Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution

Bryan A. Moyers1 and Jianzhi Zhang*,2

1Department of Computational Medicine and Bioinformatics, University of Michigan
2Department of Ecology and Evolutionary Biology, University of Michigan

*Corresponding author: E-mail: jianzhi@umich.edu.
Associate editor: Sudhir Kumar

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.

Key words: BLAST, gene age, new genes, phylostratigraphy, proto-gene, yeast.

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 (Hattori 2005; Ezkurdia et al. 2014). 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 (Nei 1969; Ohno 1970; Wolfe 2001; Long et al. 2003; Zhang 2003, 2013; KAESSMANN et al. 2009). 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. Although the first two mechanisms and their contributions to organismal adaptation have been abundantly documented (Koonin et al. 2001; Pál et al. 2005; Zhang 2013; QIAN and Zhang 2014), the arising of genes from nongenic material through de novo gene birth (Tautz and 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 (Levine et al. 2006; Begun et al. 2007; Cai et al. 2008; Heinen et al. 2009; Knowles and McLysaght 2009; Xiao et al. 2009; Li, Zhang, et al. 2010; Wu et al. 2011; Yang and 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 whereas de novo gene birth is thought to have a minimal contribution.

The above view was recently challenged by Carvunis et al. (2012), who claimed that de novo gene birth is common in evolution and is a larger source of new genes than gene duplication. Specifically, they proposed that nongenic 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 nongenic 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 and 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).

Although 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 random sequences, it is likely that a de novo gene is short and will increase in length 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 My 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 origin.

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 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 (fig. 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 (fig. 1B). We subsequently computed age distributions of various properties of S. cerevisiae proteins using the above estimated ages (figs. 2 and 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 5,261 published orthologous protein sequence alignments from five sensu stricto yeast species (S. cerevisiae, S. paradox, S. mikatae, S. kudriavzevi, 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 5,878 proteins present in the Carvunis et al. data set. The genetic distance of simulated orthologous proteins matches well that of real proteins (supplementary fig. S1, Supplementary Material online).

Because the true ages are 10 for all genes in the simulation (fig. 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) (fig. 1A), which was estimated to diverge from S. cerevisiae approximately 788 Ma (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 (fig. 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 Sc. pombe. Nevertheless, the number of genes with estimated age 1–9 is greater in the actual data than in the simulated data (fig. 1B). Although 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 Distributions 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 (Carvunis et al. 2012). These trends are significant increases in ORF length (fig. 2A), mRNA abundance (fig. 2B), proportion of genes in proximity of transcription factor (TF)-binding sites (fig. 2C), proportion of genes under significant purifying selection (fig. 2D), proportion of genes with optimal AUG context (fig. 2E), and codon adaptation index (CAI) (fig. 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, although qualitative appearances differed between the real and simulated data in these age distributions (fig. 2), statistical trends, quantified by Kendall’s $\tau$ as in (Carvunis et al. 2012), were almost identical between the two (table 1). Kendall’s $\tau$ was used following Carvunis et al. Using Spearman’s $\rho$ 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 approximately 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 nt 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 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 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
Fig. 2. Age distributions of six gene properties in real and simulated proteins. (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 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 CAI for each age bin.
**Fig. 3.** Age distributions of four additional gene properties in real and simulated proteins. (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.

**Table 1.** Correlations (Kendall’s τ) between Estimated Gene Age and Various Gene Properties for Real and Simulated Proteins.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>ORF Length</th>
<th>RNA Abundance</th>
<th>Proximity of TF-Binding Sites or Not</th>
<th>CAI</th>
<th>Purifying Selection or Not</th>
<th>Optimal AUG Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age groups 0–10a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real proteins</td>
<td>0.31**</td>
<td>0.27**</td>
<td>0.11**</td>
<td>0.12**</td>
<td>0.45**</td>
<td>0.14**</td>
</tr>
<tr>
<td>Simulated proteins</td>
<td>0.31**</td>
<td>0.27**</td>
<td>0.11**</td>
<td>0.12**</td>
<td>0.45**</td>
<td>0.14**</td>
</tr>
<tr>
<td>Simulated proteins (assuming oldest paralog agesb)</td>
<td>0.31**</td>
<td>0.27**</td>
<td>0.11**</td>
<td>0.12**</td>
<td>0.45**</td>
<td>0.14**</td>
</tr>
<tr>
<td>Age groups 1–10c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real proteins</td>
<td>0.39**</td>
<td>0.26**</td>
<td>0.08*</td>
<td>0.31**</td>
<td>0.32**</td>
<td>0.13**</td>
</tr>
<tr>
<td>Simulated proteins</td>
<td>0.33**</td>
<td>0.26**</td>
<td>0.06*</td>
<td>0.21**</td>
<td>0.27**</td>
<td>0.12**</td>
</tr>
<tr>
<td>Simulated proteins (assuming oldest paralog agesb)</td>
<td>0.31**</td>
<td>0.21**</td>
<td>0.04*</td>
<td>0.22**</td>
<td>0.26**</td>
<td>0.13**</td>
</tr>
</tbody>
</table>

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

aAnalysis includes all smORFs.
bThe age of a gene is assumed to equal that of the oldest gene in the same gene family. See main text for details.
cAnalysis excludes all smORFs.
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 (fig. 3A–C). A notable exception is the patterns found in amino acid usage, where simulated data match real data quite closely (fig. 3D).

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). 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). 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. cerevisiae*-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 1–2078, 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 et al. 2010), so it is not surprising that an orthologous region could not be found. Additionally, when
BLASTed against the \textit{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 (\textit{Ingolia et al. 2009}). Several papers report changes in the transcript concentration of YOL166C is expressed at the protein level. Based on these analyses, 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 \textit{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 sequences collected from different species and is inappropriate for testing selection in sequences from the same species, because, for intraspecific 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 (\textit{Wei and Zhang 2015}). So, in the cases of these genes, only their nonoverlapped portions should be used to infer selection. To increase the accuracy and power of selection detection, we used 38 \textit{S. cerevisiae} strains in the SGRP (\textit{Cherry et al. 2012}) and counted the number of synonymous and nonsynonymous polymorphisms in the region of a gene that is nonoverlapping with other genes (table 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 (\textit{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 in the set. For the nonoverlapped region of an average gene in table 3, the probability of detecting significant purifying selection is about 2% even when all nonsynonymous mutations are strongly deleterious. In other words, there is virtually no power to detect purifying selection acting on such short sequences.

**Discussion**

The origin of new protein-coding genes from noncoding sequences is a fascinating hypothesis that has been supported by the discoveries of dozens of cases of de novo gene birth in human, \textit{Drosophila}, yeast, and other species (\textit{Levine et al. 2006}; \textit{Clark et al. 2007}; \textit{Cai et al. 2008}; \textit{Heinen et al. 2009}; \textit{Knowles and McLysaght 2009}; \textit{Xiao et al. 2009}; \textit{Li, Zhang, et al. 2010}; \textit{Wu et al. 2011}; \textit{Yang and 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 noncoding (\textit{Knowles and 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.

**Table 3. Reexamining Purported \textit{Saccharomyces cerevisiae}-Specific Selected Genes.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Age Based on Full Sequence</th>
<th>Nonoverlapped Length in Nucleotides (full length)</th>
<th>No. of Synonymous Polymorphisms in Nonoverlapped Region</th>
<th>No. of Nonsynonymous Polymorphisms in Nonoverlapped Region</th>
<th>P value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>YBR232C</td>
<td>6</td>
<td>55 (360)</td>
<td>1</td>
<td>0</td>
<td>0.29</td>
</tr>
<tr>
<td>YCL046W</td>
<td>2</td>
<td>58 (324)</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>YDR537C</td>
<td>7</td>
<td>47 (606)</td>
<td>0</td>
<td>2</td>
<td>0.57</td>
</tr>
<tr>
<td>YER087C-A</td>
<td>7</td>
<td>62 (552)</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>YFL013W-A</td>
<td>5</td>
<td>53 (804)</td>
<td>1</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>YGL152C</td>
<td>6</td>
<td>71 (678)</td>
<td>2</td>
<td>2</td>
<td>0.58</td>
</tr>
<tr>
<td>YHL030W-A</td>
<td>9</td>
<td>49 (462)</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>YIL071W-A</td>
<td>6</td>
<td>111 (477)</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>YLR232W</td>
<td>9</td>
<td>58 (348)</td>
<td>1</td>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>YLR358C</td>
<td>6</td>
<td>50 (564)</td>
<td>0</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>YNL105W</td>
<td>8</td>
<td>88 (429)</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>YNL109W</td>
<td>8</td>
<td>50 (546)</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>YOL150C</td>
<td>8</td>
<td>62 (312)</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>YOL166C</td>
<td>1</td>
<td>339 (339)</td>
<td>3</td>
<td>3</td>
<td>0.37</td>
</tr>
<tr>
<td>YOR055W</td>
<td>6</td>
<td>55 (435)</td>
<td>0</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>YOR135C</td>
<td>10</td>
<td>91 (342)</td>
<td>1</td>
<td>0</td>
<td>0.30</td>
</tr>
</tbody>
</table>

*Based on two-tailed Fisher’s exact test of the neutral hypothesis.
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 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) is 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 and 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 2B, one can see a gray peak at age 7, indicating that genes of age 7 have unusually high expression. 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 2, all four properties in figure 3, genetic coregulation, % alpha helices, and % beta sheets in table 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 hydropathy 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 CAI), 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 (fig. 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 (fig. 2D). Because protein evolutionary rate is strongly negatively correlated with its mRNA expression level (Zhang and Yang 2015), mRNA expression level must also impact phylostratigraphic error, as seen in our simulated data (table 1). Hence, a positive correlation between gene age and expression level (fig. 2B) reflects an expected bias of phylostratigraphy. Phylostratigraphic error is also expected to create a positive correlation between gene age and CAI (fig. 2F), because CAI is positively correlated with gene expression level (Sharp and 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) (P < 1E-300), phylostratigraphic error also causes a positive correlation between gene age and proportion in proximity of TF-binding sites (fig. 2C). It was reported (Miyasaka et al. 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 (P < 1E-300), potentially explaining why a positive correlation between gene age and proportion in optimal AUG context is created by phylostratigraphic error (fig. 2E). Amino acid usage is known to be correlated with gene expression level (Akashi and Gojobori 2002), potentially explaining the observed trends in figure 2E. 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 4 and supplementary table S1, Supplementary Material online).

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 with a suggested five genes formed by duplication in the same time period (Gao and Innan 2004), though this duplicate gene number has since been challenged.
sequences in eukaryotes (Johnson et al. 2005) and widespread selection. Given the widespread transcription of intergenic a gene by demonstrating that it is under purifying or positive model discussed here, what is meant is a functional, protein- and translated. For example, over 100 human pseudogenes random noncoding sequence to be spuriously transcribed (Ingolia et al. 2014), it is probably not rare for a translation of noncoding RNAs (at least based on ribosome profiling data) (Ingolia et al. 2014). 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 (Cai et al. 2008; Li, Dong, 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 et al. 2005). Although these estimates may not be precise, gene duplication appears to surpass de novo gene birth by 2 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 transcription of intergenic sequences in eukaryotes (Johnson et al. 2005) and widespread translation of noncoding RNAs (at least based on ribosome profiling data) (Ingolia et al. 2014), it is probably not rare for a random noncoding 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 2016). 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 ancestor of a few closely related species such as that of S. cerevisiae and S. paradoxus rather than in S. cerevisiae. Although Carvunis et al. and this study focused on protein-coding genes, noncoding RNAs may also play important biological functions. It is possible that the larger part of genetic novelty in evolution is in the aspect of noncoding 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.

### Materials and Methods

#### Yeast Genes

For simulation of sequence evolution, we acquired 5,261 orthologous sequence alignments in protein format from

![Table 4. Correlations (Kendall’s τ) between Various Gene Properties and Three Properties Known to Bias Phylostratigraphy, Using Genes in Age Groups 1–10.](https://academic.oup.com/mbe/article-abstract/33/5/1245/2579728/doi)
the sensu stricto group of yeast species from http://www.
saccharomycessensustricto.org/current//aligns/coding_all
files.fasta.tgz last accessed January 25, 2016 (Scannell et al.
2011). Except for two alignments, all contain five ortholo-
gous sequences from five sensu stricto yeast species. The
simulation of the 5,259 genes that have alignments of five
sequences used parameters estimated from the alignments.
The simulation of other genes in S. cerevisiae used param-
eters estimated from a set of sensu stricto restricted genes.

To identify sensu stricto restricted genes, we acquired pro-
tein databases of four yeast species outside of the sensu
stricto group. These species were S. castellii and S. kluveri,
downloaded from the Saccharomyces Genome Database
(SGD) at http://www.yeastgenome.org/download-data/se-
quence last accessed January 25, 2016 (Cherry et al. 2012),
as well as Kluyveromyces thermoderans and
Zygosaccharomyces rouxii, acquired from the Genolevures
Consortium (Souciet et al. 2009). Using the alignments ac-
cquired from Scannell et al. (2011), we created five databases,
one for each of the sensu stricto species. We then per-
formed 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. kluveri, K.
thermoderans, 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. Although homology detection
error may explain the apparent restriction of these genes to
the sensu stricto group, this is not a problem for our simul-
ation, 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 et al. 2007). For each of the 5,259 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 esti-
mated 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 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 func-
tion. 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 (Jones, Taylor, and Thornton)-f model in
the ROSE simulation of protein sequence evolution, where “f”
refers to the amino acid compositions of the protein con-
cerned (Nei and Kumar 2000). Each site along the protein has
a particular relative rate. The relative rate for a site is multi-
plied 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 evo-
uation 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 informa-
tion, we determined the number of indels per amino acid,
averaged over all proteins. This indel threshold was then ap-
plied 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 (ac-
counting 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
determine the rate of 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 selected sensu stricto limited genes to simulate these
proteins. To do this, we took each protein in this group and
and multiplied the relative rates of all sites by the average evolution-
ary rate for the protein. This gave us an absolute evolution-
ary rate for each site. We then concatenated these
numerical vectors into a single vector from which we could
sample rates for each protein (supplementary fig. S2,
Supplementary Material online). 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 esti-
mate 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 cor-
responding 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 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 nonredundant 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 ORF. 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 last accessed January 25, 2016. We extracted the SNPs of 38 strains present in both the SGRP data and the phylogeny in Liti et al. (2009). We extracted only those SNPs for which quality score was 55 or greater, following Carvunis et al.'s 16 genes of age group 1 that were purported to be under purifying selection. 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 last accessed January 25, 2016. We extracted the SNPs of 38 strains present in both the SGRP data and the phylogeny in 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 (Zhang et al. 1997). In no case was the result significantly different from the neutral expectation.

The 38 strains used are as follows: DBVPG6040, NCYC361, S288c, W303, 578064X, YJM789, Y52, Y54, Y59, 273614N, Ylkc17_E5, RM11_1A, YJM975, YJM978, YJM981, DBVPG1853, 322134S, BC187, DBVPG6765, DBVPG1878, 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 Data Sets
We were provided with various gene properties from Carvunis et al. through email communication. We downloaded data sets used by Abrusán (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; Abrusán 2013).

Supplementary Material
Supplementary figures S1 and S2 and table S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
The authors thank Anne-Ruxandra Carvunis and Marc Vidal for supplying their data reported in Carvunis et al. (2012) and Anne-Ruxandra Carvunis for sharing unpublished results, members of the Zhang lab for discussion, and Jian-Rong Yang and Zhengtong Zou for valuable comments on the manuscript. This work was supported in part by the US National Institute of Health (NIH) grant R01GM103232 to J.Z. and by the NIH training grant in genome sciences (T32HG00040) to B.A.M.

References
