No Evidence for Phylostratigraphic Bias Impacting Inferences on Patterns of Gene Emergence and Evolution

Tomislav Domazet-Lošo,[†,1,2] Anne-Ruxandra Carvunis,[*,†,3] M. Mar Albà,[4,5] Martin Sebastijan Šestak,[1] Robert Bakarić,[1] Rafik Neme,[6] and Diethard Tautz[*,6]

[1]Laboratory of Evolutionary Genetics, Division of Molecular Biology, Ruđer Bošković Institute, Zagreb, Croatia

[2]Catholic University of Croatia, Zagreb, Croatia

[3]Department of Medicine, University of California, San Diego, CA

[4]Evolutionary Genomics Group, Research Programme on Biomedical Informatics, Hospital del Mar Research Institute, Universitat Pompeu Fabra, Barcelona, Spain

[5]Catalan Institution for Research and Advanced Studies, Barcelona, Spain

[6]Max-Planck Institute for Evolutionary Biology, Plön, Germany

[†]These authors contributed equally to this work.

***Corresponding authors:** E-mails: tautz@evolbio.mpg.de; carvunis@gmail.com.

Associate editor: Yuseob Kim

Abstract

Phylostratigraphy is a computational framework for dating the emergence of DNA and protein sequences in a phylogeny. It has been extensively applied to make inferences on patterns of genome evolution, including patterns of disease gene evolution, ontogeny and de novo gene origination. Phylostratigraphy typically relies on BLAST searches along a species tree, but new simulation studies have raised concerns about the ability of BLAST to detect remote homologues and its impact on phylostratigraphic inferences. Here, we re-assessed these simulations. We found that, even with a possible overall BLAST false negative rate between 11–15%, the large majority of sequences assigned to a recent evolutionary origin by phylostratigraphy is unaffected by technical concerns about BLAST. Where the results of the simulations did cast doubt on previously reported findings, we repeated the original analyses but now excluded all questionable sequences. The originally described patterns remained essentially unchanged. These new analyses strongly support phylostratigraphic inferences, including: genes that emerged after the origin of eukaryotes are more likely to be expressed in the ectoderm than in the endoderm or mesoderm in *Drosophila*, and the de novo emergence of protein-coding genes from non-genic sequences occurs through proto-gene intermediates in yeast. We conclude that BLAST is an appropriate and sufficiently sensitive tool in phylostratigraphic analysis that does not appear to introduce significant biases into evolutionary pattern inferences.

Key words: **genome analysis, phylostratigraphy, BLAST, gene age estimation.**

Article

Introduction

Correlating the emergence of particular DNA or protein sequences with molecular and phenotypic features is one way to harness the information that we obtain from genome sequencing projects. Phylostratigraphy is a framework in which this can be done in a phylogeny aware context (Domazet-Lošo et al. 2007). Starting from the genome of a focal species, phylostratigraphy infers the emergence of novel sequences at a particular phylogenetic node, usually by using the similarity search algorithm BLAST (Altschul et al. 1990) on a set of genomes that represent the nodes. Each sequence in the focal genome is thereby assigned an "evolutionary age" corresponding to the most distant node in the phylogeny where BLAST could detect a homologue for this sequence. This age classification, also referred to as "phylostrata" or "conservation level" classification, enables to distinguish younger sequences, for which homologues can only be found in closely related species [often called orphans or taxonomically restricted (Khalturin et al. 2009)], from older sequences that are conserved in very distant species (Tautz and Domazet-Lošo 2011). While phylostratigraphy is a general evolutionary framework that in theory applies to any type of sequence, it has mostly been exploited to study the evolution of genes, transcripts or open reading frames (ORFs).

It is important to note that genes with apparent young sequences may have evolved through two different mechanisms. One is de novo evolution, which has only relatively recently been recognized as an important mechanism for evolution of novelty (Levine et al. 2006; Zhou et al. 2008; Heinen et al. 2009; Knowles and McLysaght 2009; Toll-Riera et al. 2009; Carvunis

et al. 2012; Neme and Tautz 2014). The other is divergence from an ancestral gene followed by a phase of large sequence divergence (Domazet-Lošo and Tautz 2003). The concept of phylostratigraphy was originally based on this latter mechanism and proposed the idea of a punctuated evolution of protein-coding genes and their descendant families (Domazet-Lošo et al. 2007; Domazet-Lošo and Tautz 2010a). Punctuated evolution assumes that a gene originates by duplication from an existing gene followed by divergence with a subsequent slow-down in sequence evolution. Such slow evolving orphan genes were first detected in *Drosophila* (Domazet-Lošo and Tautz 2003). The shifts in sequence space generated by phases of large evolutionary divergence after gene duplication may indicate new adaptive functions and phylostratigraphy aims to trace such events and to statistically correlate them to biological patterns (Domazet-Lošo and Tautz 2010b; Quint et al. 2012; Mendoza et al. 2013; Šestak et al. 2013; Šestak and Domazet-Lošo 2015; Drost et al. 2016). In this case, phylostratigraphy aims to capture the time when the sequence divergence took place, not necessarily the time of origin of the ancestral gene.

De novo emergence from a previously non-genic sequence can be equally detected by phylostratigraphy. For a long time, de novo emergence was considered to be very unlikely (Tautz 2014) and had therefore initially not been seriously considered as a model of origin of orphan genes (Domazet-Lošo and Tautz 2003). However, it is now clear that de novo gene birth is in fact another important process that can be traced by phylostratigraphy (Tautz and Domazet-Lošo 2011). Accordingly, phylostratigraphy has also been used in later studies specifically focusing on the patterns and mechanisms of de novo evolution (Carvunis et al. 2012; Abrusán 2013; Neme and Tautz 2013). However, from the results of phylostratigraphy alone, one cannot know with certainty if a young sequence corresponds to a case of de novo birth or a case of divergence from an ancestral gene. An unequivocal demonstration of de novo evolution requires also invoking synteny information and reconstruction of the mutational events that have led to the novel sequence (Tautz et al. 2013; McLysaght and Hurst 2016). Since this is not always possible at the genome scale, it is often assumed as a proxy that sequences for which BLAST cannot find homologues among even closely related species are most likely to be enriched in de novo cases (Tautz and Domazet-Lošo 2011; Carvunis et al. 2012).

Although BLAST is very powerful in detecting homologues in large databases, it has known limitations when sequences are highly diverged. In particular, it was observed that BLAST has problems to detect remote homologues of short and fast-evolving sequences (Elhaik et al. 2006; Moyers and Zhang 2015). These limitations do not much affect evolutionary inferences related to punctuated evolution of proteins and their descendant families, where the existence of possible remote homologues is not the primary question (Domazet-Lošo et al. 2007; Domazet-Lošo and Tautz 2010a). If anything, BLAST could be too sensitive in this context, and find an older origin for a protein, although it has gone through a recent shift in sequence space. For example, transcription factors that have arisen to regulate a specific function in a young lineage may become placed into a much older node because of a match within their DNA binding domain (Capra et al. 2013). BLAST could also overestimate a sequence's evolutionary age by yielding spurious hits that do not reflect true homology, especially if used with permissive statistical cutoffs.

On the other hand, the difficulty of BLAST searches to find remote homologues could be problematic in the context of making cases for de novo emergence, versus divergence from an ancient gene (Schlötterer 2015). Ancient genes that have diverged too much for BLAST to detect them in the genomes of distant species may then be erroneously categorized as too young by phylostratigraphy. These BLAST limitations have motivated the development of further refined search methods, such as PSI-BLAST (Altschul et al. 1997), HHMER3 (Finn et al. 2011) or HHblits (Remmert et al. 2012). Although these refined methods can detect more remote homologues, they are partially computationally more costly, require similarity profiles from well-populated gene families and are therefore less generally applicable. Another approach is to use orthology detection algorithms to estimate gene age (Liebeskind et al. 2016), but the properties of this as well as the above approaches have still to be further explored. Hence, BLAST remains currently the workhorse for obtaining initial phylostratigraphic information and it is therefore important to understand its advantages, as well as its limitations and possible error margins.

In an attempt to estimate the false negative error rate of BLAST and its impact on evolutionary inferences, Elhaik et al. (2006) simulated DNA sequence evolution and used BLAST to look for homologues of these simulated sequences. They found in these simulations that fast-evolving DNA sequences tended to appear younger than they were, and suggested that the "Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes" previously reported (Albà and Castresana 2005) may have been an artifact. This suggestion was rapidly mitigated when Albà and Castresana (2007) pointed out a problem in the simulation framework used by Elhaik et al. (2006). BLAST uses a two-step search algorithm that starts by finding matches on short motifs and extending the alignment based on these (Altschul et al. 1990). Proteins that evolve homogeneously along their whole sequence are thus more difficult to trace than proteins that include at least one or more slowly evolving domains. Real proteins fall mostly into this latter class, allowing BLAST to find homologues even when the rest of a protein sequence evolves very fast. Therefore, Albà and Castresana (2007) argued that simulating protein evolution to assess the power of BLAST needs to take natural among-site rate heterogeneity into account.

Using this controlled approach, Albà and Castresana (2007) have shown that less than 5% of simulated homologues of mammalian genes are misclassified as recently evolved (i.e. too young) when rate heterogeneity is taken into account. Applying an orthogonal approach, Carvunis et al. (2012) found that only 5% of ORFs appearing young in a phylostratigraphy of Ascomycota fungi actually had ancient homologues revealed when searching the entire non-redundant protein sequence database of NCBI. The false negatives in BLAST searches were therefore considered to occur at an acceptable rate, similar to most genome-scale analyses.

Still, the question re-emerged recently when Moyers and Zhang (2015; 2016a) sought to quantify the power of BLAST to detect remote homologues, and to assess the possible implications for trends and patterns inferred from phylostratigraphic analysis. In the first study (Moyers and Zhang 2015), they make the point that Albà and Castresana (2007) used rate heterogeneity models that were derived from only 14 genes conserved across vertebrates. Hence, they used a much larger set of genes derived from *Drosophila melanogaster* and calculated among-site rate heterogeneity and average divergence rate for each gene based on an alignment among 12 *Drosophila* species, which represent a similar relative conservation level as vertebrates. These actual genes and their associated divergence rates were then used to simulate their possible ancestors at the origin of life and ask which percentage of such ancestors can be traced by BLAST. They find that BLAST makes an incorrect assignment for 14% of the sequences simulated. While this still implies that the large majority of sequences are not affected by problems with BLAST, the authors propose that this level of error could lead to systematic biases in gene evolution patterns.

In their second paper, Moyers and Zhang (2016a) addressed the question of de novo evolution of genes in yeast species. Starting from protein sequence alignments between yeast species closely related to the focal species *Saccharomyces cerevisiae*, Moyers and Zhang (2016a) measured among-site rate heterogeneity and average divergence rates, and simulated possible ancestors throughout the Ascomycota phylogeny based on the measured rates. They report that BLAST missed 11% of the simulated ancient homologues. They show that the corresponding ORFs, which may erroneously appear young in phylostratigraphy, despite potentially being ancient, share many physical and functional properties with the ORFs deemed young by Carvunis et al. (2012) and Abrusán et al. (2013). Based on these observations, Moyers and Zhang (2016a) question the validity of genome-wide phylostratigraphic analyses for deriving models of de novo gene birth.

In summary, Moyers and Zhang (2015; 2016a) have revived several important technical and conceptual issues pertaining to an older debate on the limitations of BLAST (Albà and Castresana 2005; Elhaik et al. 2006; Albà and Castresana 2007). Here, we show that the previously published inferences on gene emergence and evolution that were questioned by Moyers and Zhang (2015; 2016a) are in fact robust to BLAST limitations, even if error rates were as estimated by Moyers and Zhang (2015; 2016a). We argue that Moyers and Zhang's (2015; 2016a) simulations have underestimated the power of BLAST in phylostratigraphy. We conclude that the alleged evidence for a systematic phylostratigraphic bias cannot be reproduced.

Results

Most Young Phylostratigraphic Age Assignments Cannot Be Attributed to BLAST Errors

The simulations performed by Moyers and Zhang suggested that up to 14% of *Drosophila melanogaster* sequences (2015) and up to 11% of *Saccharomyces cerevisiae* sequences (2016) may erroneously appear to have originated recently due to the limitations of BLAST. While these are higher fractions than the 5% found by Albà and Castresana (2007), there is no reason to claim an "extreme" problem of age underestimation and the large majority of assignments in phylostratigraphic studies is still not in doubt. Here, we updated the *Drosophila* phylostratigraphy for over 13K *Drosophila melanogaster* real sequences (see supplementary table S1, Supplementary Material online). We then compared the simulated and the real data obtained for 6,629 of these sequences, where simulated age assignments were available (see supplementary table S1, Supplementary Material online). The resulting distributions (fig. 1*A*) show that the large number of sequences lacking remote homologues in real data cannot be recapitulated by simulations (fig. 1*B*), as Moyers and Zhang (2015) found as well (compare fig. 5 in the respective paper). Similarly in yeast, comparing the age distributions of 5,878 *S. cerevisiae* sequences inferred from real (Carvunis et al. 2012) and simulated (Moyers and Zhang 2016a) data reveals striking differences (fig. 1*C*). In stark contrast with the 11% estmate of misplaced sequences by Moyers and Zhang (2016a), ~40% of sequences lack a detectable homologue in the distant species *S. pombe* (fig. 1*D*). Focusing on the three youngest age classes for example, Moyers and Zhang (2016a) find 14 misplaced sequences in their simulations while Carvunis et al. (2012) found 445, i.e. over 30 times more (fig. 1*D*). The number of sequences found young in real phylostratigraphy analyses thus dwarfs the number of error-prone sequences expected to appear young because of BLAST false negatives as estimated by Moyers and Zhang (2015 and 2016a).

Since simulations are by nature stochastic, the list of sequences found error-prone in a given simulation run is expected to vary somewhat each time a new simulation run is performed. Therefore, the low number of sequences found error-prone could potentially increase towards values equal or superior to the values observed in real data if the union of multiple simulation runs was considered. We thus investigated whether increasing the number of simulation runs could eventually yield as many error-prone sequences as young sequences found in real phylostratigraphy. To this aim, we performed a saturation analysis on a series of 10 independent runs simulated by Moyers and Zhang (2015) on *Drosophila* sequences. Starting from 3,840 sequences that were classified as restricted to Eukaryota in the real phylostratigraphy (fig. 1*B*), we asked how many of these sequences are found susceptible to BLAST limitations in the union of up to ten successive independent simulation runs. We found that, while on average a single simulation run identifies 866 error-prone sequences, this number increases only to 1,006 when the union of ten simulation runs is considered (figs. 1B and 2). The number of times a simulation is re-ran thus barely affects estimates of BLAST limitations. Therefore, while phylostratigraphic methods should be improved to reduce an already low false negative rate of 5–15%, technical BLAST artifacts cannot explain the large numbers of sequences lacking recognizable homologues across species.

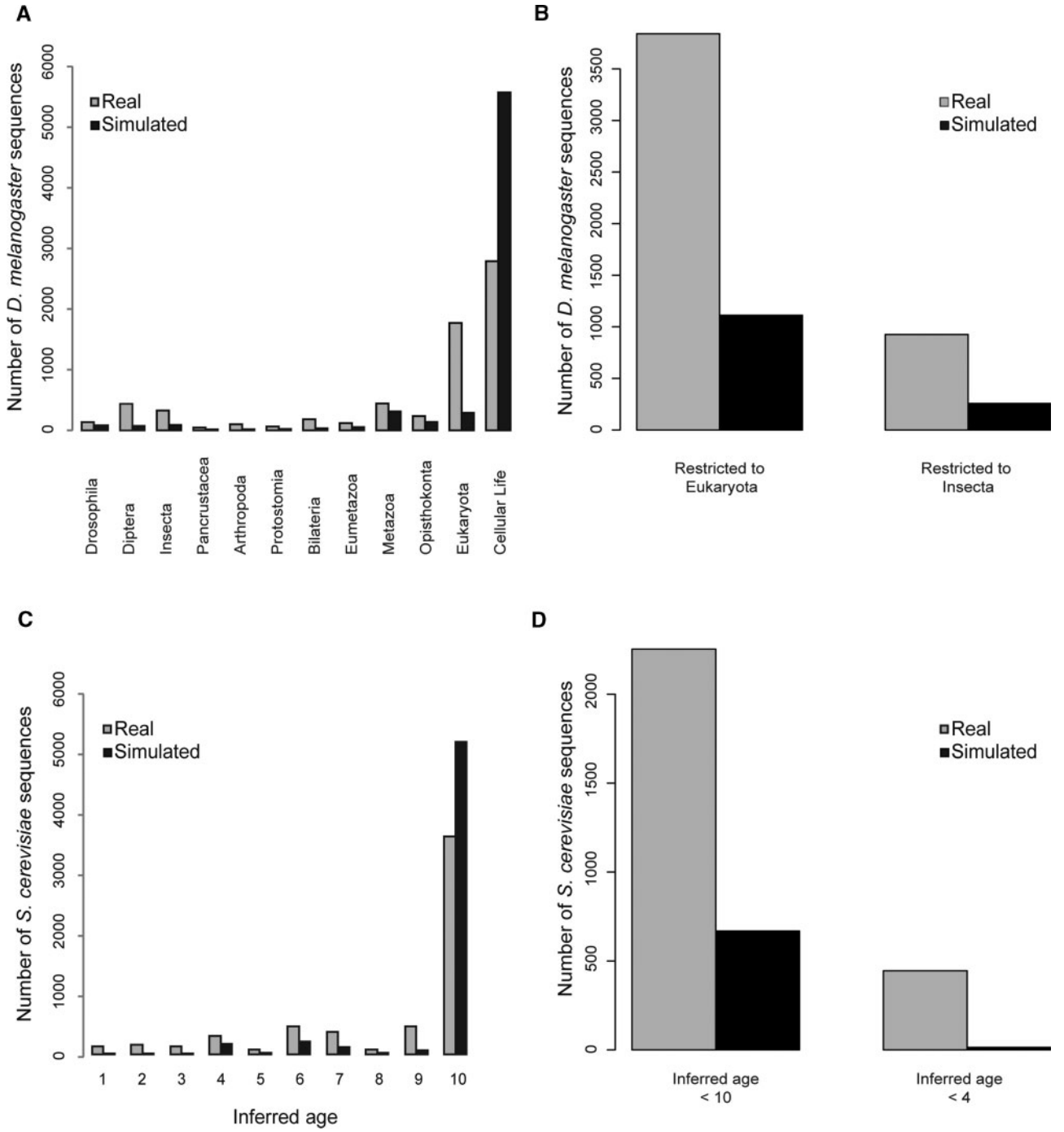

Fig. 1. The majority of phylostratigraphy-based young age assignments cannot be attributed to BLAST limitations for *D. melanogaster* or *S. cerevisiae*. (*A*) Phylostratigraphic assignments for the subset of *D. melanogaster* sequences chosen by Moyers and Zhang (2015) using real and simulated sequences. (*B*) Bar graph comparing the number of *D. melanogaster* sequences found young by phylostratigraphy using real and simulated sequences, when young is restricted to Eukaryota, or to the youngest three phylostrata (*Drosophila*, Diptera, Insecta). (*C*) Distribution is redrawn from Figure 1B in Moyers and Zhang (2016a), using a linear scale, rather than a log scale. Numbers indicate groups of *S. cerevisiae* ORFs of increasing conservation level within the Ascomycota, from *S. cerevisiae*-specific (1) to conserved *in S. pombe* (10). (*D*) Bar graph comparing the number of *S. cerevisiae* sequences found young by phylostratigraphy using real and simulated sequences, when young is considered to include all yeast species used for analyses except for *S. pombe* (inferred age < 10), or to the youngest three phylostrata (inferred age < 4). Note that the simulated results for *D. melanogaster* sequences represent the average number of sequences assigned to each phylostrata over ten runs.

No "Spurious" Patterns of Phylostratigraphy

We next asked if BLAST limitations, whatever their magnitude, could have influenced the published correlations between phylostratigraphic and evolutionary patterns. This was attempted by Moyers and Zhang (2015; 2016a) who, although they admittedly could not reproduce the exact patterns that were found in real data, claimed that the simulated sequences also yielded evolutionary patterns that appear

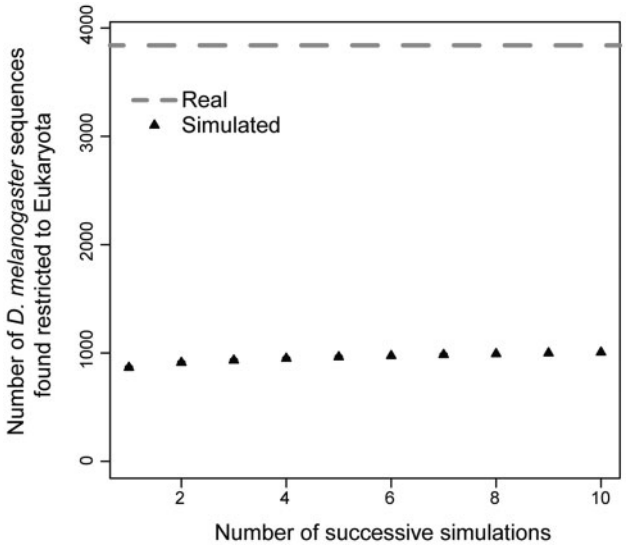

FIG. 2. Saturation analysis of *D. melanogaster* genes that are found error-prone by Moyers and Zhang's simulations (2015) (black triangles). Gray dashed line marks 3,840 sequences found restricted to Eukaryota in the real phylostratigraphy (Figure 1B). The average of 15 random permutations of 10 successive simulations is shown; standard errors of the mean are not shown because they are shorter than the height of the triangles.

interesting and significant, and that one would have no possibility to tell which ones are correct. Specifically, Moyers and Zhang criticize three series of results that we previously published: 1) Domazet-Lošo et al. (2007) reported that the genes expressed in ectoderm, mesoderm and endoderm during *Drosophila* development show a non-random distribution of phylostratigraphic conservation levels; 2) Domazet-Lošo and Tautz (2008) showed that human disease genes tend to be more ancient than expected; 3) Carvunis et al. (2012) and Abrusán (2013) found that many structural and functional characteristics of ORFs sequences (such as length, expression level or hydropathicity) correlate with their date of emergence in the Ascomycota fungal phylogeny. Moyers and Zhang (2015) also re-investigated the finding that new gene origination peaked in the common ancestor of Bilateria (Domazet-Lošo et al. 2007) but they could not recapitulate this pattern in their simulations.

First, we re-examined the claim according to which simulated *D. melanogaster* sequences may yield significant over- and underrepresentation of genes from certain age groups that are expressed in ectoderm, mesoderm, and endoderm (Moyers and Zhang 2015). We obtained the simulated sequence sets from the authors and reproduced the patterns shown in Moyers and Zhang (2015). However, we could not reproduce the corresponding significance values (fig. 3*A*), even without Bonferroni correction (see supplementary table S2, Supplementary Material online). The magnitude of the log-odds ratio obtained in their simulations (fig. 3*A*) is much lower than reported in the original study (Domazet-Lošo et al. 2007). Moyers and Zhang (2015) have combined ten simulation runs to obtain this pattern and its significance, but we show that the individual runs have no common trend and that none are significant (fig. 3*B*). After having been alerted to this problem by us while the present manuscript was under review, Moyers and Zhang re-evaluated their algorithms and found a mistake. A corresponding erratum has been published in MBE (Moyers and Zhang 2016b).

Nevertheless, to further evaluate the robustness of the central original finding that the genes emerging after the origin of eukaryotes tend to be expressed more in ectodermal than in endodermal and mesodermal tissues (Domazet-Lošo et al. 2007) we repeated the analysis of *Drosophila* germ layers using the most recent expression and sequence databases. The input dataset we use here was much better populated compared to the datasets in the original study (see Methods). This analysis confirmed the initial finding that the ectoderm is expressing evolutionary younger genes than the mesoderm and endoderm (fig. 4*A*). However, some of the fluctuations seen in the original data (i.e. fig. 2*A* in Domazet-Lošo et al. 2007) appear to be more smoothed out in the current analysis, likely due to the more extensive data available. When we removed from the analysis genes that Moyers and Zhang found susceptible to the BLAST error in their simulations (192 out of 4157 genes with expressions) the general profiles remained largely unaffected (fig. 4*B*), i.e. such potentially misplaced genes do not distort the major results.

Second, we observed another statistical problem in Moyers and Zhang's (2015) critique of our finding that human disease genes are enriched in ancient genes relative to young ones, which was originally shown by assessing the significance of log-odds ratio per phylostratum (Domazet-Lošo and Tautz 2008). In Moyers and Zhang's analyses, a set of human genes was simulated and they reported "a positive correlation between the inferred age of a gene and its probability of being a disease gene (Spearman's $\rho = 0.623$, $P = 0.004$; fig. 4)." [quote from (Moyers and Zhang 2015)]. This statement is actually different from our finding that two phylostrata (origin of life and origin of metazoans) show a significant enrichment of disease genes and that young genes are significantly under-represented (Domazet-Lošo and Tautz 2008). We tested Moyers and Zhang's simulated dataset using the original log-odds ratio approach (Domazet-Lošo and Tautz 2008) and found only non-significant under- and over-representations (fig. 5*A*). We then removed from the original dataset genes that Moyers and Zhang found error-prone in their simulations (571 out of 5217 simulated genes) and found that the general profiles remained largely unchanged (fig. 5*A*). In the simulated data, the distribution of disease genes on the phylostratigraphy map is not different from the distribution of the total set of genes, opposite of what we found in the original study (fig. 5*B*). These results together suggest that Moyers and Zhang's simulations did not mimic real phylostratigraphic maps in humans.

Third, we investigated whether the evolutionary continuum of structural and functional features in fungal ORFs reported by Carvunis et al. (2012) and Abrusán (2013) could be attributed to false negatives in BLAST, as claimed by Moyers and Zhang (2016a). In the original study, Carvunis et al. (2012) had included several technical controls. They showed that a significant correlation between conservation level and ORF length could be reproduced even when limiting

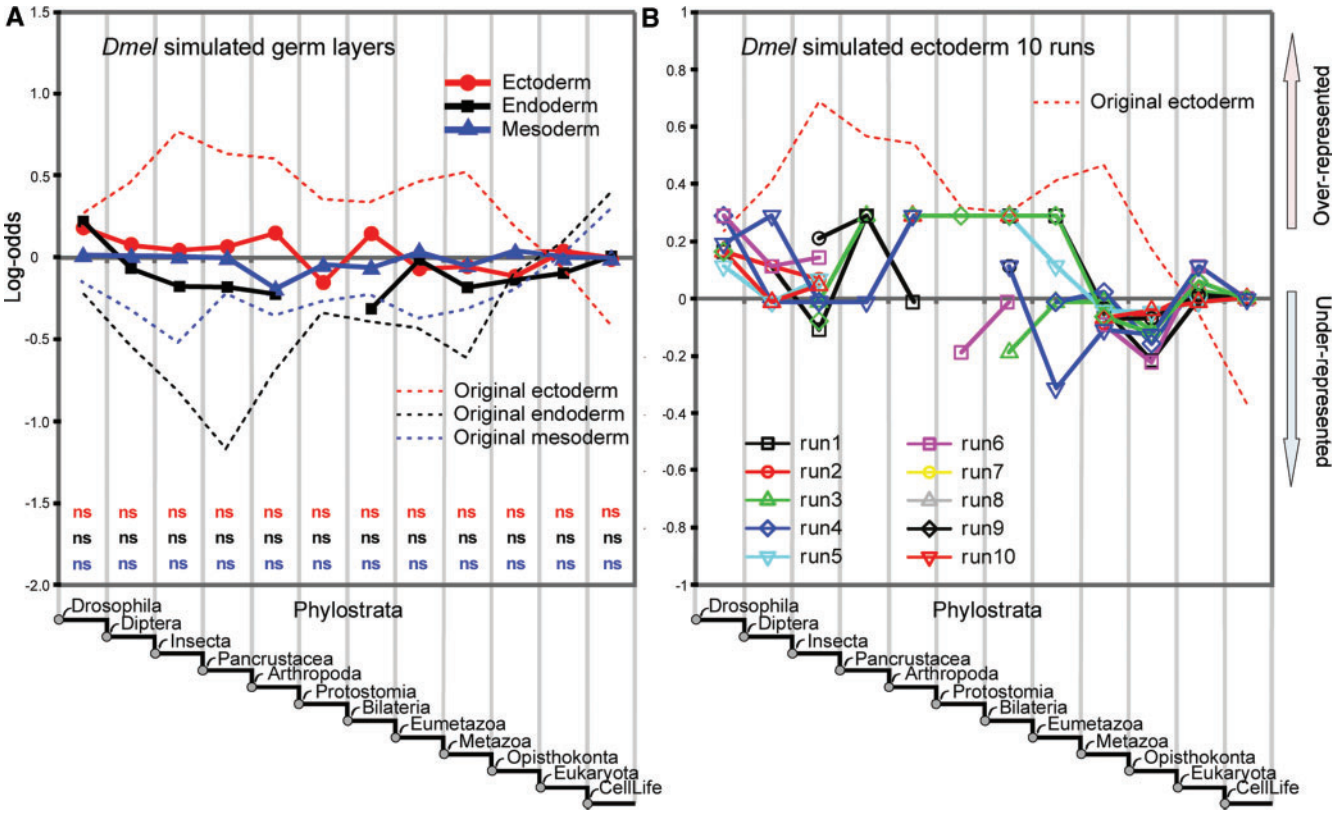

FIG. 3. Phylostratigraphic analyses of gene expressions in fruit fly germ layers are not attributable to false negatives in BLAST. (*A*) Overrepresentation profiles averaged over 10 simulated datasets reported by Moyers and Zhang (2015) in their figure 3c. None of the deviations is significant by hypergeometric test (ns) with Bonferroni correction. For comparison real phylostratigraphy profiles for germ layers are shown (dashed lines). (*B*) Overrepresentation profiles in ectoderm for 10 replicated simulations. Note the instability of profiles across the replicates and number of phylostrata without any expressed genes. None of the deviations at any phylostrata is significant by hypergeometric tests (ns). For comparison real phylostratigraphic profile for ectoderm are shown (dashed lines).

analysis to ORF sequences with BLAST hits covering at least 80% of sequence length. This ensured that the correlation was not solely driven by the higher probability of having a small conserved domain in a long ORF relative to a short one. They also implemented a series of partial correlations to control for the known cross-correlations between length, expression level, and evolution rates. Hence, even if BLAST were biased by length or evolution rates as observed by Moyers and Zhang, the correlation between conservation level and expression level would not be overly affected. Furthermore, all correlations reported by Carvunis et al. (2012) were checked for robustness by verifying that significance was also observed when excluding very young ORFs and when sampling only 50 ORFs from each phylostratum (100 bootstrap simulations per correlation statistics). This ensured that the signals were not solely driven by differences between ORFs of conservation level 10, which constitute the majority of the annotated genome, and other ORFs. Moyers and Zhang (2016a) did not reproduce any of these controls as they focused exclusively on the BLAST false negative rate.

To determine whether BLAST false negatives as estimated by Moyers and Zhang (2016a) may nevertheless explain the observed evolutionary continuum, we revisited the original analyses after excluding all ORFs deemed error-prone by Moyers and Zhang (2016a). All trends reported by Carvunis et al. (2012) were qualitatively and statistically robust to the BLAST false negative rate, although the values of Kendall *taus* decreased slightly when using this smaller subset of ORFs relative to the original study (fig. 6, table 1). We performed Kruskal–Wallis tests within each age group to quantify the significance of differences between the original, simulated, and reduced original ORF sets (see supplementary table S3, Supplementary Material online). The *P*-value of the Kruskal–Wallis test was smaller when comparing the original and simulated sets than when comparing the original and reduced original sets in the large majority of cases. Hence, rather than undermining the original conclusions, the simulation approach actually strengthens them.

Putative *De Novo* Genes and Proto-Genes Identified by Phylostratigraphy

Having shown that false negatives in BLAST searches barely affect phylostratigraphic outcomes, we next investigated whether phylostratigraphy can indeed detect de novo genes. As stated in a recent opinion piece (McLysaght and Hurst 2016), de novo genes like all genes must be under functional selection. Phylostratigraphy in this regard can only identify "putative" de novo genes since it evaluates sequence

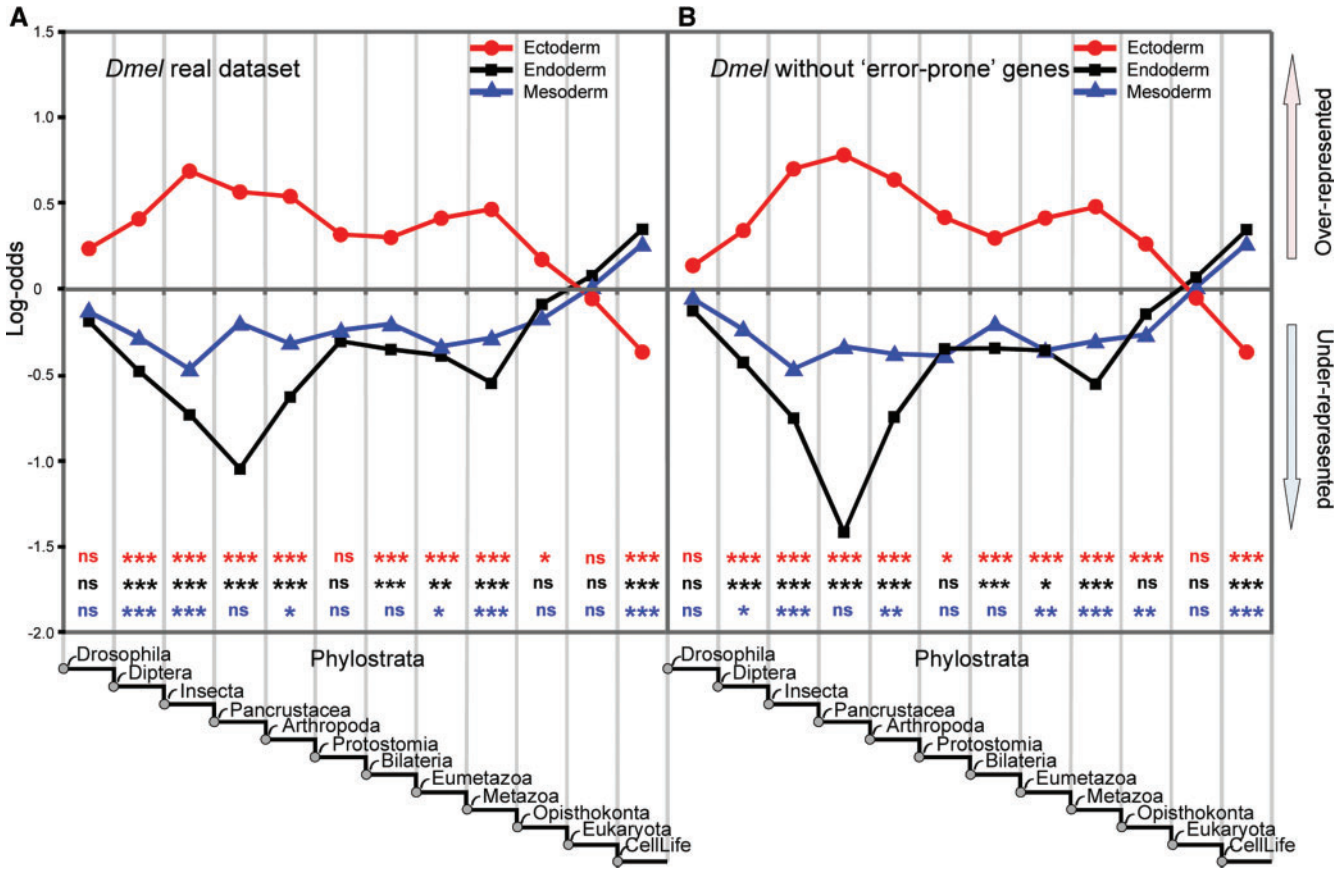

FIG. 4. Updated phylostratigraphic analyses of gene expression in fruit fly germ layers from Domazet-Lošo et al. 2007. (*A*) Real phylostratigraphic map using the latest sequence and expression databases. (*B*) Real phylostratigraphic map after the removal of genes that are found error-prone by Moyers and Zhang (2015). Note that the profiles remain largely unaffected. Stars represent significances after hypergeometric test with Bonferroni correction (* at 0.05 level, ** at 0.01 level and *** at 0.001 level).

conservation in a function-agnostic way. Carvunis et al. (2012) proposed to model recently-evolved ORF sequences as intermediate "proto-gene" stages that may harbor valuable information to study the mechanisms leading to the emergence of new genes even if they are not per se functional. Thus, they identified 143 *S. cerevisiae*-specific ORFs on the basis of sequence conservation alone. Among these, 16 exhibited evidence of selection at the intra-species level, suggestive of function even though no protein product has been reported. These 16 ORFs may thus cautiously be considered de novo genes rather than proto-genes. Moyers and Zhang (2016a) reanalyzed the evolutionary properties of these ORFs and argued that 15/16 of them were neither species-specific nor under selection. Based on their re-analyses these ORFs would be neither proto-genes nor de novo genes.

We found that these discrepancies stem from the use of differing methodologies between the two studies, both to determine species-specificity and to estimate selection. The 15 ORFs under consideration partially overlap a more conserved gene on another reading frame, as is frequent in the compact yeast genome. To estimate selection, Carvunis et al. (2012) calculated dN/dS on the full-length sequences based on the assumption that the codon-level evolution of the alternative reading frames would not overly influence estimations of dN/dS of the ORF sequences of interest. They found significant evidence for purifying selection. Moyers and Zhang (2016a) challenged these assumptions and focused on the overlap-free regions of the sequences to re-estimate dN/dS, finding no evidence of selection. However, they reported only between 0 and 3 SNPs per region. These low numbers prevent any statistical assessment of whether the number of non-synonymous SNPs compared to the synonymous ones is more or less than what is expected under neutrality using a Fisher test. There was, however, enough power to find evidence of selection based on the full length ORFs.

Carvunis et al. (2012) established the *S. cerevisiae*-specificity of these ORFs using phylostratigraphy on the overlap-free regions of the sequences. In contrast, Moyers and Zhang (2016a) this time considered the full-length ORF sequences and found them to be more conserved. This is not surprising, since the full-length sequences include sequences pertaining to other genes on alternative reading frames that are indeed more conserved. To determine unambiguously whether these 15 ORFs are species-specific or not, we performed here a synteny analysis based *Saccharomyces sensu stricto* alignments (Cliften et al. 2003; Kellis et al. 2003). We inferred the expected location of potential homologues for the 15 ORFs by virtue of the presence of homologues for the conserved genes partially

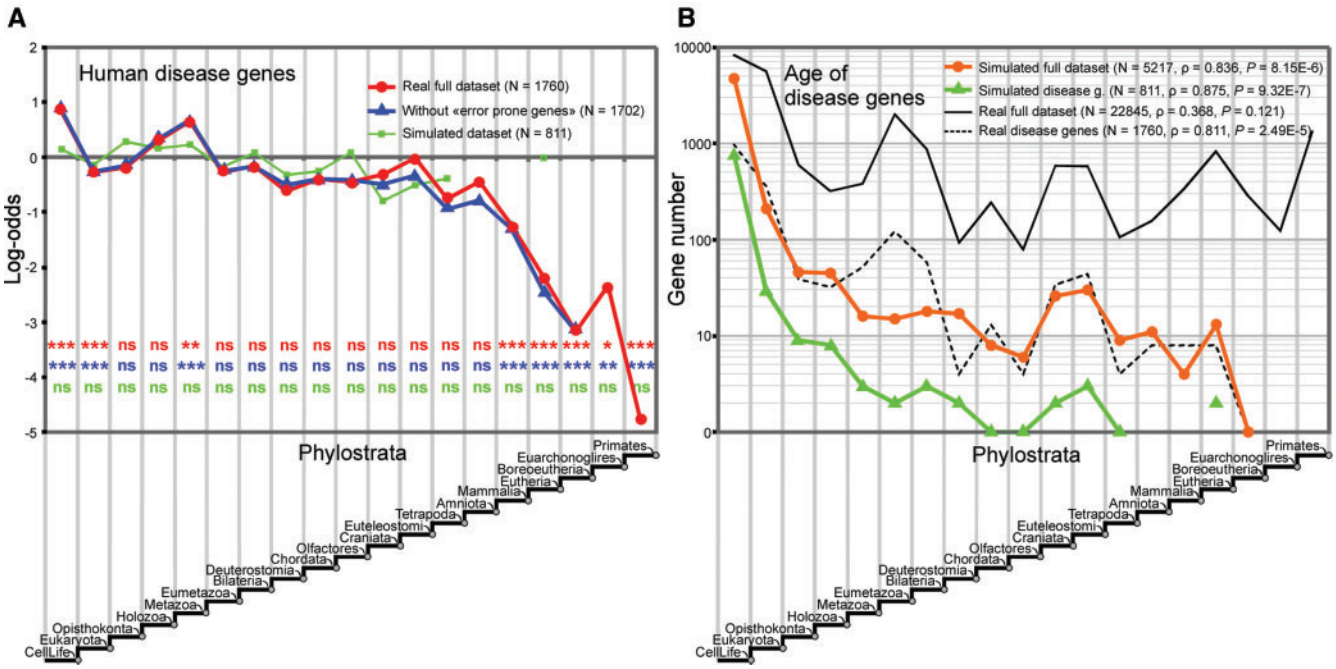

Fig. 5. Repeated phylostratigraphic analyses of disease genes in humans from Domazet-Lošo and Tautz (2008). (*A*) Real phylostratigraphic map after the removal of genes that are found error-prone by Moyers and Zhang (2015). Note that the profiles remain largely unchanged. The profile of Moyers and Zhang (2015) simulated dataset (green line) is completely non-significant. Stars represent significances after hypergeometric test with Bonferroni correction (* at 0.05 level, ** at 0.01 level and *** at 0.001 level). (*B*) Reanalyses of correlation patterns in Moyers and Zhang simulated data. The correlation coefficients (Spearman's rho) and associated *P*-values between gene count and ranked evolutionary time are in brackets. Note that the total set of simulated genes as well as simulated disease genes negatively correlate with evolutionary time.

overlapping them in another reading frame. ORFs were found in syntenic locations in 5 cases (YCL046W, YLR232W, YNL105W, YOR055W, and YOR135C). The remaining ten cases were confirmed as species-specific ORFs. Based on the SNPs distribution on their full-length sequences, these ten ORFs would thus be considered de novo genes. In absence of additional functional evidence, however, it is more prudent to consider them putative de novo genes or proto-genes.

In the original publication, Carvunis et al. (2012) had evaluated whether apparent novel sequences identified in their phylostratigraphy were more likely to represent recent de novo emergence versus sequence divergence (see supplementary fig. S3, Supplementary Material online in the original publication). They focused on sequences that appeared species-specific and overlapped a gene that appeared ancient and had one or more paralogues. They found that only 4 of 145 paralogues of these ancient genes also overlapped another ORF. Although the apparent species-specific ORFs could have recently and independently lost their paralogues as well as all of their homologues in the Ascomycota phylogeny, the most parsimonious scenario is that they did emerge de novo, after and independently from the duplication events. The prevalence of de novo emergence in the young phylostrata was also supported by a positive correlation between conservation level and number of paralogues per ORF. Altogether, these analyses confirm that apparent novel sequences detected by phylostratigraphy are enriched in putative de novo genes and proto-genes, at least when applied to closely related species.

Discussion

Estimating the power of BLAST from computer simulations is a difficult task, since simulations can never capture the complexity of real evolution. Nevertheless, Moyers and Zhang (2015; 2016a) attempted to do so and reported that phylostratigraphic analyses have a false negative rate of 11–14%. In this manuscript, we have taken the approach of assuming that Moyers and Zhang had correctly identified the sequences that are susceptible to appear young erroneously due to being too short or to evolving too fast for BLAST to find significant hits at the base of the phylogenetic tree. Under these assumptions, we showed that these sequences represent a negligible fraction of the sequences that do in reality appear young in phylostratigraphy (fig. 1). We also showed that the number of error-prone sequences identified by multiple re-runs of simulations saturates rapidly (fig. 2), demonstrating that the majority of sequences found young in phylostratigraphy cannot be attributed to BLAST problems, contrary to what has been suggested (Moyers and Zhang 2015; 2016a; McLysaght and Hurst 2016). Furthermore, previously reported patterns of gene emergence and evolution reanalyzed here and by Moyers and Zhang (2015; 2016a) are virtually unaffected by removing the error-prone sequences (figs. 4,5 and 6). Some of the proposed significant phylostratigraphic patterns observed in the Moyers and Zhang's simulations have turned out to be attributable to an error in their statistical analysis (fig. 3). Altogether, the studies by Moyers and Zhang have revisited previously discussed important issues but have failed to provide evidence for the existence

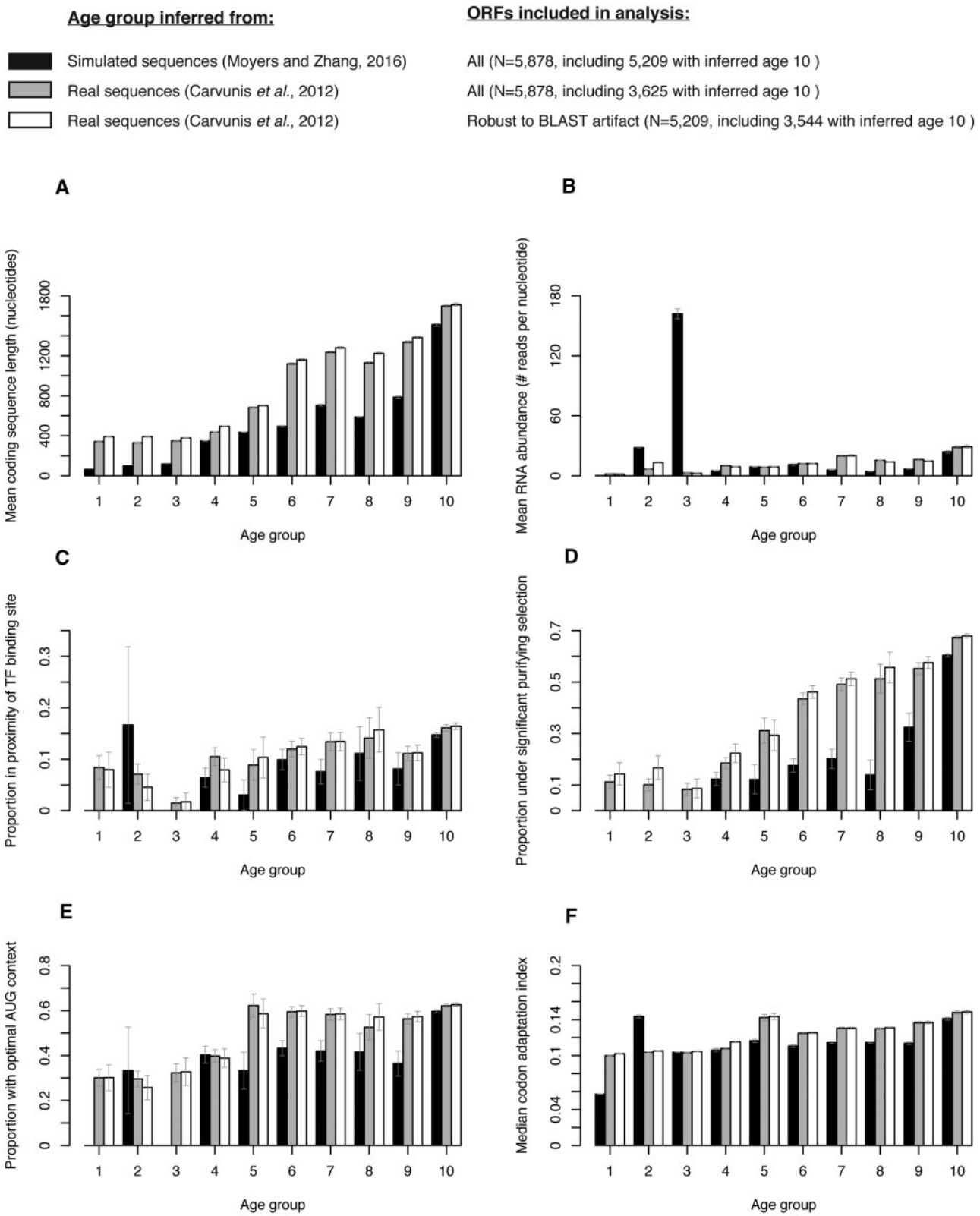

Fig. 6. Distribution of six biological features for 5,878 *S. cerevisiae* ORF sequences with age inferred from real data (grey), for the same 5,878 ORF sequences with age inferred in simulations (black) and for 5,209 ORF sequences shown to be robust to potential BLAST artifact because they are assigned to the oldest age group in the simulation, with age inferred from real data (white). Vertical error bars represent standard error of the mean (*A* and *B*), standard error of the proportion (*C*, *D* and *E*) or standard error of the median (*F*).

Table 1. Correlations (Kendall's *tau*) Between Inferred ORF Age and Various Biological Features.

Comparison	ORF Length	RNA Abundance	Proximity to TFBS	CAI	Purifying Selection	Optimal AUG Context
All real ORFs	**0.39****	**0.26****	**0.08***	**0.31****	**0.32****	**0.13****
Real ORFs, without error-prone ones	**0.33****	**0.17****	**0.07***	**0.25****	**0.23****	**0.09***
Simulated ORFs	**0.28****	**0.26****	**0.06***	**0.21****	**0.27****	**0.12****

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

Note that the conservation levels in the original Carvunis et al. (2012) paper and the first half of table 1 from Moyers and Zhang (2016a) comprised level 0, which corresponds to non-annotated *S. cerevisiae* ORFs, plus levels 1–10, estimated by phylostratigraphy on real or simulated sequences. Here, only levels 1–10 are considered.

of a hypothetical phylostratigraphic bias due to the use of BLAST.

Circularity in the Simulations Explains Similarities between Simulated and Real Sequences

We sought to understand why the simulation approach yielded in some cases results that were somewhat comparable to the real data. In their simulations, Moyers and Zhang (2015; 2016a) started with the real sequences–rather than in silico generated random sequences–and let them evolve randomly according to rate parameters inferred from real alignments among closely related species. Hence, the true features of these sequences are inherently still implied in the model, i.e. the same sequences that are short or fast-evolving in reality are also short or fast-evolving in the simulations. These sequences in turn are most likely to be misclassified in the simulations since length and evolution rate affect the performance of BLAST. Other features associated with the sequences, such as RNA expression levels or AUG context, were then taken from the real data, without any shuffling. This leads to circularity, since it has been well established that recently emerged genes are short and evolve rapidly, in part through studies of closely related species where no BLAST error could reasonably be invoked (Reinhardt et al. 2013; Ruiz-Orera et al. 2015). The effect becomes very evident when one looks at the overlap between the real sequences placed at particular nodes and the simulated equivalents (fig. 7). The vast majority of sequences assigned a young age group in the simulated phylostratigraphy (76% and 88% for *D. melanogaster* and *S. cerevisiae*, respectively) were also assigned a young age group in the original phylostratigraphies (although usually not the same one). Given these overlaps, it is evident that the characteristics of sequences of any given age group will be somewhat comparable between simulated and real data, since the sequences appearing young in the simulated data comprises mostly of the same sequences appearing young in the real data, with noise added and without dissociating the age-influencing features (length and divergence rate) from other features such as expression level or AUG context.

It is this circularity, rather than the false negative rate of BLAST per se (the alleged "phylostratigraphy bias"), that leads to seemingly similar patterns in the real and simulated data. Indeed, AUG context, proximity to TF binding sites or expression levels are associated with, but not contained within ORF sequences. How could these features technically affect BLAST similarity searches in any way? Moyers and Zhang (2016a) and others (McLysaght and Hurst 2016) have proposed that indirect cross-correlations between the different features could explain how a BLAST artifact would generate all these trends. For example, expression level is known to be inversely correlated with evolution rate (Pal et al. 2001; Drummond et al. 2006). It was argued that if evolution rate induces an ascertainment bias in age estimation, this bias would transcend to expression levels and explain why young sequences tend to have low RNA abundance (Moyers and Zhang 2016a; McLysaght and Hurst 2016). However, such argument would ignore another known fact, i.e. length and expression level are also inversely correlated (Jansen and Gerstein 2000). Thus, under the same reasoning, the ascertainment bias due to evolution rate would predict that young sequences are poorly expressed, but the bias due to length would predict the opposite. Such reasoning is thus rather uninformative. If one wanted to assess whether the false negative rate of BLAST per se would give rise to such significant patterns, one could randomly distribute the rate parameters across genome sequences to simulate their evolution along the phylogeny in a manner that would be independent of their associated features.

The Power of BLAST in Phylostratigraphic Analysis

We argue that Moyer and Zhang's estimates are likely to be exaggerated. In particular, the phylostratigraphy methodology used by Moyers and Zhang (2016a) to search for remote homologues among their simulated yeast sequences is less sensitive than the one deployed in the original analysis of real sequences (Carvunis et al. 2012). In the original analyses, the authors assigned to each ORF sequence the conservation level of its most conserved paralogue, in an effort to avoid underestimating conservation (Carvunis et al. 2012). Moyers and Zhang (2016a) did not implement this "oldest paralogue age" approach except in a single analysis, for which they did not report the corresponding BLAST false negative rate. Furthermore, where Moyers and Zhang (2016a) used the program BLASTP, Carvunis et al. (2012) used three BLAST programs: BLASTP, TBLASTX, TBLASTN. The use of three BLAST programs necessarily results in a lower false negative rate than the use of a single program, especially by enabling comparisons against whole-genome databases rather than against databases containing only annotated transcripts and protein-coding genes. It is thus evident that the false negative rates of the original phylostratigraphic analyses must be lower than those estimated by Moyers and Zhang.

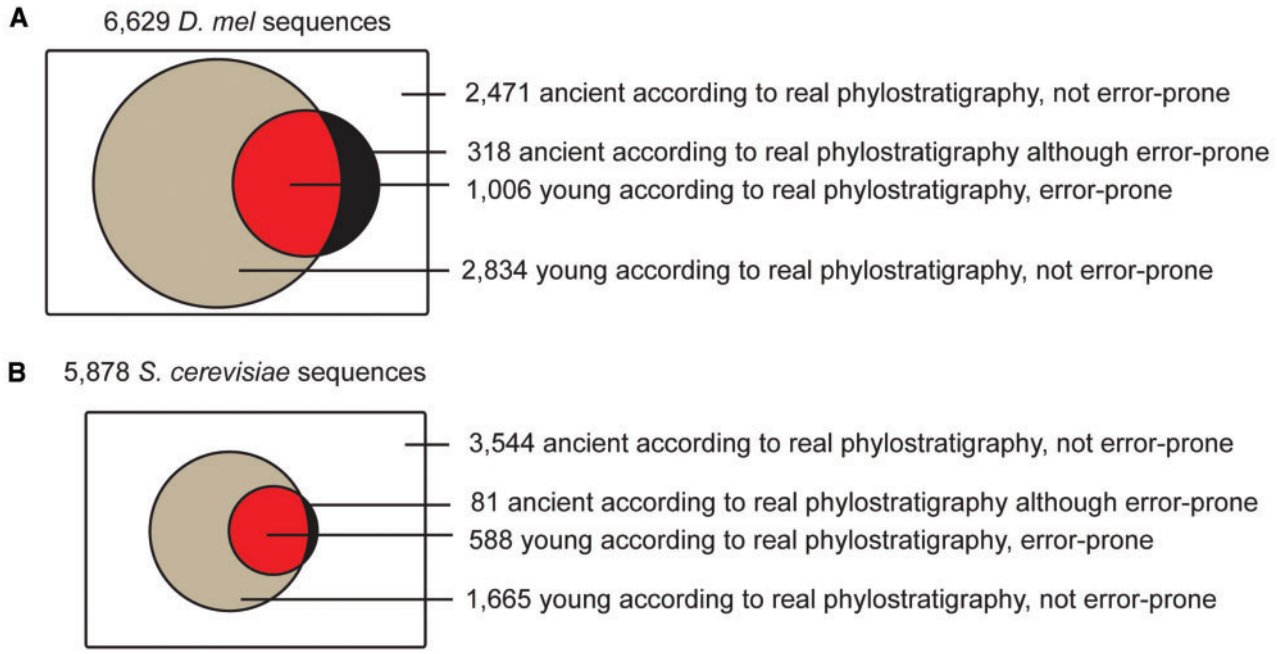

Fig. 7. Pie charts representing sequences in the real phylostratigraphy and their relation to the sequences found error-prone in the Moyers and Zhang simulations for *D. melanogaster* (*A*) and *S. cerevisiae* (*B*). The majority of sequences found young in real data are robust to BLAST artifact (grey). Some sequences are found ancient in the real data but not in the simulated data (black), indicating that the phylostratigraphic methods used in the real data were more sensitive than those used on the simulated data. The only sequences whose phylostratum may be have been underestimated due to BLAST errors are in red. For *Drosophila*, a conservative approach was taken where we counted as susceptible to BLAST artifact all sequences found young in at least one of ten simulation runs. For yeast, a single run was performed and analyzed. Note that the proportion of sequences found young is larger in *Drosophila* (*A*) than in yeast (*B*) because the species tree considered is much deeper.

We also note that Moyers and Zhang (2016a) misinterpreted Abrusán (2013) by stating he "used Carvunis et al.'s data to examine a number of additional gene properties that he proposed to reflect the gradual genetic integrations of de novo genes into cellular networks or maturation of protein structures" [quote from (Moyers and Zhang 2016a)]. However, Abrusán (2013) only used the classification of very young ORFs from Carvunis et al. ("proto-genes") but drew from the orthology classification provided by Wapinski et al. (2007) to classify all more conserved genes, which constitute the majority of annotated ORFs in the *S. cerevisiae* genome. The false negative rate associated with the methodology used by Abrusán (2013) was not estimated by Moyers and Zhang (2016a).

Moyers and Zhang (2015) claimed that their estimates of BLAST detection errors are conservative, in particular due to not taking into account variations in rate heterogeneities across time. Such changes are indeed well known in phylogenetic analysis under the term covarion pattern of protein evolution (Penny et al. 2001). Moyers and Zhang (2015) simulate such a covarion pattern to assess BLAST performance in an attempt to provide an even more realistic framework of protein evolution. They find that BLAST performs indeed less well under these conditions, with up to 67% error rate in finding the oldest assignments. However, to obtain such a high rate of misplacement, they had to assume unrealistic parameters. This should already be evident from the fact that such a high misplacement rate is not compatible with real data, since most genes are actually mapped to the basal nodes in all phylostratigraphies (e.g. Domazet-Lošo and Tautz 2008; Tautz and Domazet-Lošo 2011). In their covarion model they shuffle over time the rates of up to 5% of sites per 50My and state that shuffling 1% of sites per 50My is a "tiny amount of covarion evolution" [quote from (Moyers and Zhang 2015)]. However, when 2,500My of evolution are simulated, 1% per 50My amounts to 50% of the protein in total. Actual covarion proportions in well-studied real proteins of this age were found to be around 10% (Wang et al. 2009). Hence, even 1% of sites per 50My is already beyond the realistic parameter space, let alone the 5% where they find the highest error rate. Even at the exaggerated 1% rate, the BLAST error is only around 18% [compare table 2 in (Moyers and Zhang 2015)]. The actual interpretation should therefore be that BLAST, when used in the phylostratigraphic framework, is very robust with respect to the rate heterogeneities found in real data.

Phylostratigraphy and *De Novo* Evolution

Moyers and Zhang (2016a) concede that "nothing is wrong with the theoretical model of de novo gene birth." Their fundamental point of contention with Carvunis et al. (2012) and Abrusán (2013), which goes beyond a mere supposed 11% false negatives, is that the original publications did not explicitly state why the observed trends would be expected from the de novo gene birth model. For example, Moyers and Zhang wonder "why the refinement of biological function of an ORF has to occur by increasing the ORF length rather than by decreasing the length," why "the mean hydropathicity should decrease" etc. They are particularly surprised

to see that many of the trends continue even for older phylostrata, "as if the maturation of de novo genes takes more than 500 Myrs." Let us here clarify these questions.

The prediction of the proto-gene model for de novo gene birth is actually broader than any single descriptor of genes such as length or hydropathicity: it is that the functional and structural characteristic of ORFs should follow an evolutionary continuum between non-genic sequences and genes (Carvunis et al. 2012). For example in the case of *S. cerevisiae*, non-genic sequences are rippled with short ORFs thought to appear and disappear by chance through random mutations. In contrast, canonical protein coding genes with established biological functions are on average much longer. Thus, the continuum prediction of the de novo gene birth model is that, in Ascomycota, ORF length should increase on average with evolutionary conservation. This is not meant to imply that ORF length would continuously increase, for all ORFs, over extended periods of evolutionary time. Rather, the statement simply indicates that, since the randomly appearing ORFs are virtually all short, only those that have been maintained over longer periods of time can be long. This mathematically leads to an increase of average length over time-since-emergence. This trend is indeed also seen in studies of vertebrate taxa (Toll-Riera et al. 2009; Neme and Tautz 2014). This by no means implies that established genes cannot shorten due to the action of natural selection at some point during their evolution, or that all proto-genes lengthen and keep lengthening continuously even after their function is established. The low values of correlations coefficients (table 1) illustrate well that these are merely statistical trends supporting the existence of a continuum. One could imagine that the continuum prediction would actually predict the opposite trends in species where randomly appearing ORFs would tend to be longer, as may be the case in the Mycoplasmataceae lineage, which uses only two stop codons (Tatarinova et al. 2016). Because the continuum prediction is so general, it allows investigators to discover evolutionary trends without a priori suppositions of how de novo proteins should evolve. Rather, the data can be examined with an open mind thanks to the power of phylostratigraphy.

Moyers and Zhang (2016a) discuss also whether there is a prevalence of origination of new genes via gene duplication or de novo evolution. Carvunis et al. (2012) have discussed long-term trends and concluded that de novo evolution may be more frequent. This was also the finding of Neme and Tautz (2013) in vertebrates. The overall pattern of extensions of transcript length, number of exons, length of ORFs and acquisition of domains makes it more likely that new genes are initially short. If one would want to explain such trends through a duplication-divergence model, one would have to assume either that short genes are more likely to be duplicated, or that genes become shorter after duplication. Neither of these trends have so far been reported.

The Future of Phylostratigraphy

The data presented here demonstrate that phylostratigraphic analyses of patterns of gene emergence and evolution are robust to the false negative rate of BLAST, whether it is in the range of 5% or 15%. Still, future research is needed to improve existing methods and date the emergence of sequences with even higher accuracy. Research in this direction should consider not only BLAST false negatives, but also BLAST false positives, where BLAST hits are spurious rather than true homologues of the query sequences. A promising approach is also to derive error estimates for the placement at particular nodes (Liebeskind et al. 2016). Phylogenetic comparative methods, which account for phylogenetic structure in the data, are helpful when one aims to correlate phenotypes between multiple species on the phylogeny (Hejnol and Dunn, 2016).

There is now overwhelming evidence that de novo gene birth has occurred repeatedly in many lineages, where possible deficiencies of detection via BLAST play no practical role. There is no reason to assume that the proven high rate of de novo evolution has not occurred throughout evolutionary history. Although the turnover of proto-genes seems very high (Palmieri et al. 2014; Neme and Tautz 2016), some have been retained, in particular at times of major radiations and evolution of new lineages (Tautz and Domazet-Lošo 2011). We concur with Moyers and Zhańgs (2016a) suggestions that gene-by-gene studies will provide deeper insights into these questions. In particular, coupling phylostratigraphy with synteny analyses may in the future enable to distinguish between duplication-divergence and de novo evolution, at least for sequences with traceable genomic locations across species (McLysaght and Hurst 2016).

Methods

Scripts and data files necessary to reproduce our figures are provided at: https://github.com/annerux/Domazet-Loso_MBE_2016 (last accessed December 16, 2016).

Reanalysis of Moyers and Zhang 2015 Dataset and Statistics

Moyers and Zhang kindly sent us their *Drosophila* dataset with the list of genes that contained ectoderm, endoderm, and mesoderm and their simulated phylostrata over ten simulation runs. For our saturation analysis (fig. 2), we generated 15 random permutations of these ten simulation runs. For each permutation, we calculated the number of *Drosophila melanogaster* genes found young in the real phylostratigraphy (lacking a detected ancestor at Cellular Life) that could have been misplaced when considering the union of 1 simulation, 2 simulations, . . ., 10 simulations. We then averaged the numbers over the 15 random permutations. We also repeated their ontogeny analysis and calculated hypergeometric tests with Bonferroni correction for all ten runs and three germ layers (see supplementary table S2, Supplementary Material online). We created our figure 3A to match their figure 3C by using average values of ten runs. To be able to calculate significances by hypergeometric tests we rounded rational numbers obtained by averaging to integers. We also obtained from Moyers and Zhang their human dataset with the list of 809 disease genes and simulated phylostrata for 5217 human genes over ten simulation runs. To

calculate statistics and visualize simulated disease genes we averaged ten runs and rounded the obtained numbers to integers. This procedure changed the number of disease genes in the calculations to 811 (fig. 5 and in see supplementary table S2, Supplementary Material online). Using this simulated dataset we performed overrepresentation and correlation analyses as in Domazet-Lošo and Tautz (2008) (fig. 5). Due to changes in gene annotations we were able to link 571 out of 585 error-prone genes to our previous study.

Phylostratigraphic Reanalysis of *Drosophila melanogaster*

To allow broad sequence similarity searches we first built a custom built protein database by combining complete genomes from National Center for Biotechnology Information (NCBI), Ensembl and Joint Genome Institute (JGI). In total, we collected 113,834,351 protein sequences from 25,223 genomes. To reduce the large redundancy of prokaryotic sequences (23,675 prokaryotic genomes) we clustered prokaryotic parts of the database with the CD-HIT at 90% identity (Li and Godzik 2006). After this procedure our database contained 43,899,817 protein sequences. For comparison, in the original study we used a database that comprised 2,777,855 protein sequences (only around 2% of the present database size).

We compared 13,389 protein sequences of *Drosophila melanogaster* retrieved from the Ensembl database (Yates et al. 2016) against the protein database by using the similarity search algorithm BLASTP (Altschul et al. 1997) at E-value cutoff of 1e-03 (Domazet-Lošo et al. 2007). Using the obtained BLAST output we mapped the fruit fly genes onto a consensus phylogeny (12 phylostrata) using the most-distant BLAST match above the significance threshold (BLAST E-value less than 1e-03) as described in the original study (Domazet-Lošo et al. 2007). This updated *Drosophila* phylostratigraphy is provided in see supplementary table S1, Supplementary Material online.

Drosophila Expression Data and Statistics

We retrieved in situ hybridization expression data for 4,157 fruit fly genes that show tissue-specific expression during ontogeny from the Berkeley *Drosophila* Genome Project (Tomancak et al. 2002). In total, this set of genes contributes to 38,627 expression domains expressed over multiple tissues and the different stages of the ontogeny. In the original study we had used 1,967 genes with 10,432 expression annotations (only around 27% of the present expression dataset). We divided the fruit fly expression dataset into subsets corresponding to the specific germ layer (either ectoderm, endoderm, or mesoderm). For every germ layer, we performed an over-representation analysis by comparing a frequency of expression domains in a phylostratum to a frequency in the total dataset (expected frequency) (Domazet-Lošo et al. 2007; Domazet-Lošo and Tautz 2008; Domazet-Lošo and Tautz 2010b; Šestak et al. 2013; Šestak and Domazet-Lošo 2015). Obtained deviations, i.e. more or less expression than expected, are depicted in the figures by log-odds ratios and their significance was tested by two-tailed hypergeometric tests (Rivals et al. 2007) controlled for multiple comparisons via a Bonferroni correction.

Fungal Data and Statistics

The conservation levels of *S. cerevisiae* ORFs was estimated by Carvunis et al. (2012) and simulated by Moyers and Zhang (2016a). Moyers and Zhang kindly provided us with the results of their simulations. Only 5,878 ORFs that were assigned a conservation level by both studies are included here. ORF characteristics (length, expression level etc.) were taken as in Carvunis et al. (2012). Distributions, error bars and *P*-values were computed using R scripts available at https://github.com/annerux/Domazet-Loso_MBE_2016. Synteny analysis for 15 ORFs was performed using the synteny viewer and fungal alignment resources provided by SGD (Cliften et al. 2003; Kellis et al. 2003).

Supplementary Material

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

Acknowledgments

We thank B. Moyers for discussion and for providing analysis files, as well as G. Abrusán, M. Calderwood, H. Carter, J. Castresana, B. Charloteaux, J. Kreisberg, A. McLysaght and M. Domazet-Lošo for discussion, comments, and suggestions on the manuscript. We thank the following funding organizations for support of our work: TD-L: City of Zagreb and Adris Foundation grants; A-RC: National Institute of Health (NIH) grant K99 GM108865; MA: grant BFU2015-65235-P from MINECO/FEDER, EU; DT: ERC grant NewGenes, 322564.

References

Abrusán G. 2013. Integration of new genes into cellular networks, and their structural maturation. *Genetics* 195:1407–1417.

Albà MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol*. 22:598–606.

Albà MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evol Biol*. 7:53.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403–410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acids Res*. 25:3389–3402.

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

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* 487:370–374.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterson R, Cohen BA, Johnston M. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301(5629):71–76.

Domazet-Lošo T, Tautz D. 2003. An evolutionary analysis of orphan genes in drosophila. *Genome Res*. 13:2213–2219.

Domazet-Lošo T, Brajkovic J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet*. 23:533–539.

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

Domazet-Lošo T, Tautz D. 2010a. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468:815–818.

Domazet-Lošo T, Tautz D. 2010b. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol*. 8:66.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*. 23:327–337.

Drost H-G, Bellstädt J, Ó'Maoiléidigh DS, Silva AT, Gabel A, Weinholdt C, Ryan PT, Dekkers BJW, Bentsink L, Hilhorst HWM, et al. 2016. Post-embryonic hourglass patterns mark ontogenetic transitions in plant development. *Mol Biol Evol*. 33:1158–1163.

Elhaik E, Sabath N, Graur D. 2006. The "Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes" is an artifact of increased genetic distance with rate of evolution and time of divergence. *Mol Biol Evol*. 23:1–3.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucl Acids Res*. 39:W29–W37.

Heinen TJ, Staubach F, Häming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. *Curr Biol*. 19:1527–1531.

Hejnol A, Dunn CW. 2016. Animal evolution: are phyla real? *Curr. Biol*. 26:R424–R426.

Jansen R, Gerstein M. 2000. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucl Acids Res*. 28:1481–1488.

Kellis M, Patterson N, Endrissi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241–254.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet*. 25:404–413.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res*. 19:1752–1759.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci*. 103:9935–9939.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.

Liebeskind BJ, McWhite CD, Marcotte EM. 2016. Towards consensus gene ages. *Genome Biol Evol*. 8:1812–1823.

McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat Rev Genet*. 17:567–578.

Mendoza AD, Sebé-Pedrós A, Šestak MS, Matejčić M, Torruella G, Domazet-Lošo T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc Natl Acad Sci* 110:E4858–E4866.

Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol*. 32:258–267.

Moyers BA, Zhang J. 2016a. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. *Mol Biol Evol*. 33:1245–1256.

Moyers BA, Zhang J. 2016b. Erratum. *Mol Biol Evol*. 33(11):3031.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genom*. 14:117.

Neme R, Tautz D. 2014. Evolution: dynamics of de novo gene emergence. *Curr Biol*. 24:R238–R240.

Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife* 5:e09977.

Pal C, Papp B, Hurst L. 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.

Penny D, McComish BJ, Charleston MA, Hendy MD. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J Mol Evol*. 53:711–723.

Quint M, Drost H-G, Gabel A, Ullrich KK, Bönn M, Grosse I. 2012. A transcriptomic hourglass in plant embryogenesis. *Nature* 490:98–101.

Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. De novo ORFs in drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLOS Genet*. 9:e1003860.

Remmert M, Biegert A, Hauser A, Söding J. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*. 9:173–175.

Rivals I, Personnaz L, Taing L, Potier M-C. 2007. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* 23:401–407.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. 2015. Origins of de novo genes in human and chimpanzee. *PLOS Genet*. 11:e1005721.

Schlötterer C. 2015. Genes from scratch–the evolutionary fate of de novo genes. *Trends Genet*. 31:215–219.

Šestak MS, Božičević V, Bakarić R, Dunjko V, Domazet-Lošo T. 2013. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Front Zool*. 10:18.

Šestak MS, Domazet-Lošo T. 2015. Phylostratigraphic profiles in zebrafish uncover chordate origins of the vertebrate brain. *Mol Biol Evol*. 32:299–312.

Tatarinova TV, Lysnyansky I, Nikolsky YV, Bolshoy A. 2016. The mysterious orphans of Mycoplasmataceae. *Biol Direct*. 11:2.

Tautz D. 2014. The discovery of de novo gene evolution. *Perspect Biol Med*. 57:149–161.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet*. 12:692–702.

Tautz D, Neme R, Domazet-Lošo T. 2013. Evolutionary origin of orphan genes. In: eLS. Chichester: Wiley. DOI: 10.1002/9780470015902.a0024601

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà MM. 2009. Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol*. 26:603–612.

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis S, Richards S, Ashburner M, Hartenstein V, Celniker S, et al. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*. 3:research0088.1–research0088.14.

Wang H-C, Susko E, Roger AJ. 2009. PROCOV: maximum likelihood estimation of protein phylogeny under covarion models and site-specific covarion pattern analysis. *BMC Evol Biol*. 9:225.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, et al. 2016. Ensembl 2016. *Nucl Acids Res*. 44:D710–D716.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in drosophila. *Genome Res*. 18:1446–1455.