. (2010). A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Research*, *20*(4), 408–420. http://doi.org/10.1038/cr.2010.31

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. a, … Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature*, *458*(7236), 337–341. http://doi.org/10.1038/nature07743

Long, M., Betrán, E., Thornton, K., & Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, *4*(11), 865–75. http://doi.org/10.1038/nrg1204

Misra, S., Crosby, M. A., Mungall, C. J., Matthews, B. B., Campbell, K. S., Hradecky, P., …

Lewis, S. E. (2002). Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biology*, *3*(12), 1–22. http://doi.org/10.1186/gb-2002-3-12-research0083

Miyasaka, H., Kanai, S., Tanaka, S., Akiyama, H., & Hirano, M. (2002). Statistical analysis of the relationship between translation initiation AUG context and gene expression level in humans. *Bioscience, Biotechnology, and Biochemistry*, *66*(3), 667–669. http://doi.org/10.1271/bbb.66.667

Moyers, B. A., & Zhang, J. (2014). Phylostratigraphic bias creates spurious patterns of genome evolution. *Molecular Biology and Evolution*, *32*(1), 258–267. http://doi.org/10.1093/molbev/msu286

Nei, M. (1969). Gene duplication and nucleotide substitution in evolution. *Nature*, *224*, 177–8.

Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, *14*(117). http://doi.org/10.1186/1471-2164-14-117

Ohno, S. (1970). *Evolution by gene duplication*. Berlin: Springer-Verlag.

Pál, C., Papp, B., & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, *37*(12), 1372–1375. http://doi.org/10.1038/ng1686

Pavlidis, P., Jensen, J. D., Stephan, W., & Stamatakis, A. (2012). A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, *29*(10), 3237–48. http://doi.org/10.1093/molbev/mss136

Penny, D., McComish, B. J., Charleston, M. A., & Hendy, M. D. (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of Molecular Evolution*, *53*(6), 711–23. http://doi.org/10.1007/s002390010258

Qian, W., & Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. *Genome Research*, *24*(8), 1356–1362. http://doi.org/10.1101/gr.172098.114

Scannell, D. R., Zill, O. A., Rokas, A., Payen, C., Dunham, M. J., Eisen, M. B., … Hittinger, C. T. (2011). The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the saccharomyces sensu stricto genus. *G3*, *1*(1), 11–25. http://doi.org/10.1534/g3.111.000273

Schmidt, H. A., Strimmer, K., Vingron, M., & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, *18*(3), 502–504. http://doi.org/10.1093/bioinformatics/18.3.502

Sharp, P. M., & Li, W.-H. (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295.

Souciet, J.-L., Dujon, B., & Gaillardin, C. (2009). Comparative genomics of protoploid Saccharomycetaceae. *Genome Research*, *19*(10), 1696–1709. http://doi.org/10.1101/gr.091546.109

Stoye, J., Evers, D., & Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics*, *14*(2), 157–163.

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, *12*(10), 692–702. http://doi.org/10.1038/nrg3053

Wapinski, I., Pfeffer, A., Friedman, N., & Regev, A. (2007). Natural history and evolutionary

principles of gene duplication in fungi. *Nature*, *449*(7158), 54–61. http://doi.org/10.1038/nature06107

Wei, X., & Zhang, J. (2015). A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biology and Evolution*, *7*(1), 381–390. http://doi.org/10.1093/gbe/evu294

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, *2*(May), 333–341.

Wong, E. S., Thybert, D., Schmitt, B. M., Stefflova, K., Odom, D. T., Flicek, P., … Campus, G. (2015). Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Research*, *25*, 167–178.

Wu, D. D., Irwin, D. M., & Zhang, Y. P. (2011). De novo origin of human protein-coding genes. *PLoS Genetics*, *7*(11). http://doi.org/10.1371/journal.pgen.1002379

Xiao, W., Liu, H., Li, Y., Li, X., Xu, C., Long, M., & Wang, S. (2009). A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PLoS ONE*, *4*(2), 1–12. http://doi.org/10.1371/journal.pone.0004603

Yang, Z., & Huang, J. (2011). De novo origin of new genes with introns in Plasmodium vivax. *FEBS Letters*, *585*(4), 641–644. http://doi.org/10.1016/j.febslet.2011.01.017

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, *18*(6), 292–298. http://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, J. (2013). Gene duplication. In J. Losos (Ed.), *The Princeton Guide to Evolution* (pp. 397–405). Princeton, New Jersey.

Zhang, J., Kumar, S., & Nei, M. (1997). Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Molecular Biology and Evolution*, *14*(12), 1335–1338.

http://doi.org/10.1080/13518040701205365

Zhang, J., & Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, *16*(7), 409–420. http://doi.org/10.1038/nrg3950

Zou, Z., & Zhang, J. (2015). Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? *Molecular Biology and Evolution*, 1–21.

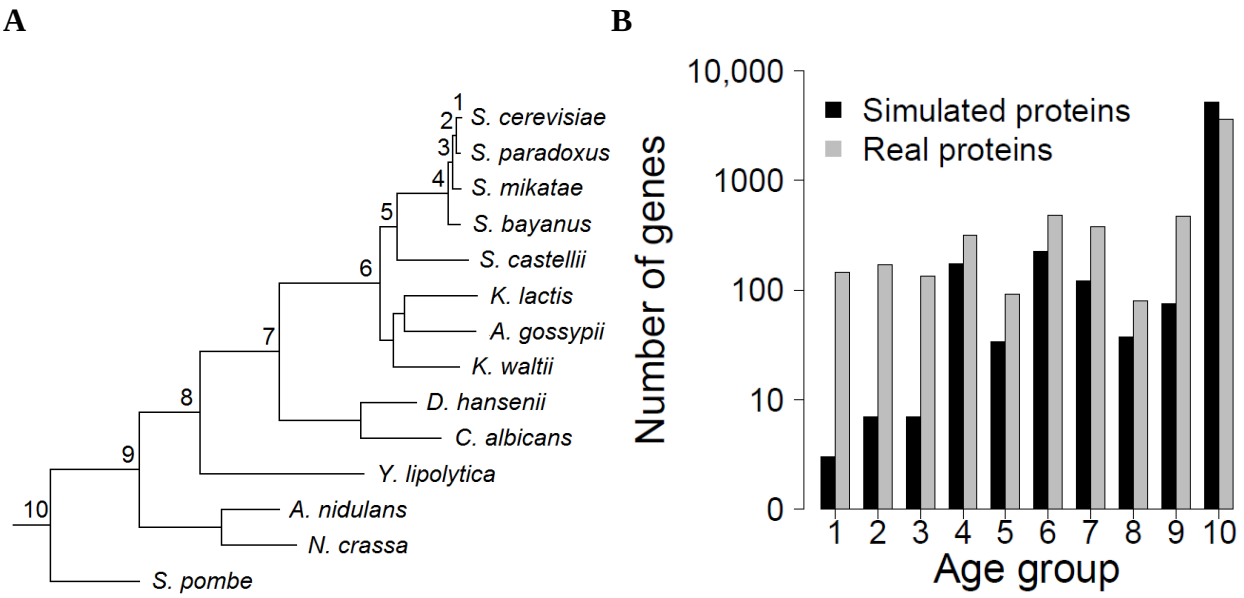

Figure 3- 1 Computer simulation for examining phylostratigraphic errors

(A) Tree used in the simulation of protein sequence evolution. The tree, including relative branch lengths, follows Wapinski et al. (2007). Node label refers to the age group corresponding to that node. (B) Numbers of genes estimated to belong to each age bin for real and simulated protein data. Numbers of genes in bins 1-10 for simulated protein data are 2, 6, 6, 171, 33, 222, 119, 36, 74, and 5209, respectively. Numbers of genes in bins 1-10 for real data, as provided by Carvunis et al., are 143, 169, 133, 314, 90, 476, 381, 78, 469, and 3625, respectively. Carvunis et al. arbitrarily assigned 107,425 smORFs to bin 0, which is not shown here.

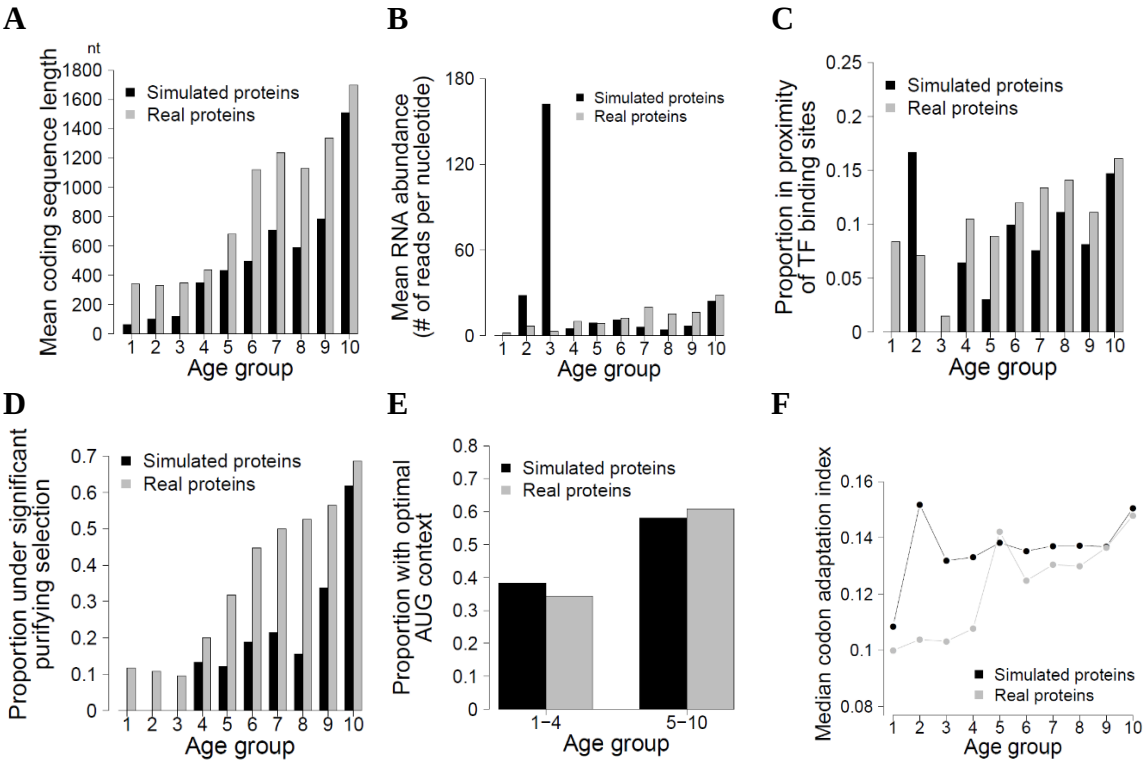

Figure 3- 2 Age distributions of six gene properties

(A) Average coding sequence length of genes in each age bin. Interestingly, although the same lengths are used for the real and simulated proteins, mean length is lower for simulated than real proteins in each bin. This is an example of Simpson's paradox in statistics and is not due to mistakes in our analysis. (B) Mean expression level of genes in each age bin. (C) Proportion of genes having a transcription factor (TF) binding site within 200 bp of the translation start site for each age bin. (D) Proportion of genes under purifying selection for each age bin. (E) Proportion of genes with optimal AUG context for each age bin. (F) Median codon adaptation index (CAI) for each age bin.

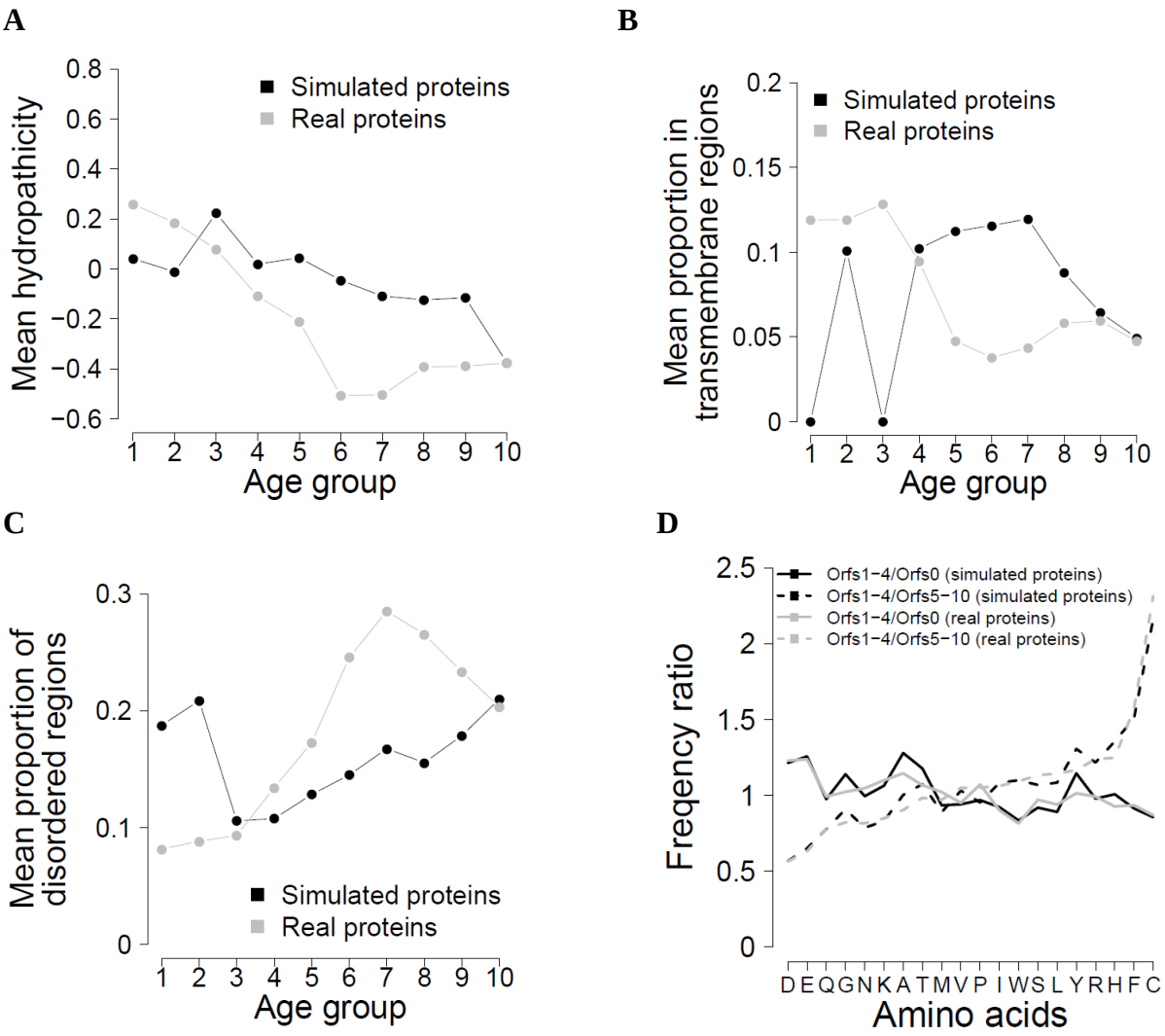

Figure 3- 3 Age distributions of four additional gene properties

(A) Mean hydropathicity value for each age bin. (B) Mean proportion of transmembrane regions for each age bin. (C) Mean proportion of disordered regions for each age bin. (D) Amino acid frequency ratios between age groups.

Comparison	ORF length	RNA abundance	Proximity of TF binding sites or not	Codon adaptation index	Purifying selection or not	Optimal AUG Context
Age groups 0-10						
Real proteins	0.31**	0.27**	0.11**	0.12**	0.45**	0.14**
Simulated proteins	0.31**	0.27**	0.11**	0.12**	0.45**	0.14**
Simulated proteins (ass paralog ages)	0.31**	0.27**	0.11**	0.12**	0.45**	0.14**
Age groups 1-10						
Real proteins	0.39**	0.26**	0.08*	0.31**	0.32**	0.13**
Simulated proteins	0.33**	0.26**	0.06*	0.21**	0.27**	0.12**
Simulated proteins (assuming oldest paralog a	0.31**	0.21**	0.04*	0.22**	0.26**	0.13**

Table 3- 1 Correlations (Kendall's tau) between estimated gene age and various gene properties for real and simulated proteins

*, $P < 0.05$; **, $P <$ 1E-16; Note that two analyses ensure that these trends are not due to the dubious age group 0, i.e. smORFs. See main text for further explanation.

Table 3- 2 Correlations (Kendall's tau) between estimated gene age and gene properties purported to reflect genetic integration or protein structural maturation

	Real Proteins	Simulated Proteins
Genetic coregulation	0.05*	0.06*
% in alpha helices	0.04*	-0.01
% in beta sheets	-0.08*	-0.11**
Aggregation propensity	-0.14**	-0.15**
Protein-protein interactions	0.22**	0.11**
Genetic interactions	0.14**	0.08*
Magnitude of epistasis	0.13**	0.08*
Feed-forward loops	0.02	0.03*
Number of transcription factors	0.02*	0.03*

*, $P < 0.05$; **, $P <$ 1E-16.

Table 3- 3 Reexamining purported S. cerevisiae-specific selected genes

Gene	Age based on full sequence	Non-overlapped length in nucleotides (full length)	No. of synonymous polymorphisms in non-overlapped region	No. of non-synonymous polymorphisms in non-overlapped region	*P*-value*
YBR232C	6	55 (360)	1	0	0.29
YCL046W	2	58 (324)	0	0	1.00
YDR537C	7	47 (606)	0	2	0.57
YER087C-A	7	62 (552)	0	0	1.00
YFL013W-A	5	53 (804)	1	1	1.00
YGL152C	6	71 (678)	2	2	0.58
YHL030W-A	9	49 (462)	0	2	0.57
YIL071W-A	6	111 (477)	0	0	1.00
YLR232W	9	58 (348)	1	2	1.00
YLR358C	6	50 (564)	0	1	1.00
YNL105W	10	88 (429)	0	0	1.00
YNL109W	8	50 (546)	0	0	1.00
YOL150C	8	62 (312)	0	0	1.00
YOL166C	1	339 (339)	3	3	0.37
YOR055W	6	55 (435)	0	0	1.00
YOR135C	10	91 (342)	1	0	0.30

*Based on two-tailed Fisher's exact test of the neutral hypothesis.

Table 3- 4 Correlations (Kendall's tau) between various gene properties and three properties known to bias phylostratigraphy

	Evolutionary rate	ORF length	Expression level
Transcription factor binding sites	-0.09*	0.02*	0.08*
Codon adaptation index	-0.33**	0.15**	0.26**
Optimal AUG context	-0.14**	0.05*	0.14**
Purifying selection	-0.22**	0.37**	0.09**
Mean hydropathicity	0.03*	-0.14**	-0.10**
Percent in disordered regions	0.05*	0.13**	0.01
Percent in transmembrane regions	0.07*	-0.07*	-0.07*
Genetic coregulation	-0.10**	0.03*	0.07*
Number of transcription factors	-0.07*	0.02*	0.02*
Feed-forward loops	-0.07*	0.02	0.03*
Percent alpha helices	-0.05*	-0.07*	0.09**
Percent beta sheets	-0.01	-0.22**	0.03*
Aggregation propensity	0.05*	-0.06*	-0.11**
Protein-protein interactions	-0.23**	0.11**	0.15**
Genetic interactions	-0.11**	0.11**	0.04*
Average magnitude of epistasis	-0.12**	0.05*	0.10**

*, $P < 0.05$; **, $P <$ 1E-16.

Chapter 4

Defense of the Role of Error in Phylostratigraphic Trends

Abstract

We have previously demonstrated that some phylostratigraphic trends can be attributable, at least partially, to homology detection error. Recently, these findings have been questioned, weaknesses have been suggested, and it has been argued that error plays virtually no role in creating spurious trends in phylostratigraphy. Here, we present results which contradict this argument. We also discuss problems with the theory of novel sequences and the future of phylostratigraphic analysis.

Introduction

Phylostratigraphy is a method for dating the origin of extant sequences, whether they have been generated through *de novo* gene birth or through some form of sequence divergence between two homologs. The method uses homology detection programs, typically the BLAST suite of algorithms, to identify homologs between query sequences and a target database, most often some subset of the NCBI non-redundant database which is sometimes combined with additional sequence data (Domazet-Lošo et al., 2016; Domazet-Lošo, Brajkovic, & Tautz, 2007; Domazet-Lošo & Tautz, 2003; Neme & Tautz, 2013). After identifying the most distant homolog as measured by divergence time, the date of a novel sequence's emergence is taken to be

approximately the time of the most recent common ancestor between the query species and target species of the hit.

It is clear that such a method, being based on sequence similarity, can create the appearance of a novel sequence even under a model of general divergence. It has been argued (Domazet-Lošo et al., 2016, 2007; Domazet-Lošo & Tautz, 2003) that this method detects novel sequences which arise through a rapid shift in sequence space, presumably due to some novel functional requirement. This can most easily be understood in the example of a duplication-divergence model, in which a given gene is duplicated and undergoes a short period of rapid evolution followed by a subsequent slowing of the evolutionary rate (Pegueroles, Laurie, & Alba, 2013), though this is not the only example when such a burst of evolution might occur. However, the method of phylostratigraphy itself cannot say anything about the existence of such bursts, only the sequence similarity of different genes. It has previously been demonstrated that no such burst of evolution is necessary for a sequence to appear to be novel in a recent node, despite being much older than that node (Albà & Castresana, 2007; Elhaik, Sabath, & Graur, 2006; Moyers & Zhang, 2015, 2016). Studies suggest that this mechanism for apparent novel sequence emergence, which we refer to as homology detection error, occurs in 5% to 14% of genes. However, because these methods have required the existence of some conservation, this is likely an underestimate of the rate.

Two of these studies suggested that because homology detection error is biased it can reproduce phylostratigraphic trends (Moyers & Zhang, 2015, 2016). If this is true, then the relative contribution of homology detection error and real biological signal to any phylostratigraphic

trend must be determined. These assertions have received several criticisms from well-established phylostratigraphic researchers (Domazet-Lošo et al., 2016). Several particular claims were leveled at the previous work of Moyers and Zhang, including: (1) using real sequences, real rates, and real heterogeneity patterns as starting sequences for the simulations is circular, and will by necessity recreate phylostratigraphic trends, (2) Associating gene features which are not simulated is inappropriate and circular, and cannot but reproduce known phylostratigraphic trends, (3) Homology detection error is virtually non-existent in some contexts, and if trends are robust in these contexts it promotes the efficacy of phylostratigraphic findings, (4) some parameters used in prior simulations, particularly those of covariation, are unrealistic, and (5) even in spite of all of these objections, when error-prone genes are removed from phylostratigraphic studies, the results remain unchanged.

Here, we respond to these criticisms. We first investigate homology detection error under randomization of various properties. We further reanalyze the data of Domazet-Loso *et al*. under more appropriate constraints. We demonstrate that error-prone genes do contribute significantly to homology detection error. Finally, we offer several important questions and concerns for the current framing of "novel sequences" and their biological meaning. We hope that this dialogue can continue, and phylostratigraphic theory and analysis can be further refined and improved.

Methods

Randomization of evolutionary properties

We used the ROSE files for the 5217 human proteins described in Moyers and Zhang (2015). For each of sequence content, evolutionary rate, and rate heterogeneity patterns, we randomized

the properties either individually or combined. To randomize sequence content, we shuffled the order of amino acids in each gene. To randomize average evolutionary rate, we shuffled the guide trees among genes. These guide trees provide the branch lengths that ROSE uses to determine how many substitutions and indels occur throughout evolution. To randomize rate heterogeneity patterns, we concatenated the relative rates of all proteins, and then sampled contiguous strings of this vector of appropriate size to assign to each protein.

Simulation of Evolution

We simulated sequence evolution using ROSE (Stoye, Evers, & Meyer, 1998), which allows 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 4942 human sequences, similar to the methodology described in Moyers and Zhang 2016. For each protein in all simulations, we simulated evolution using a JTT-f matrix with observed amino acid frequencies from the alignment.

Phylostratigraphy of simulated sequences

Phylostratigraphy was performed using default BLASTP parameters with an e-value of 0.001. The collection of simulated human sequences as the query and the sequences of all other simulated species as the target.

Human disease data

For data on human disease, we used the list of disease genes described in Moyers and Zhang 2015. We acquired phylostratigraphic ages for human genes from Domazet-Loso and Tautz 2008 (Domazet-Lošo & Tautz, 2008). We acquired error-prone status of human genes from Moyers and Zhang 2015.

Drosophila developmental data

We used the developmental expression status of genes in *Drosophila melanogaster* and phylostratigraphic ages for drosophila genes from Tautz and Colleagues (2016). We acquired error-prone status of drosophila genes from Moyers and Zhang 2015.

Statistical analyses

All statistical analyses were performed using R version 3.2.3.

Results

Phylostratigraphy with randomized evolutionary properties

In the work of Tautz and colleagues, they suggested that sequence evolution parameters should be randomized, generated *in silico* rather than taken from extant sequences. This was highlighted as a fundamental error in our simulations. We therefore began by randomizing evolutionary properties for our simulations, and performing phylostratigraphy (Figure 4-1).

We find that when sequence is randomized, error increases slightly but significantly (mean number of error prone genes increases from 474.78 to 500.22, p=4641E-9, n=9, t.test). This may be due to some degree of paralogy between the incipient human sequences which survived the simulation. Because we began the simulation of each genes evolution with the extant human sequence, rather than first randomizing sequence content, if two human genes had any paralogy between them hey maintain some sequence similarity throughout the simulation of evolution. This sequence similarity will slightly increase the chances of finding a homolog during phylostratigraphy, because there are a greater number of potential targets. Upon randomizing sequences, this paralogy was lost. Therefore, on this measure, the randomized assignment of evolutionary properties increases error and makes its problematic contributions to phylostratigraphy worse.

For all other randomizations, error was decreased. The greatest decrease in error was when randomizing the relative rate of proteins. This observation caused us to question whether or not we were observing a decrease in error because we were making the simulation less realistic. Because it is known that certain gene properties are correlated, we investigated the relationships between length, rate, and longest block of conserved sites in each of our simulations (Table 4-1). Unsurprisingly, we find that in the base and randomized sequence simulations, there are significant associations between length and evolutionary rate, length and longest block of conserved sites, and rate and longest block of conserved sites. However, whenever a given property is randomized, its association with other properties is either destroyed or reduced (Table 4-1). This destruction of associations is what influences error. While length is negatively correlated with error, a slowly-evolving short protein will have less error than a quickly-evolving

short protein. Similarly, a slowly-evolving long protein will have less error than a quickly-evolving long protein. Because these and other features are correlated in real sequences, randomization of sequence parameters gives unrealistic estimates of actual homology detection error.

These findings emphasize the need to retain real sequence evolutionary parameters when simulating and estimating error, as opposed to "randomization" of these parameters, as suggested by Tautz and colleagues. The idea that making simulated proteins less like real proteins will give more realistic views of the influence of error is both counterintuitive and wishful thinking. By destroying these sequence parameter associations, one destroys the relationship between simulations and reality. While this will produce lower error rates, these lower error rate estimates are unrealistic, and breed false complacency.

Error Influences Phylostratigraphic Findings

In the work of Domazet-Loso *et al*., they claimed that the effects of error are not influencing phylostratigraphic trends. However, in doing so, they only removed those genes which we both (1) simulated and (2) found to be subject to error. This methodology makes the assumption that those genes which we were unable to simulate are inherently not error-prone. We strongly disagree with that assumption. Our simulation was based on genes for which there was a homolog conserved in several species (5 to 12 species) diverged many millions of years (up to 92MY). Those genes which were not so conserved might be so for two reasons. Either they are truly young, or they have lost detectable homology in the species of interest. In either case, these genes are expected to have at least two properties: they are expected to be short, and fast-

evolving (Carvunis et al., 2012; Moyers & Zhang, 2015). These properties are expected to be associated with homology detection error. Therefore, these genes are likely to be enriched with error-prone genes. Therefore, in assessing phylostratigraphic trends in the absence of homology detection error, it is inappropriate to include genes for which there was insufficient information to simulate evolution.

In order to assess this problem, we reanalyzed the data of Domazet-Loso et al (2016) restricting to only genes which were both simulated and found to be non-error-prone. Tautz and colleagues specifically reanalyzed three arguments from our previous publications: disease-prone status of human genes, drosophila gene expression during development, and trends of sequence properties with age in yeast.

We began by reanalyzing human disease genes. Previously, Domazet-Loso and Tautz had demonstrated that the number of genes in each phylostratum was not correlated with age, but that the number of genes associated with disease was correlated with phylostratum, with older phylostrata having more disease-associated genes (Domazet-Lošo & Tautz, 2008). We had demonstrated that homology detection error alone could produce a correlation between age and the proportion of genes which were associated with disease (Moyers & Zhang, 2015). In their recent paper, Tautz and colleagues argued that our analysis was not the same as theirs, and was therefore an inappropriate comparison, and further demonstrated that when error-prone genes were removed from the dataset their trend was unaffected. Because the central point of Domazet-Loso and Tautz (2008) was that disease genes have an ancient origin in humans, we plotted the proportion of genes which were disease-associated as a function of age as reported by

Domazet-Loso and Tautz (Figure 4-2A, black line). We regard this as essentially the same experiment and, as expected from the original paper, there was a significant correlation between the age and proportion of disease genes (Rho=-0.85, p<2.2E-316). However, when we restricted this gene set to those genes which were both simulated and found to be non-error-prone, we found that this correlation disappeared (Rho=-0.29, p=0.28), despite having a significant number of genes remaining (4587 genes total, 565 disease-associated genes, compared to 22845 and 1760 prior to correction). This suggests that the majority of the trend is found in genes which are fast-evolving and prone to losing detectable homologs. Indeed, when we instead restrict to error-prone genes (those which we could not simulate plus those which we simulated but found to be error-prone), the trends between all genes and error-prone genes match almost exactly (Figure 4-2B). This trend holds true when we instead plot the absolute number of genes in each phylostratum and the number of disease genes in each phylostratum (Figure B-1A, B-1B, B-1C). When error is not accounted for, there is a significant correlation between age and number in disease genes but not for absolute number of genes. When genes are restricted to non-error-prone sets, both sets are significantly correlated with age. When genes are restricted to error-prone sets, the trends are as with all genes.

We next reanalyzed drosophila developmental data. Previously, Domazet-Loso and Tautz had demonstrated that genes dated to the emergence of Eukaryota or younger were overrepresented in ectodermal expression during development (Domazet-Lošo et al., 2007). In our previous analysis, we had demonstrated that homology detection error alone could produce significant peaks in this kind of analysis (Moyers & Zhang, 2015). In their recent critique of our work, Tautz and Colleagues correctly identified that we had made a statistical error in this analysis, and

none of our findings for this analysis were significant. It is therefore interesting to reanalyze their work, because presumably there should be virtually no effect when removing error-prone genes, as our simulation did not suggest that there was a bias in the error. We used data as provided by Domazet-Loso *et al*. to reconstruct the expression of drosophila genes during development. We first ensured that we replicated their peaks and significance using all data (Figure B-2A), noting that we successfully reconstructed peaks and significance values. We also noted, though, that the numbers now made public by Tautz and colleagues suggested that they were counting genes multiple times during development. This is true—if a gene was expressed in different regions of an ectodermal tissue, or in the same region but different timepoints, the gene was counted as "expressed" multiple times, providing numbers far greater than the actual number of genes analyzed, and greatly inflating the power of their analysis. It is more appropriate to consider each gene as being either expressed or not expressed during development in a particular tissue. We therefore reanalyzed their data under this methodology (Figure 4-3A). We find that while some trends are similar to their initial publication, significance of their trends are greatly deflated.

However, this criticism is separate from the consideration of whether or not homology detection error contributes to phylostratigraphic trends. To this end, we reanalyzed their data using only those genes whose evolution was possible to simulate, and which were not found to be error-prone. We found that while some peaks and significance changed, the broad strokes of the analysis was not largely affected (Figure 4-3B). This was also true using the original methodology of Tautz and colleagues (Figure B-2B). This emphasizes a point that we made in our initial criticisms of phylostratigraphy (Moyers and Zhang 2015), that not all

phylostratigraphic trends are attributable to error, but some are. It also demonstrates that when simulation of genes does not produce significant trends, it is more likely that real phylostratigraphic observations can be trusted.

Finally, we reanalyzed the data presented in Carvunis *et al*. In their original work, Carvunis and colleagues analyzed various sequence properties of yeast genes and properties of these genes' products and surrounding sequences, such as the proximity of transcription factor binding sites, expression levels, etc. (Carvunis et al., 2012). In our previous analysis, we demonstrated that homology detection error could reproduce their statistical trends quite closely (Moyers & Zhang, 2016). In addition to analyzing their data with error-prone genes removed, they commented that our association of gene age with various properties was inappropriate. We disagree, for reasons discussed elsewhere in this paper. In this analysis, we removed the genes for which we had randomly assigned evolutionary rate and rate heterogeneity parameters. This is because, as mentioned and observed, it destroys the observable associations between evolutionary parameters, and was therefore a poor method to simulate genes (Figure 4-1, Table 4-1). Additionally, these genes were necessarily faster-evolving and had shorter conserved blocks in reality compared to simulation, as predicted by their phylostratigraphic theory or homology detection error.

We find that restricting to non-error-prone gene sets tends to reduce the apparent effect size and significance of many apparent trends (Table 4-2). However, it is important to note that these trends still exist, and are significant. Therefore, there may yet be phylostratigraphic support for the proto-gene model of gene birth. However, when observing only those genes which are error-

prone (i.e. genes for which evolutionary parameters could not be inferred, or genes which were simulated and found to be error-prone), we find that the effect sizes and significance are generally greatly increased (Table 4-2). Furthermore, in all associations except for proximity to a transcription factor binding site the trend observed in all genes is much closer to the trend observed for error-prone genes. This is true despite the fact that there are over 3.6 times as many genes which are non-error-prone as opposed to error-prone genes. Error-prone genes in this case have a disproportionate influence on phylostratigraphic trends. This further emphasizes that phylostratigraphy is biased, and that error-prone genes influene observed trends in ways that cannot be ignored.

Discussion

We have here thoroughly responded to the criticisms of our work by Domazet-Loso *et al* (2016). We have demonstrated that error has non-negligible impact on phylostratigraphic trends, and that, though sometimes the minority, error-prone genes disproportionately impact trends. For clearer phylostratigraphic findings and more accurate evolutionary theory, phylostratigraphy must be performed in an error-aware context.

Efficacy of simulations

It has been suggested that our previous simulations were inappropriate, as they associated too many real genetic properties with the simulated genes (Domazet-Lošo et al., 2016). This is a confusion assertion, as a major point of Alba and Castresanna in 2007 was that one needs to respect these more realistic gene qualities of sequence content, length, evolutionary rate, and rate

heterogeneity patterns. Nonetheless, at the urging of us to use "simulated" properties which have some correspondence to real genetic properties, we have run new simulations. We do find that randomization of evolutionary parameters can reduce error, but only because known associations which, when combined, compound error, are broken in such randomizations. Short, fast-evolving genes have greater error rates than short-slow-evolving genes, and destroying observable associations between length and evolutionary rate destroys these real trends in homology detection error. In short, less realistic simulations produce lower and less realistic estimates of error.

Indeed, we expect the percentage of error to be higher in more realistic simulations. Even ignoring potential impacts of covariation or rate heterogeneity among branches, it is expected that those genes which we are unable to simulate are likely to have higher degrees of error. Such genes are conserved in fewer species. Under a model of homology detection error, this is likely because they are short, fast-evolving, and have short blocks of conserved sites. Therefore, simulations of these genes are likely to show that they are error-prone. If, instead, they are truly young, they are predicted by trends produced from phylostratigraphy to be short and fast-evolving. Thus, simulating them is likely to show that they are error-prone. In any case, those genes which are not represented in our simulations are likely to be more error-prone, and thus estimates of 5-15% are necessarily underestimates of the true influence of phylostratigraphic error. The only sequence property which seems appropriate to randomize is the sequence content, as it destroys any latent relationship between simulated sequence. Randomizing this property actually increases error.

It was also suggested that associating genetic properties with our simulations is inappropriate, because these simulations do not include models which incorporate these extra-sequence properties, such as AUG context, developmental expression, expression level, etc. This is an unfair criticism, as it suggests that the purpose of our simulation is to fully reproduce biological properties through a simulation of evolution. However, that is not the case. Our simulations are meant to determine the propensity for error that a given gene is subject to. It is therefore not the case that we are associating these biological features with our simulation, but we are measuring an evolutionary biological property—namely, error propensity—of existing genes. This requires an assessment of existing genes and their evolutionary properties without somehow randomizing these properties.

It may be argued that because such simulations are based on extant genetic properties, they cannot accurately assess the true propensity of error for a given gene, because estimates of evolutionary rate based on 90MY of conservation may suggest that a given gene is evolving quickly, whereas such fast evolution may only be the result of a temporary burst of evolution. Similarly, our simulations may suggest that a gene evolves slowly based on this time range, whereas it is possible that the rate of evolution for this sequence may have decreased and was faster in the past. These are reasonable concerns, but no reason to entirely ignore the error propensity of genes. There are numerous evolutionary traits which we are unable to investigate in many contexts, including the occurrence of temporary bursts in evolutionary rates. But we still make inferences based on extant properties until such additional properties can be investigated. Additionally, by requiring a moderate amount of conservation and inferring average evolutionary rate over that time, we can partially account for such changes in

evolutionary rate. In total, we expect that these simulations generally underestimate propensity for error, as the species from which evolutionary rate information is inferred tend to be slower-evolving than many clades, such as Bacterial or Fungal clades.

The definition of novel sequences

Tautz and colleagues (2016) have spent substantial space clarifying their meaning of "novel sequences" in phylostratigraphy in their recent criticism of our work. This kind of clarification is of course paramount to the discussion. They suggest that a loss of detectable homology (presumably only through the work-horse of phylostratigraphy, BLAST) constitutes the emergence of a new sequence. They note two primary ways in which a new sequence arises: (1) through a rapid burst of evolution or sudden shift in sequence space, presumably due to a change in functional constraints, and (2) through the *de novo* birth of a gene. While this is a reasonable model, like any nascent model it requires substantial revision to be intelligible.

First, this separates the definition of novel sequences from that of historical homology—i.e. that genetic homologs are those sequences which came from the same ancestral sequence. Under the model of phylostratigraphic novel sequences, novel sequences may indeed have historical homologs. Given that the purpose of BLAST and other homology detection tools is to attempt to recapitulate historical homology, it is not clear that the tool under use is appropriate for the purpose it is being set to.

Second, we demonstrate that, if homology detection programs are to be the measure of novel sequence emergence, there is clearly a third mechanism of the emergence of novel sequences:

false negative error due to steady sequence divergence. We characterize this as "error" as opposed to novel sequence emergence as it is clear that no new sequences or sudden shifts in sequence space are necessary to miss a homolog. Therefore, any phylostratigraphic conclusions under current methodology must recognize steady sequence divergence as a third method for novel sequences to emerge. That is, they must be error-aware.

Third, it has been suggested that current phylostratigraphic methodology may be "too sensitive" (Domazet-Lošo et al., 2016). An example given by these researchers was the story of two historical homologs, one of which has undergone a rapid and temporary shift in sequence space, but for which BLASTP or another homology detection program identifies the historically homologous relationship due to a conserved domain between the proteins. This argument and example wholly undermines the proposed method for detecting novel sequences. It appears that the formal definition of a novel sequence is based on homology detection tools. However, the suggestion that when a homology detection program detects a true historical homology between two proteins that this is an error is in direct contradiction to the methodology. If this is truly a case in which phylostratigraphy should identify a novel sequence as opposed to two homologs, then researchers must propose a formal definition and measurement of novel sequences independent of homology detection tools like BLAST. If no such formal definition and measure of novel sequence exists, then phylostratigraphic researchers must accept a situation like that proposed above as a case where a novel sequence has indeed not emerged. One cannot rely on intuitive definitions of "novelty" in this case—a specific, numerical methodology has been proposed. Allowing ad-hoc acceptance or rejection of this numerical methodology based on intuitions about what is or isn't "novel" is hand-waving.

The Future of Phylostratigraphy Reconsidered

Despite its current problems, phylostratigraphy may be a useful technique for future researchers. However, its current methodology must change if it is to produce reliable results. In addition to the above discussion of clarifying the theoretical explanation of phylostratigraphy, we give here several recommendations.

Other homology detection methods must be investigated to identify those techniques which have the lowest rates of false positive and false negative error. This is a complex and computationally intensive topic. There are numerous homology detection programs such as PSIBLAST (Altschul et al., 1997), PHMMER (Söding, 2005), HMMER (Finn, Clements, & Eddy, 2011), the MEME suite of algorithms (Bailey et al., 2009), PSIPRED (Buchan, Minneci, Nugent, Bryson, & Jones, 2013), HHSEARCH (Söding, Biegert, & Lupas, 2005), and many other tools. Each of these tools and BLAST have several parameters which might be altered to produce more accurate results. This is further complicated by the fact that some tools cannot be reasonably or accurately assessed under our current simulation methodology, as they rely on structural and extra-sequence properties which are not a part of our simulation.

In addition to raw concerns about the amount of homology detection error occurring is the more important problem of biased error. If error exists, this may not be a major problem for phylostratigraphy if it is not biased with relation to biological properties. With BLAST at least we have thoroughly demonstrated that error is biased, depending on at least length, evolutionary rate, and rate heterogeneity parameters. There may be other as yet undetected biases, and there

is no reason to think that other homology detection tools which rely upon sequence similarity will not be biased. These tools should therefore be investigated to determine whether or not they produce biased phylostratigraphic trends.

Most importantly, unless and until the above two goals can be addressed and achieved, phylostratigraphy must be performed in an error-aware manner. More precise and effective methodology is needed to evaluate the error-prone status of genes. While prior studies have used far fewer genes than those in our methodology, and genes which are far older and therefore inappropriate for determining error rates (Albà & Castresana, 2007), our studies have greatly reduced this problem by incorporating orders of magnitude more proteins, and less conserved proteins (Moyers & Zhang, 2015, 2016). But they still restrict analyses to a relatively small number of genes (<1/4 of human proteins in this study). Methodology which can correctly assess the error-prone status of greater numbers of proteins will improve phylostratigraphic research and increase power to find meaningful results.

Finally, while phylostratigraphy is meant to identify novel sequences, it is unable to make statements about how these sequences emerged. The relative contributions of homology detection error, rapid divergence, and *de novo* gene birth must be more fully elucidated. It is quite possible that these three mechanisms undergo different evolutionary dynamics, and a greater understanding of their contributions will shed light on phylostratigraphic findings.

References

Albà, M. M., & Castresana, J. (2007). On homology searches by protein Blast and the

characterization of the age of genes. *BMC Evolutionary Biology*, *7*(53). http://doi.org/10.1186/1471-2148-7-53

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME SUITE : tools for motif discovery and searching. *Nucleic Acids Research*, *37*(May), 202–208. http://doi.org/10.1093/nar/gkp335

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., & Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research*, *41*(June), 349–357. http://doi.org/10.1093/nar/gkt381

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*, 370–374. http://doi.org/10.1038/nature11184

Domazet-Lošo, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics : TIG*, *23*(11), 531–3. http://doi.org/10.1016/j.tig.2007.07.007

Domazet-Lošo, T., Carvunis, A., Alba, M. M., Sestak, M. S., Bakarić, R., Neme, R., & Tautz, D. (2016). No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Biorxiv*, 1–24.

Domazet-Lošo, T., & Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila.

Genome Research, *13*, 2213–9. http://doi.org/10.1101/gr.1311003

Domazet-Lošo, T., & Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular Biology and Evolution*, *25*(12), 2699–707. http://doi.org/10.1093/molbev/msn214

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. *Molecular Biology and Evolution*, *23*(1), 1–3. http://doi.org/10.1093/molbev/msj006

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(SUPPL. 2), 29–37. http://doi.org/10.1093/nar/gkr367

Moyers, B. A., & Zhang, J. (2015). Phylostratigraphic bias creates spurious patterns of genome evolution. *Molecular Biology and Evolution*, *32*(1), 258–267. http://doi.org/10.1093/molbev/msu286

Moyers, B. A., & Zhang, J. (2016). Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Molecular Biology and Evolution*, *33*(5), 1245–1256. http://doi.org/10.1093/molbev/msw008

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, *14*(117). http://doi.org/10.1186/1471-2164-14-117

Pegueroles, C., Laurie, S., & Alba, M. M. (2013). Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology and Evolution*, *30*(8), 1830–1842. http://doi.org/10.1093/molbev/mst083

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, *21*(7), 951–960. http://doi.org/10.1093/bioinformatics/bti125

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(SUPPL. 2), 244–248. http://doi.org/10.1093/nar/gki408

Stoye, J., Evers, D., & Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics*, *14*(2), 157–163.

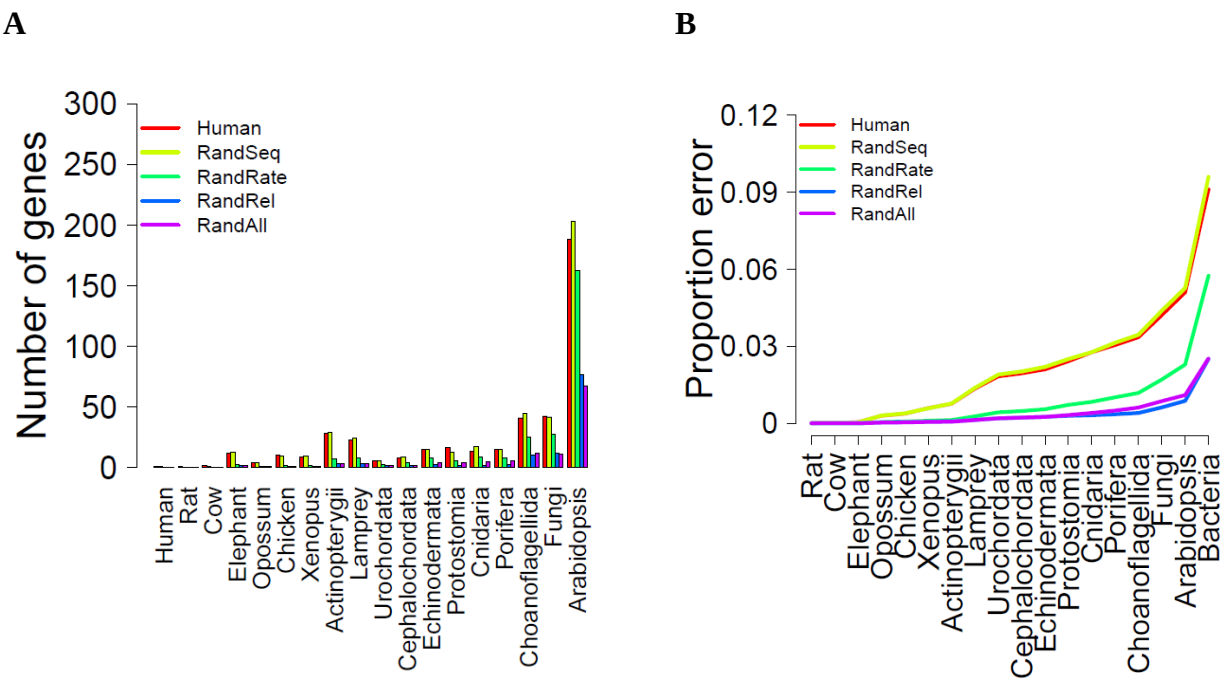

Figure 4- 1 Homology detection error under randomization of evolutionary properties

In both panels, red lines refer to the base simulation with no parameters randomized; yellow lines refer to the simulation with amino acid sequence randomized; green lines refer to the simulation with average evolutionary rate randomized; blue lines refer to the simulation with the relative rates of sites along the protein randomized; purple lines refer to the simulation with all three of these sequence properties randomized. (A) The number of genes which fall into each clade under various conditions. Note that a bacterial clade is not shown for the sake of greater resolution. (B) The percent of genes which miss a homolog at each age.

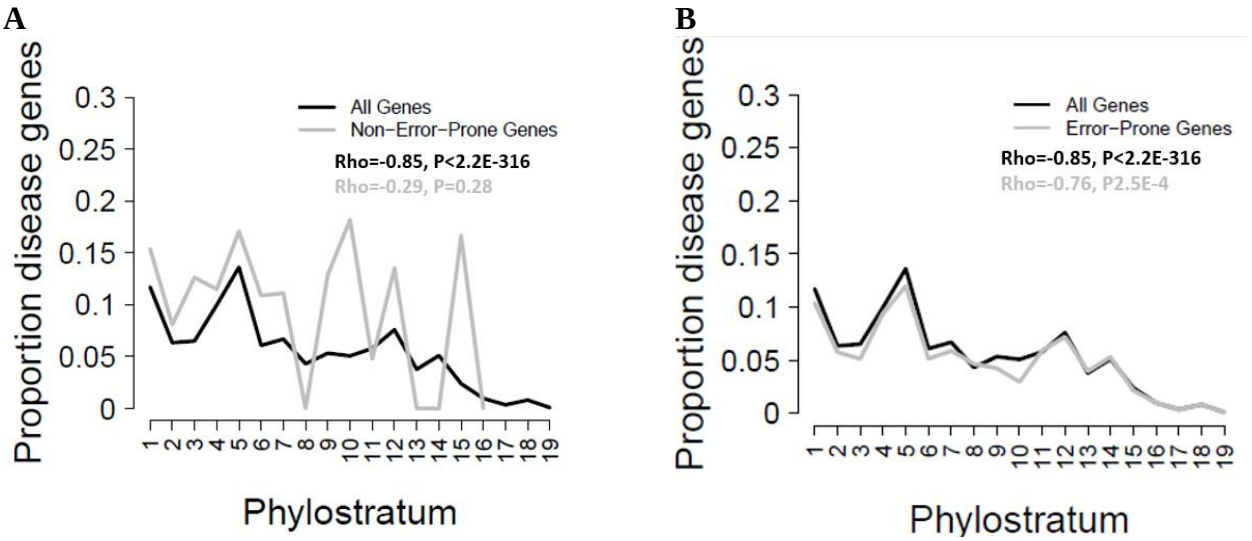

Figure 4- 2 Phylostratigraphic findings in human disease genes when restricting to certain gene sets

(A) The proportion of genes in each phylostratum which are disease-causing. When all genes are considered, there is a clear correlation which implies that the older a gene is the more likely it is to cause disease. However, when genes are restricted to those which are demonstrated to be non-error-prone (n = 4632), this correlation disappears. (B) As (A), but restricting only to those genes which were either not simulated or found to be error-prone (n=18258). Under this condition, both all genes and error-prone genes were found to have a significant correlation with age, implying that older genes are more likely to be disease-causing.

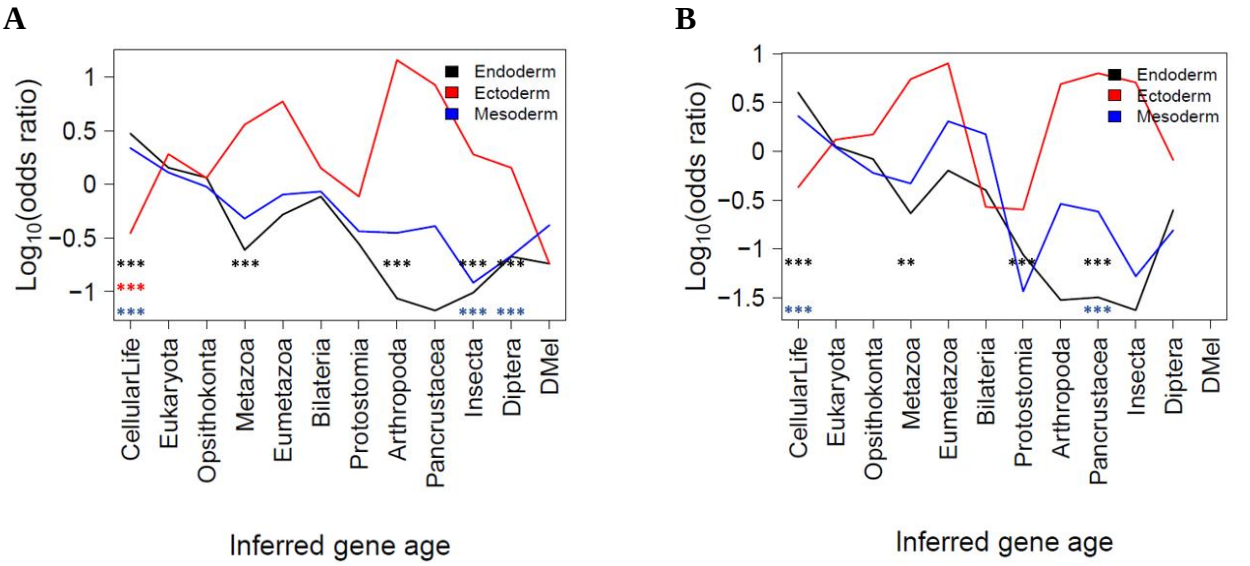

Figure 4- 3 Phylostratigraphic findings in drosophila developmental genes when restricting to certain gene sets

(A) Results when all genes in the dataset are considered, but considering each gene only once in the developmental dataset rather than counting genes multiple times based on expression at different time points or different regions of the same germ layer (n=4157). (B) Results as presented in (A), but when genes are restricted to only those genes for which evolution can be simulated and no error-prone status was found (n=1973).

Table 4- 1 Spearman's rho correlation between evolutionary properties under different randomizations

	Rate	Longest Block
Length		
Base	-0.034*	0.353***
Randomized Seq	-0.034*	0.353***
Randomized Rate	-0.014	0.353***
Randomized Sites	-0.034*	0.254**
Randomized All	-0.014	0.254**
Rate		
Base		-0.766***
Randomized Seq		-0.766***
Randomized Rate		-0.013
Randomized Sites		-0.004
Randomized All		-0.004

* P < 0.05, ** P < 1E-10, *** P < 1E-100

Table 4- 2 Kendall's Tau correlation between gene properties and age in non-error-prone and error-prone sets

	Length	RNA Abundance	Proximity to TFBS	CAI	Purifying Selection	Optimal AUG Context
All real ORFs (n = 5878)	0.386***	0.261***	0.077**	0.312***	0.316***	0.133***
Non-error-prone (n = 4620)	0.179***	0.093**	0.050*	0.208***	0.166***	0.045*
Only error-prone (n= 1258)	0.429***	0.163**	-0.002	0.324***	0.331***	0.212***

* P < 0.05, ** P < 1E-10, *** P < 1E-100

Chapter 5

Toward an Improved Phylostratigraphic Analysis

Abstract

We have previously demonstrated that some phylostratigraphic trends can be attributable, at least partially, to homology detection error. There has been a call for better methodology to reduce false negative error and a call to investigate the contributions of false positive error to such trends. Here, we perform a diverse set of simulations to further explore homology detection error and avenues for reducing it. We investigate both false positive and false negative error under BLASTP, PSIBLAST, PHMMER, HMMER, and GLAM2Scan. We further explore a large number of parameter sets for each program to determine if false negatives and false positives can be reduced compared to the default parameter sets. We generally find that using default BLASTP parameters for homology detection cannot be improved upon in a reasonably implementable way. In an attempt to explore other methods for reducing homology detection error, we explore some machine learning algorithms to identify error-prone genes, and find that such methods are neither accurate nor precise. We propose a simulation-based phylostratigraphic framework in which error is addressed and removed. We find that when error is appropriately accounted for, some phylostratigraphic trends disappear and some are reversed. Finally, we give recommendations for the future of phylostratigraphy.

Introduction

Phylostratigraphy is a method for identifying novel sequences, whether they have been generated through *de novo* gene birth or through sequence divergence between two homologs. The method uses homology detection programs, typically the BLAST suite of algorithms, to identify homologs between query sequences and a target database, most often some subset of the NCBI non-redundant database which is sometimes combined with additional sequence data (Domazet-Lošo et al., 2016; Domazet-Lošo, Brajkovic, & Tautz, 2007; Domazet-Lošo & Tautz, 2003; Neme & Tautz, 2013). After identifying the most distant homolog as measured by divergence time, the date of a novel sequence's emergence is taken to be approximately the time of the most recent common ancestor between the query species and target species of the hit.

Because the definition of novel sequences is based on sequence similarity, novel sequences can be a product of false negative error in homology detection. It has previously been demonstrated that this kind of error can cause a sequence to appear to be novel in a recent node, despite being much older than that node (Albà & Castresana, 2007; Elhaik, Sabath, & Graur, 2006; Moyers & Zhang, 2015, 2016). Studies suggest that this mechanism for apparent novel sequence emergence, which we refer to as homology detection error, occurs in 5% to 14% of genes. But even these estimates are based on the qualities of genes which have moderate to extreme conservation (60 million to 450 million years of conservation). It stands to reason that genes with less apparent conservation have qualities—a fast evolutionary rate or short sequence—which promote greater error (Moyers & Zhang, 2015).

Because this error is non-random, it can produce phylostratigraphic trends (Moyers & Zhang, 2015, 2016). By controlling for error, one may reduce the influence of this bias (Chapter 4).

There are at least four potential methods for reducing homology detection error in phylostratigraphy. First, there is an abundance of tools for homology detection. The BLASTP (Stephen F Altschul, Gish, Miller, Myers, & Lipman, 1990) algorithm is used most commonly in phylostratigraphy. But aside from BLAST there are a number of other tools including PSIBLAST (S F Altschul et al., 1997), PHMMER (Söding, 2005), HMMER (Finn, Clements, & Eddy, 2011), the MEME suite of algorithms (Bailey et al., 2009), PSIPRED (Buchan, Minneci, Nugent, Bryson, & Jones, 2013), HHSEARCH (Söding, Biegert, & Lupas, 2005), and many other tools. 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 our simulated sequences and identify an ideal set of parameters for each. We further assess the false positive and false negative rates of each when using ideal parameters. Aside from the precise homology detection rates, it is also important to determine whether or not these programs have the same biases as BLAST, so we assess the correlation between homology detection error and various sequence features to determine if any of the methods is unbiased.

Second, it is possible and has been suggested that phylostratigraphy is not subject to homology detection error under certain contexts, most notably among closely-related species (Domazet-Lošo et al., 2016). We perform two simulations using a tree consisting of primates and mouse to assess whether or not homology detection error exists in these contexts, and whether or not it is biased to preferentially create trends with apparent age.

Third, one could develop some algorithm for assessing, *a priori*, a sequence's propensity for homology detection error based on its sequence features. We investigate multiple machine

learning methods to determine if there is a sufficiently sensitive and precise method to identify error-prone sequences.

Fourth, one could assess the error-prone status of genes through direct simulation and remove any such genes from phylostratigraphic analyses. While previous researchers have claimed to do this (Domazet-Lošo et al., 2016) we demonstrated that their method for control is insufficient (chapter 4). Those genes which we are unable to simulate are likely to have properties which make them more error-prone (Carvunis et al., 2012; Moyers & Zhang, 2015; chapter 4).

Finally, through real and error-aware phylostratigraphic analysis, we demonstrate that some phylostratigraphic trends disappear under error-aware phylostratigraphy, while others reverse direction. We conclude with recommendations for error-aware phylostratigraphic methodology and comments on challenges for the future.

Methods

Sequence acquisition

We acquired 4942 human sequences with 1-to-1 orthologs in 14 mammalian species diverged approximately 90MYA (Hedges, Dudley, & Kumar, 2006) from OrthoMaM (Ranwez et al., 2007). The specific species in question were: *Homo, Pan, Gorilla, Pongo, Nomascus, Macaca, Callithrix, Tarsius, Otolemur, Microcebus, Rattus, Mus, Dipodomys, Cavia*. Separately, we acquired a full database of human protein sequences from Ensemble, current as of 20 September

2016, available at the following web address:

http://ftp.ensembl.org/pub/current_fasta/homo_sapiens/pep/.

Inferring evolutionary rate information

From the orthologs of 14 mammalian species, we used TreePuzzle (Schmidt, Strimmer, Vingron, & von Haeseler, 2002) to infer evolutionary rate information including average evolutionary rate and rate heterogeneity patterns of each of the 4942 human proteins. We used the JTT-f matrix (Jones, Taylor, & Thornton, 1992) with a discrete gamma model with 16 rate heterogeneity categories.

Simulated sequence properties

We created three sets of proteins for later simulation. In our first set, we assigned to the 4942 human proteins the exact evolutionary rate and rate heterogeneity parameters of the protein as determined by TreePuzzle, but shuffled the amino acid content of each protein so as to destroy any remaining parology between proteins, ensuring a set of truly unrelated sequences. This is referred to as our Base set.

In our second set, we randomly generated a set of 10,000 protein lengths by sampling the actual distribution of protein lengths found in the set of human proteins downloaded from Ensemble. Amino acid sequence was assigned randomly based upon the frequency of each amino acid in the 4942 human proteins downloaded from OrthoMam. We assigned evolutionary rate information and rate heterogeneity information using a sampling method similar to Moyers and

Zhang 2016. Briefly, for each of the 4942 proteins in our first set we multiplied the relative rate of each site by the absolute evolutionary rate of the protein. We then concatenated each of the 4942 evolutionary rate strings into a large ring structure. Then, for each of the 10,000 proteins, we sampled a continuous string of sites equal to the length of the protein in question, requiring that the sampled string not have all sites equal to the same rate. We then determined the average of this string as the average evolutionary rate of the protein, and we divided the string by the average rate to determine the rate heterogeneity pattern of the simulated protein. Thus, in this simulation we have created a set of proteins whose length, rates, and rate heterogeneities are simulated and independent. This is referred to as our Size Distribution set.

In our third set of proteins, we sought to investigate more extreme models of evolution. Because all rates were sampled from 4942 proteins with full conservation to a moderately old ancestor, it is highly likely that these evolutionary rates are not representative of the average evolutionary rates of all proteins. We therefore created a set of proteins with faster evolutionary rates by using the exact same methodology as described for our second set, but multiplying the average evolutionary rate by a factor of 5. This set represents a set of proteins with randomly and independently assigned lengths, evolutionary rates, and rate heterogeneity patterns, but under a more extreme model of evolution. This is referred to as our Size Distribution Fast set.

Construction of trees for simulation

We evolved our sets of proteins through two trees (Figure 5-1A, and Figure 5-3B). We simulated evolution of all three sets through the first tree, and the second and third sets through

the second tree. In both cases, the trees were constructed based on average divergence times as listed by TimeTree (Hedges et al., 2006).

Simulation of Evolution

We simulated sequence evolution using ROSE (Stoye, Evers, & Meyer, 1998), which allows 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 4942 human sequences, similar to the methodology described in Moyers and Zhang 2016. For each protein in all simulations, we simulated evolution using a JTT-f matrix with observed amino acid frequencies from the alignment.

Comparison of simulated and real genetic distances

We determined genetic distances between Human and Mouse sequences using TreePuzzle in both real and simulated sequences. Comparison was done by plotting the real versus simulated genetic distances for each of the 4942 proteins.

Phylostratigraphy of simulated sequences

Phylostratigraphy using simulations along our first tree (Figure 5-1A) was performed using several programs, including BLASTP, PSIBLAST, PHMMER, HMMER, and GLAM2Scan. In all cases, the "Human" simulated sequences were used as the query, whereas all other species were combined into a single target database.

For BLASTP, in addition to using the default parameters of the program, we also performed phylostratigraphic runs wherein we varied independently several parameters including Gap Extention and Gap Opening (using all possible combinations allowed by the program), 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 evalue 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 also varied gap open 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. For all runs we set the evalue to 100, which allowed us to progressively restrict E-value from 100 to 1E-10 for each run and observe the results.

For each of PSIBLAST and HMMER, we ran the initial BLASTP and PHMMER searches using the ideal parameters as determined from each of BLASTP and PHMMER. Using these starting points, we tested default parameters for each of BLASTP 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. For all runs we set the evalue to 100, which allowed us to progressively restrict E-value from 100 to 1E-10 for each run and observe the results.

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 1 phylostratigraphic run using GLAM2Scan for each of the three protein sets.

In addition, we performed phylostratigraphy using the results of simulation through our second tree (Figure 5-3B). In this case, we used default BLASTP settings.

Identification of ideal parameters

In order to identify the ideal parameters under a particular simulation and homology detection program, we first determined for each phylostratigraphic run the minimum false positive rate as a function of e-value based on a program's ability to detect the Bacterial false positive, and removed from consideration parameter sets that had an unusually high false positive rate (Figure C-4 through C-8). We then identified the largest evalue for which the false positive rate was minimum. Then, for all runs of all parameters, we compared the false negative rates at the e-

value with minimum false positive rates, based on the program's ability to detect the Bacterial homolog. Whichever parameter set had the lowest degree of false negatives was selected as the ideal parameter set.

Real phylostratigraphy

We performed real phylostratigraphic analysis using two separate protein sets. First, we performed phylostratigraphy using the 4942 human proteins acquired from OrthoMaM (Ranwez et al., 2007). Second, we performed phylostratigraphy using 4942 randomly-chosen proteins from the Ensemble collection of human proteins. For both phylostratigraphic runs we used the BLASTP algorithm using default parameters and an e-value of 0.001. We converted GI numbers of hits to corresponding taxon names. We then acquired taxon lists corresponding to the following classifications: Primate, Euarchontoglires, Boreouthera, Eutheria, Mammalia, Amniotes, Tetrapoda, Gnathostomata, Vertebrata, Chordata, Bilateria, Eumetazoa, Opisthokonta, and Eukaryota. We determined the number of genes that fell into each clade in each run.

Statistical analyses

All statistical analyses were performed using R version 3.2.3.

For the creation of support vector machines, we used the R packages "MASS" and "e1071" (Venables & Ripley, 2002). For the creation of random forests, we used the R package "randomForest" (Liaw & Wiener, 2002). For these models, we calculated sensitivity as the number of correctly identified error prone genes divided by the total number of error prone

genes. We calculated specificity as the number of correctly ignored non-error-prone genes over the total number of gene which were not error prone. We calculated precision as the number of correctly identified error-prone genes over the total number of genes which were identified as being error-prone. Hypergeometric tests were performed using the methodology provided in Rivals et al 2006 (Rivals, Personnaz, Taing, & Potier, 2007).

Results

Identifying an idealized parameter set

We created three sets of sequences for simulation (see methods). There are three properties which are known to be relevant to homology detection error in BLASTP: sequence length, evolutionary rate, and the longest conserved block of sites (Moyers & Zhang, 2015). The simulations varied in two of these properties (Figure C-1, C-2, and C-3). We simulated evolution through a guide-tree (Figure 5-1A), and confirmed that the genetic distances generated by our simulation were comparable to real genetic distances (Figure 5-1B). We then performed phylostratigraphy using a number of different programs to assess their relative performance. A brief description of each is given below.

We first assessed BLASTP, as the BLAST suite of algorithms is "the workhorse of phylostratigraphy" (Domazet-Lošo et al., 2016). BLAST (Stephen F Altschul et al., 1990) 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 programming algorithm to score the alignment. Once the score is determined, the algorithm compares the realized score versus 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.

PSIBLAST 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 BLASTP uses (S F Altschul et al., 1997). 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 historical homolog into the PSSM. This has the risk of inflating the false positive rate.

PHMMER is typically used as a sequence similarity search tool that generates homologs which can then be used as inputs to the HMMER algorithm, described below. PHMMER searches a target database for matches to a query using a substitution matrix to determine the score of an alignment. 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 is an iterative, profile-based algorithm which 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 hidden markov model and the algorithm can be run again with a new query.

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 was 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.

There are numerous other homology detection tools which we did not test. Most notably, tools which compare profiles with profiles, such as HHSearch (Söding et al., 2005) and PSIPRED

(Buchan et al., 2013). These are of particular interest in reducing false negative error in phylostratigraphy, as they tend to detect a greater number of homologs. For instance, HHSearch was found to detect 4.2 times the number of homologs as HMMER (Söding, 2005). In particular, it is important to evaluate the false positive rate of these programs, as they may be greatly inflated. However, we were unable to assess these methods under our simulation paradigm. These programs require comparison of query and target sequences to established databases or which structural information is available. However, there is no reason to think that our simulations respect structural constraints on protein evolution, and the simulations further have destroyed amino acid sequence conservation between real sequences and simulated sequences, so any hits between simulated proteins and these databases must necessarily be a case of false positive error.

We therefore applied each of these five homology detection programs to the simulated results of our three protein sets, separately. For each set and each program, we determined the ideal set of parameters which minimizes first False Positives and then False Negatives (Figures C-4, C-5, C-6, C-7, and C-8). In attempting to identify ideal parameter sets and comparing the runs of various homology detection methods, we noticed several interesting patterns. We next compared the ideal parameter sets and the default BLASTP parameters based on their ability to detect homologs in each phylostratum (Figure 5-2). We note first that GLAM2Scan appears to be unsuited for this kind of analysis, which is not surprising given that it was never intended for this purpose. Among the other programs, we note that the particular dynamics depends at least partially on the qualities of the protein set under consideration. We also note that, generally speaking, HMMER and PSIBLAST tend to outcompete BLASTP in terms of false negative rate,

whereas BLASTP tends to have the lowest false positive rates. However, these differences tend to be generally marginal, suggesting that default BLASTP may be sufficient under most conditions. 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 (Figure 5-2C). Finally, we find that false positive rate, while it differs among programs, is generally negligible (less than 1%), except when using PSIBLAST in the case of fast-evolving proteins (Figure 5-2F).

In terms of the absolute error rate of BLASTP, we note that for our Base set of 4942 proteins, false negative error rate falls between 5 and 7%, depending on the program used (Figure 5-2A). However, for simulated protein lengths with a realistic Size Distribution (Figure 5-2C) false negative error approaches 15%, and when considering a realistic size distribution with faster evolutionary rates (generally expected for apparently species-specific genes), false negative error can approach as high as 30% (Figure 5-2E).

Bias of homology detection

Based on the above idealized results, we note that false negative error cannot be wholly eliminated and generally is not largely reduced by deviating from the standard practice of using default BLASTP. However, there is a separate question of whether or not the error of these programs is biased with sequence properties. We therefore determined the correlation of simulated age with sequence properties of length, evolutionary rate, and the maximum length of conserved block for each of BLASTP, PSIBLAST, PHMMER, and HMMER in each of our three simulation sets. GLAM2Scan was excluded because its false positive and false negative error rates were high enough to be disregarded as a potential tool for phylostratigraphy. We find

that except for a few cases homology detection error creates spurious correlations with age (Table 5-1). No program is without this bias.

Error in species-restricted contexts

It was suggested that error may be negligible when performing phylostratigraphy in contexts where species are not very far diverged (Domazet-Lošo et al., 2016). While no particular measure has been given for what constitutes such a context, we sought to investigate whether or not this claim had support. First, we took the 4942 human proteins from which we inferred evolutionary rate information and performed real phylostratigraphy using BLASTP with an E-value cutoff of 0.001, and sorted the results into 15 strata. For each protein, we then paired its age by real phylostratigraphy to whether or not it was subject to error in simulation under default BLASTP settings. We found that the younger a protein is found to be by real phylostratigraphy, the greater its propensity for error (Figure 5-3A). One might argue, however, that our simulation is insufficiently connected to the real properties of these proteins to make such a claim. While that argument effectively relegates error propensity of any given gene to the realm of the unobservable, it may yet be true. We therefore sought to investigate via simulation whether homology detection error was found in closely-related clades. We simulated our Size Distribution and Size Distribution Fast sets of proteins through a tree containing 13 primate species plus rat as an outgroup (Figure 5-3B). We did not simulate our 4942 genes through this tree as these genes are sufficiently long and have properties associated with conserved genes so as to virtually ensure that error in these contexts would be rare. The other two sets, though, have a more realistic length distribution and the Size Distribution Fast set is arguably more representative of the evolutionary rates of primate-specific genes.

After completion of the simulation, we performed phylostratigraphy with the human proteins as query and all other species' proteins as the target database. As is standard in these studies, we next removed those genes which had a hit in the outgroup, Rat. This left us with 103 proteins in the Size Distribution simulation and 273 proteins in the Size Distribution Fast simulation. We then plotted the number of genes in each set which did not have homologs in each target species (Figure 5-3C and 5-3D). We find that, among genes without a homolog in rat, error rates of approximately 90% can be observed. While this corresponds to only ~1.0% and ~2.7% of all proteins in their respective simulations, the measure of error which is appropriate for such studies is the one provided in Figure 5-3, as studies in these contexts first restrict genes to those which are not found outside the clade of interest. It is therefore clear that error is present and prevalent among genes in closely-restricted contexts. This is corroborated by the finding that 5 of 15 genes which had previously been classified as *S. cerevisiae* specific in Carvunis (2016) were found to be non-species specific upon application of syntenic methods (Domazet-Lošo et al., 2016).

We further demonstrate that error even within this context is still biased with gene properties (Table 5-2). We find significant correlation between gene ages and length and evolutionary rate for both simulations, and with the maximum length of conserved blocks for the Size Distribution Fast simulation. This finding contradicts the assertion that phylostratigraphic trends in closely-related clades is not influenced by homology detection error. We do note, however, that there is an unexpected reversal in the correlation between age and evolutionary rate in the case of the Size Distribution simulation. This is likely due to the fact that in this simulated gene set with

randomly-assigned properties and restricted to a relatively small amount, the effect of size overrides the effect of evolutionary rate, particularly given the large and highly-significant effect of length in this context. However, it may be the case that for such closely-related species the dynamics of homology detection and the interrelation between length and rate in this process are more complicated than we currently understand.

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 which identifies *a prior* those genes which are likely to be subject to homology detection error before performing phylostratigraphy. We reasoned that if we could construct a model which was able to correctly identify 90% or more of error-prone genes correctly without removing a substantial proportion of non-error-prone genes, this would be an effective model. We therefore used BLASTP simulation results of Base, Size Distribution, and Size Distribution Fast gene sets simulated through the tree in Figure 5-1A to construct support vector machine (SVM) and random forest models. We used ten-fold cross-validation to determine the average sensitivity, specificity, and precision of each model. We tried as many combinations of the parameters length, evolutionary rate, and maximum length of conserved block for each model, using error (as measured by a missed bacterial homolog) as a response variable. We then determined which of the predictor variable sets produced the model with the greatest sensitivity.

We found that all models were insufficiently sensitive, though random forests performed better than SVM models (Table 5-3). We reasoned that a less stringent definition of error might create

better-performing models, and thus created models wherein the response variable was whether or not a homolog was found in Fungi, a much less distantly removed homolog. This did not change the results (Table C-1).

An error-aware framework of phylostratigraphy using simulation results

Having investigated several methods to remove the effects of error and found none, we were left with only one remaining method: removal of error-prone genes from real phylostratigraphic results. In this context, error-prone genes are identified as those genes which experience any amount of homology detection error in a simulation of evolution. We have previously demonstrated using this method that homology detection error significantly affects the outcome of phylostratigraphic results (Chapter 4).

We sought to investigate whether or not this methodology might offer new insight into biological trends over evolutionary time. To that end, we performed real phylostratigraphy against the NCBI non-redundant protein database on the 4942 human sequences acquired from OrthoMam and 4942 randomly-selected sequences from the Ensemble collection of human proteins. We then removed from the first of these sets those genes which were found in our base simulation any homology detection error. We plotted the ages of these genes (Figure 5-4). We note that while there are hundreds of genes which are dated to the common ancestor of humans and elephants or younger in the randomly-chosen gene set, the youngest genes in the non-error-prone gene dates to the common ancestor of all mammals. It is not surprising that this is the case, given that we required these genes to have orthologs in all 14 mammalian species and then removed any which were error-prone. This does highlight, however, that it is certainly possible

to have a range of gene ages when restricting genes to a non-error-prone set. However, it seems almost certain that the youngest clades will necessarily be left out of such analyses.

Finally, we investigated previously-reported trends in these datasets. Not having evolutionary rate information for the randomly-chosen set, we could only evaluate the relationship between length and phylostratigraphic age. But for the 4619 non-error-prone genes which we used to determine simulation parameters, we were able to investigate relationships between age and three parameters: length, evolutionary rate, and the maximum length of conserved blocks in the protein (Table 5-4). We find that the previously-reported association of older proteins generally being longer retained in our random phylostratigraphic set. However, once restricted to a non-error-prone set, we find that this trend is reversed, such that older genes are actually shorter than younger genes with a weak effect commensurate with the weak positive effect previously reported. This is combined with the finding that evolutionary rate is not significantly associated with age, and that the older a gene is the shorter its conserved blocks tend to be, with relatively weak effect. These findings provide insight into evolutionary dynamics of proteins that have moderate conservation and are not error-prone, and provide further insight into phylostratigraphic theory.

Discussion

We have demonstrated here that false negative error is prevalent in phylostratigraphy. While the 4942 human genes which we began simulation with has only a marginal degree of error (6.5% of genes missed their bacterial homolog in our simulation), this is expected. These genes have necessarily been conserved for at least 90MY among 14 mammals. Their lengths, evolutionary

rates, and rate heterogeneity patterns are therefore representative of genes which have no propensity for error at least out to 90MY. However, even changing length distributions to be more realistic (Figure C-1) without substantially changing rate or rate heterogeneity properties (Figures C-2 and C-3) produces greatly increased error. We observed that 14.3% of genes could not find a bacterial homolog in our Size Distribution set of genes. When faster evolutionary rates are introduced, we find that this error rate can be greatly increased, with 33.4% missing a bacterial homolog. Clearly, those genes which are reported to be the youngest in phylostratigraphy—short, fast-evolving genes—are most likely to be subject to homology detection error.

This error is further not entirely accounted for by investigating only young clades. It is obviously true that missing a bacterial homolog does not guarantee that a human gene will also miss a homolog in a closely-related species, this kind of error still does occur. The contention that there are contexts "where no BLAST error could be reasonably invoked" (Domazet-Lošo et al., 2016) is demonstrably false, as such a context depends upon the particular genes under investigation and their properties. In fact, we find that all efforts to decrease false-negative error through changing context or tool are ineffective (Figure 5-2). This is in concordance with the results of Domazet-Loso (2016), who re-evaluated the species-specific status of 15 ORFs (as assigned by phylostratigraphy) and found that 1/3 of them were falsely classified as species-specific.

Comparatively, the concerns surrounding False Positive error are not well-supported by our results. While we find that profile-based homology detection programs (PSIBLAST, HMMER)

generally have a higher degree of false-positive error than does BLASTP, we find that this error is small (<1% of genes) except for the case of small, fast-evolving genes using PSIBLAST. We therefore cannot recommend the use of PSIBLAST as a reasonable tool for phylostratigraphic analysis. 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, as any bias in the trends will be randomly distributed throughout time.

Future work should investigate alternative methods for identifying error-prone genes. As has been previously mentioned (Chapter 4) and here, our simulation set is inappropriate for use with certain homology detection methods. If a new simulation set can be performed which captures such features as structural evolution and similarity as well as sequence evolution constraints, homology detection error may be reduced by more sensitive tools. The error-prone status of genes might be further probed by using a larger number of genes for simulation. There is an inherent problem here, as 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. Additionally, for those genes with fewer 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 have here demonstrated that error-aware phylostratigraphy is not merely a conservative approach to phylostratigraphy, but can provide novel biological insight. We hope that prior

phylostratigraphic findings will be re-evaluated in this context, and that future work will consider error in inferring evolutionary trends.

References

Albà, M. M., & Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology*, *7*(53). http://doi.org/10.1186/1471-2148-7-53

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Jounal of Molecular Biology*, *215*(3), 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., … Noble, W. S. (2009). MEME SUITE : tools for motif discovery and searching. *Nucleic Acids Research*, *37*(May), 202–208. http://doi.org/10.1093/nar/gkp335

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., & Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research*, *41*(June), 349–357. http://doi.org/10.1093/nar/gkt381

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*, 370–374. http://doi.org/10.1038/nature11184

Domazet-Lošo, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics : TIG*, *23*(11), 531–3. http://doi.org/10.1016/j.tig.2007.07.007

Domazet-Lošo, T., Carvunis, A., Alba, M. M., Sestak, M. S., Bakarić, R., Neme, R., & Tautz, D. (2016). No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Biorxiv*, 1–24.

Domazet-Lošo, T., & Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome Research*, *13*, 2213–9. http://doi.org/10.1101/gr.1311003

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. *Molecular Biology and Evolution*, *23*(1), 1–3. http://doi.org/10.1093/molbev/msj006

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research*, *39*(SUPPL. 2), 29–37. http://doi.org/10.1093/nar/gkr367

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, *22*(23), 2971–2. http://doi.org/10.1093/bioinformatics/btl505

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275–282. http://doi.org/10.1093/bioinformatics/8.3.275

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(December), 18–22.

Moyers, B. A., & Zhang, J. (2015). Phylostratigraphic bias creates spurious patterns of genome evolution. *Molecular Biology and Evolution*, *32*(1), 258–267. http://doi.org/10.1093/molbev/msu286

Moyers, B. A., & Zhang, J. (2016). Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Molecular Biology and Evolution*, *33*(5), 1245–1256. http://doi.org/10.1093/molbev/msw008

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, *14*(117). http://doi.org/10.1186/1471-2164-14-117

Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.-K., & Douzery, E. J. (2007). OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evolutionary Biology*, *7*(1), 241. http://doi.org/10.1186/1471-2148-7-241

Rivals, I., Personnaz, L., Taing, L., & Potier, M.-C. (2007). Databases and ontologies Enrichment or depletion of a GO category within a class of genes : which test ? *Bioinformatics*, *23*(4), 401–407. http://doi.org/10.1093/bioinformatics/btl633

Schmidt, H. A., Strimmer, K., Vingron, M., & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, *18*(3), 502–504. http://doi.org/10.1093/bioinformatics/18.3.502

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, *21*(7), 951–960. http://doi.org/10.1093/bioinformatics/bti125

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(SUPPL. 2), 244–

248. http://doi.org/10.1093/nar/gki408

Stoye, J., Evers, D., & Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics*, *14*(2), 157–163.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. (Springer, Ed.) (4th ed.). New York.

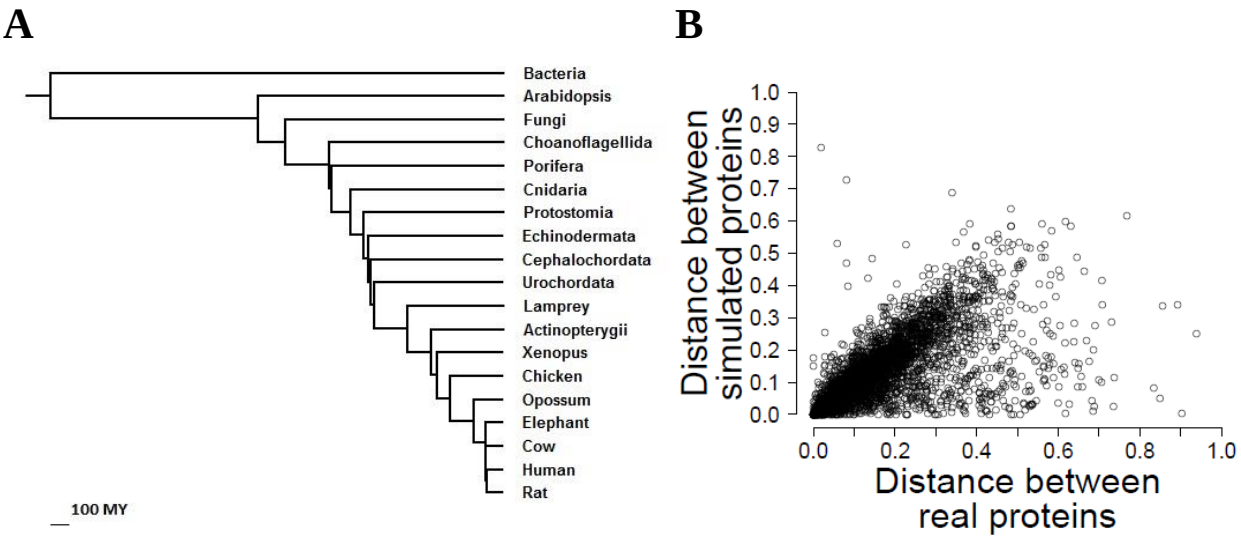

Figure 5- 1 Simulation for the assessment of homology detection error

(A) Tree through which simulation was performed. Branch lengths were determined by TimeTree estimation of divergence time of a given species from Humans. (B) Comparison of genetic distance between humans and mouse in real and simulated proteins (R=0.6130, p=2.2E-316, Rho=0.7013, p=2.2E-316). Though the correlation is only moderately strong, we note that there is a clear skew toward our simulation under-evolving the sequences, supporting the idea that this methodology is conservative.

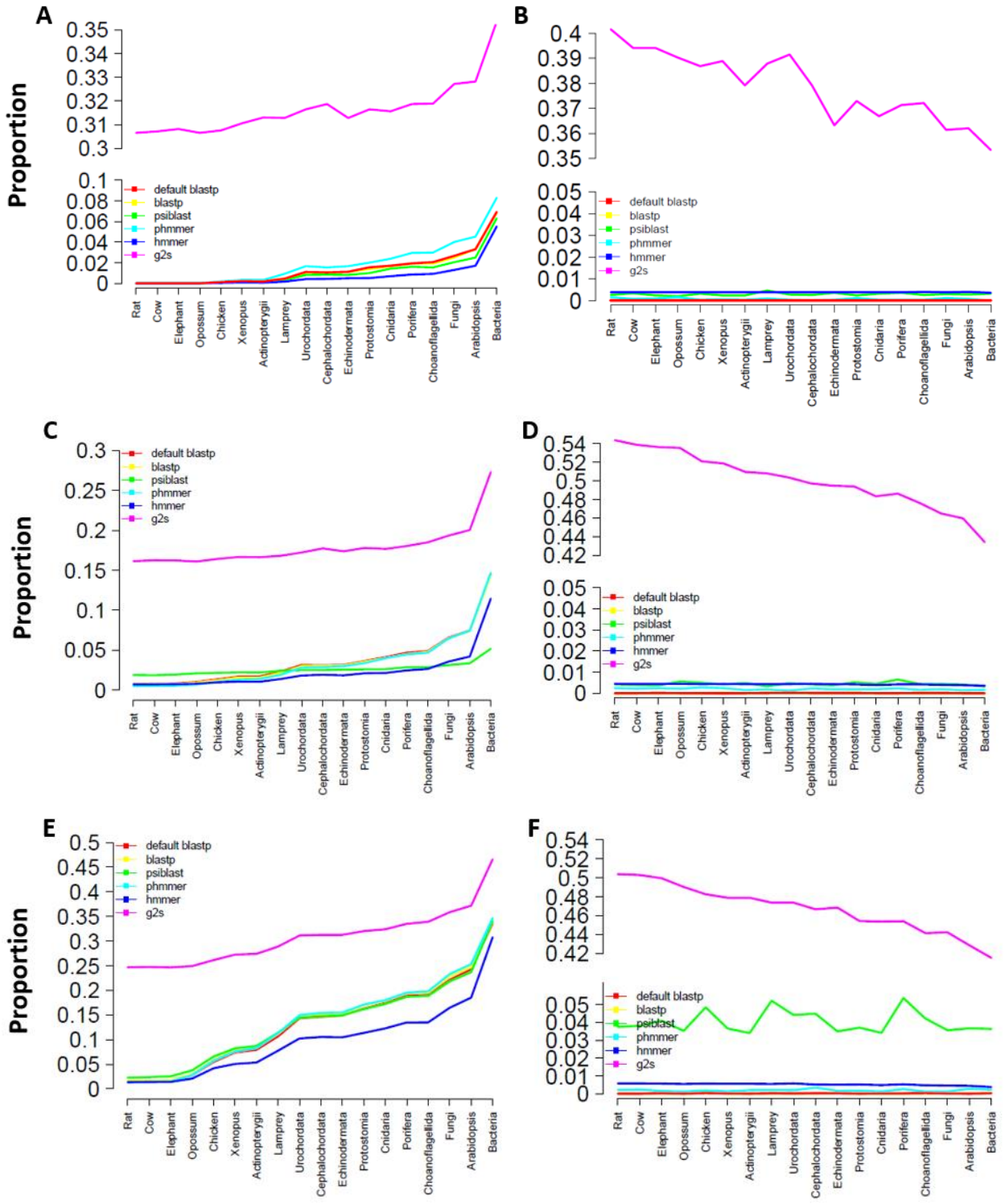

Figure 5- 2 False negatives and positives by phylostrata

False negative and false positive rates in detecting by phylostrata for default BLASTP parameters and the ideal parameters for all five programs. The left column shows false negative rates while the right column shows false positive rates. The first row shows the results of our Base set, the second row shows the results of our Size Distribution set, and the third row shows the results of our Size Distribution Fast set. (A) False negative rates for Base set. (B)

False positive rates for Base set. (C) False negative rates for Size Distribution set. (D) False positive rates for Size Distribution set. (E) False negative rates for Size Distribution Fast set. (F) False positive rates for Size Distribution Fast set.

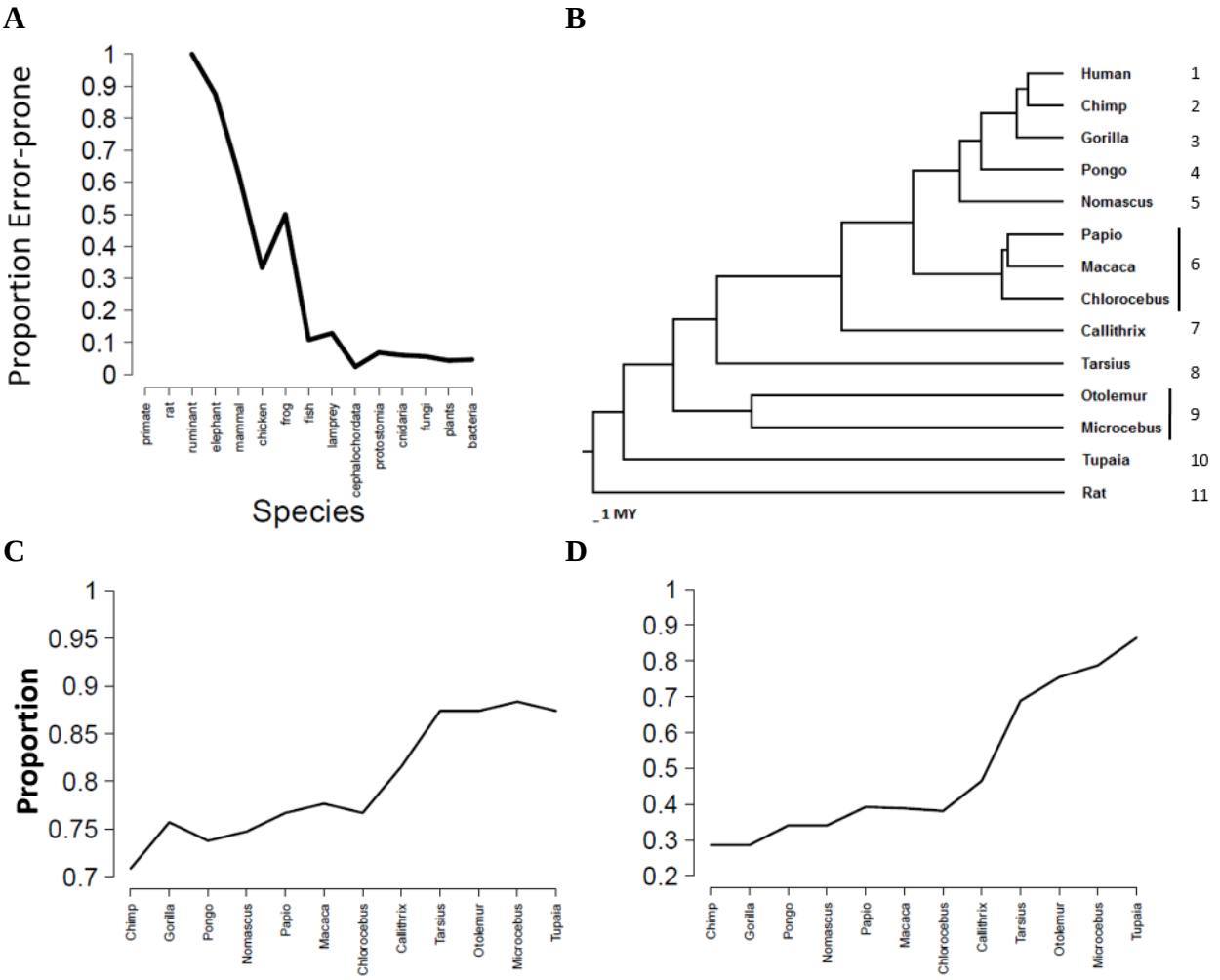

Figure 5- 3 Homology detection error in closely-related species

Homology detection error in close-related species. (A) Proportion of genes in each phylostratum which, in simulation, were subject to homology detection error. (B) Tree through which simulation was performed for Size Distribution and Size Distribution Fast gene sets. Branch lengths were determined by TimeTree estimation of divergence time of a given species from Humans. (C and D) False negative error rate for each species in simulation for the Size Distribution (C, N=103) and Size Distribution Fast (D, N=273) sets.

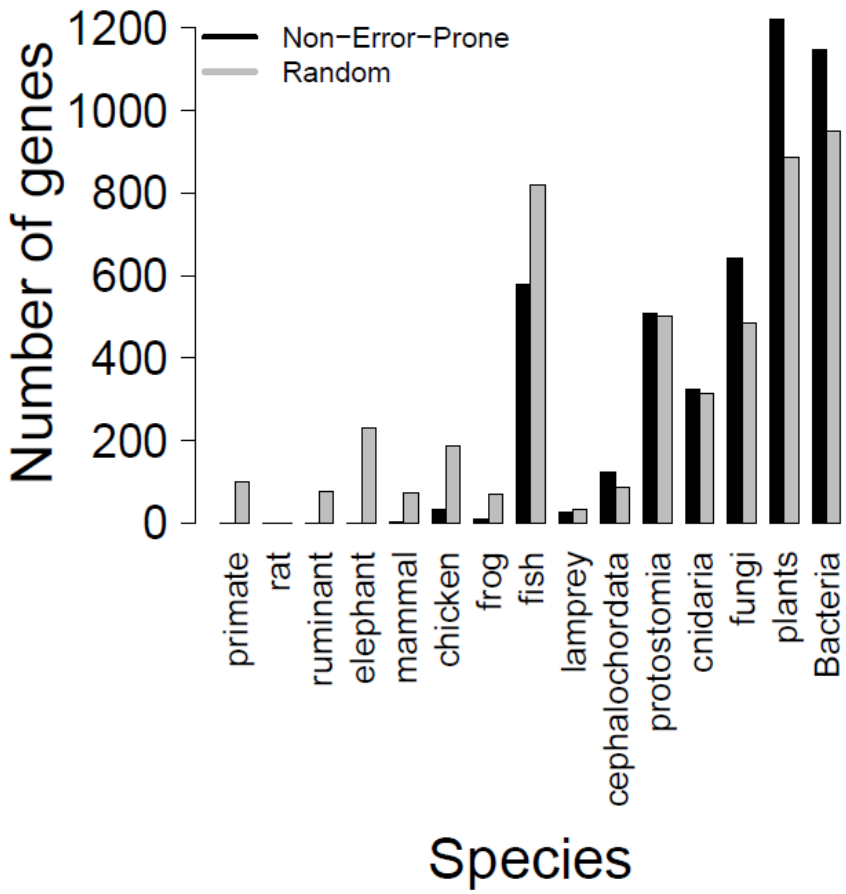

Figure 5- 4 Ages of two distinct gene sets in real phylostratigraphy

Age of two distinct gene sets in real phylostratigraphy. Black bars represent the 4619 genes which were simulated and found to be non-error-prone. Grey bars represent the ages of a randomly-selected 4942 genes.

Table 5- 1 Spurious correlations between age and biological features

	BLASTP	PSIBLAST	PHMMER	HMMER
Base Simulation				
Length	0.14**	0.16**	0.03	0.11**
Rate	-0.37***	-0.36***	-0.02	-0.34***
Block	0.35***	0.35***	0.03*	0.32***
Size Dist. Simulation				
Length	0.31***	0.30***	0.28***	0.27***
Rate	-022***	-0.08**	-0.22***	-0.22***
Block	0.37***	0.28***	0.37***	0.33***
Size Dist. Fast Simulation				
Length	0.32***	0.36***	0.29***	0.28***
Rate	-0.12**	-0.12**	-0.13**	-0.13**
Block	0.41***	0.44***	0.42***	0.38***

*P<0.05, **P<1E-10, ***P<1E-100

Table 5- 2 Correlation between gene properties and phylostratigraphic error in closely-related species

	Length	Rate	Block
Size Distribution	0.78**	0.28*	0.14
Size Distribution Fast	0.75**	-0.15*	0.58**

*P<0.05, **P<1E-10, ***P<1E-100

Table 5- 3 Performance of machine learning algorithms for identifying error-prone genes

	Base SVM	Size Dist. SVM	Size Dist. Fast SVM	Base RF	Size Dist. RF	Size Dist. Fast RF
Model*	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

*L=length, E=evolutionary rate, B=maximum length of conserved block

Table 5- 4 Spearman's rho correlation between age and gene properties in real phylostratigraphy

	Random 4942 proteins	Non-error-prone proteins
Length v Age	0.16**	-0.12**
Evolutionary Rate v Age	NA	0.002
Block length v Age	NA	-0.09**

*P<0.05, **P<1E-10, ***P<1E-100

Chapter 6

De novo genes contribute significantly to novel sequence formation

Abstract

Phylostratigraphy is a method for identifying novel sequences based on homology detection. A novel sequence may arise in at least three ways, *de novo* gene birth, homology detection error, or a sudden shift in sequence space, as expected after a gene duplication. The relative contributions of these three mechanisms to novel sequence formation is still not known, and whether or not *de novo* gene birth accounts for a significant portion of novel sequences is not clear. Here, we investigate the relative contributions of these sequences through an error-aware phylostratigraphic analysis. We simulate the evolution of sequences and investigate the phylostratigraphic dynamics of several models of duplication. We find that, even under extreme models of duplication, phylostratigraphy suggests that alternative sources contribute a non-negligible number of novel sequences. We also find that homology detection error contributes approximately twice as many sequences to novel sequence formation as does duplication and divergence, even under an extreme model.

Introduction

Lineage specific genes, or orphan genes, are genes which are restricted to a particular taxon. These genes are identified through the use of phylostratigraphy: that is, searching for homologs

in a set of species using BLAST and determining where homologs are and are not found. Phylostratigraphy is a method for identifying "novel sequences." Based on this analysis, a set of genes is found which is present only in a particular lineage. There have been numerous studies associated with lineage-specific genes (Domazet-Lošo, Brajkovic, & Tautz, 2007; Domazet-Loso & Tautz, 2010; Domazet-Lošo & Tautz, 2010; Neme & Tautz, 2013; M. Sestak & Domazet-Lo o, 2014; M. S. Sestak, Bozicevic, Bakaric, Dunjko, & Domazet-Loso, 2013), and estimates of the rate of gene fixation have been performed based on these analyses (Domazet-Lošo et al., 2007).

A major limitation of using phylostratigraphic analyses is that it cannot distinguish between the different mechanisms of novel sequence formation. There are likely three major sources: *de novo* gene birth, rapid but short-lived periods of sequence divergence due to change in functional constraints or duplication, and homology detection error (or general divergence). A major unanswered question is the relative contribution of each of these. Some investigations have suggested that *de novo* gene birth is frequent, placing it as being more frequent than duplication (Carvunis et al., 2012). However, these have been subject to biased homology detection error (Moyers & Zhang, 2016), and this error is non-negligible, even for closely-related species (Chapter 4, Domazet-Loso et al. 2016).

Another major question regarding the formation of orphan genes is whether or not their restriction corresponds to a biologically meaningful age, as defined by historical homology. It has been previously established that homology detection error can make genes appear younger than they truly are on the basis of historical homology (Albà & Castresana, 2007; Elhaik, Sabath, & Graur, 2006; Moyers & Zhang, 2015, 2016). However, the dynamics of duplication-

divergence may exacerbate this problem, or create different kinds of problems, such as sequences mapping to dates older than the time of duplication. This would not be surprising, as it has been demonstrated that sequence similarity measures as a measure for phylogeny reconstruction can produce incorrect phylogenies (Smith & Pease, 2016). If this is the case, then studies of orphan genes are studying something entirely different than historical homology. While identifying orphan genes may provide interesting avenues of research, their precise ages are evolutionarily and biologically meaningless.

These issues might be approached by a simulation of evolution which incorporates models of divergence and duplication-divergence. Simulations have been applied to study rates of homology detection error (Elhaik et al 2006, Alba and Castresana 2007, Moyers and Zhang 2015, Moyers and Zhang 2016, Chapter 4, Chapter 5). It may be possible to study the contributions of duplication-divergence through simulation, though there are major questions about the models of duplication that might be used (Lynch & Conery, 2000; Zhang, 2013) and the frequency with which gene duplication and whole genome duplication occurs and how many of these genes survive (Carvunis et al., 2012; Cliften, Fulton, Wilson, & Johnston, 2006; De Smet et al., 2013; Gao & Innan, 2004; Moyers & Zhang, 2016). However, *de novo* gene birth presents a special problem, as it is unknown how frequently it occurs despite some suggestions that it is common (Carvunis et al., 2012; Moyers & Zhang, 2016; Neme & Tautz, 2013), and there is scant information about how *de novo* sequences evolve in their early stages. Additionally, divergence in the absence of duplication due to supposed changes in functional constraints are ill-defined, and there is not a clear suggested model by which this kind of

evolution would occur. Therefore, these contributions of these mechanisms can be studied only indirectly.

Here, we assess these problems by simulating the evolution of protein sequences, including several models of gene duplication. We also perform real phylostratigraphy on the same gene set from which we draw evolutionary parameters to compare the number of lineage-specific genes at each clade, both for real and simulated proteins. We find that *de novo* gene birth is likely to account for a non-negligible portion of apparent orphan genes, as general homology detection error and models of duplication-divergence cannot fully account for the number of orphan genes identified in real phylostratigraphy. When duplications do produce "novel sequences", we find that they are not dated to the correct time of emergence—they are often said to be older or younger than the time of actual duplication-divergence, creating strange situations in which a "novel sequence" is said to have emerged prior to when a duplication occurred.

We conclude that methods other than divergence and duplication-divergence are major contributors to orphan gene formation, but identification of orphan genes is insufficient as a measure of the contributions of these mechanisms to gene birth, as the number of orphan genes is also substantially influenced by homology detection error and duplication-divergence. Phylostratigraphy is, therefore, an important first step in identifying rates of ancient or recent *de novo* gene birth. We also challenge the paradigm of phylostratigraphy as a measure for "gene age," as it places the dates of emergence for novel sequences at times when sequences did not exist.

Methods

Sequences

For simulation of sequence evolution, we acquired 4942 orthologous sequence alignments in protein format from 14 mammalian species through OrthoMaM (Ranwez et al., 2007). The specific species used were Pan, Homo, Gorilla, Pongo, Nomascus, Macaca, Callithrix, Tarsius, Otolemur, Microcebus, Rattus, Mus, Dipodomys, and Cavia. All sequences had at least these 14 species, and any other species included in the alignments were removed.

Simulation of evolution

The evolutionary tree was constructed from general species divergence times acquired from TimeTree (Hedges, Dudley, & Kumar, 2006). For each of the 4942 proteins with alignments of fourteen sequences, we used TreePuzzle (Schmidt, Strimmer, Vingron, & von Haeseler, 2002) to classify all sites into 16 rate bins according to a discrete gamma model of among-site rate heterogeneity and estimated the relative rates of the 16 bins. We also inferred the mean evolutionary rate across all sites of the protein between *H. sapiens* and *M. myoxinus* (Microcebus). Using all of these parameters, we simulated the evolution of these proteins using ROSE (Stoye, Evers, & Meyer, 1998), which allows the evolutionary rate for each site to be specified by the user, along the tree in Fig. 1A. ROSE evolves sequences through amino acid substitutions and insertions and deletions (indels). For each branch of the tree, ROSE first performs the amino acid substitution function, and then performs the indel function. If the branch is an internal branch in the tree, it then copies the resulting amino acid sequence to the base of each of the two branches after the split.

For amino acid substitution, ROSE uses an amino acid substitution matrix provided by the user. We used a JTT-f matrix for the amino acid substitution model (Nei & Kumar, 2000). Each site along the protein has a particular relative rate. The relative rate for a site is multiplied by the length of the branch to obtain the expected amount of evolution along the branch at the site. ROSE makes substitutions based on this expected amount of evolution and the substitution matrix supplied. This is repeated for all sites along the amino acid sequence.

For indels, there are two parameters that determine indel formation in ROSE, the indel threshold and the indel function. The indel threshold measures how frequently indels occur and was determined in the following manner. Taking the alignments of the orthologs acquired from OrthoMaM and using a custom script, we determined the minimum number of indels necessary to produce the observed gapped alignments. From this information, we determined the number of indels per amino acid, averaged over all proteins. This indel threshold was then applied to all proteins in simulation. The indel function is a vector that sums to 1 and gives, at each vector site *i*, the probability of an indel of size *i*, given that an indel is occurring. For the indel function, we took the observed frequencies of indel sizes from 1 amino acid to 30 amino acids long (accounting for the majority of observed indels), and adjusted these frequencies to sum to 1. Sequence simulation was performed once for each protein. We confirmed that our methods of simulating sequence evolution were conservative by comparing evolutionary distances for real and simulated proteins (Fig. 1B).

Models of Duplication

For duplication models, we identified the sequence which existed immediately following the Human-Urochordata split in the simulation of each gene, or for later simulations the sequence extant at each 200MY point. We copied this gene, all evolutionary parameters, and the remaining tree branches to either one (baseline, neofunctionalization) or two (subfunctionalization) new files and manipulated heterogeneity parameters as appropriate for each simulation (Figure D-1).

For the baseline simulation (Figure D-1A), no modifications were made to any parameters, and the two daughter genes continued evolution independently.

For the first set of neofunctionalization models (Figure D-1B) in the copied Rose File we selected as much as 90% of the protein sequence to be set at the fastest rate category. This was done by first selecting all sites with the most conserved category and determining whether or not this accounted for at least 10% of the sequence. If not, we iteratively added the next most conserved rate category until at least 10% of the sequence was selected. We then split all other sites into two equal categories and set their evolutionary rate to either the fastest rate category or twice the fastest rate category to simulate positive selection. We then evolved this sequence along the branch for 0, 5, 10, or 20 million years. After this "burst" of evolution, we set all sites back to their original rates. Sites which were a result of insertions throughout this process were assigned the average rate of the protein, which is the rule of the ROSE program. Once this burst of evolution was complete, we then randomly selected 2/3 of sites by randomly selecting 5-amino-acid chunks of the gene. The remaining 1/3 of sites then had their relative rates shuffled,

allowing some conserved sites to become non-conserved, and visa-versa. After this shuffling, the simulation of evolution was allowed to complete.

For the second set of neofunctionalization models (Figure D-1C), termed "neofunctionalization, all sites", we made two changes. First, we selected all sites as opposed to <=90% of sites for the burst of evolution which lasted 0, 5, 10, or 20 million years. Second, after the burst of evolution was completed and all sites had been set back to their original rate categories, we shuffled the rate categories of all sites, rather than just a select 1/3 of sites.

For the subfunctionalization models (Figure D-1D), we made two copies of the sequence immediately after the human-urochordata split, and these two copies were the only two considered afterwards (i.e. we did not consider the original third copy from the base simulation). In each of the two copies, we selected 1/3 of sites and set these sites to the maximum rate of the protein, to simulate neutral evolution of those sites. There were four simulations in total, because we could either allow overlap between the sites selected in the two genes, or not, and we could also allow the most conserved rate category sites to be selected or not. When disallowing the most conserved rate category from being selected, we selected 1/3 of the remaining sites, not all sites. This is because for some genes the most conserved rate category makes up more than 1/3 of the gene's length, meaning that the two genes could not select 1/3 of all sites in a non-overlapping manner. When selection of sites was allowed to overlap between the two genes, no amount of overlap was forced, but random selection allowed for overlap to occur. After setting the appropriate sites to the average rate of the protein, the simulation of evolution was allowed to complete.

Phylostratigraphy of simulated sequences

After completing the simulation of evolution, we constructed two databases of sequences for subsequent phylostratigraphy in each of our 13 simulations. The first of the two databases contained all genes that belonged to the human group, two copies of each gene. The second database contained the genes for all other species. In the species "Bacteria", "Arabidopsis", "Fungi", "Choanoflagellida", "Porifera", "Cnidaria", "Protostomia", "Echinodermata", "Cephalochordata", and "Urochordata", this was one copy of each gene. For the species "Lamprey", "Actinopterygii", "Xenopus", "Chicken", "Opossum", "Elephant", "Cow", and "Rat", this was two daughter copies for each gene. All genes were labelled with their appropriate species name to keep clear which species each particular gene came from during further analysis. We used BLASTP (Madden & Morgulis, 2009) with default settings and an evalue of 0.001 to search for human homologs among the species. A gene's phylostratigraphic age was assigned as the age of the most recent common ancestor between human and the furthest species in which a homolog was found.

Paralog control

We performed a BLASTP search using the simulated human sequences as both query and target with an E-value of 0.001, and evaluated whether genes identified paralogs. When reciprocal hits (not necessarily reciprocal best-hits) were identified, these genes were considered paralogs. While phylostratigraphic analysis does not commonly control for paralogs, it is necessary to

consider the potential controls available. We considered three methods for dealing with paralogs, and the results of each (Figure D-2).

First, one can take no actions based on paralogy relationships, and assign all genes an age based on where they found hits in the target database. This is identical to performing no paralog search at all. Second, one could assign both sequences to the oldest age of the two paralogous hits, counting the emergence of two sequences, but equating their age. Finally, one could count all sequences as only one sequence emergence, based on their detectable paralogy, and assign that one sequence to the oldest age among the paralogs. We regard this final method as the ideal method for controlling for paralogs, as it respects the idea of novel sequence emergence (based on detectable homologs) to the greatest degree.

Phylostratigraphy of real sequences

We took the 4942 human sequences from orthomam and performed a BLASTP search against the NCBI non-redundant protein database with an e-value cutoff of 0.001. We converted GI numbers of hits to corresponding taxon names. We then acquired taxon lists corresponding to the following classifications: Primate, Euarchontoglires, Boreouthera, Eutheria, Mammalia, Amniotes, Tetrapoda, Gnathostomata, Vertebrata, Chordata, Bilateria, Eumetazoa, Opisthokonta, and Eukaryota. We identified the number of genes which were restricted to each of these categories, cumulatively (that is, those genes which are primate-specific are also, by definition, eukaryote-specific). We also performed a paralog correction for these genes, as described for simulated genes, with paralogs being placed in the oldest age category among reciprocal hits. However, we counted the age of each of the two as a separate sequence birth, because we were

unable to distinguish the precise relationship of these two genes (as we were able to do in our simulation). Thus, the number of orphan genes for the real dataset is potentially inflated, making our results conservative.

Comparison of real and simulated phylostratigraphy

We took our simulated gene ages (after one of three paralog corrections), and classified them into these bins as well, based upon the taxa we simulated. The following classifications were made: Human was classified as Primate; Human to Rat was classified as Euarchontoglires; Human to Cow was classified as Boreouthera; Human to Elephant was classified as Eutheria; Human to Opossum was classified as Mammalia; Human to Chicken was classified as Amniotes; Human to Xenopus was classified as Tetrapoda; Human to Actinopterygii was classified as Gnathostomata; Human to Urochordata was classified as Vertebrata; Human to Echinodermata was classified as Chrodata; Human to Protostomia was classified as Bilateria; Human to Choanoflagellida was classified as Eumetazoa; Human to Fungi was classified as Opisthokonta; Human to Arabidopsis was classified as Eukaryota. We then compared the numbers, which gave us an estimate of the number of taxonomically-restricted genes attributable to error and duplication among this gene set. Because the result of our duplications in simulation and the paralogy correction of each set cause different absolute numbers of genes in each of the real and simulated data, with the real data always having fewer effective genes than simulated data, we normalized the total number of sequences in our simulation (after any paralog correction) to the total number of sequences in real phylostratigraphy (after any paralog correction).

Results

Simulation of gene duplication

We acquired 4942 aligned gene sequences from 14 mammalian species diverged approximately 90 MYA (Hedges et al., 2006; Ranwez et al., 2007). From this information, we were able to infer evolutionary rate and rate heterogeneity information using TreePuzzle (Schmidt et al., 2002). We then simulated the evolution of each gene according to previously-described methods through a subset of the tree of life (Figure 6-1A) (Moyers & Zhang, 2015, 2016). In order to assure that our results were approximately conservative, we compared the genetic distance between human and mouse orthologs in both real and simulated proteins (Figure 6-1B) and saw that our simulation of evolution correlated well with observed data.

Just after the split between Humans and Urochordata, we simulated a duplication for each gene in the dataset (Figure 6-1A, asterisk). While it is known that after duplication one copy typically evolves more quickly than the other (Pegueroles, Laurie, & Alba, 2013), the precise molecular dynamics are not clear. Because the molecular dynamics of evolution after a duplication are not well understood, we performed 13 simulations of duplication, each with different assumptions (Figure D-1A, D-1B, D-1C, D-1D).

In the first of these simulations, dubbed "baseline", we created two protein copies at the node of duplication with the sequence and evolutionary rate patterns that existed at that node. We did not change any of the molecular dynamics for either copy (Figure D-1A). We regard this as a

baseline duplication as it makes very few assumptions about the molecular dynamics after gene duplication—it assumes that they do not change.

In the next set of simulations, dubbed "Neofuctionalization", we first duplicate the gene, and then modify the dynamics of one of the two copies (Figure D-1B). First, we create a burst of evolution for 20, 10, 5, or 0 MY. This burst is performed by selecting at most 90% of sites (progressively excluding the most conserved sites until 90% or fewer of sites are excluded). We then set half of these sites to the maximum rate, and half to twice the maximum rate for the duration of the burst. After the burst of evolution, the sites are then returned to their original rate. We then select 1/3 of sites, excluding the most conserved category, and shuffle their rate category to simulate a change in function.

In the above set of neofunctionalization simulations, it might be argued that the simulations are too conservative. And, given that we are interested in probing a range of molecular dynamics for duplication, including more extreme simulations seems reasonable. We therefore performed a separate set of neofunctionalization simulations (Figure D-1C). In these simulations, the pattern followed the above description, but during the burst of evolution all sites were selected for the burst. Then, after the burst of evolution, the rate categories were returned to their original and all sites were shuffled. This simulation is meant to represent an extreme case in which all function is entirely lost and a new, unrelated function is then gained after the burst of evolution.

Our last category of simulations corresponds to "subfuctionalization." In this simulation, we duplicated the gene, and then selected a subset of sites in each of the two daughter genes, setting

them to the maximum evolutionary rate of the genes (Figure D-1D). We performed four modifications of this: one in which conserved sites could not be selected, and the two genes did not overlap in their selected sites; one in which conserved sites could not be selected and the two genes were allowed to overlap in their selected sites; one in which conserved sites could be selected and the two genes were not allowed to overlap in their selected sites; and one in which conserved sites could be selected and the two genes were allowed to overlap in their selected sites. This simulation mimics the idea that upon duplication restrictions on some sites are loosened to allow specialization for an alternate function and relaxation of other functional constraints.

We then performed phylostratigraphy (Domazet-Lošo et al., 2007) on these simulated genes, using all human genes as the queries and a database consisting of all other genes as the target. As in typical phylostratigraphic analyses, we did not distinguish between true positive and false positive hits, as it is difficult or impossible to determine them through phylostratigraphy in real contexts, and false positives contribute minimally if at all to such analyses (Chapter 5). If a significant hit was found, it was considered to be true. Depending on the particular duplication model used, we found that among our 4942 (9884 after duplication) genes, simulation of duplication produced 632-1020 genes out of 9884 genes (6.4%-10.3%) which were counted as novel either due to general divergence (homology detection error) or duplication-divergence (Figure 6-2) when no corrections for paralogs are made. Dynamics of error were dependent upon the particular model of duplication, but were consistent with expectations—those models which had more extreme divergence produced a greater numbers of novel sequences. For the remainder of the paper, we consider only the most extreme version of each of the four models

(baseline, neofuctionalization with 20MY burst, neofunctionalization all sites with 20MY burst, subfunctionalization allowing conserved sites and no overlap between the sites selected in the two paralogs).

Identification of Novel Sequences

In order to determine the relative contributions of homology detection error and duplication-divergence models, we identified novel sequences. In phylostratigraphy, novel sequences are those that do not have a homolog in the oldest age category. When duplication has occurred, paralog correction can have an effect on the inferred number of inferred novel sequences (Figure D-2). We therefore investigated the number of novel sequences in our duplication under no correction (method 1, Figure D-3), a strictly age-based correction (method 2, Figure D-4), and a full correction of both age and number (method 3, Figure 6-3). We note that in the neofunctionalization models, we expect one paralog to behave in a substantially different way than the other, whereas in the baseline and subfunctionalization models we expect the two paralogs to have similar dynamics (Figure 6-2). We therefore analyzed all four models one of the two sets of paralogs, and in the neofunctionalization cases we analyzed those paralogs that had undergone a special model of evolution with a burst and shuffling of sites.

Looking at the two paralog sets together, we identify 334-832 novel sequences (Figure 6-3). The true age of the novel sequences generated by duplicated genes should be in Lamprey, based on the time of our duplication (Figure 6-1). It could be argued that some of these novel sequences require time to arise beyond that due to divergence after duplication, so we can consider any gene whose age is younger than lamprey as a reasonable novel sequence arising as a result of a

duplication. However, we found that many sequences were dated to an age older than the time of sequence divergence even after a full paralog correction; they could not detect a paralog and yet were dated as older than lamprey. For the second paralog, this corresponded to 8% of novel sequences in the baseline simulation, 16% of genes in the neofunctionalized simulation, 12% of genes in the neofunctionalized all sites simulation, and 79% of genes in the subfunctionalized simulation. These sequences are therefore dated as being older than the event which created them.

Comparison to Real Phylostratigraphy

Our goal was to assess the relative contribution of divergence and duplication-divergence to novel sequences, and thus indirectly approach the contribution of *de novo* gene birth to novel sequence formation. This required performing real phylostratigraphy on the same set of 4942 genes in our dataset, and a more realistic distribution of duplication in our simulation.

First, we ran another simulation with a whole-genome duplication every 200 MY using the subfunctionalization (condition 3), and performed phylostratigraphy separately for each duplication event. It is known that not all genes survive duplication, and many duplicates are removed from the genome or pseudogenized. So, to assess the percentage of genes which are duplicated and survive every 200 MY, we downloaded protein databases for human, mouse, and chicken. We restricted each database to only the longest of all alternative splice forms. We then performed two BLAST searches using Human as query and the combination of either human and mouse or human and chicken as the target. For each of the BLAST searches, we reasoned that if the top (non-self) hit was a human protein, this was indicative of a duplication having occurred.

We were thus able to determine how many duplications had occurred since the split of human and mouse (~90MYA) and since the split of human and chicken (~320MYA). We found that 5618/22109 proteins were duplicates after the human-mouse divergence, and 8086/22109 proteins were duplicates after the human-chicken split. Because a single duplication is likely to result in two genes that have this property, we divided each of these numbers by 2. The percentage of genes which were duplicated every 230MY was therefore 1234/22109 genes. This implies that every 200M, the number is approximately 1073/22109, or 4.85%. We therefore determined the number of novel sequences from duplications in each age by summing the number of novel sequences attributable to duplication from all of our sequential 200MY duplication events and taking 5% of their total. For novel sequences not attributable to duplications, we simply took the average over the 12 duplications (Figure 6-4A).

Next, we wished to compare these numbers of novel sequences to the number of true novel sequences in each age, as determined by phylostratigraphy. We took the original human sequences of the 4942 genes we began with and performed a BLASTP search against the NCBI non-redundant database. We further performed a self-BLAST of the original human sequence database, as we had done with our simulated sequences, which allowed us to perform paralog corrections (Figure D-2, method 3). We then grouped these genes as being specific to primates, mammals, chordates, etc. (see methods). We similarly grouped the genes which fell into each age category based on our simulations, either excluding (red line) or including (blue line) genes from duplication (Figure 6-4B, Table 1). Note that after paralog correction, our simulation produced many more genes than real phylostratigraphy. To account for this, we normalized our

simulations to contain in total the same number of genes after paralog correction as real phylostratigraphy (see methods).

We find that for closely-restricted taxa (primates to boreouthera), all species-specific genes are attributable to error. However this is because no genes in the dataset are considered lineage specific under real phylostratigraphy until the taxonomy is considered out to boreouthera, at which point one of the 4942 genes is considered lineage specific. Therefore, in this particular dataset, it is best to consider estimates only for eutherians (105 MYA) and beyond.

We find that when not including duplications, the percent of taxonomically-restricted genes attributable to homology detection error ranges from 8.7% to 18.9%, with a typical value being somewhat stably around 14%, congruent with previous estimates (Chapters 2, 3, 4). When considering error and duplications in conjunction, we find that these numbers range from 11.1% to as high as 37.5%, with a typical value being around 21%. However, it is clear from this data that a non-negligible portion of orphan genes are not attributable to the combined effects of homology detection error or duplication-divergence. When using other methods for paralog correction, these results do not differ substantially (Figures D-5 and D-6, Tables D-1 and D-2).

The ratio of orphan genes attributable to duplication and those due to error remains relatively stable around 0.5 (Table 1). This suggests that under our conditions duplications consistently contribute half as many novel sequences as general divergence. Of course, changes in the assumptions of duplication model and frequency would affect this estimate. But the fact that a somewhat extreme model of duplication cannot account for as many sequences as general

divergence (homology detection error) suggests that homology detection error is a major contributor to phylostratigraphic findings and trends, and should be seriously considered in any phylostratigraphic study.

Discussion

Estimates of the lineage specific genes are influenced by several factors, including duplication rate, changes in functional constraint, *de novo* gene birth, and homology detection error. Because the specific dynamics of *de novo* gene birth and supposed sudden sequence shifts due to change in functional constraint are unknown, it is difficult to directly estimate their contribution. However, because divergence and duplications are better understood, it is possible to investigate these rates directly and other rates indirectly. Additionally, the relationship between orphan-gene status and gene age is unclear.

Here, we have performed the first investigation into the relative contributions of these three sources to orphan gene formation. We have simulated the evolution of 4942 human genes, including a duplication of each gene individually. We demonstrate that while homology detection error and duplication-divergence are non-negligible contributors to the number of orphan genes observed, other mechanisms appar to contribute a non-negligible proportion of orphan genes. However, the contribution of homology detection error and duplication vary depending on the particular taxon restricted to. It is therefore impossible to precisely determine the contribution of each of these sources at any particular internal branch, and thus estimating rates of *de novo* gene formation and fixation via phylostratigraphy is difficult, as there are few

ways to distinguish this mechanism from others. Still, the long-term contributions appear to be relatively stable.

We also find that, whether or not any attempt is made to control for paralogs, homology detection error contributes more to orphan genes than does a duplication-divergence mechanism (Table 6-1, Tables D-1 and D-2). Given that other researchers have pointed to rapid divergence as being a major contributor to phylostratigraphy (Domazet-Lošo et al., 2016), this implies that homology detection error cannot be ignored as a contributor to novel sequence formation and therefore phylostratigraphic trends. Duplications would need to be over twice as frequent or substantially more extreme in order for them to be a greater contributor to novel sequence formation than homology detection error.

We have also established that there is little relationship between age based on historical homology and age based on orphan gene status. It was previously known that homology detection error can create the appearance that genes are phylostratigraphically younger than they are on the basis of historical homology. In this study, we have demonstrated that "novel sequences" can actually appear to be older than their date of emergence, based on the fact that genetic distance does not correctly recapitulate phylogeny (Smith & Pease, 2016). Though they are unable to detect their human paralogs, human duplicate genes were found to be older than their time of duplication on the basis of phylostratigraphy (Figure 6-3). Therefore, phylostratigraphic age is biologically and evolutionarily meaningless, as the age of novel sequences has little bearing on their actual age and sequences can be dated as being both much older and much younger than their actual age. While we were unable to simulate *de novo* gene

birth due to a lack of information about the dynamics of their evolution, we note that these are highly likely to be dated as younger than their actual time of emergence, as they are generally thought to be short and fast-evolving, two traits which are associated with homology detection error (Moyers and Zhang 2015, chapter 5). It is also likely that "novel genes" formed through sudden shifts in sequence space would demonstrate the same problem as some duplicate genes, i.e. finding detectable homologs which are older than the timing of the proposed sudden shift in sequence space.

There are a number of limitations to our study, but we do not regard these as problems for our arguments. It might be argued that we have only simulated the evolution of one quarter of human genes, and therefore our dataset is not representative of actual biological patterns. This is not a meaningful argument, as we restricted our comparison of simulated and real phylostratigraphy to the same set of genes, and therefore the results are representative of the dataset in question. Furthermore, there is no reason to think that our results would change were it possible to simulate the remaining three quarters of human genes (Chapter 5). Those genes that we did not simulate the evolution of were not simulated because there was insufficient data—i.e., they did not have sufficient detectable homologs in the species in question (See methods). In theory, genes with fewer detectable homologs are expected to be shorter and faster-evolving, which gives them a greater propensity for error. Therefore, arguments that a more complete study of orphan genes under our methodology would change the broad strokes of our conclusions are likely incorrect. In fact, it is likely that the contribution of homology detection error to novel sequence formation in these cases would increase, as previously suggested (Chapter 5).

One might also argue that our simulation of duplication is inaccurate. First, our simulations could be inaccurate in the model of duplication used, either being too extreme or too conservative. We have simulated a wide range of models, and compared the novel sequences produced by the most extreme model with real phylostratigraphy to estimate the contributions of duplications. Under even this extreme model, other mechanisms appear to be large contributors. Therefore, our main conclusion that *de novo* gene birth is a non-negligible contributor to novel sequence formation is unaffected. Second, we could be simulating too much duplication or too little. While we estimated the number of duplications per 200 million years, it is possible that this rate is variable over time. Of course, greater or lower amounts of gene duplication have predictable effects on our results. Our method also did not incorporate the effects of whole genome duplication (Figure 6-4), which is known to occur. While there are estimates of the numbers of genes that survive after such an event they can vary widely by species, even within the same event (Mcgrath, Gout, Johri, Doak, & Lynch, 2014). Therefore, further study into the frequency of duplications and whole genome duplications can be included in later studies. However, based on these considerations it is possible that we underestimate the contribution of duplications.

It may be claimed that our results are in congruence with those of Carvunis *et al* (2012), despite our previous study contradicting their results (Moyers and Zhang 2016). These results are not in conflict with our previous study. The methods used in Carvunis *et al* 2012 are biased to produce patterns according to their expectations as they do not account for error (Chapter 5). Because the primary support for their model of frequent *de* novo gene birth was fundamentally the trends

with gene age, the bias created by phylostratigraphic error confounds their findings. In contrast, we here perform an error-aware analysis using models that consider various contributions to apparent orphan gene formation, and compare these to real orphan genes. Though the conclusions of the present study and Carvunis are similar, they use fundamentally different evidences, and our previous criticisms of the work of Carvunis *et al* remain.

The study of orphan genes is an interesting subject, because it may provide information about how species are specialized. While it is clear from our experiment that *de novo* gene birth is a major contributor to novel sequences, further analyses are required to identify a sequence as *de novo* (Knowels and McLysaght 2009). Additionally, ages assigned by phylostratigraphic study of orphan genes are not evolutionarily meaningful, whatever their use in comparative studies of active genes. This study further delineates areas in which phylostratigraphy is and is not a useful and meaningful tool for biological analyses.

References

Albà, M. M., & Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology*, *7*(53). http://doi.org/10.1186/1471-2148-7-53

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*, 370–374. http://doi.org/10.1038/nature11184

Cliften, P. F., Fulton, R. S., Wilson, R. K., & Johnston, M. (2006). After the duplication: Gene loss and adaptation in saccharomyces genomes. *Genetics*, *172*(2), 863–872.

http://doi.org/10.1534/genetics.105.048900

De Smet, R., Adams, K. L., Vandepoele, K., Van Montagu, M. C. E., Maere, S., & Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 2898–2903. http://doi.org/10.1073/pnas.1300127110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1300127110

Domazet-Lošo, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics : TIG*, *23*(11), 531–3. http://doi.org/10.1016/j.tig.2007.07.007

Domazet-Lošo, T., Carvunis, A., Alba, M. M., Sestak, M. S., Bakarić, R., Neme, R., & Tautz, D. (2016). No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Biorxiv*, 1–24.

Domazet-Loso, T., & Tautz, D. (2010). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biology*, *8*(66). http://doi.org/10.1186/1741-7007-8-66

Domazet-Lošo, T., & Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature*, *468*, 815–8. http://doi.org/10.1038/nature09632

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. *Molecular Biology and Evolution*, *23*(1), 1–3. http://doi.org/10.1093/molbev/msj006

Gao, L.-Z., & Innan, H. (2004). Very low gene duplication rate in the yeast genome. *Science*,

306, 1367–1370. http://doi.org/10.1126/science.1102033

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, *22*(23), 2971–2. http://doi.org/10.1093/bioinformatics/btl505

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, *290*(5494), 1151–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11073452

Madden, T., & Morgulis, A. (2009). BLAST Command Line Applications User Manual, *1997*, 1–42.

Mcgrath, C. L., Gout, J., Johri, P., Doak, T. G., & Lynch, M. (2014). Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research*, *24*, 1665–1675. http://doi.org/10.1101/gr.173740.114.24

Moyers, B. A., & Zhang, J. (2015). Phylostratigraphic bias creates spurious patterns of genome evolution. *Molecular Biology and Evolution*, *32*(1), 258–267. http://doi.org/10.1093/molbev/msu286

Moyers, B. A., & Zhang, J. (2016). Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Molecular Biology and Evolution*, *33*(5), 1245–1256. http://doi.org/10.1093/molbev/msw008

Nei, M., & Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, *14*(117). http://doi.org/10.1186/1471-2164-14-117

Pegueroles, C., Laurie, S., & Alba, M. M. (2013). Accelerated evolution after gene duplication: A time-dependent process affecting just one copy. *Molecular Biology and Evolution*, *30*(8), 1830–1842. http://doi.org/10.1093/molbev/mst083

Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.-K., & Douzery, E. J. (2007). OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evolutionary Biology*, *7*(1), 241. http://doi.org/10.1186/1471-2148-7-241

Schmidt, H. A., Strimmer, K., Vingron, M., & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, *18*(3), 502–504. http://doi.org/10.1093/bioinformatics/18.3.502

Sestak, M., & Domazet-Lo o, T. (2014). Phylostratigraphic Profiles in Zebrafish Uncover Chordate Origins of the Vertebrate Brain. *Molecular Biology and Evolution*, *32*(2), 299–312. http://doi.org/10.1093/molbev/msu319

Sestak, M. S., Bozicevic, V., Bakaric, R., Dunjko, V., & Domazet-Loso, T. (2013). Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Front Zool*, *10*(1), 18. http://doi.org/10.1186/1742-9994-10-18

Smith, S. A., & Pease, J. B. (2016). Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. *Briefings in Bioinformatics*, (January), 1–7. http://doi.org/10.1093/bib/bbw034

Stoye, J., Evers, D., & Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics*, *14*(2), 157–163.

Zhang, J. (2013). Gene duplication. In J. Losos (Ed.), *The Princeton Guide to Evolution* (pp. 397–405). Princeton, New Jersey.

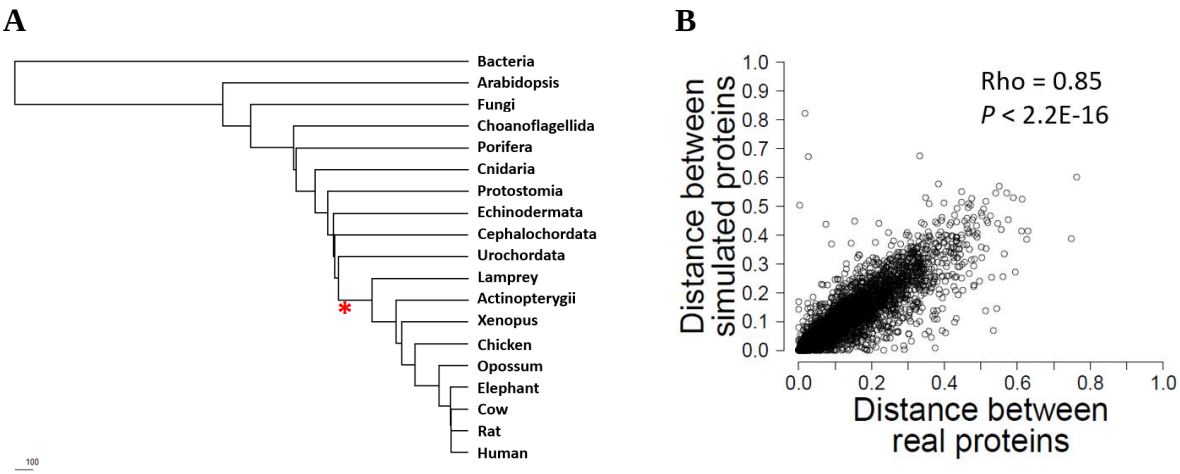

Figure 6- 1 Simulation of evolution

(A) The tree through which we simulated evolution, with proportional branch lengths. Asterisk denotes the time of duplication event. (B) Comparison of real and simulated sequence divergence between human and rat proteins for all proteins in the simulation. Spearman's Rho and the associated p-value are reported.

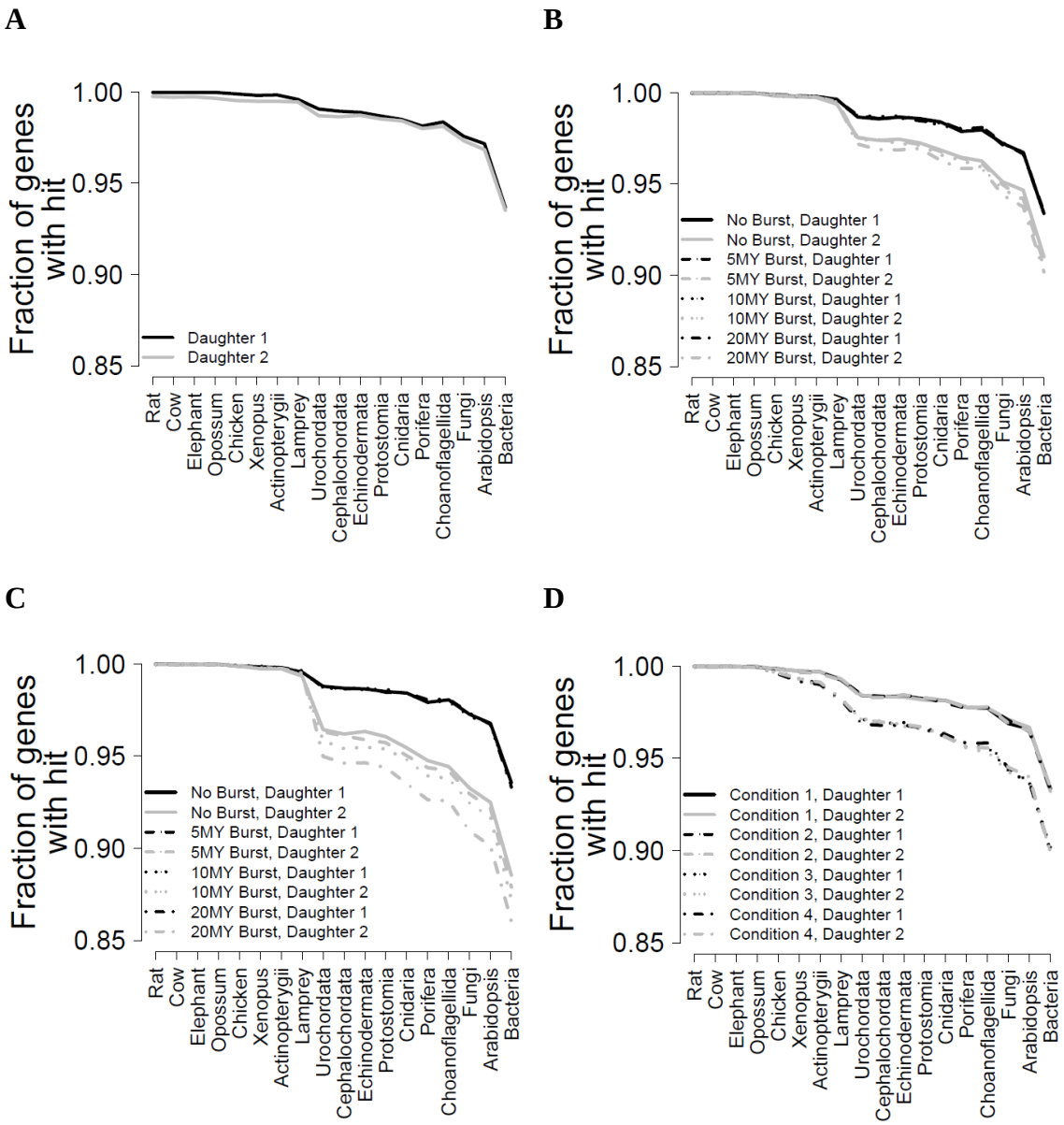

Figure 6- 2 Percentage of genes lacking a homolog in each phylostratum

(A) Baseline, (B) Neofunctionalized, (C) Neofunctionalized, All Sites, and (D) Subfunctionalized.

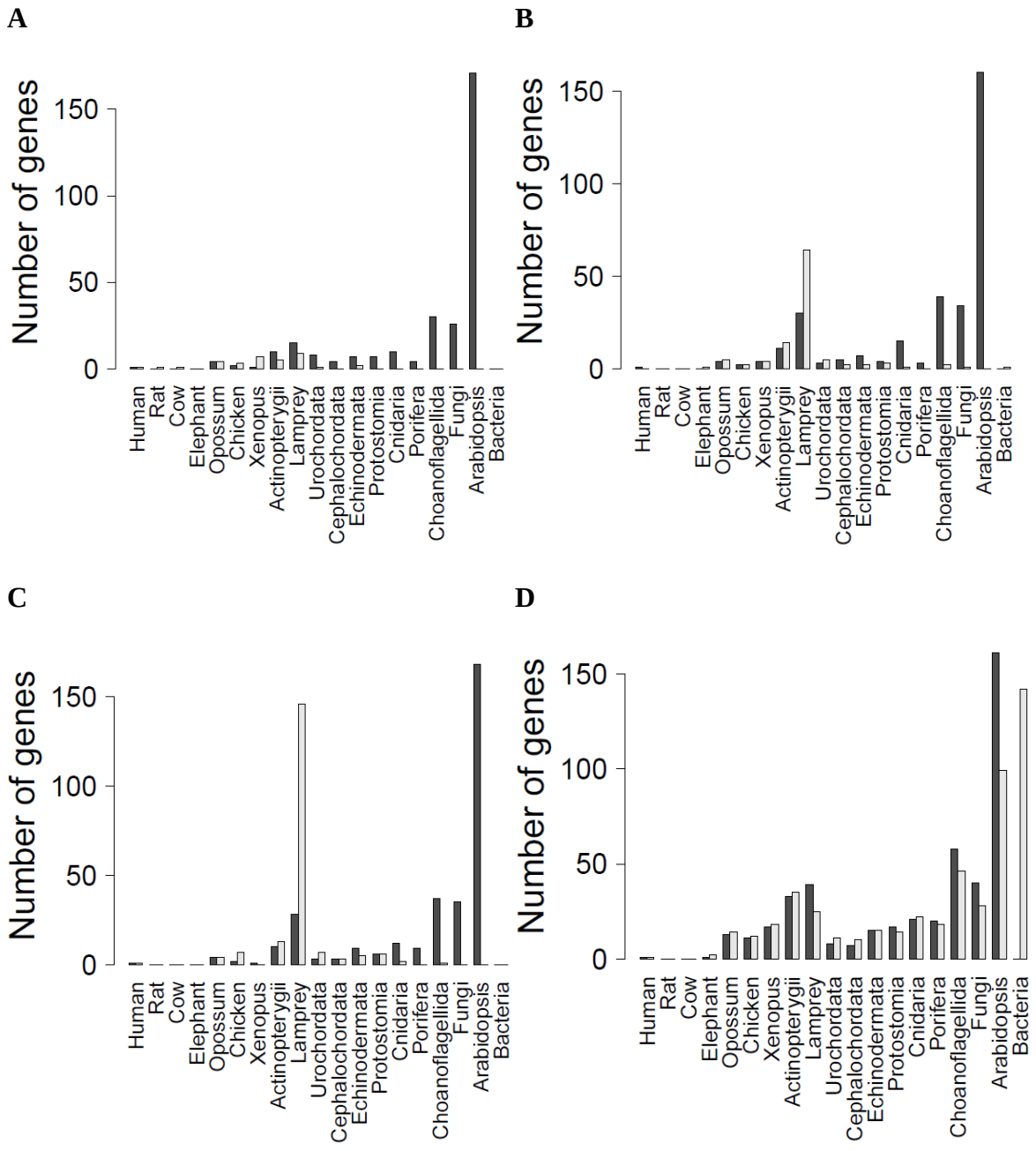

Figure 6- 3 Number of novel sequences at each age

Number of novel sequences at each apparent age when the age and number of sequences have been corrected with respect to paralogs (method 3). Dark grey bars denote the first of the two paralogs, while light grey bars denote the second of the two paralogs. Note that we do not display genes mapped to bacteria for dark grey bars, for scaling purposes. In the two neofunctionalization simulations, the second of the two paralogs is the paralog which underwent a burst of evolution and subsequent shuffling of rates. We include here a count for the number of genes at each age, for each of the two paralogs. (A) Baseline; paralog 1: c(1, 0, 0, 0, 4, 2, 1, 10, 15, 8, 4, 7, 7, 10, 4, 30, 26, 171, 4642), paralog 2: c(1, 1, 1, 0, 4, 3, 7, 5, 9, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0). (B) Neofunctionalization; paralog 1: c(1, 0, 0, 0, 4, 2, 4, 11, 30, 3, 5, 7, 4, 15, 3, 39, 34, 160, 4620), paralog2: c(0, 0, 0, 1, 5, 2, 4, 14, 64, 5, 2, 2, 3, 1, 0, 2, 1, 0, 1). (C) Neofunctionalization, all sites; paralog 1: c(1, 0, 0, 0, 4, 2, 1, 10, 28, 3, 3, 9, 6, 12, 9, 37, 35, 168, 4614), paralog 2: c(1, 0, 0, 0, 4, 7, 0, 13, 146, 7, 3, 5, 6, 2, 0, 1, 0, 0, 0). (D) Subfunctionalization; paralog 1: c(1, 0, 0, 1,

13, 11, 17, 33, 39, 8, 7, 15, 17, 21, 20, 58, 40, 161, 4480), paralog 2: c(1, 0, 0, 2, 14, 12, 18, 35, 25, 11, 10, 15, 14, 22, 18, 46, 28, 99, 142).

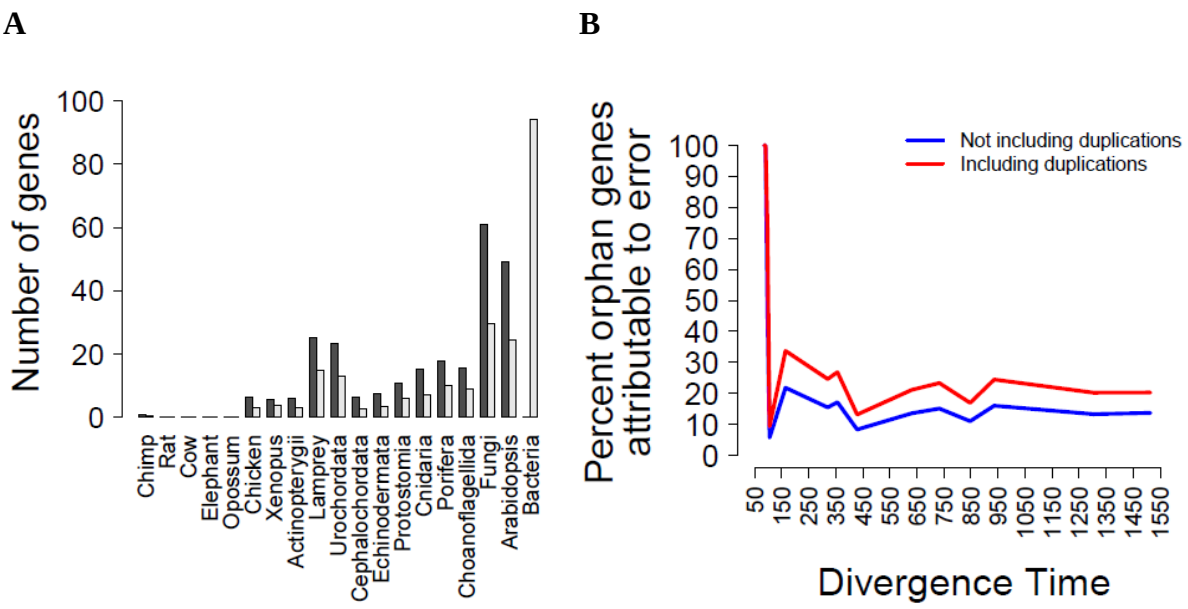

Figure 6- 4 Results of regular small duplications under a subfunctionalization model

(A) Number of genes in each age category in the simulation after paralog correction (method 3). The numbers of genes in each bin are as follows, rounded up to the nearest whole number. Paralog1: c(1, 0, 1, 1, 1, 7, 6, 7, 26, 24, 7, 8, 11, 16, 18, 16, 61, 50, 217, 4474). Paralog2: c(1, 0, 0, 1, 1, 4, 4, 4, 15, 14, 3, 4, 6, 8, 11, 10, 30, 25, 95, 221) (B) The percent of orphan genes which are attributable to general error and to duplication under this model.

Table 6- 1 Numbers of genes in real and simulated phylostratigraphy, rounded to the nearest whole gene

Div Time (MYA)	Real orphans	Simulated Orphans (with dups)	Percent attributable to error (error+dups)	Dups/Error
85	0	1 (1)	100 (100)	0
90	0	1 (1)	100 (100)	0
97	1	1 (1)	100 (100)	0
105	9	1 (1)	11.11 (11.11)	0
164	16	4 (6)	25 (37.5)	0.5
320	40	7 (10)	17.5 (25)	0.43
356	53	10 (15)	18.87 (28.30)	0.5
429	253	22 (34)	8.70 (13.44)	0.55
631	260	36 (55)	13.85 (21.15)	0.53
733	292	45 (69)	15.41 (23.63)	0.53
847	466	52 (79)	11.16 (16.95)	0.52
936	601	97 (147)	16.14 (24.46)	0.52
1303	901	120 (183)	13.32 (20.31)	0.53
1514	1628	223 (330)	13.70 (20.27)	0.48

Chapter 7

Conclusions and Future Directions

"I have no doubt that in reality the future will be vastly more surprising than anything I can imagine. Now my own suspicion is that the Universe is not only queerer than we suppose, but queerer than we can *suppose."*
- J. B. S. Haldane, 1927

Phylostratigraphy as a method has given us a new tool to investigate the molecular relationships between species and the nature of biological diversity. However, because it is based on tools with known limitations (Albà & Castresana, 2007; Elhaik, Sabath, & Graur, 2006; Moyers & Zhang, 2015, 2016; Rost, 1999; Smith & Pease, 2016), this method does have a problem with homology detection error. Whether this error is seen as a problem with the method (Chapters 2 and 3) or as a contributor to novel sequences and biological innovation (Chapters 4 and 5), it is clear that the influence of these errors on phylostratigraphic analysis cannot be ignored. This is made clear by the fact that the errors are nonrandom (Chapters 2 and 5) and can substantially influence phylostratigraphic findings (Chapters 2, 3, and 4).

It is generally understood that in science a theory or method is only overturned when it can be replaced with a theory of greater accuracy and explanatory power (Kuhn, 1996). I have here developed and applied a method for improving upon phylostratigraphic analysis to produce new biological insights (Chapters 5 and 6). In even minor scientific revolutions, the replacement of a paradigm opens up new avenues for research and discovery. Below I discuss the major outstanding problems for phylostratigraphy which should be addressed in the future.

Clarification of the definition of novels sequences

It has been argued that phylostratigraphy is a method for identifying novel sequences (Domazet-Lošo et al., 2016; Domazet-Lošo, Brajkovic, & Tautz, 2007). This definition differs in an important way from the concept of a historical homolog, i.e. sequences which are derived from a common ancestral sequence. The accepted way of identifying such sequences is through the use of a homology detection program, such as BLASTP. However, the developers of this method recognize that there are at least two ways in which these tools can fail to reproduce their idea of a novel sequence. They can either 1) fail to identify a homologous relationship which they consider to not represent a novel sequence or 2) identify a homologous relationship between two historically homologous proteins which they regard as two novel sequences. This highlights a key failure in the definition: there is as of yet no independent measure and definition for novel sequences. Future directions will require clarification of this concept.

Improved simulation of molecular evolution

Our assessment of the accuracy of phylostratigraphy and the capabilities of homology detection tools has been based on ROSE (Random mOdel of Sequence Evolution) (Stoye, Evers, & Meyer, 1998), because it allows precise assignment of rate heterogeneity parameters along a sequence and modification of indel frequency and size. However, this tool has several limitations for application to this problem. It is unable to respect functional constraints in nucleotide sequence evolution, allowing the creation of stop codons and destruction of start codons. It also does not simulate characteristics which are important for the use of some homology detection tools, such as the structure of a resulting protein. Future work will require more accurate models of

sequence and structural evolution for a better understanding of both evolutionary trends and the assessment of homology detection tools. This is an active area of research, with many exciting avenues being explored (Arenas, 2012; Carvajal-rodríguez, 2010).

Related to the ability to simulate diverse evolutionary parameters is the necessary ability to determine those parameters accurately. Most pointedly, understanding the dynamics of how *de novo* genes mature is a key problem which will require the collection of large numbers of well-curated examples of *de novo* gene birth of varying ages. While some phylostratigraphic studies have purported to do this (Carvunis et al., 2012; Neme & Tautz, 2013), we have demonstrated that these studies are subject to error (Chapters 4 and 5) and that these studies cannot distinguish between different contributors of novel sequences (Chapter 6). Important reviews have been written outlining some of the challenges on this front (Mclysaght & Hurst, 2016; Schlotterer, 2015).

In addition to this problem, there are several other evolutionary events, parameters, and trends which require further elucidation. Though we have performed an initial probe into this realm (Chapter 6), further research is needed into the relative contributions of homology detection error, duplication, and *de novo* gene birth to novel sequence formation, as well as consideration of other potential mechanisms for novel sequence formation. Because our work has been criticized for not respecting lineage-specific evolutionary rates (Domazet-Lošo et al., 2016), further characterization of rate heterogeneity among branches is required. Additionally, Domazet-Loso *et al* (2016) noted that changes in functional constraints may be a driver for novel

sequence formation, but it is not clear how frequently such events occur nor how drastic their effects on sequence divergence are.

Conservative methodology and novel biological insight

Our error-aware phylostratigraphic methods are clearly a generally conservative method, as they require evidence of the non-error-prone status of a gene for inclusion in phylostratigraphic study. One might argue that this method is too stringent, and that it discards many sequences which are not error-prone for lack of evidence. That may well be true, but conservative methodology is generally accepted as a positive quality of scientific tests. It is much better to fail to reject the null when an effect exists than reject the null when an effect does not exist. This is therefore a more appropriate method. As our ability to simulate molecular evolution improves, we expect that more sequences can be confidently included in phylostratigraphy, and biological signal can be rescued.

It bears emphasizing that this conservative methodology has produced novel biological insight, as demonstrated in chapters 5 and 6. In chapter 5, we demonstrated that two well-established phylostratigraphic trends were not found to be true—the relationship between age and length was reversed compared to previous findings, and the relationship between age and evolutionary rate was found to be non-existent. One might argue that by being so restrictive with our dataset, we have removed real biological signal and these results are only due to the particular kinds of genes that we are able to simulate. Even if this argument is true, this still provides an interesting biological insight. By binning data in relevant ways, one can show that an average trend in the data may not hold true for all subsets of the data. If that is occurring in this case, we might find

that increases of length with gene age are not true for all subsets. It could be the case, for instance, that truly young genes do have a relatively rapid increase in length to allow specialization for their selected function. In this hypothetical scenario, once a gene has reached a length sufficient to allow it to specialize for its function, selection could drive the gene toward pruning its length—by having a shorter length, there are fewer potential sites for mutation which could cause the protein to lose its function, and a shorter length of a gene also has less of an energetic burden to express the gene. Therefore, even if one argues that our methodology is too conservative, it is still the case that this approach can provide novel biological insight by binning of genes into relevant categories—i.e. those that have some baseline level of conservation, and those which do not.

Broader Implications

The research here conducted has bearing only for dating the emergence of sequences based on homology, and downstream analyses. However, its logic is potentially further reaching. Fundamentally, we note that the use of homology to determine the age of sequences can produce a false estimate of that sequence's date of emergence, typically an underestimate due to a false negative error. Evolutionary biologists are interested in the emergence of several kinds of features, though, including large structural phenotypes. If one uses homology of some phenotype to estimate the date of emergence for that phenotype, one may similarly estimate the phenotype as having emerged more recently than it actually did. As an instructive example, if one were to estimate the date of the emergence of feathers, they might use the most recent common ancestor of birds and the approximate time that it existed to date the emergence of feathers. However, recent evidence has suggested that feathers existed outside of the dinosaur

lineage which lead to birds (Godefroit et al., 2014), suggesting a significantly older date for feathers may exist. This emphasizes that the issues approached in this thesis have broad, if tenuous, implications for larger areas of evolutionary biology.

Reassessment of phylostratigraphic trends

We have demonstrated clearly the homology detection error, because it is biased, influences phylostratigraphic trends (Chapters 2 and 3), and that this kind of error disproportionately affects reported trends (Chapter 4). We have further offered a method which accounts for this error, and demonstrated that it can offer novel biological insight (Chapters 5 and 6). We encourage the community of researchers using phylostratigraphy to reassess previous findings in light of this new method, and to improve upon it. The current method of assessing the error-prone status of genes can assess far fewer than all genes in a given species. This means that large numbers of genes cannot be considered in error-aware phylostratigraphy. Future work should focus on attempting to assess the error-prone status of these genes for the clearer inference of evolutionary trends.

References

Albà, M. M., & Castresana, J. (2007). On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology*, *7*(53). http://doi.org/10.1186/1471-2148-7-53

Arenas, M. (2012). Simulation of Molecular Data under Diverse Evolutionary Scenarios. *PLoS Computational Biology*, *8*(5). http://doi.org/10.1371/journal.pcbi.1002495

Carvajal-rodríguez, A. (2010). Simulation of Genes and Genomes Forward in Time. *Current*

Genomics, *11*, 58–61.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*, 370–374. http://doi.org/10.1038/nature11184

Domazet-Lošo, T., Brajkovic, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics : TIG*, *23*(11), 531–3. http://doi.org/10.1016/j.tig.2007.07.007

Domazet-Lošo, T., Carvunis, A., Alba, M. M., Sestak, M. S., Bakarić, R., Neme, R., & Tautz, D. (2016). No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Biorxiv*, 1–24.

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. *Molecular Biology and Evolution*, *23*(1), 1–3. http://doi.org/10.1093/molbev/msj006

Godefroit, P., Sinitsa, S. M., Dhouailly, D., Bolotsky, Y. L., Sizov, A. V, Mcnamara, M. E., … Spagna, P. (2014). A Jurassic ornithischian dinosaur from Siberia with both feathers and scales. *Science*, *345*(6195).

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd ed.). University of Chicago Press.

Mclysaght, A., & Hurst, L. D. (2016). Open questions in the study of de novo genes: what, how and why. *Nature Reviews Genetics*, *17*(9), 567–578. http://doi.org/10.1038/nrg.2016.78

Moyers, B. A., & Zhang, J. (2015). Phylostratigraphic bias creates spurious patterns of genome evolution. *Molecular Biology and Evolution*, *32*(1), 258–267.

http://doi.org/10.1093/molbev/msu286

Moyers, B. A., & Zhang, J. (2016). Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Molecular Biology and Evolution*, *33*(5), 1245–1256. http://doi.org/10.1093/molbev/msw008

Neme, R., & Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics*, *14*(117). http://doi.org/10.1186/1471-2164-14-117

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94. http://doi.org/10.1093/protein/12.2.85

Schlotterer, C. (2015). Genes from scratch – the evolutionary fate of de novo genes. *Trends in Genetics*, *31*(4), 215–219. http://doi.org/10.1016/j.tig.2015.02.007

Smith, S. A., & Pease, J. B. (2016). Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. *Briefings in Bioinformatics*, (January), 1–7. http://doi.org/10.1093/bib/bbw034

Stoye, J., Evers, D., & Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics*, *14*(2), 157–163.

Appendix A

Supplementary Figures and Tables for Chapter 3

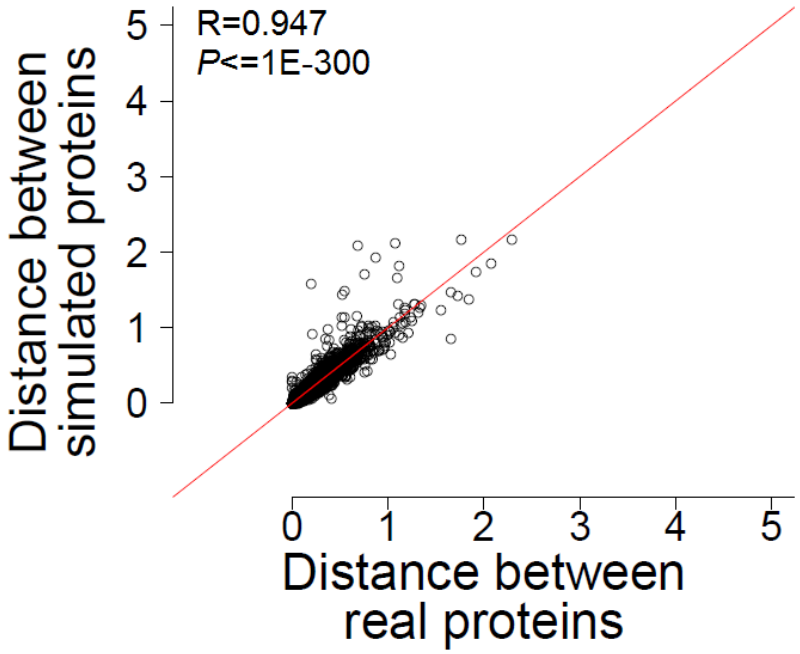

Figure A- 1 Comparison of real and simulated genetic distances

Correlation between distances estimated from real sequences and estimated from simulated sequences for 5259 genes between *S. cerevisiae* and *S. bayanus*. Genetic distance was estimated by the maximum likelihood method in TreePuzzle. Pearson's correlation coefficient (*R*) and associate *P*-value are indicated. The diagonal is shown by a red line.

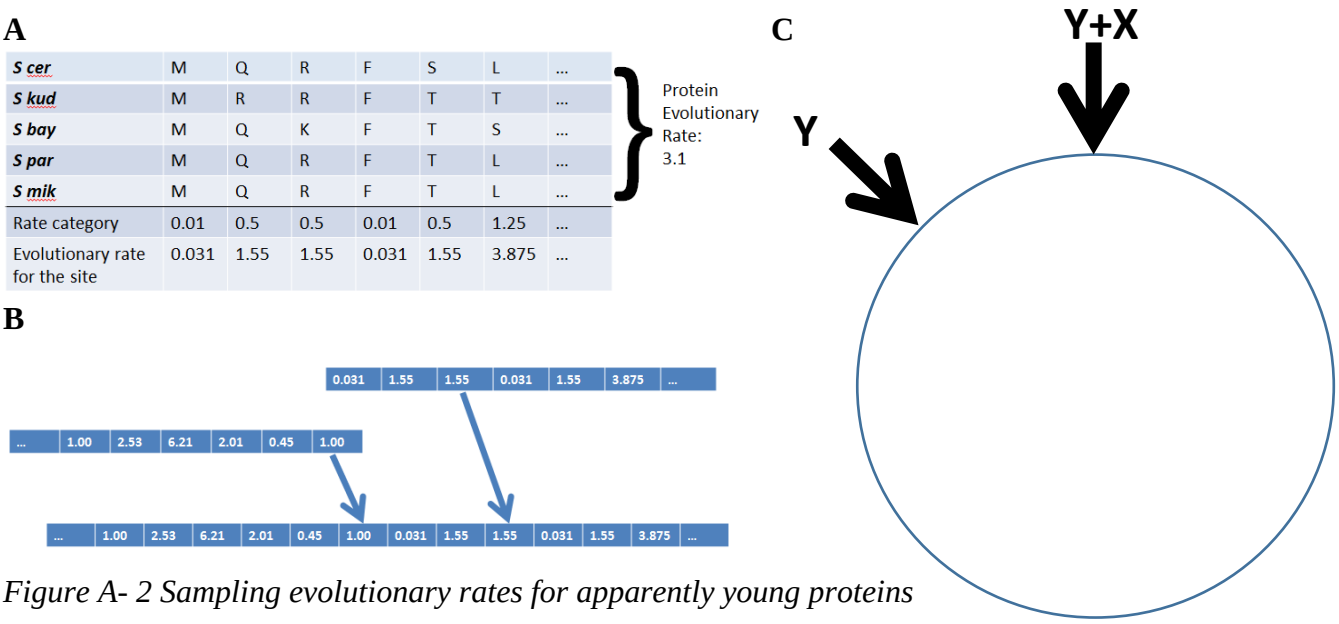

Figure A- 2 Sampling evolutionary rates for apparently young proteins

(A) Hypothetical creation of vectors corresponding to evolutionary rates and rate heterogeneity of a protein found in all *sensu stricto* yeast species but not in other yeast species. Evolutionary rate for a given site is the product of the rate category of that site and the protein evolutionary rate. (B) Concatenation of vectors corresponding to site-specific evolutionary rates of proteins which are found in the five *sensu stricto* species of yeast, but not in other species. This vector is then made into a ring by connecting the two ends. (C) Example of sampling of site-specific evolutionary rates for a protein of length X. A random location (Y) along the ring vector is selected as the start of this protein. Afterwards, a location Y+X is identified. All sites in the vector are then copied to serve as the evolutionary rate information for the apparently young protein of interest.

Table A- 1 Correlations between various gene properties known to bias phylostratigraphy using gene ages 0-10

	Evolutionary rate	ORF length	Expression level
Transcription factor binding sites	-0.09*	0.03**	0.09**
Codon adaptation index	-0.33**	0.04**	0.03**
Optimal AUG context	-0.14**	0.04**	0.06**
Purifying selection	-0.22**	0.47**	0.17**
Mean hydropathicity	0.03*	-0.14**	-0.10**
Percent in disordered regions	0.05*	0.13**	0.01
Percent in transmembrane regions	0.07*	-0.07*	-0.07*
Genetic coregulation	-0.10**	0.03*	0.07*
Number of transcription factors	-0.07*	0.02*	0.02*
Feed-forward loops	-0.07*	0.02	0.03*
Percent alpha helices	-0.05*	-0.07*	0.09**
Percent beta sheets	-0.01	-0.22**	0.02*
Aggregation propensity	0.05*	-0.06*	-0.11**
Protein-protein interactions	-0.23**	0.11**	0.15**
Genetic interactions	-0.11**	0.11**	0.04*
Average epistasis	-0.12**	0.05*	0.10**

* $P < 0.05$; ** $P <$ 1E-16.

Appendix B

Supplementary Figures and Tables for Chapter 4

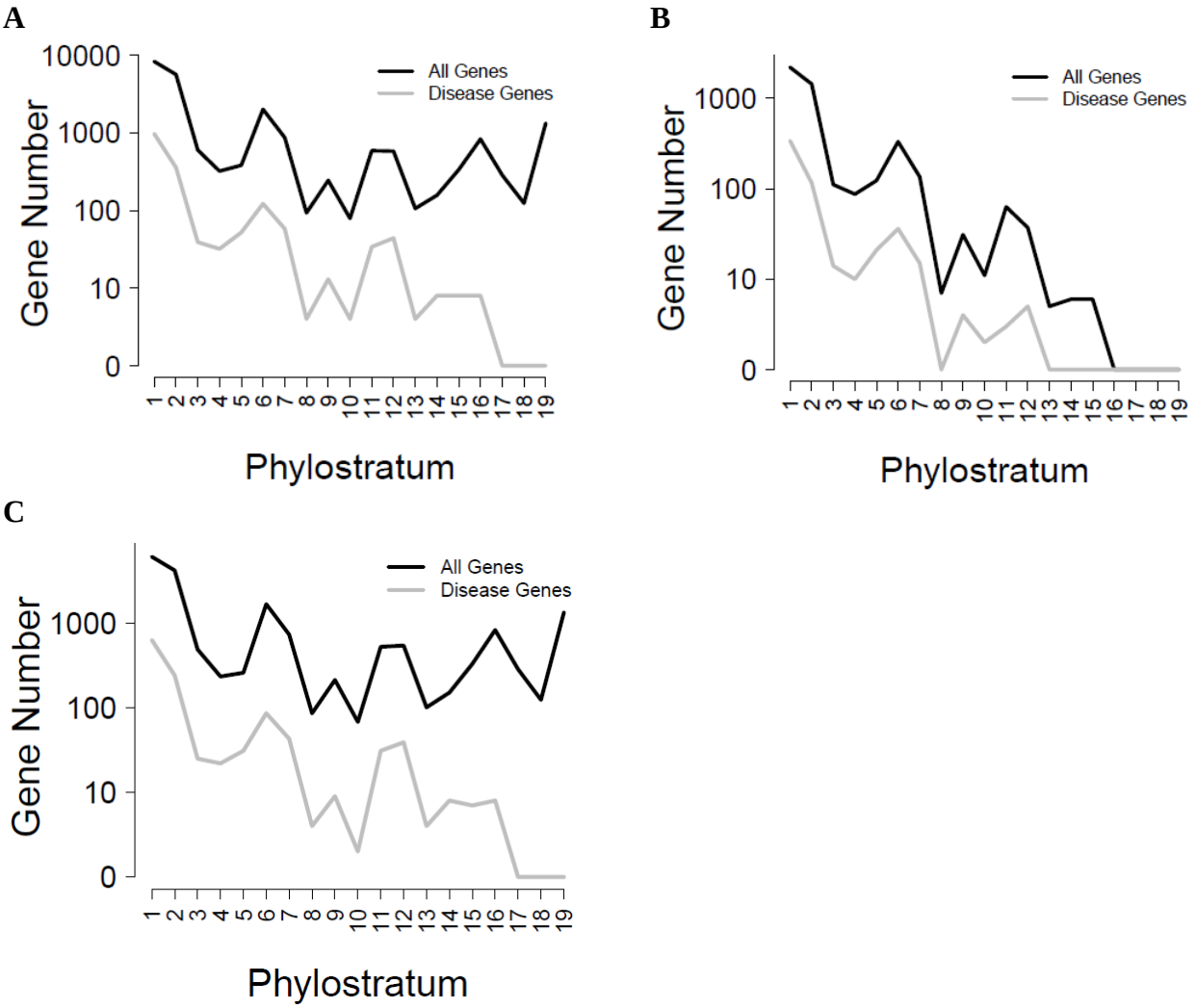

Figure B- 1 Gene number in each phylostratum by disease status

Note that the y-axis is plotted in log10 scale. (A) All genes (black, Rho=-0.37, p=0.121) and disease-associated genes (grey, Rho=-0.81, p=2.49E-5) when no correction for error has been made. (B) All genes (black, rho=-0.93, p=9.08E-9) and disease-associated genes (grey, rho=-0.865, p=1.72E-6) when genes have been restricted to non-error-prone genes. (C) All genes (black, Rho=-0.26, p=0.282) and disease-associated genes (grey, rho=-0.78, p=9.11E-5)

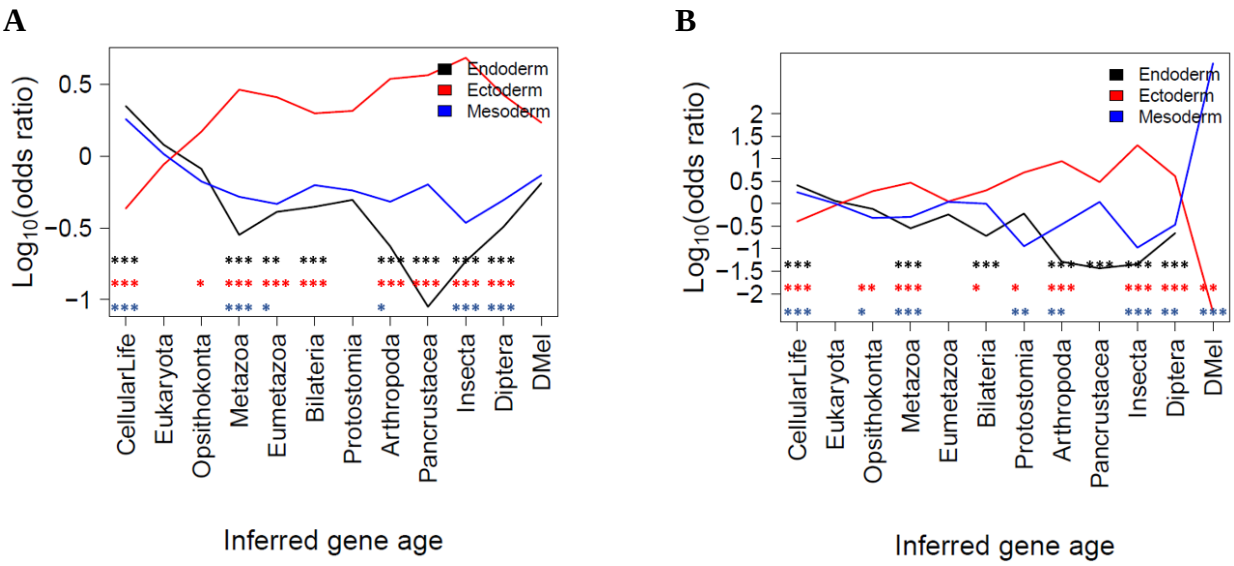

Figure B- 2 Reconstruction of drosophila developmental figures

Reconstruction of figures using (A) all genes or (B) only genes which were simulated and not found to be error-prone.

Appendix C

Supplementary Figures and Tables for Chapter 5

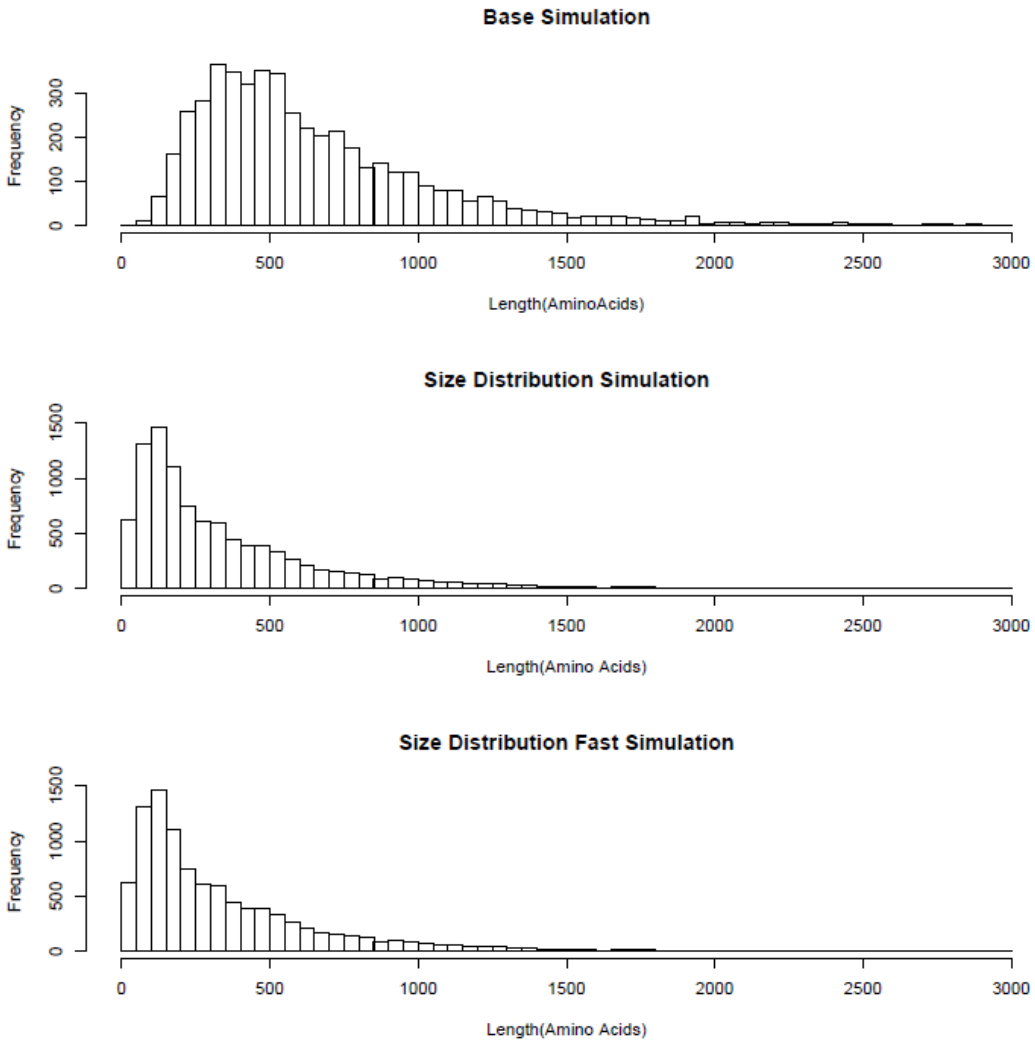

Figure C- 1 Length distribution of three protein sets prior to simulation

The first row shows the length distribution of our Base set. The second row shows the length distribution of our Size Distribution set. The third row shows the length distribution of our Size Distribution Fast set.

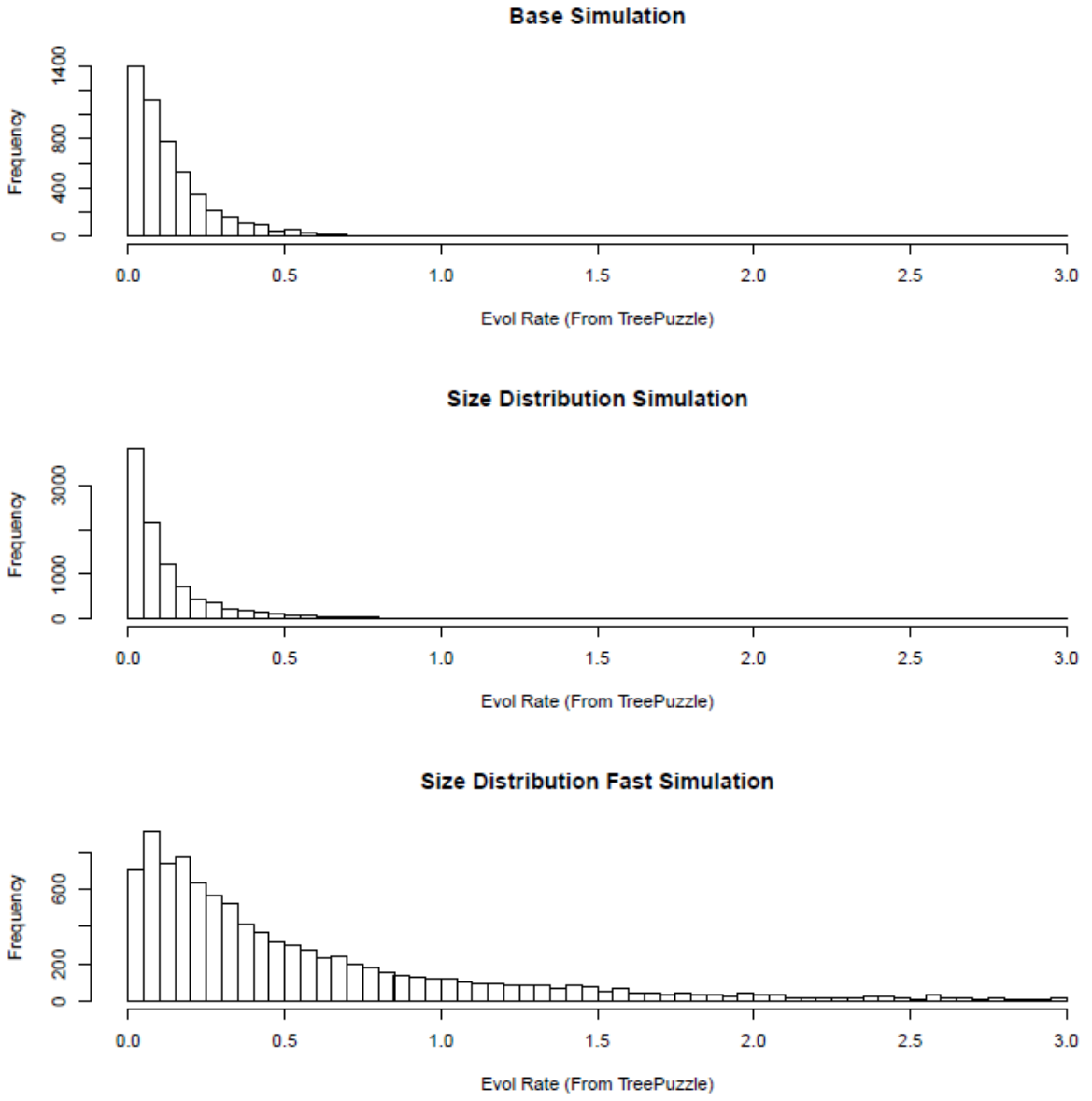

Figure C- 2 Evolutionary rate distribution of three protein sets prior to simulation

The first row shows the evolutionary rate distribution of our Base set. The second row shows the evolutionary rate distribution of our Size Distribution set. The third row shows the evolutionary rate distribution of our Size Distribution Fast set.

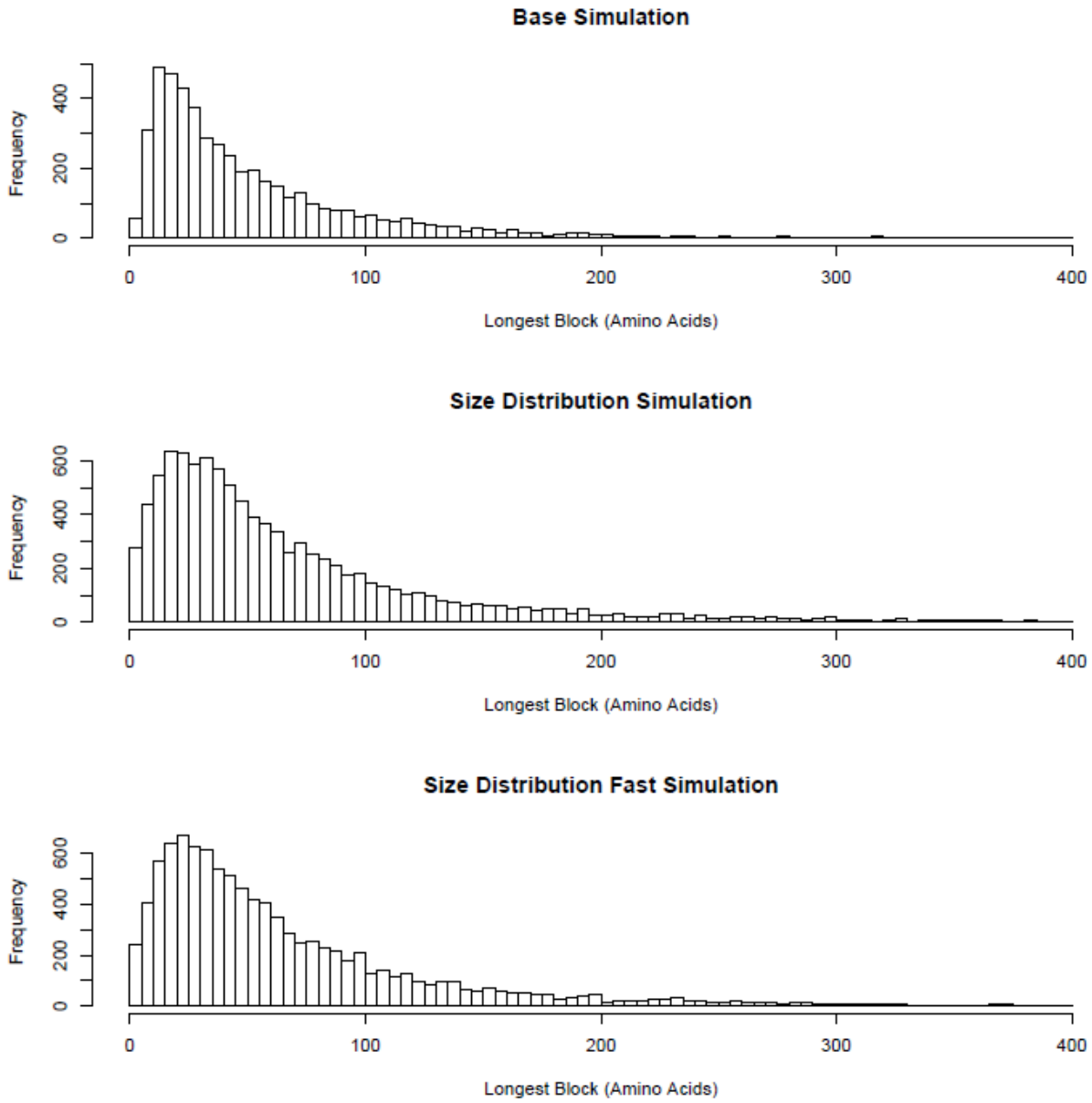

Figure C- 3 Conserved block size distribution of three protein sets prior to simulation

The first row shows the conserved block size distribution of our Base set. The second row shows the conserved block size distribution of our Size Distribution set. The third row shows the conserved block size distribution of our Size Distribution Fast set.

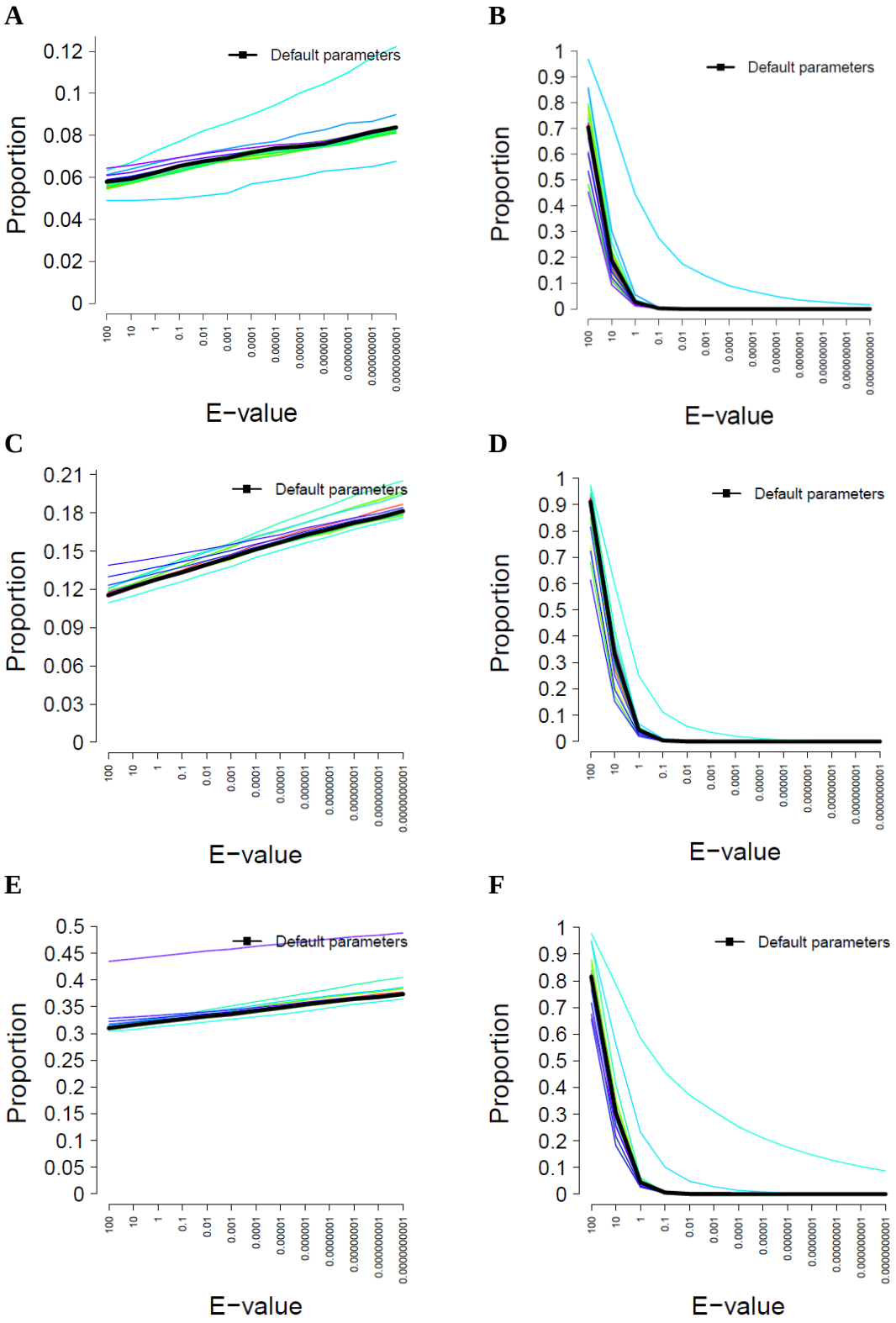

Figure C- 4 False negative and false positive rates in detecting bacterial homologs for BLASTP

The left column shows false negative rates while the right column shows false positive rates. The first row shows the results of our Base set, the second row shows the results of our Size Distribution set, and the third row shows the results of our Size Distribution Fast set. (A) False negative rates for Base set. (B) False positive rates for Base set.

(C) False negative rates for Size Distribution set. (D) False positive rates for Size Distribution set. (E) False negative rates for Size Distribution Fast set. (F) False positive rates for Size Distribution Fast set.

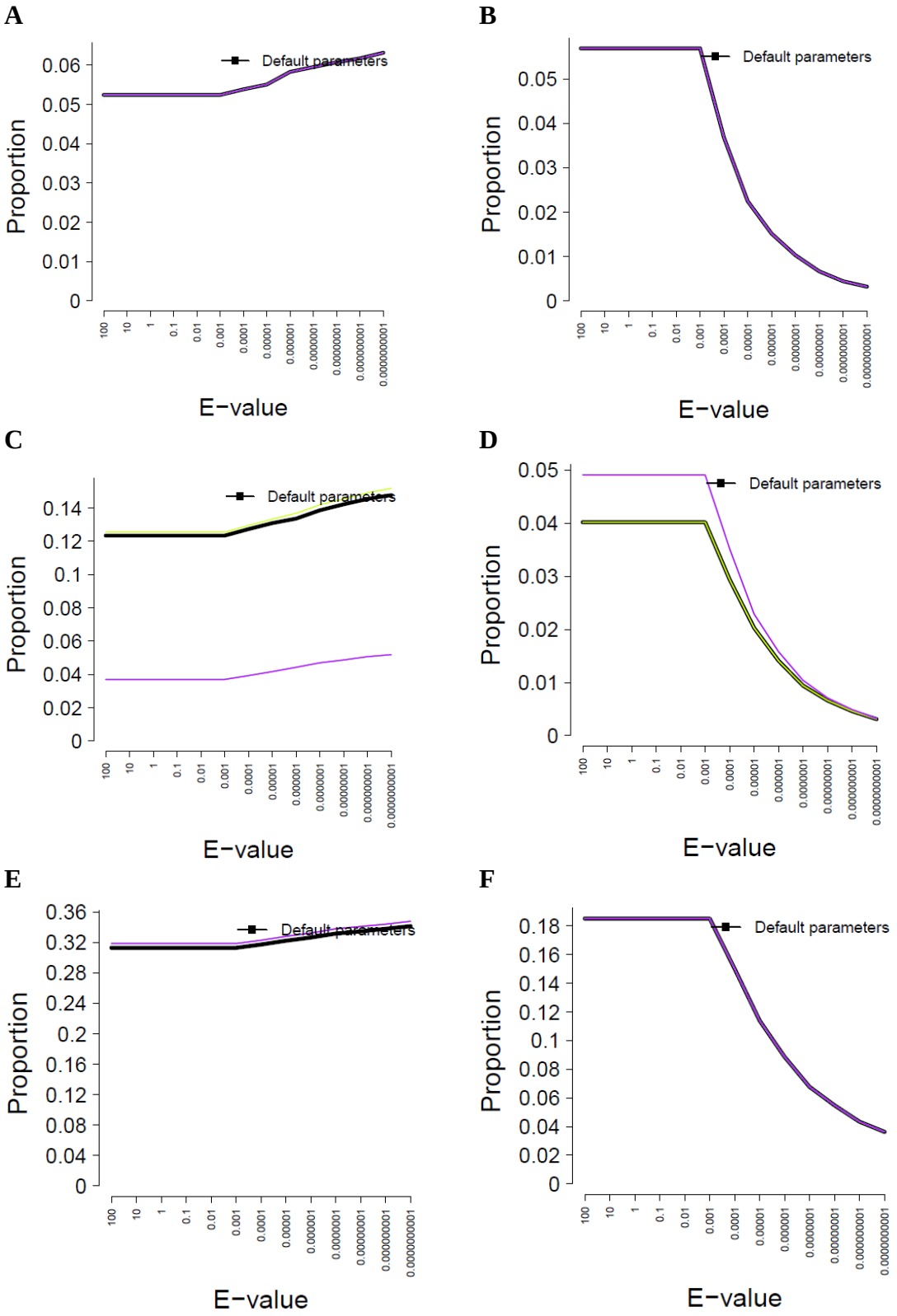

Figure C- 5 False negative and false positive rates in detecting bacterial homologs for PSIBLAST

False negative and false positive rates in detecting Bacterial homologs for all PSIBLAST parameter sets in our three simulated sets of proteins. The left column shows false negative rates while the right column shows false positive rates. The first row shows the results of our Base set, the second row shows the results of our Size Distribution set, and the third row shows the results of our Size Distribution Fast set. (A) False negative rates for Base set. (B) False positive rates for Base set. (C) False negative rates for Size Distribution set. (D) False positive rates for Size Distribution set. (E) False negative rates for Size Distribution Fast set. (F) False positive rates for Size Distribution Fast set.

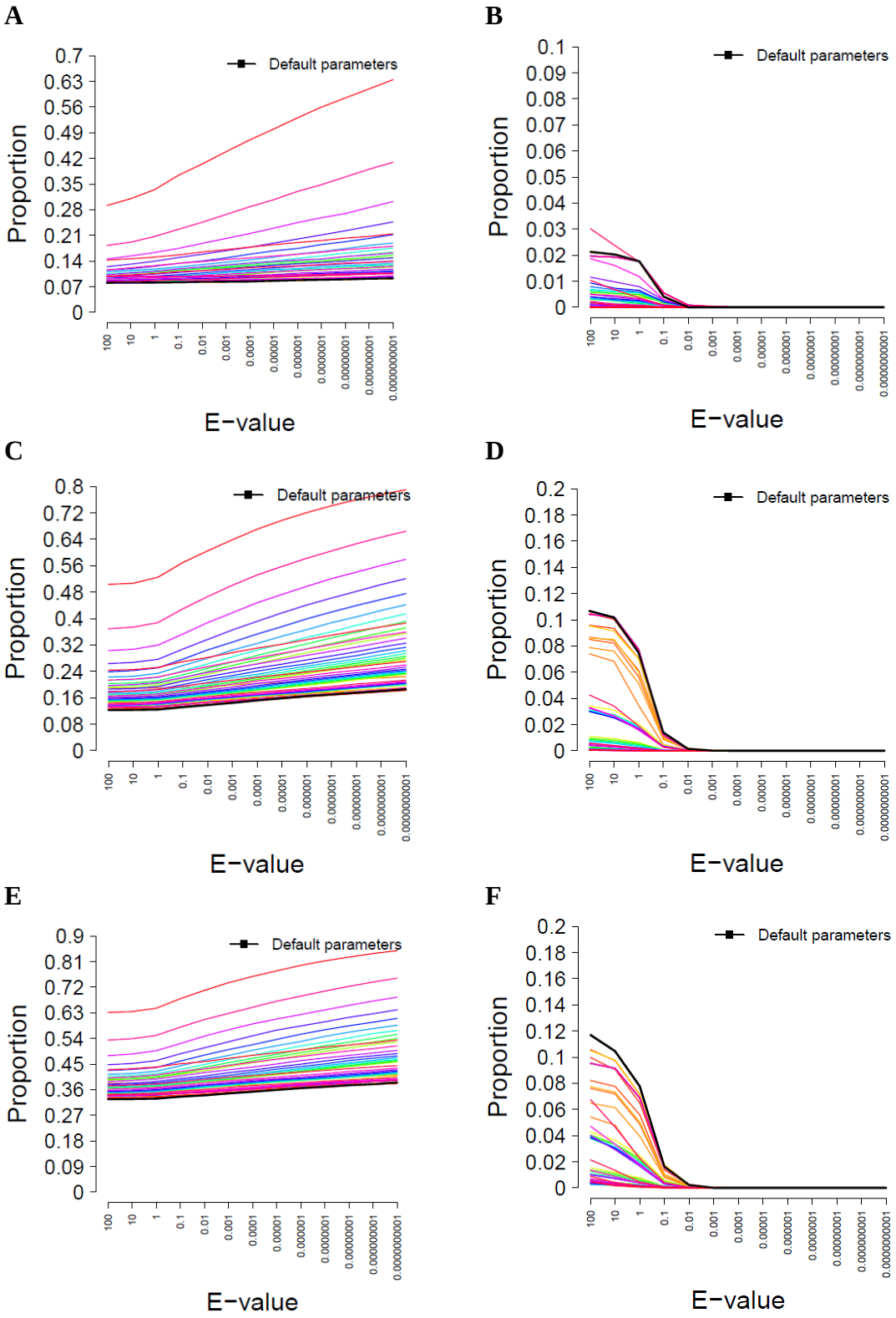

Figure C- 6 False negative and fals positive rates in detecting bacterial homologs for PHMMER

The left column shows false negative rates while the right column shows false positive rates. The first row shows the results of our Base set, the second row shows the results of our Size Distribution set, and the third row shows the

results of our Size Distribution Fast set. (A) False negative rates for Base set. (B) False positive rates for Base set. (C) False negative rates for Size Distribution set. (D) False positive rates for Size Distribution set. (E) False negative rates for Size Distribution Fast set. (F) False positive rates for Size Distribution Fast set.

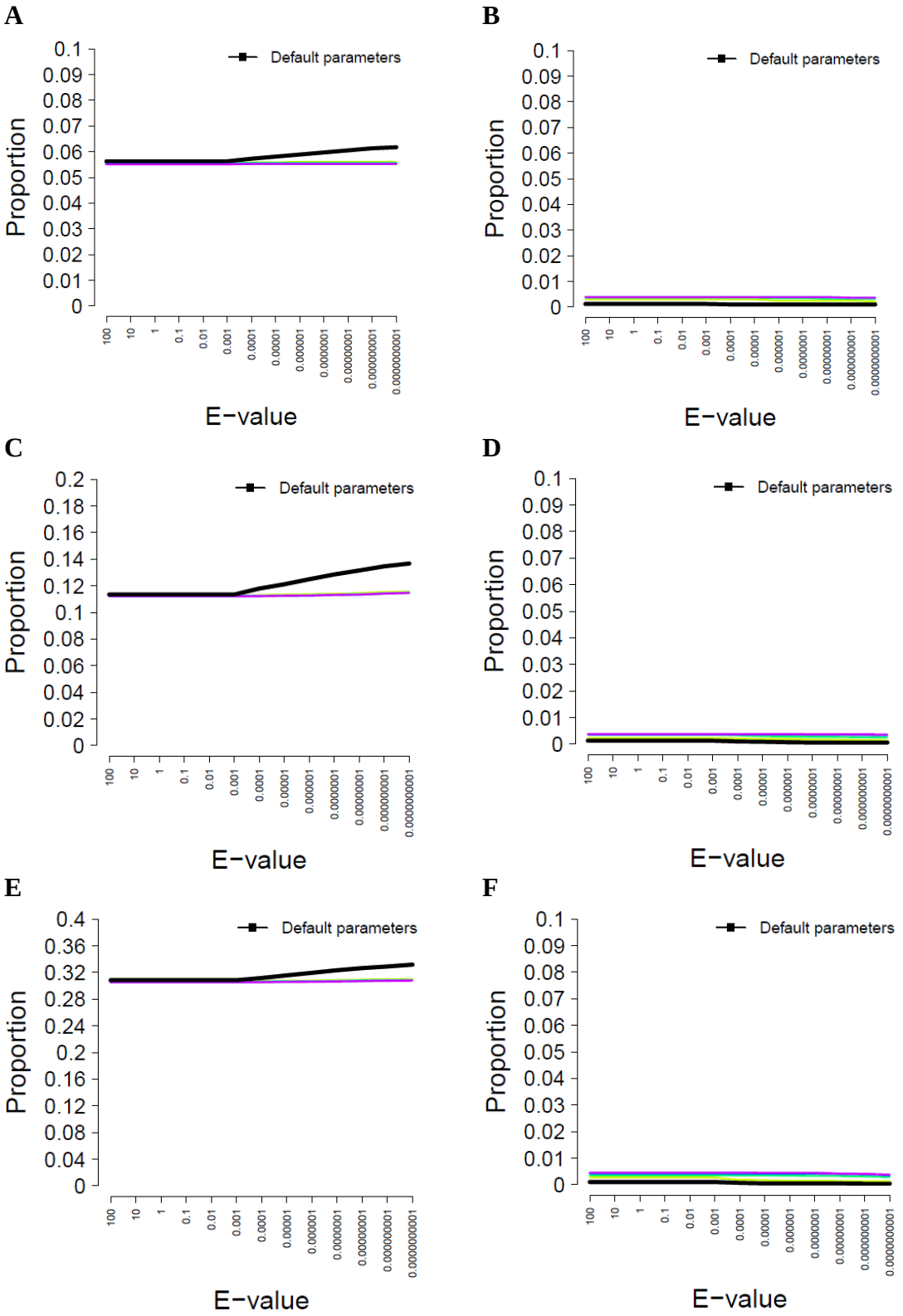

Figure C- 7 False negative and false positive rates in detecting bacterial homologs for HMMER

The left column shows false negative rates while the right column shows false positive rates. The first row shows the results of our Base set, the second row shows the results of our Size Distribution set, and the third row shows the

results of our Size Distribution Fast set. (A) False negative rates for Base set. (B) False positive rates for Base set. (C) False negative rates for Size Distribution set. (D) False positive rates for Size Distribution set. (E) False negative rates for Size Distribution Fast set. (F) False positive rates for Size Distribution Fast set.

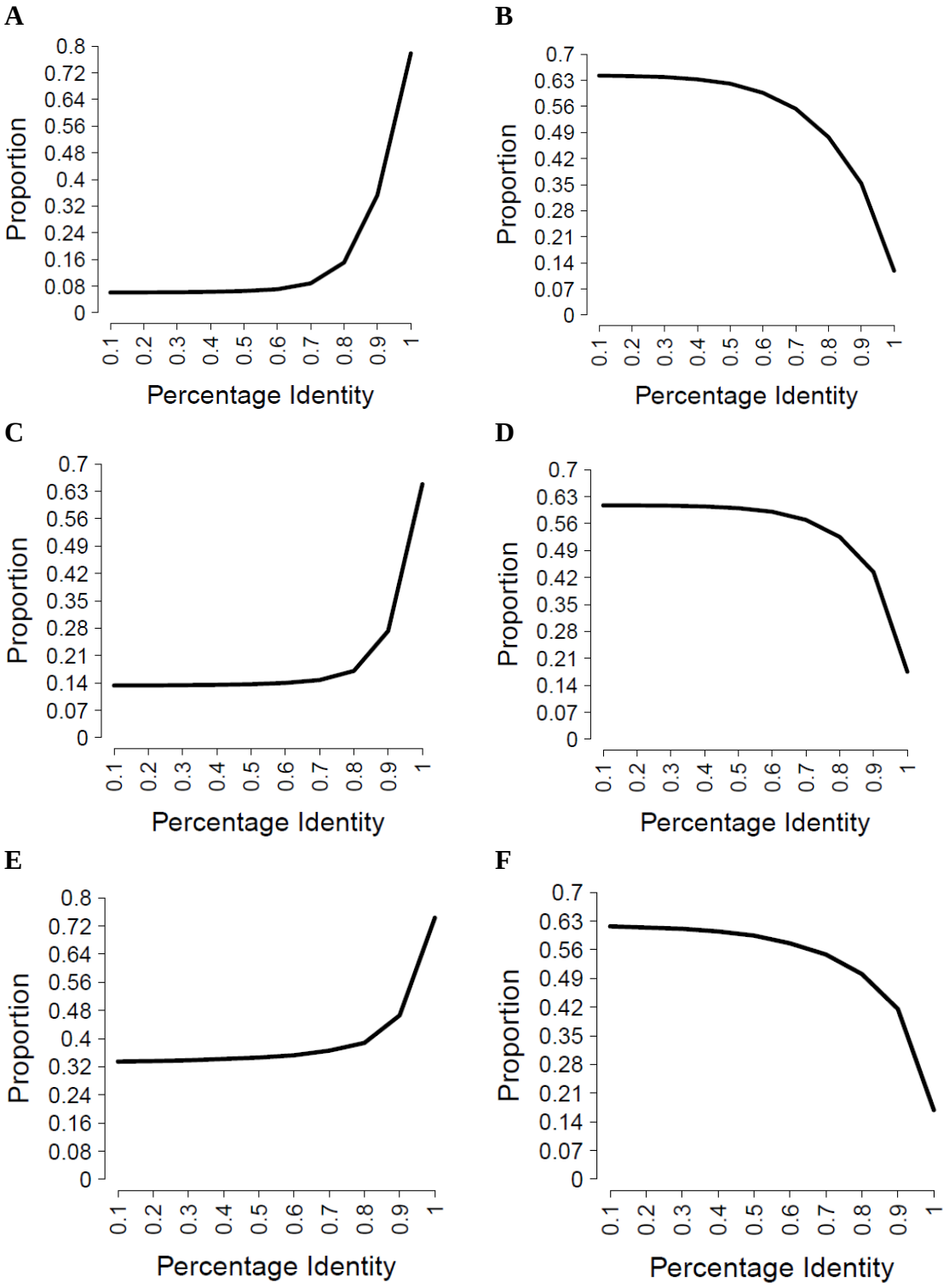

Figure C- 8 False negative and false positive rates in detecting bacterial homologs for GLAM2Scan

False negative and false positive rates in detecting Bacterial homologs for GLAM2Scan in our three simulated sets of proteins. The left column shows false negative rates while the right column shows false positive rates. The first row shows the results of our Base set, the second row shows the results of our Size Distribution set, and the third row

shows the results of our Size Distribution Fast set. (A) False negative rates for Base set. (B) False positive rates for Base set. (C) False negative rates for Size Distribution set. (D) False positive rates for Size Distribution set. (E) False negative rates for Size Distribution Fast set. (F) False positive rates for Size Distribution Fast set.

Table C- 1 Performance of machine learning algorithms for identification of error-prone genes with less strict criteria for error

	Base SVM	Size Dist. SVM	Size Dist. Fast SVM	Base RF	Size Dist. RF	Size Dist. Fast RF
Model*	Error ~ L+E+B	Error ~ L*E*B	Error ~ L*E*B	Error ~ B	Error ~ L+E+B	Error ~ L+E+B
Sensitivity	0.344	0.412	0.333	0.838	0.717	0.644
Specificity	0.994	0.985	0.964	0.978	0.951	0.802
Precision	0.624	0.675	0.706	0.301	0.354	0.247

*L=length, E=evolutionary rate, B=maximum length of conserved block

Appendix D

Supplementary Figures and Tables for Chapter 6

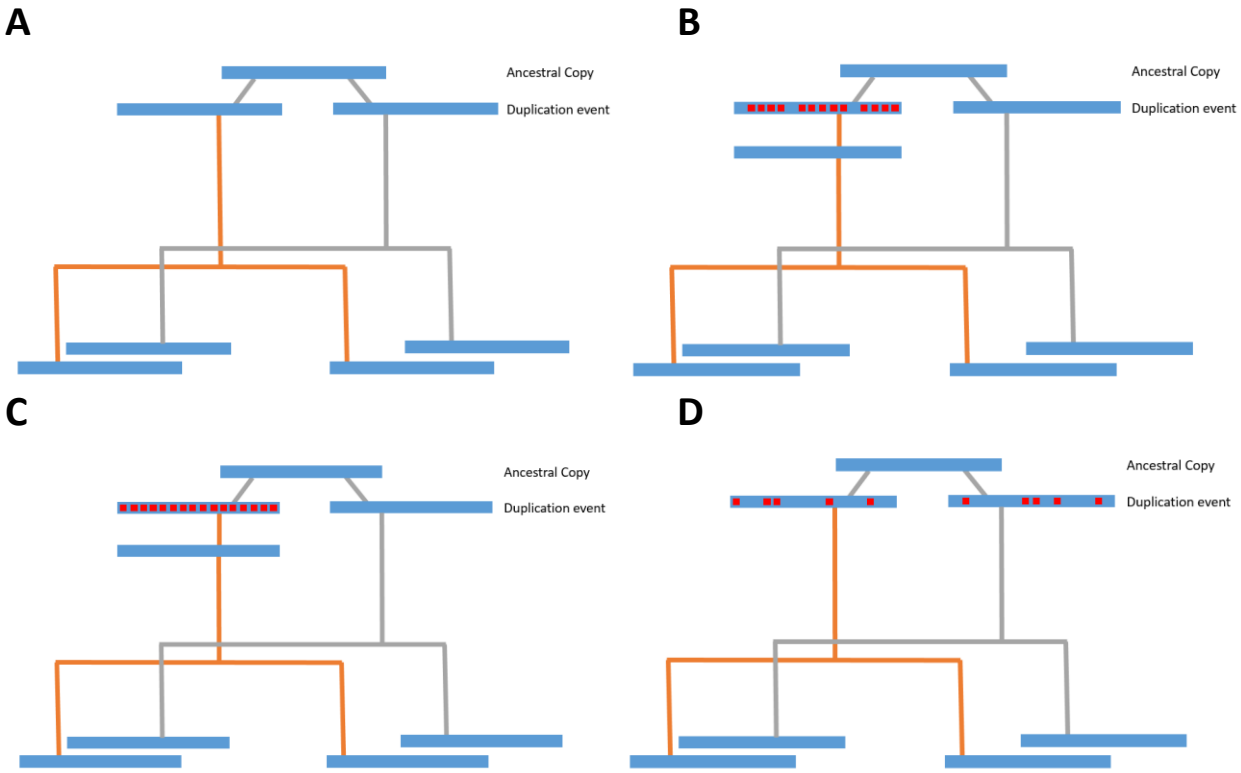

Figure D- 1 Models of duplication

Red dots indicate examples of sites being selected for changes in relative rate. (A) Baseline. (B) Neofunctionalization. (C) Neofunctionalization All Sites. (D) Subfunctionalization.

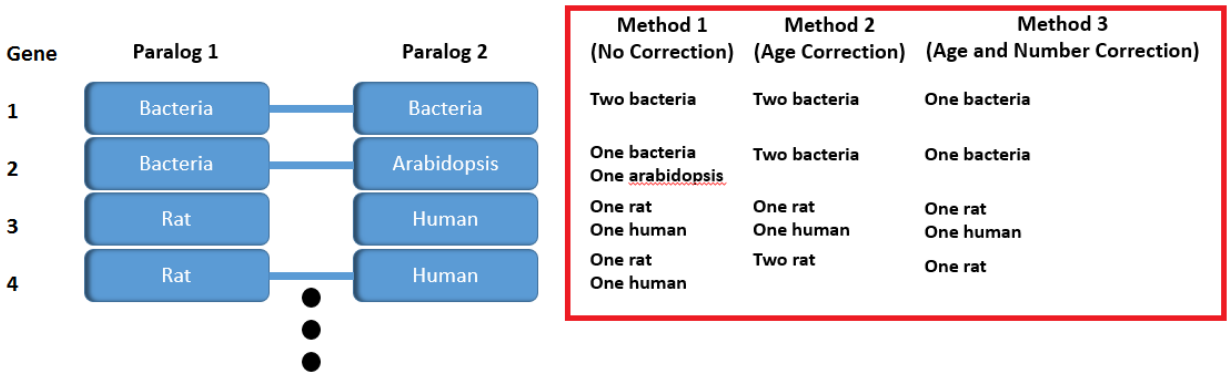

Figure D- 2 Different methods of correcting for paralogs

For each gene, consider two paralogs (Paralog 1, Paralog 2) with a common ancestral sequence. This paralogy is either detectable (blue connections between paralogs) or not detectable (no connection between paralogs) via standard homology detection through BLASTP. Under each of three potential paralog correction methods, different numbers of sequences will be counted among different nodes (red box). Under method 1, no correction, all genes will be assigned to a given age based solely on homologs detected in the target database in phylostratigraphy. Under method 2, age correction, all confirmed paralogs will be assigned the same age but will be considered as separate sequences. Under method 3, for detectable paralogs only one sequence will be considered, and its age will be the oldest of all detectable paralogs.

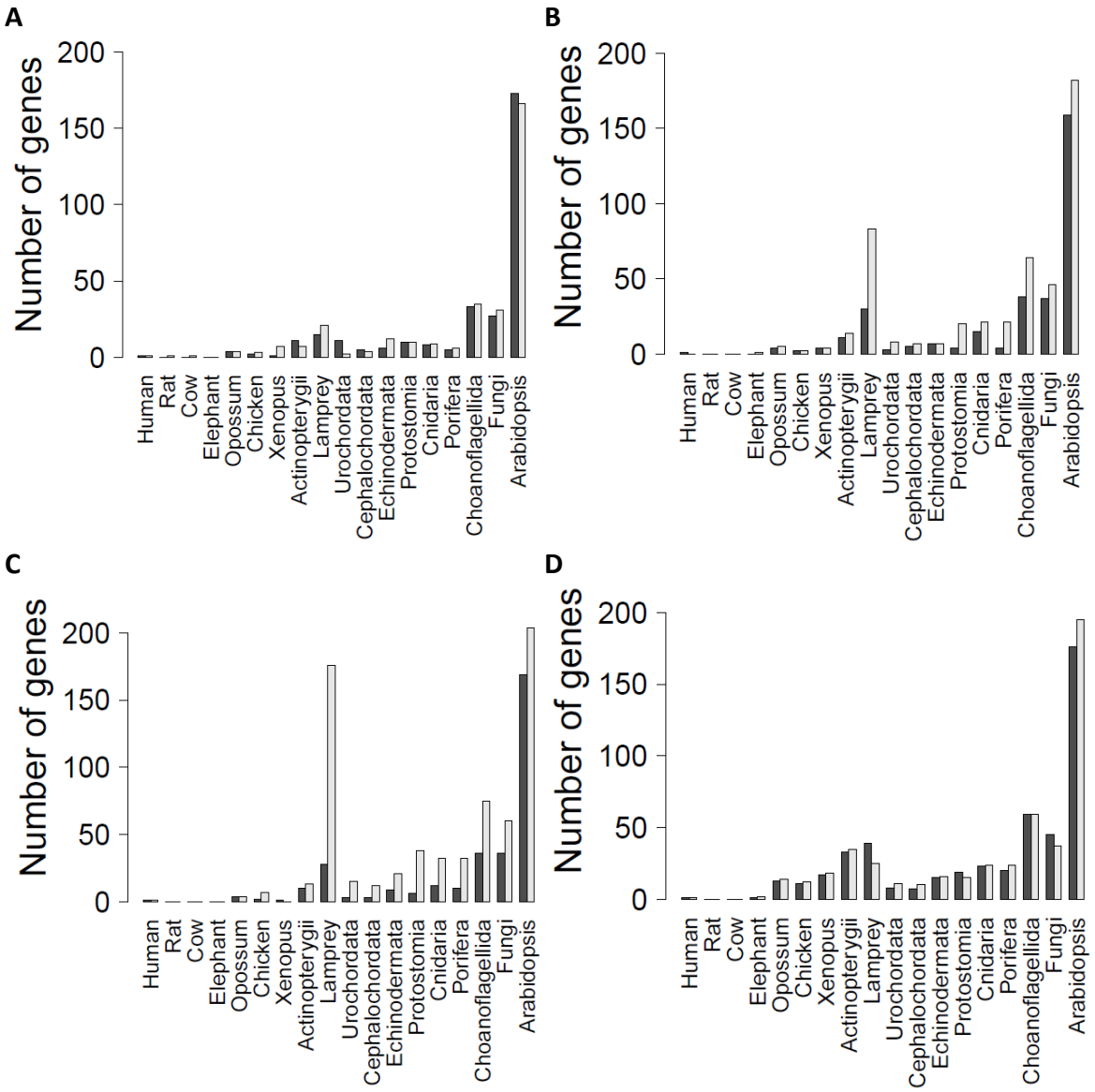

Figure D- 3 Number of novel sequences at each age when no correction for paralogs is made (method 1)

Dark grey bars denote the first of the two paralogs, while light grey bars denote the second of the two paralogs. Note that we do not display genes mapped to bacteria, for scaling purposes. In the two neofunctionalization simulations, the second of the two paralogs is the paralog which underwent a burst of evolution and subsequent shuffling of rates. We include here a count for the number of genes at each age, for each of the two paralogs. (A) Baseline; paralog 1: c(1, 0, 0, 0, 4, 2, 1, 11, 15, 11, 5, 6, 10, 8, 5, 33, 27, 173, 4630), paralog 2: c(1, 1, 1, 0, 4, 3, 7, 7, 21, 2, 4, 12, 10, 9, 6, 35, 31, 166, 4622). (B) Neofunctionalization; paralog 1: c(1, 0, 0, 0, 4, 2, 4, 11, 30, 3, 5, 7, 4, 15, 4, 38, 37, 159, 4618), paralog2: c(0, 0, 0, 1, 5, 2, 4, 14, 83, 8, 7, 7, 20, 21, 21, 64, 46, 182, 4457). (C) Neofunctionalization, all sites; paralog 1: c(1, 0, 0, 0, 4, 2, 1, 10, 28, 3, 3, 9, 6, 12, 10, 36, 36, 169, 4612), paralog 2: c(1, 0, 0, 0, 4, 7, 0, 13, 176, 15, 12, 21, 38, 32, 32, 75, 60, 204, 4252). (D) Subfunctionalization; paralog 1: c(1, 0, 0, 1, 13, 11, 17, 33, 39, 8, 7, 15, 19, 23, 20, 59, 45, 176, 4455), paralog 2: c(1, 0, 0, 2, 14, 12, 18, 35, 25, 11, 10, 16, 15, 24, 24, 59, 37, 195, 4444).

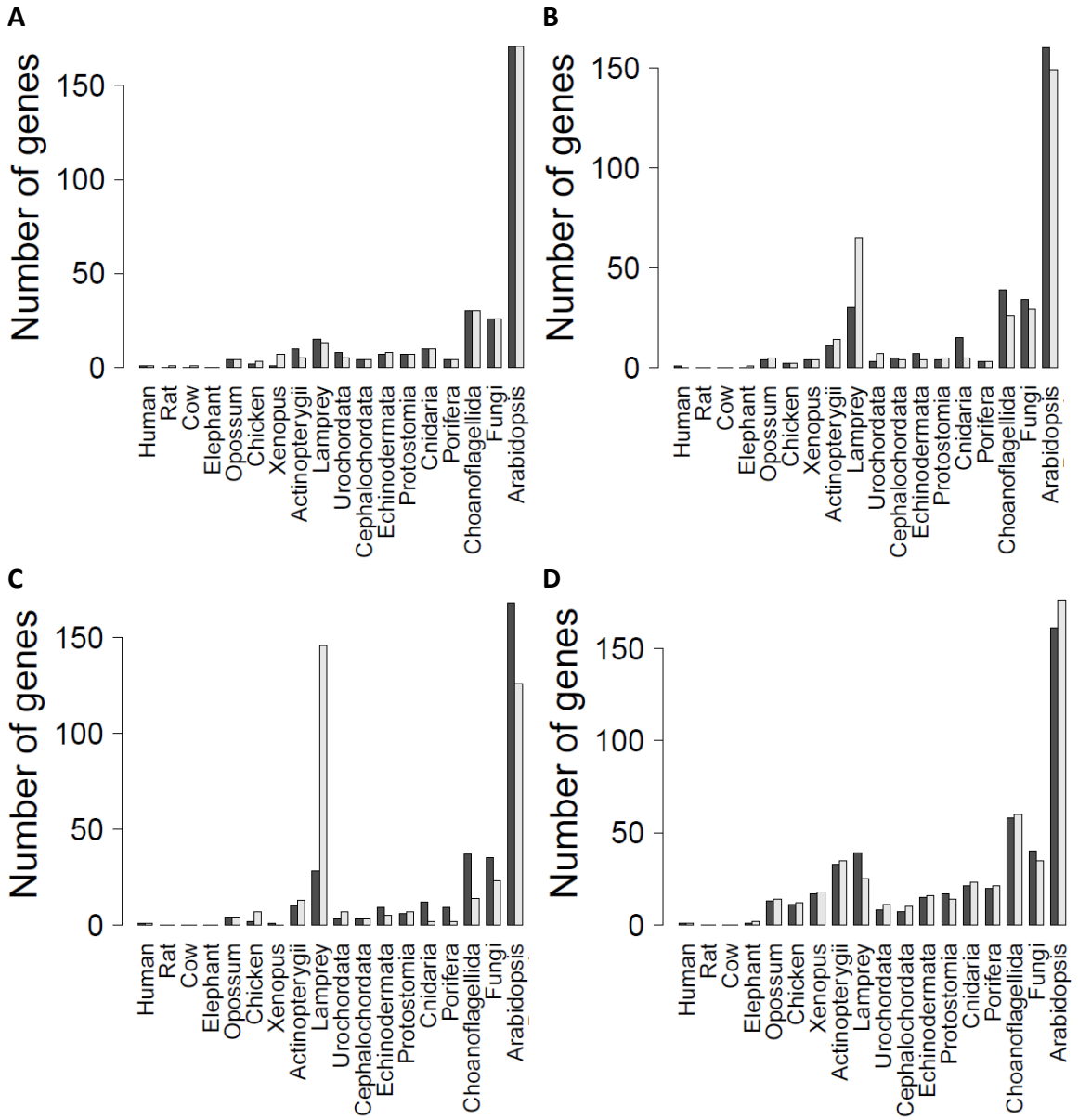

Figure D- 4 Number of novel sequences at each age when only the age of paralogs is corrected (method 2)

Dark grey bars denote the first of the two paralogs, while light grey bars denote the second of the two paralogs. Note that we do not display genes mapped to bacteria, for scaling purposes. In the two neofunctionalization simulations, the second of the two paralogs is the paralog which underwent a burst of evolution and subsequent shuffling of rates. We include here a count for the number of genes at each age, for each of the two paralogs. (A) Baseline; paralog 1: c(1, 0, 0, 0, 4, 2, 1, 11, 15, 11, 5, 6, 10, 8, 5, 33, 27, 173, 4630), paralog 2: c(1, 1, 1, 0, 4, 3, 7, 7, 21, 2, 4, 12, 10, 9, 6, 35, 31, 166, 4622). (B) Neofunctionalization; paralog 1: c(1, 0, 0, 0, 4, 2, 4, 11, 30, 3, 5, 7, 4, 15, 4, 38, 37, 159, 4618), paralog2: c(0, 0, 0, 1, 5, 2, 4, 14, 83, 8, 7, 7, 20, 21, 21, 64, 46, 182, 4457). (C) Neofunctionalization, all sites; paralog 1: c(1, 0, 0, 0, 4, 2, 1, 10, 28, 3, 3, 9, 6, 12, 10, 36, 36, 169, 4612), paralog 2: c(1, 0, 0, 0, 4, 7, 0, 13, 176, 15, 12, 21, 38, 32, 32, 75, 60, 204, 4252). (D) Subfunctionalization; paralog 1: c(1, 0, 0, 1, 13, 11, 17, 33, 39, 8, 7, 15, 19, 23, 20, 59, 45, 176, 4455), paralog 2: c(1, 0, 0, 2, 14, 12, 18, 35, 25, 11, 10, 16, 15, 24, 24, 59, 37, 195, 4444).

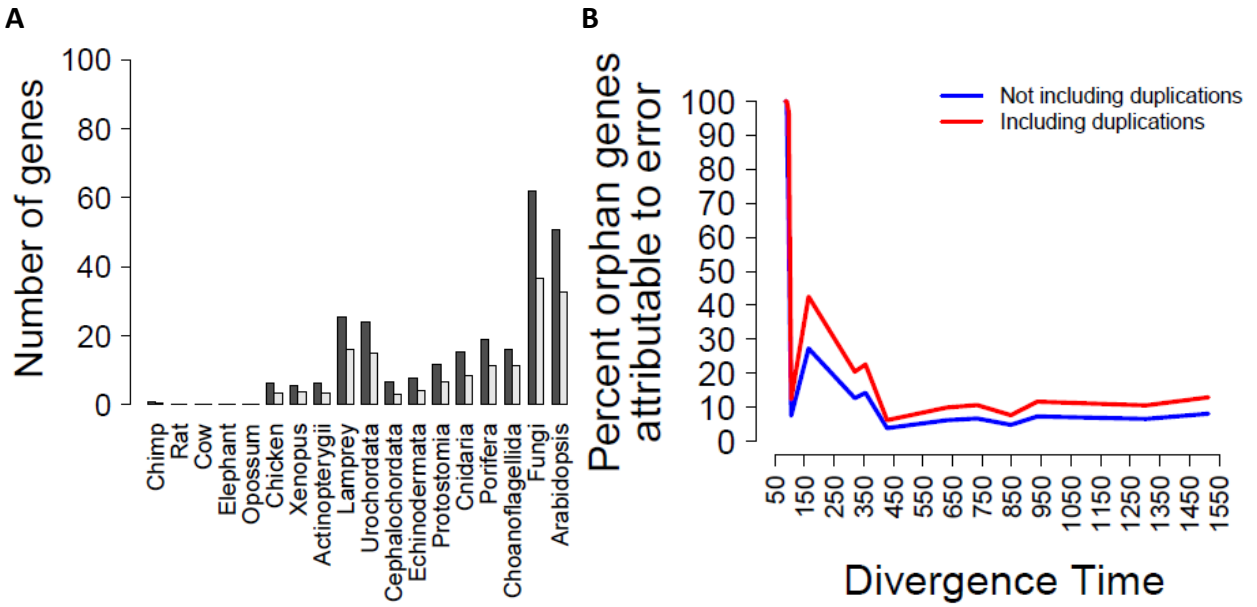

Figure D- 5 Phylostratigraphic results under a model of regular small duplications (method 1)

As figure 6-4, but using method 1 for paralog correction. (A) Number of genes in each age category in the simulation after paralog correction (method 3). The numbers of genes in each bin are as follows, rounded up to the nearest whole number. Paralog1: c(1, 0, 1, 1, 1, 7, 6, 7, 26, 24, 7, 8, 12, 16, 19, 17, 62, 51, 225, 4460). Paralog2: c(1, 0, 0, 1, 1, 4, 4, 4, 17, 15, 4, 4, 7, 9, 12, 12, 37, 33, 131, 2678) (B) The percent of orphan genes which are attributable to general error and to duplication under this model.

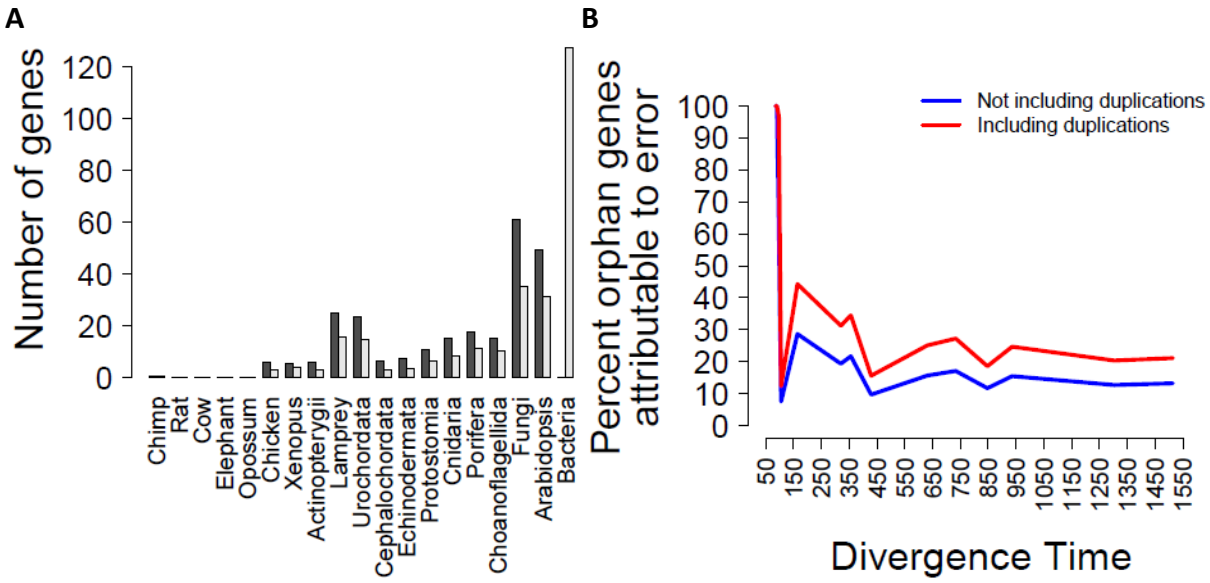

Figure D- 6 Phylostratigraphic results under a model of regular small duplications (method 2)

As figure 6-4, but using method 2 for paralog correction. (A) Number of genes in each age category in the simulation after paralog correction (method 3). The numbers of genes in each bin are as follows, rounded up to the nearest whole number. Paralog1: c(1, 0, 1, 1, 1, 7, 6, 7, 26, 24, 7, 8, 11, 16, 18, 16, 61, 50, 217, 4474). Paralog2: c(1, 0 ,0, 1, 1, 4, 4, 4, 16, 15, 3, 4, 7, 9, 12, 11, 36, 32, 128, 2686) (B) The percent of orphan genes which are attributable to general error and to duplication under this model.

Table D- 1 Number of genes in each age category in real and simulated data under regular small duplications (method 1)

Div Time (MYA)	Real orphans	Simulated Orphans (with dups)	Percent attributable to error (error+dups)	Dups/Error
85	0	1 (1)	100 (100)	0
90	0	1 (1)	100 (100)	0
97	1	1 (1)	100 (100)	0
105	9	1 (2)	11.11 (22.22)	1
164	17	5 (8)	29.41 (47.06)	0.6
320	65	9 (14)	13.85 (21.54)	0.56
356	85	13 (20)	15.29 (23.53)	0.54
429	732	29 (46)	2.96 (6.38)	0.59
631	763	48 (76)	6.29 (9.96)	0.58
733	888	60 (95)	6.76 (10.70)	0.58
847	1430	69 (110)	4.83 (7.69)	0.59
936	1781	130 (207)	7.30 (11.62)	0.59
1303	2460	162 (260)	6.59 (10.57)	0.60
1514	3730	302 (481)	8.10 (12.90)	0.59

Table D- 2 Number of genes in each age category in real and simulated data under regular small duplications (method 2)

Div Time (MYA)	Real orphans	Simulated Orphans (with dups)	Percent attributable to error (error+dups)	Dups/Error
85	0	1 (1)	100 (100)	0
90	0	1 (1)	100 (100)	0
97	1	1 (1)	100 (100)	0
105	9	1 (2)	11.11 (22.22)	1
164	14	5 (8)	31.25 (50)	0.6
320	42	9 (14)	21.43 (33.33)	0.56
356	55	12 (19)	21.82 (34.55)	0.58
429	286	28 (45)	9.79 (15.73)	0.61
631	296	47 (75)	15.88 (25.34)	0.60
733	339	58 (93)	17.11 (27.43)	0.60
847	577	68 (108)	11.79 (18.72)	0.59
936	818	127 (202)	15.53 (24.69)	0.59
1303	1240	158 (253)	12.74 (20.40)	0.60
1514	2215	293 (468)	13.23 (21.13)	0.60