Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth

Benjamin A. Wilson1, Scott G. Foy1,31, Rafik Neme2,3* and Joanna Masel1†

The phenomenon of de novo gene birth from junk DNA is surprising, because random polypeptides are expected to be toxic. There are two conflicting views about how de novo gene birth is nevertheless possible: the continuum hypothesis invokes a gradual gene birth process, whereas the preadaptation hypothesis predicts that young genes will show extreme levels of gene-like traits. We show that intrinsic structural disorder conforms to the predictions of the preadaptation hypothesis and falsifies the continuum hypothesis, with all genes having higher levels than translated junk DNA, but young genes having the highest level of all. Results are robust to homology detection bias, to the non-independence of multiple members of the same gene family and to the false positive annotation of protein-coding genes.

It has become clear that protein-coding genes can originate de novo from non-coding sequences1. This is surprising, because amyloid is a generic structural form of any polypeptide, making the expression of random polypeptides a dangerous affair2. De novo gene birth is a radical evolutionary transition, and two hypotheses have previously been presented to explain how it is possible. The ‘continuum’ view posits that there is a series of intermediate stages, or ‘proto-genes’, between non-genes and genes3. In contrast, the ‘preadaptation’ theory posits that de novo birth is an all-or-nothing transition to functionality, and the key to successful innovation is the imperative to avoid the most toxic ‘hopeless monster’ options in favour of ‘hopeful monsters’4–6; given a marker for such avoidance, newborn genes will therefore have exaggerated, rather than intermediate, gene-like characteristics. In other words, newborn gene birth occurs only from sequences that happen to be pre-adapted to ‘first, do no harm’ in the most direct way possible. Only later do they adapt to protect themselves against risks in subtler ways, increasing tolerance with respect to other more subtle characteristics7.

These two theories can be empirically distinguished, given a simple trait that systematically makes proteins less likely to be harmful. According to the continuum theory, the trait should be strongest in old genes, intermediate in young genes and weakest in non-coding sequences. According to the preadaptation theory, the trait should be strongest in young genes, intermediate in old genes and weakest in non-coding sequences (Fig. 1).

A good candidate trait is the degree of intrinsic structural disorder (ISD), that is, the degree to which a given peptide folds as a stable three-dimensional structure (ordered) versus as a rather flexible and unstructured entity (disordered)8. Predicted levels of ISD can conveniently be calculated from sequence alone9. Natural protein sequences are more intrinsically disordered than random sequences10. Meanwhile, there is conflicting evidence as to whether young genes are more or less disordered than old genes. Increased ISD is found in orphan domains and recent extensions to domains11–13. Elevated ISD was found in complete orphan genes in Leishmania15 and in genes created de novo in alternative reading frames of existing viral genes16, but low ISD was found in Saccharomyces orphan genes13.

Here, we stratify genes in two distantly related eukaryotic organisms, the house mouse and baker’s yeast, by age, and use predicted ISD values to show that young genes do not behave like intermediates between non-genic sequences and older genes, as predicted by a continuum theory. Instead, young genes have exaggerated gene-like structural properties, as expected from sequences biased by the demands of preadaptation.

Results and discussion

We rigorously determine how ISD depends on gene age in mice using a phylostratigraphy approach17 to assign ages to genes. We exclude genes unique to a single species, as these are likely to be contaminated with many false positives that are not protein-coding genes at all. Mouse is a good choice of taxon because of the quality of its gene annotation and the large number of genome sequences available for closely related species18, which together provide temporal resolution among young protein-coding genes. The reality of horizontal gene transfer in mouse ancestors is also an important consideration.

We find that young mouse genes have higher ISD (Fig. 2). Our statistical analysis avoids pseudoreplication in the form of phylogenetic confounding among multiple members of the same gene family, by controlling for gene family as a random effect in a linear model; this is not a technique we have previously encountered in related literature.

The validity of phylostratigraphy has been challenged on the grounds of homology detection bias19–21. Disordered proteins evolve faster13,22, if this makes homology undetectable, then high ISD could cause young gene status, rather than young gene status being the cause of high ISD23. Homology detection bias is minimized by focusing on the youngest genes26, we therefore collapsed all prevertebrate phylostra into a single ‘old gene’ category, to focus on gene ages with the least homology detection bias. What is more, correcting for the influence of evolutionary rate, via a linear regression analysis, had no effect on the predictive power of gene age (P > 0.05), despite the fact that evolutionary rate and age are correlated (Fig. 3a). Both proponents27 and detractors18,19 of phylostratigraphy have used simulations of protein evolution to justify their position on the impact of homology detection bias; unfortunately,
this line of argument relies on our ability to model protein evolution realistically. Our more empirical approach suggests low impact, in a more direct and less model-dependent fashion.

Homology is also easier to detect for longer genes and, in agreement with previous findings, length and age are correlated (Fig. 3b). However, longer genes have higher ISD as scored with IUPred, with quite large effect sizes—we find a Pearson correlation coefficient (following transformation for normality) of 0.17 for old genes and in the range of 0.32–0.44 in newer phylostrata. Using a linear model to correct for the length–ISD relationship therefore makes our ISD–age relationship stronger, not weaker (Fig. 2, green).

Note that our ISD scores strongly reflect amino acid composition, with low ISD representing hydrophobicity. In this light, our findings agree with previous results on length-dependent frequencies of particular amino acids25, but contradict previous studies, restricted to single-domain globular proteins, that showed no length-dependence for hydrophobicity as a whole26–27. Applying the same hydrophobicity measure26 to our more comprehensive protein set, we continue to find that long genes have more hydrophilic amino acids (Pearson correlation coefficients are slightly weaker, at —0.13 for old genes, and in the range of —0.24 to —0.45 in new phylostrata).

The high ISD of young genes is primarily due to amino acid composition (calculated as the ISD of scrambled versions of a gene; Fig. 4a, orange) rather than the exact order of the amino acids (calculated as the difference between the ISD of the real gene and that of a scrambled control; Fig. 4b). This suggests that genes are born with high ISD, driven by amino acid composition. Amino acid composition can therefore be seen as a preadaptation28, or ‘nonaptation’ in the terminology of ref. 29, for de novo gene birth. This does not imply any pre-adaptive process of the kind postulated elsewhere1–6,10. Instead, the term preadaptation simply refers to backward-time conditional probability; given that gene birth occurred, the non-coding sequence from which the gene was born is likely to have had more favourable characteristics for gene birth than the average non-coding sequence. Note that while higher guanine–cytosine (GC) content leads to higher ISD31, GC content is not the primary driver of the high ISD-promoting amino acid composition of young genes (Fig. 4a).

In contrast, the contribution of amino acid order to ISD seems to be an adaptation rather than a preadaptation in young genes (those born in vertebrates), because it is not initially present but appears only after some time. This contributes an independent line of evidence that high ISD values are favoured in young genes. Now that we have determined how ISD depends on gene age, the non-coding sequences from which de novo proteins must be born allow us to distinguish between the continuum and preadaptation hypotheses. The continuum theory predicts that non-coding sequences will resemble exaggerated versions of young genes, and hence have the highest ISD. In contrast, the preadaptation theory expects young genes to show the most extreme deviation from random sequences, predicting that non-coding sequences will have the lowest ISD.

We sampled intergenic sequences near each mouse gene in our analysis as representative of the raw material from which de novo genes are born, rather than analysing randomly generated sequences matching only a subset of known variables, such as GC content31. Our intergenic controls reflect the subleties found in a real genome with a complex evolutionary history, such as the avoidance of CpG sites. We find strong evidence refuting the continuum hypothesis and supporting the preadaptation hypothesis (Fig. 2). This result is not attributable to repetitive sequences; results are nearly identical when RepeatMasker is used to filter the intergenic control sequences (Fig. 2). The size of the gap between the average ISD of a translated intergenic sequence and that of young genes strikingly illustrates the nature of the filter applied during de novo gene birth, and the relevance of ISD to the process.

Why, then, did a previous yeast study find that young genes and proto-genes have low ISD? We believe that the difference lies in the
our methods, young yeast genes, like young mouse genes, have higher ISD. To pinpoint the source of the discrepancy, we used gene age classifications from the previous study (A.-R. Carvunis, personal communication of dataset) and, omitting gene families from the analysis, reproduced the previously reported trend of low ISD in young proto-genes (Fig. 6b, black). Details such as the treatment of disulfide bonds made no substantive difference (Fig. 6b, dark blue). However, filtering out potentially non-coding sequences eliminated the previously reported trend (Fig. 6b, light blue). Details of the elimination criteria are shown in Table 1; the preferential elimination of younger phylostrata is consistent with the operation of Simpson's paradox.

Figure 6 shows that the differences between our annotation procedure. Our study included only genes with BLASTp homology across at least two species, whereas the previous study accepted BLASTn homology, which might result in homologous non-coding sequences being scored as protein-coding genes. When there is a mixture of true protein-coding genes combined with sequences that do not encode a functional polypeptide, the mean of the entire mixture will of course be intermediate between the means of the two groups. The overall mean will depend strongly on the ratio between the two components. Specifically, if the proportion of non-functional open reading frames (ORFs) decreases with conservation level, then a continuum will automatically be observed, and in the wrong direction as a function of apparent gene age. Figure 5 illustrates, in context, the statistical problem known as Simpson’s paradox32, which drives this effect.

To verify that this is responsible for the discrepancy, we repeated our mouse pipeline on yeast (Fig. 6a) and confirmed that following
conclusions and those of a previous study are due to the categoriza-
tion of what is a gene, not to the details of the ISD calculations or of how genes are assigned to phylostrata. In both our mouse and yeast analyses, we were careful to discard all possible non-genes, leaving us looking at a single group and not a mixture in each phylostratum. Mouse has more well-verified young protein-coding genes, allowing for clearer resolution of ISD on shorter timescales.

Note that if a continuum of increasing ISD with age was to take place only on short timescales (a less parsimonious hypothesis than ours, and one that Simpson's paradox would make difficult to confirm), our mouse analysis restricts it to, at most, the last 21–82 million years before the split of mouse and rat, and after the split of mouse and rabbit. In contrast, the continuum of ISD scores reported in ref. (3) is claimed to go all the way back to the split between Saccharomyces cerevisiae and Candida albicans (~300 million years), despite the much shorter generation times of yeast.

Figure 5 shows that the existence of intermediate proto-genes is not necessary to explain data on trends in mean properties. What is more, the very concept of a proto-gene as intermediate between gene and non-gene is problematic, with inappropriately teleological connotations. However, as a non-teleological definition, it may be useful to refer to slightly expressed but non-functional ORFs as proto-genes. Stretches between a start codon and a stop codon (ORFs) occur frequently by chance. ORFs can be used as the raw materials for proto-genes defined in this fashion. Once proteins are born with a given ISD, evolutionary tinkering and differential loss seem to change ISD only slowly, resulting in the consistent trend seen over hundreds of millions of years in Fig. 2. While gene birth is a sudden transition to functionality, subsequent descent with modification can generate extraordinarily slow trends.

Figure 5 | Putative evidence for the continuum hypothesis can be explained as a statistical artefact known as Simpson's paradox. a. The continuum view posits the existence of proto-genes that have characteristics intermediate between non-genic ORFs and genes. Candidate proto-genes were classified on the basis of being annotated as ORFs and having detectable sequence homology in sister species (without necessarily retention of approximate ORF boundaries), and the authors of ref. claimed to show a continuum of properties as a function of conservation level, shown as a greyscale. b. The same data can be explained without resorting to the existence of such intermediates. Sequence homology for ORFs that are not protein-coding genes (white circles) becomes more difficult to detect as a function of age, such that the proportion of true genes (black circles) increases with age, giving rise to the same observations as a. The downward trend in ISD arises as an example of Simpson's paradox. c. By carefully excluding all non-genes, we see the true relationship between gene age and ISD, and compare it with intergenic control sequences that are definitely not protein-coding genes. Note that if true protein-coding genes were excluded in b (rather than excluding non-genes as in c), there would be no relationship with conservation levels.

Methods
Mus musculus proteins from Ensembl (v75) were subjected to BLASTp search with an E-value threshold of 0.001 against the National Center for Biotechnology Information (NCBI) non-redundant protein sequences (nr) database (June 2014). The most phylogenetically distant hit was used to place the gene into one of the 20 phylostrata (gene ages) listed in Supplementary Table 1, following O'dilios parsimony and neglecting horizontal gene transfer. From 22,778 available protein sequences, 126 could not be successfully assigned to any phylostratum due to BLASTp-related problems such as too short queries, majority of query composed of low-complexity sequences, or a combination of both. These sequences were not considered for further analysis.

To remove dubious genes and perform evolutionary rate-controlled ISD estimates, dN/dS values were downloaded for all mouse proteins from the Ensembl Biomart (accessed on 18 February 2016) and mapped to our dataset using the Ensembl protein ID. Evolutionary rates were calculated using PAML by comparing all mouse proteins with their orthologues in rats. Genes with no rat orthologue and leading to non-functional translation provide the raw materials for proto-genes defined in this fashion. However, it must be noted that even harmless ORFs, in the absence of selection for some beneficial property, are rapidly disrupted by mutation. There is therefore a discrete dichotomy between the states of ‘gene’ and ‘non-gene’, determined by whether the selection coefficient is greater than zero, and hence capable of sustaining their continued existence in the face of mutational onslaught.

A dichotomy of functionality (defined in evolutionary terms) is still compatible with the idea that some non-genes are under selection that weeds out the most deleterious of options, in a manner that promotes evolvability.

In terms of the adaptive potential of non-coding sequences, and leading to non-functional translation provide the raw materials for proto-genes defined in this fashion. However, it must be noted that even harmless ORFs, in the absence of selection for some beneficial property, are rapidly disrupted by mutation. There is therefore a discrete dichotomy between the states of ‘gene’ and ‘non-gene’, determined by whether the selection coefficient is greater than zero, and hence capable of sustaining their continued existence in the face of mutational onslaught.

A dichotomy of functionality (defined in evolutionary terms) is still compatible with the idea that some non-genes are under selection that weeds out the most deleterious of options, in a manner that promotes evolvability.
was taken (to prevent any further exclusion of genes with high evolutionary rate, beyond the low bar of detectable mouse–rat homology). Restricting analysis to one-to-one orthologues did not qualitatively change the results.

Pairwise paralogue information among non-orthor-mouse genes was taken from Ensembl, from which gene families were constructed via a single-link cluster analysis. This yielded 8,124 gene families, 7,113 of which showed complete agreement among their member genes regarding age. Of the remainder, 824 gene families contained genes assigned to exactly two different phylostra; 526 of these had only a single gene in the younger phylostratum, which was reassigned back to the older phylostratum. Of the remainder split across exactly two phylostra, 150 had only a single older member, which we reassigned to be younger, leaving 148 gene families unclassified. In addition, 187 gene families contained genes split among more than two phylostra, of which 86 could similarly easily be reconciled by discounting the gene age status of singletons, leaving another 101 gene families unclassified. Most of the 249 gene families with unclassified ages are split between multiple old phylostra; as we had no shortage of data in these older phylostra, and as this group includes complex scenarios such as gene fusion and repetitive sequences, these gene families were excluded from further analysis, leaving 15,347 total genes for our analysis.

Twenty-six gene families, consisting of 29 M. musculus genes, did not originally return NCBI nr BLAST hits outside their own species, yet Ensembl reported dN/dS values relative to a rat orthologue that met our sequence identity filter. Fourteen of those that did not return the rat orthologue nevertheless passed manual inspection of the protein-coding status of the Ensembl-identified orthologue. These were therefore assigned to the Rodentia phylostratum.

Our single-link clustering plus cleanup procedure to construct gene families produced a much better fit to the data than treating genes as being independent, explaining far more variance than phylostratum, the property of interest (ΔAkaike’s information criterion (AIC) = 114 removing phylostratum from the model versus ΔAIC = 9.928 removing the random effect of gene family from the model).

We calculated ISD using IUPred, after first excising all cysteines from the protein sequence (from Ensembl v73) because of uncertainty about their disulfide bond status combined with a profound impact of disulfide bond status on ISD. For each gene, we averaged the ISD across all other amino acids and performed a Box–Cox transformation (Box–Cox exponent λ = 0.66, λ optimized using only coding genes not controls) before linear model analysis. Central tendency estimates and confidence intervals were then back-transformed for the plots.

Protein lengths were approximately log-transformed (Box–Cox λ = −0.0432). Hydrophobicities were calculated first for amino acids: leucine, isoleucine, valine, phenylalanine, methionine and tryptophan were scored as +1, and all other amino acids were scored as −1. Then the mean hydrophobicity for a protein was used to examine the length-dependence of amino acid composition.

For each gene, we generated scrambled controls by resampling amino acids without replacement. To generate GC-matched controls, the numbers of GC and adenine–thymine (AT) nucleotides were calculated excluding the stop codon, then GC versus AT identity was resampled without replacement, and then G versus C and A versus T were assigned at 50% probability. If a premature stop codon arose, one of the three stop codon nucleotides was switched with another nucleotide position chosen at random. This process was iterated until no premature stop codons remained, and then a stop codon was appended to the end.
To generate one intergenic control per gene, we took one intergenic sequence 100 nucleotides downstream from the end of the 3’ end of the Ensembl v80 annotation of the transcript, and progressed further, excising stop codons along the way, until a length match to the neighbouring protein-coding gene was obtained. We then obtained a second control sequence near each gene by repeating the process after starting a search 100 nucleotides farther downstream. For the RepeatMasked controls, intergenic sequences farther downstream were used as necessary to extract the control sequence from a contiguous non-masked intergenic sequence.

S. cerevisiae. Genes taken in June 2014 from the Saccharomyces Genome Database (SGD) were assigned gene ages according to the procedure described above for mouse. We supplemented our phylostratigraphic analysis of species supported by the NCBI taxonomy browser with a selection of more closely related yeast species; our youngest phylostratum contains any other age assignments within their gene family. Genes classified by us as specific were assigned gene ages according to the procedure described above for SGD. ‘Conservation level’ (an alternative phylostratigraphy that includes BLASTs homology detection) was provided (A.-R. Carvunis, personal communication) to reproduce the classification presented in ref. 1. ISD values were calculated as for mouse except with Box–Cox λ = 0.554.

Data availability. Source data for the statistical analyses and figures are provided in Supplementary Tables 2–7. Code associated with generating and analysing these tables is publicly available at https://github.com/Masellpab.

Received 23 September 2016; accepted 16 March 2017; published 24 April 2017

References
1. McLoskey, J. E. & Guerzoni, D. N. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. Phil. Trans. R. Soc. B 370, 20110332 (2015).
2. Monegolli, E. & Chiti, F. Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep. 8, 737–742 (2007).
3. Carvunis, A.-R. et al. Proto-genes and de novo gene birth. Nature 487, 370–374 (2012).
4. Masel, J. Cryptic genetic variation is enriched for potential adaptations. Genetics 172, 1985–1991 (2006).
5. Rajon, E. & Masel, J. The evolution of molecular error rates and the consequences for evolvability. Proc. Natl Acad. Sci. USA 108, 1082–1087 (2011).
6. Wilson, B. A. & Masel, J. Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol. Evol. 3, 1245–1252 (2011).
7. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics 14, 117 (2013).
8. Romero, P. et al. Thousands of proteins likely to have long disordered regions. Pac. Symp. Bio Comput. 1998, 437–448 (1998).
9. Dostáňová, Z., Ciszková, V., Tompa, P. & Simon, I. UPRpred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21, 3433–3434 (2005).
10. Yu, J.-F. et al. Natural protein sequences are more intrinsically disordered than random sequences. Cell. Mol. Life Sci. 73, 2949–2957 (2016).
11. Buljan, M., Frankish, A. & Bateman, A. Quantifying the mechanisms of domain gain in animal proteins. Genome Biol. 11, R74 (2010).
12. Moore, A. D. & Bornberg-Bauer, E. The dynamics and evolutionary potential of domain loss and emergence. Mol. Biol. Evol. 29, 787–796 (2012).
13. Ekman, D. & Elofsson, A. Identifying and quantifying orphan protein sequences in fungi. J. Mol. Biol. 396, 396–405 (2010).
14. Bornberg-Bauer, E. & Alba, M. M. Dynamics and adaptive benefits of modular protein evolution. Curr. Opin. Struct. Biol. 23, 459–466 (2013).
15. Mükherjee, S., Panda, A. & Ghosh, T. C. Elucidating evolutionary features and functional implications of orphan genes in Leishmania major. Infect. Genet. Evol. 32, 330–337 (2015).
16. Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R. & Karlin, D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. J. Virol. 83, 10719–10736 (2009).
17. Domazet-Loso, T., Brajkovic, J. & Tautz, D. A phylostratigraphic approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23, 533–539 (2007).
18. Moyers, B. A. & Zhang, J. Phylostratigraphic bias creates spurious patterns of genome evolution. Mol. Biol. Evol. 32, 258–267 (2015).
19. Moyers, B. A. & Zhang, J. Evaluating phylostratigraphic evidence for widespread de novo gene birth in genome evolution. Mol. Biol. Evol. 33, 1245–1256 (2016).