More benign random peptides in *E. coli* are enriched for similar amino acids as young animal but not plant genes

Luke Kosinski^{*,††,1}, Nathan Aviles^{†,††}, Kevin Gomez[‡], Joanna Masel^{§,2}

* Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721

† Graduate Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ 85721

‡Graduate Interdisciplinary Program in Applied Math, University of Arizona, Tucson, AZ 85721

§ Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721

†† These authors contributed equally to this work.

¹Present address: Critical Path Institute, Tucson, AZ 85718

ORCID IDs: 0000-0002-8146-5955 (L. K.); 0000-0001-9998-4406 (N. A.); 0000-0002-6356-0318

(K. G.); 0000-0002-7398-2127 (J. M.)

14 Running title: Amino acid effects on fitness

15 Keywords: Experimental evolution, evolvability, fitness estimation, preadaptation, de novo

16 gene birth

17 ²Corresponding author: Joanna Masel, Department of Ecology and Evolutionary Biology,

18 University of Arizona, 1041 E Lowell St., Tucson, AZ 85721. E-mail: masel@email.arizona.edu

19 +1(520)626-9888

Abstract

Proteins are the workhorses of the cell, yet they carry great potential for harm via misfolding and aggregation. Despite the dangers, proteins are sometimes born *de novo* from non-coding DNA. Proteins are more likely to be born from non-coding regions that produce peptides that do little to no harm when translated than from regions that produce harmful peptides. To investigate which newborn proteins are most likely to “first, do no harm”, we estimate fitnesses from an experiment that competed *Escherichia coli* lineages that each expressed a unique random peptide. A variety of peptide metrics significantly predict lineage fitness, but almost all this predictive power stems from simple amino acid composition. Amino acids that are smaller and that promote intrinsic structural disorder have more benign fitness effects. Our amino acid composition-based predictions also have validity for an independent dataset using small random N-terminal tags. The same amino acids that predict high fitness in *E. coli* are enriched in young Pfams in animals, but not in plants. To modify Jacques Monod’s famous quote, what makes peptides benign in *E.coli* also makes them benign in elephants, but not in eucalyptus.

Introduction

Proteins are the workhorses of the cell, but they are dangerous. For example, the polypeptide backbone is the key structural feature of amyloids, putting all proteins at risk of forming insoluble aggregates (Chiti and Dobson 2017), and most proteins are expressed at or just beyond their solubility limits (Vecchi, et al. 2020). Despite these dangers, new protein-coding genes are nevertheless born *de novo* from essentially random sequences (McLysaght and Guerzoni 2015; Van Oss and Carvunis 2019; Vakirlis, Carvunis, et al. 2020). To be beneficial enough for *de novo* birth, a random peptide must first do no serious harm, i.e. it must not be detrimental to the basic functioning of a cell. Here we quantify the degree to which, and the summary statistics via which, the fitness effects of a random peptide can be predicted.

Neme et al. (2017) competed over 2 million *Escherichia coli* lineages, each containing a plasmid designed to express a unique random peptide, and tracked lineage frequencies over four days using deep DNA sequencing. This study has been criticized for providing too little support for the beneficial nature of the top candidates (Weisman and Eddy 2017; Knopp and Andersson 2018). But these criticisms do not detract from using the dataset to identify statistical predictors of serious harm versus relatively benign effect. Neme et al. (2017) used a strong promoter, so evaluation is of tolerance of high expression. Some fitness differences might be due to variation in expression e.g. due to auto-downregulation at the RNA level (Knopp and Andersson 2018) - we will return to this point in the Discussion. Here we pursue analyses based on the hypothesis that the properties of the peptides contribute to variation in fitness among lineages.

Conveniently, computational predictors from peptide sequences alone are available for some properties, such as intrinsic structural disorder (ISD) and aggregation propensity. Because insoluble proteins have been implicated in toxicity and disease (Chiti and Dobson 2017) and peptides with high ISD are less prone to forming insoluble aggregates (Linding, et al. 2004; Angyan, et al. 2012), we hypothesize that highly disordered peptides are least likely to be strongly deleterious. Random sequences with high predicted disorder are well-tolerated *in vivo* (Tretyachenko, et al. 2017). Existing mouse (Wilson, et al. 2017) and Drosophila (Heames, et al. 2020) proteins, which are the product of evolution, are predicted from their amino acid sequences to be more disordered than what would be translated from intergenic controls.

Younger protein-coding sequences should be particularly constrained to first do no harm, as they have had little time to evolve more sophisticated harm-avoidance strategies (Foy, et al. 2019). In support of the idea that high ISD is an accessible way to avoid harm, young animal and fungal domains (James, et al. 2020) and genes (James, et al. 2020), and novel overprinted viral genes (Willis and Masel 2018) have higher predicted disorder than their older counterparts. While some studies have found that putative de novo protein candidates in *Saccharomyces* yeasts have lower rather than higher ISD (Carvunis, et al. 2012; Basile, et al. 2017; Vakirlis, et al. 2018), this could be an artifact of the inclusion of non-genes. When Wilson et al. (2017) reanalyzed Carvunis et al.'s (2012) "proto-genes" of different ages, using more rigorous criteria to exclude non-genes from the data; this reversed the direction of the ISD trend. The same reversal of trend following a quality filter was also found by Vakirlis et al. (2018). This phenomenon, where analyzing a mixture of genes and non-genic "proto-genes" in varying proportions that create a correlation in the opposite direction to the true correlation

for genes, is a common statistical artifact known as Simpson’s paradox (Malinas and Bigelow 2016). That non-genic ORFs would be more common in young “proto-genes” is expected, as is the low predicted ISD of non-genic ORFs given yeast GC content (Basile, et al. 2017); these two factors drive Simpson’s paradox.

High predicted ISD in young genes is driven primarily by amino acid frequencies, with amino acid order playing a more minor role (Wilson, et al. 2017). Fortunately, the dataset of Neme et al. (2017) is large enough to look at the frequencies of each amino acid as predictors, rather than assume that existing prediction programs such as IUPred (Dosztányi, et al. 2005; Meszaros, et al. 2018) or Tango (Fernandez-Escamilla, et al. 2004; Linding, et al. 2004; Rousseau, et al. 2006) integrate amino acid composition (and more) in the best possible way. We can then test whether such programs have additional ability to predict peptide fitness, above and beyond the influence of amino acid frequencies.

How each amino acid frequency predicts peptide harm can also help resolve disputes about the still-contentious technique of phylostratigraphy, which assigns ages to genes. Detecting homologs is more difficult for fast-evolving sequences, which may be erroneously scored as young (Alba and Castresana 2007; Moyers and Zhang 2015, 2016). Disordered proteins tend to be fast evolving (Chen, et al. 2011), casting doubt over findings that young genes are more disordered. If the amino acid preferences of higher fitness random peptides match those of purportedly young protein-coding sequences, this would be evidence that the *de novo* gene birth process, rather than homology detection bias alone, causes trends in protein properties as a function of apparent gene age.

Monod and Jacob (1961) famously claimed that “anything found to be true of *E. coli* must also be true of Elephants.” We test whether amino acids that are good for random peptides in *E. coli* are enriched in young animal and plant proteins from a recent phylostratigraphy dataset (James, et al. 2020) of over 400 different eukaryotic species. We also test whether published, sequence-based estimators of disorder, aggregation propensity, water solubility, hydrophobic residue clustering etc. predict fitness information not already predicted by amino acid frequencies.

Methods

Data retrieval

Neme et al. (2017) performed seven experiments where *E. coli* lineages, each with a plasmid containing a unique random peptide, were grown and tracked using deep DNA sequencing. We downloaded sequencing counts from Dryad at <http://dx.doi.org/10.5061/dryad.6f356>, and obtained amino acid and nucleotide sequences directly from Rafik Neme. Experiment 7 was by far the largest, containing all the peptides that the other six experiments classified as “increasing” or “decreasing,” and more. We thus chose to focus on experiment 7 in our work. Experiment 7 consists of the numbers of reads of each random peptide sequence in 5 replicate populations of *E. coli* at 4 time points. We assume that fitness is identical across replicates, so we summed across all 5 replicates to obtain a total number of reads for each polypeptide at each time point.

Following Neme et al. (2017), we took the 1061 peptides out of over one million that had ≥ 5 reads across all 5 replicates of experiment 7. Neme et al. (2017) used this cutoff because achieving statistical significance for individual peptides with fewer than five reads is challenging. In our case, we similarly want fitness estimates with confidence intervals indicative of some reasonable degree of information. The dramatic nature of this data reduction is unsurprising, firstly because each initial unique peptide was present in only one copy, and secondly because most peptides are likely deleterious. We note therefore that our analyzed subset of peptides with at least five reads are certainly non-lethal, and likely less deleterious than the average random peptide. Nonetheless, we achieved enough resolution to distinguish between more and less harmful peptides, with remarkably large effect sizes considering the restricted fitness range.

We further excluded the six peptides that, while meeting the criterion of ≥ 5 reads, had all of those reads at the same timepoint, leaving 1055 peptides for analysis.

Estimating lineage fitness from random peptide sequencing counts

The expected number of reads λ_{it} of peptide i at times $t=1,2,3,4$ was modeled as:

$$\lambda_{it} = N_t p_{i0} \prod_{k=1}^t \frac{\omega_i}{W_{k-1}},$$

where N_t is the observed total number of reads, p_{i0} is the initial frequency of peptide i at the beginning of the experiment (prior to the round of selection used to produce the first measured

timepoint $t = 1$), $\frac{\omega_i}{W_t}$ is the fitness of bacteria with peptide i at time t (i.e. their propensity to contribute to the next time point), and W_k is population mean fitness at time k .

The likelihoods of observed peptide counts were estimated from this expectation and two different error models. A Poisson distribution, which captures sampling error alone, was used to generate our initial estimates of p_{io} , ω_i , and W_k (collectively yielding λ_{it}) because it is analytically tractable. Under a Poisson error function, the likelihood of observing n_{it} reads of peptide i at time t is

$$f_{\text{Pois}}(n_{it} | \lambda_{it}) = \frac{\lambda_{it}^{n_{it}} e^{-\lambda_{it}}}{n_{it}!}.$$

To also capture variance inflation κ due to PCR amplification, we used a negative binomial distribution in the Polya form:

$$f_{\text{NBP}}(n_{it} | \lambda_{i,t}, \kappa) = \left(\frac{\Gamma\left(n_{it} + \frac{\lambda_{i,t}}{\kappa - 1}\right)}{n_{it}! \Gamma\left(\frac{\lambda_{i,t}}{\kappa - 1}\right)} \right) \left(\frac{1}{\kappa}\right)^{\frac{\lambda_{i,t}}{\kappa - 1}} \left(1 - \frac{1}{\kappa}\right)^{n_{it}}$$

where $\Gamma(\cdot)$ is the gamma function. We used the initial estimates of p_{io} , ω_i , and W_k to numerically fit the negative binomial model. For the specifics of fitting the Poisson and negative binomial models, see Supporting Information. Weights were calculated, for use in downstream linear models, from this likelihood inference procedure, as the inverse of Fisher information (see Supporting Information).

An existing software package for estimating lineage fitness from sequencing counts is Fit-Seq (Li, et al. 2018), which captures the amplification of PCR error through a more

sophisticated distribution for the number of reads that is derived in the supplementary information of Levy et al. (2015). However, Fit-Seq assumes that mean fitness is a simple average of all measured lineages' fitness, requiring all individuals to be tagged and measured. But Neme et al.'s (2017) experiment included lineages carrying an empty plasmid, i.e. with the selectable marker but no random peptide. Worse, the proportion of cells with an empty vector can be presumed to increase over time. In the absence of a reliable way to quantify these, we instead consider mean population fitness over time to be a set of independent parameters to be fitted.

Clustering non-independent sequences

Upon visual inspection, we found that some peptide sequences were extremely similar, with only one or two amino acid differences; these data points will not contain independent information about the relationship between sequence and fitness. To account for non-independence, we clustered peptides by their Hamming distance, and either took only the peptide whose fitness had the highest weight within its cluster, or took weighted means within clusters, or included cluster in our regression models as a random effect term. Single-link clustering with Hamming distance cutoffs of 6 to 29 amino acids all produced an identical set of 646 clusters for our 1055 peptides. The largest cluster had 228 random peptides, and the second largest had only 13. The vast majorities of clusters contained only 1 sequence (Dataset S1). A few peptides had mutations in their non-random regions; these mutations were counted in our Hamming distance measurements.

Such similar sequences are highly unlikely to arise by chance if the peptides were truly random; $20^{50} \approx 10^{65}$ peptides are possible, far more than the $\sim 2 \times 10^6$ observed. Because we analyze only peptides with at least 5 reads, replicated sequencing error is an unlikely cause. We see the same nearly-identical sequences appearing in every experimental replicate, suggesting either that mutations occurred during Neme et al.'s (2017) initial growth phase, or that the “random” peptides synthesized for the experiment are not entirely random. We note that construction of the “random” peptide library involved ligations of a smaller set of “seed” sequences, introducing non-randomness at this stage.

Predictors of fitness

All peptides are exactly 65 amino acids long with 50 amino acids of random sequence, so there was no need to control for length.

GC content

Many amino acid sequences mapped to several possible nucleotide sequences, as part of the same problem of mutation or non-random construction discussed above. To calculate one GC content for each random peptide, we calculated a simple average of GC content across all the nucleotide sequences in the dataset that map to the peptide with the largest weight in the cluster.

To calculate GC content for the over two million peptides with at least one sequencing read, we took a simple average of the GC content from the random portion of the peptides.

Disorder

IUPred2 (Dosztányi, et al. 2005; Meszaros, et al. 2018) returns an ISD score between zero and one for each amino acid in a sequence, with higher scores indicating greater intrinsic disorder. To calculate an ISD score for each random peptide, we took the average of the scores for the whole sequence (i.e. including non-random parts). We used a square root transform because it produced a more linear relationship with fitness than no transform. IUPred2 was used for all analyses except Δ ISD, which used the original IUPred program – differences between the two are minimal (Meszaros, et al. 2018).

Aggregation propensity

Tango (Fernandez-Escamilla, et al. 2004; Linding, et al. 2004; Rousseau, et al. 2006) returns an aggregation score for each amino acid in a sequence. At least five sequential amino acids with a score greater than or equal to five indicates an aggregation-prone region. We scored peptide aggregation propensity as the number of amino acids within regions scored as aggregation-prone, including contributions from non-random regions.

Solubility

CamSol (Sormanni, et al. 2015) returns a solubility score for each amino acid in a sequence, as well as a simple average of all scores for a sequence, which CamSol calls a “solubility profile.” We used the solubility profile of the full sequences, including non-random regions.

Hydrophobic clustering

We used a Perl script developed by Scott Foy (“calculate_hydrophobic_dispersion.pl” at <https://github.com/MaselLab/A-shift-in-aggregation-avoidance-strategy-marks-a-long-term->

[direction-to-protein-evolution/tree/master/script](#)) to calculate Irbach's hydrophobicity dispersion metric (Irbach, et al. 1996; Irbach and Sandelin 2000). This metric splits a sequence into blocks of equal length and calculates dispersion as the variance of the hydrophobicity per block by the mean hydrophobicity per block, normalized to account for differences in mean hydrophobicity and length. We used blocks of 6 amino acids, considered I, L, M, F, and V hydrophobic, and scored whole peptide sequences including non-random regions.

Amino acid composition

We scored amino acid composition by counting frequencies among the 50 amino acids in the random portion of each peptide.

The values for all the above predictors for each peptide are listed in Dataset S1.

Statistics

All statistical tests were carried out in R version 3.6.3 (R Core Team 2019), with figures generated using "ggplot2" (Wickham 2016). Weighted linear mixed models were implemented using the "lmer" function from the "lme4" package (Bates, et al. 2015), with cluster as a random effect. See Supporting Information for details, including justification of a log-transform for fitness. When R^2 values were needed, we instead averaged peptides within the same cluster into a combined datapoint, allowing us to avoid the use of random effect term, and took R^2 values from the base R "lm" function. Raw P-values are reported unless otherwise noted, i.e. without correction for multiple comparisons.

234 Data and code availability

235 All code and supplemental tables are available on GitHub at
 236 <https://github.com/MaselLab/RandomPeptides>. The original Neme et al. (2017) data can be
 237 found at Dryad <http://dx.doi.org/10.5061/dryad.6f356>, and the original sequences are available
 238 at the European Nucleotide Archive (ENA) under the project number PRJEB19640.

239 Results

240 Estimating the fitness effects of random peptides

241 Assessing predictors of the fitness effects of random peptides requires those fitness effects to
 242 be measured accurately and precisely. Neme et al. (2017) tracked lineage frequencies over four
 243 days, and categorized a peptide as increasing or decreasing in frequency by comparing the DNA
 244 sequencing counts of day 4 to day 1 using DESeq2 (Love, et al. 2014). Out of the 1082 peptides
 245 with at least five sequencing reads, they found that 277 were increasing, while 436 were
 246 decreasing.

247 We reanalyze the same data, instead using a custom maximum likelihood framework
 248 (see Materials & Methods) to quantitatively estimate “fitness” and its associated confidence
 249 interval / weight. “Fitness” here refers to allele frequency changes over an entire cycle of
 250 population growth and dilution, rather than per generation. Our method classifies peptides
 251 quantitatively rather than qualitatively. It accounts for the fact that mean population fitness
 252 increases over the four days (see Materials and Methods). Our use of all available data within

an appropriate maximum likelihood framework should make our method more sensitive and specific for identifying benign vs harmful peptides. We find statistical support for increases in frequency (i.e. fitness – $1.96 \times$ standard error > 1) for 475 peptides (Dataset S1). We exclude 22 of the previously reported 277, at least some of which arose from a bug we located in Neme et al.'s (2017) pipeline. 509 peptides had statistical support for decreases in frequency, with support lost for 9 of the previously reported 436 deleterious peptides.

Note that some peptides are pseudoreplicates (see Materials & Methods). There were 646 total clusters, of which there was statistical support for increases in frequency for the highest-weighted peptide in 138, and for decreases in 488. Some of our statistics use cluster as a random effect within a linear mixed model. When interpretable R^2 values are needed, we instead collapse each cluster into a single pseudo-datapoint with value given by the weighted mean and weight given by the sum of weights, and then use fixed effect models.

Amino acid counts are the best predictor of random peptide fitness effects

We estimated peptide disorder using IUPred (Dosztányi, et al. 2005; Meszaros, et al. 2018), water solubility using CamSol (Sormanni, et al. 2015), and general aggregation propensity using Tango (Fernandez-Escamilla, et al. 2004; Linding, et al. 2004; Rousseau, et al. 2006). Fewer than 6% of the random peptides have a predicted transmembrane helix (Dataset S1) from TMHMM (Krogh, et al. 2001), so our choice of these predictors is guided by our assumption that the random peptides are predominantly located in the cytosol. Having a predicted transmembrane helix did not in itself predict random peptide fitness effects ($P = 0.2$, likelihood ratio test relative to mixed model with only the intercept as a fixed effect). In contrast, each of our cytosol-

solubility-inspired metrics significantly predicted random peptide fitness (Fig. 1A – 1C), with effects in the predicted direction (more disorder and more solubility are good, more aggregation propensity is bad). Another aggregation predictor, Waltz (Maurer-Stroh, et al. 2010), that specializes in β aggregates, was in the right direction but did not quite meet statistical significance ($P = 0.06$).

Fig. 1. Many metrics, on their own, predict peptide fitness effects, but amino acid composition predicts the most. A) ISD (on a square root scale), B) CamSol, and C) the number of amino acids in Tango-predicted aggregation-prone regions each significantly predict peptide fitness on their own ($P = 7 \times 10^{-4}$, 0.003, and 0.02, respectively, likelihood ratio test in mixed model). D) The best model is fitness predicted by amino acid composition; weighted R^2 values for predicting cluster fitness ($n = 646$) from the ISD, CamSol, Tango, and amino acid composition models are 0.029, 0.031, 0.017, and 0.15, respectively. Each point shows a cluster with area is proportional to summed weights across that cluster. Seven clusters with fitness greater than 2 are not shown here for ease of visualization; a complete y-axis is shown in supplemental fig. 1. Log-transforming fitness would remove high fitness skew, but creates systematic heteroscedasticity, and so was not done (supplemental fig. 2). The lack of systematic

heteroscedasticity can be seen here in the form of similar point size across fitness values. Blue lines come from weighted linear mixed regression models (see Materials and Methods), where a random effect term controls for pseudoreplication within clusters.

While IUPred, CamSol, and Tango are all statistically supported predictors of fitness, all pale in comparison to amino acid composition. Weighted R^2 values for IUPred, CamSol, and Tango are 0.029, 0.031, 0.017, respectively. While the R^2 values are relatively small, the effect sizes are notable: from the 10% to 90% quantiles of ISD, CamSol, and Tango, predicted fitness changes from 0.51 to 0.69, 0.53 to 0.66, and 0.65 to 0.54, respectively. However, when the frequencies of each of the twenty amino acids were also used as predictors (Fig. 1D), IUPred, CamSol, and Tango were no longer significant predictors of fitness ($P = 0.2$, 0.2 , and 0.3 , respectively, likelihood ratio test in mixed model, see Supplemental Table S1), suggesting that their predictive power in our dataset came merely from being metrics of amino acid composition. Amino acid composition also explains a much greater amount of the variation in fitness than any of the other predictors, with a weighted adjusted R^2 of 0.15. The effect size also increases, with the 10% and 90% quantiles of model-predicted fitness being 0.45 and 0.81, respectively. We are surprised that such a high proportion of fitness can be predicted merely by a linear function of the amino acid content of a single highly expressed peptide – both protein folding and organismal function are generally assumed not to be so easily reduced.

We next tried two predictors that focus exclusively on the order rather than the frequency of amino acids in a peptide. First, the contribution of amino acid order to ISD (calculated as $\Delta\text{ISD} = \text{original sequence ISD} - \text{scrambled sequence ISD}$), is higher in younger

mouse genes (Wilson, et al. 2017). However, this was not the case for the very youngest mouse genes, and including Δ ISD did not significantly improve our model ($P = 0.2$).

Second, previous work found that young proteins have their hydrophobic amino acids more clustered together along the primary sequence than expected by chance, while old proteins have them more interspersed (Foy, et al. 2019; James, et al. 2020). This hydrophobic dispersion metric (Irback, et al. 1996; Irback and Sandelin 2000), which is normalized to exclude any influence from amino acid composition, did not significantly predict the fitness of our random peptides on its own ($P = 0.2$, likelihood ratio test in mixed model), nor did including it significantly improve our model with amino acid composition ($P = 0.2$, Supplemental Table S1).

Young animal sequences are enriched for amino acids that increase fitness in random peptides

For ease of interpretation, we calculated the marginal effect of each amino acid on fitness (see supplementary text and Table S2, displayed in fig. 2, y-axis), instead of simply using its slope in a regression model. Fitness effects are correlated with both amino acid volume (Tsai, et al. 1999) and with a metric of how often an amino acid tends to be found in known disordered versus ordered proteins (Theillet, et al. 2013) (fig. 2A and 2B; $P = 0.01$ for both disorder propensity and volume, likelihood ratio test for dropping either term from weighted regression containing both). Volume and disorder propensity together explain over half the weighted variation in marginal fitness effect (weighted adjusted $R^2 = 0.52$). Other properties of amino acids, such as stickiness (Levy, et al. 2012), relative solvent accessibility (Tien, et al. 2013), amino acid cost in *E. coli* (Akashi and Gojobori 2002), and isoelectric point (Liu, et al. 2004) did not provide

significant explanatory power on top of disorder propensity and volume (all $P > 0.1$, likelihood ratio test).

Fig. 2. Amino acids that are small and are associated with ISD promote higher fitness. The y-axis shows each amino acid's marginal effect on fitness, which is the change in fitness when one amino acid of the focal type replaces one randomly chosen amino acid of a different type in a random peptide (see Supporting Information). Error bars are $\pm$ one standard error. P-values and correlation coefficients come from weighted Pearson's correlations, where weights for marginal effects are calculated as $1 / \text{s.e.}(\text{marginal fitness effect})^2$, and volume and disorder propensity are unweighted. Amino acid fitness effects are correlated with (A) amino acid volume and with (B) disorder propensity.

Tryptophan is an outlier for amino acid effects on fitness, predicting minor positive effects on fitness despite both its large volume and its underrepresentation in disordered regions. Removing tryptophan from a weighted regression model of volume and disorder

propensity predicting marginal effect changes the weighted adjusted R^2 from 0.52 to 0.68. Tryptophan, encoded only by UGG, is nearly 60% more common among peptides with at least 5 sequences reads than we expect from the 58% GC content of our dataset. With the confidence interval for its marginal fitness effect including 1, this provides further evidence that tryptophan is advantageous, making it a distinct outlier, for reasons that are not clear to us.

Isoleucine also stands out, as even more harmful than expected by its size and order propensity. Isoleucine's harmful effects may be exacerbated by its role in amyloid formation. For example, familial amyloid cardiomyopathy is most commonly caused by a valine to isoleucine mutation (Jacobson, et al. 1997; Dubrey, et al. 2015), suggesting isoleucine has potential to form dangerous amyloids where other hydrophobic amino acids do not. Isoleucine, valine, and leucine are all hydrophobic amino acids with a branched carbon, but only raised isoleucine levels are associated with a higher risk of Alzheimer's disease (Larsson and Markus 2017), further suggesting that isoleucine may be especially prone to amyloid formation.

Given the hypothesis that young genes should first do no harm, we next tested whether the amino acids that are relatively benign in our laboratory data are similar to the amino acids enriched in purportedly young protein-coding sequences in natural populations. We took the latter from James et al. (2020), who quantified how amino acid frequencies depend on the age of annotated protein domains present in at least two species, across over 400 eukaryotic species. As predicted by the de novo gene birth interpretation of phylostratigraphy trends, amino acids that are good for random peptides are enriched among the youngest animal Pfams (fig. 3A).

Fig. 3. Purportedly young animal Pfams are enriched for amino acids that predict high fitness in random peptides. The y-axis represents how the frequency of each amino acid depends on the age of the sequence in billion years (BY), estimated as a linear regression slope for non-transmembrane Pfam domains (James, et al. 2020). Frequency is in number of percentage points, e.g. a difference in glutamic acid content of 5% vs. 6% is a difference of one percentage point. The x-axis shows each amino acid's marginal effect on fitness, which is the change in fitness when one amino acid of the focal type replaces one randomly chosen amino acid of a different type in a random peptide (see Supporting Information). Error bars are +/- one standard error. Fitness effects predict A) animal, but not B) plant, or C) ancient (older than 2.1 billion years) Pfam phylostratigraphy slopes. Correlation coefficients and P-values come from weighted Pearson correlations. Note that the P-value for animal phylostratigraphy slopes vs marginal effects survives a conservative Bonferroni correction ($P = 0.002 < 0.05/3 = 0.017$).

This prediction was not, however, supported for trends among recent plant domains (fig. 3B) nor among ancient (fig. 3C) domains older than 2.1 billion years. Plant and ancient

trends reflect a *de novo* gene birth process that enriches for the most abundant amino acids in their respective lineages, such as cysteine, rather than for amino acids that promote ISD (James, et al. 2020). We find that ISD still predicts harmlessness in *E. coli*, even though we do not find evidence it shaped *de novo* gene birth in its distant ancestors. We also note that ISD does shape recent *de novo* gene birth in viruses (Willis and Masel 2018).

The correlation in fig. 3A is nevertheless an important external validation that patterns observed using phylostratigraphy can be indicative of the gene birth process, at least in animals. As Monod suggested, what makes a peptide benign in *E. coli* does also make a peptide benign in elephants – with the caveat that this might not be the case in eucalyptus.

Determining which sequences are young is difficult, and may be subject to systematic error (homology detection bias) (Alba and Castresana 2007; Moyers and Zhang 2015, 2016). Our ability to externally validate animal phylostratigraphy slopes against an independent dataset generated in *E. coli* is a powerful demonstration that these slopes represent more than mere bias. That is, if phylostratigraphy trends were due to an artifact such as homology detection bias, such an artifact would be highly unlikely to bias our random peptide analysis in the same direction. This conclusion is also supported by the fact that these slopes differ among taxa and are not correlated with amino acid changeability (James, et al. 2020).

Fitness is better predicted by amino acid composition than by GC content

Long et al. (2018) proposed that selection acts directly on GC content, perhaps due to the three hydrogen bonds of G-C pairs. Amino acids encoded by Gs and Cs tend to promote higher ISD (Angyan, et al. 2012), making it difficult to distinguish between selection for high GC content

and selection for disorder-promoting amino acids. To do this, we compare amino acids that always have G or C to those that always have A or T, at both the first and second nucleotide positions in the codon. If selection were for GC nucleotides, we would expect GC to predict high marginal amino acid fitness effects at both positions. But if results are dramatically different at the two positions, this would show that it is selection on amino acid content that drives GC as a correlated trait. Results are statistically significant in the predicted direction at the second position (fig. 4A, $P = 0.001$, weighted Welch's t-test), and in the predicted direction but not statistically significant at the first (fig. 4B, $P = 0.2$). The effect size of GC content on fitness could not be statistically distinguished between the first and second position (fig. 4C), with wide and hence inconclusive error bars.

However, amino acid composition is a far stronger predictor of fitness than GC content. While GC content is a predictor, offering a significant improvement over an intercept-only model ($P = 6 \times 10^{-11}$, likelihood ratio test for nested fixed-effect models, weighted adjusted $R^2 = 0.06$), it predicts considerably less than amino acid composition ($P = 10^{-18}$, weighted adjusted $R^2 = 0.15$). Adding GC content to the amino acid frequencies-only model offers a modest improvement ($P = 0.004$, weighted adjusted R^2 values improves from 0.15 to 0.16), while adding amino acid composition to a GC content only model offers a major improvement ($P = 10^{-11}$, weighted adjusted R^2 improves from 0.06 to 0.16). These findings suggest that while there may be some direct selection on GC content, the strength of this selection pales in comparison to selection on amino acid composition.

Fig. 4. Amino acids that are constrained to use Gs and Cs tend to have higher marginal effects

on fitness than those constrained to use As and Ts. The difference is significant for constraints

at the second nucleotide position of a codon (A) ($P = 0.001$, weighted Welch's t-test), but not at

the first (B) ($P = 0.2$). Point area is proportional to weight, which is calculated as $1 /$

$\text{s.e.}(\text{marginal fitness effect})^2$, as described in Supporting Information. The y-axis is the same as

the fig. 2 y-axis and fig. 3 x-axis. C) The mean advantage of amino acids constrained to use GC

rather than constrained to use AT is not distinguishable in size between the first and second

codon positions. Y-axis gives the difference in the two weighted means of marginal fitness

effects from A) and B). Error bars represent 95% confidence intervals on the difference

between the means (calculated as $\text{difference} \pm t_{\text{crit}} \times \text{se}$), where $t_{\text{crit}} \approx 2.1$ is the critical value of

the t-statistic with the appropriate degrees of freedom. Weighted Welch's t-test statistic and

the corresponding standard error of the difference in means were calculated using the

"wtd.t.test" function from the "weights" R package, version 1.0.1.

Our fitness predictor is validated on random N-terminal tags

In order to validate our fitness predictor against an external dataset, we tested the ability of our amino-acid-only model to predict the fitness effects of shorter random peptides present as N-terminal tags. Goodman et al. (2013) tagged the N-prime end of green fluorescent protein (GFP) with 137 different short random sequences (11 amino acids long), allowing random peptide expression level to be measured via fluorescence. Frumkin et al. (2017) measured the fitness effects of these random peptide-tagged GFPs in *E. coli* using FitSeq (Li, et al. 2018). For 89 of them, Frumkin et al. (2017) were able to calculate a “fitness residual” based on the deviation from the fitness expected from the level of GFP expression. Note that this fitness residual combines the cost of inefficient expression (as per the intention of these papers) with the fitness effect of the peptide itself. Frumkin et al. (2017) interpreted the low fitness residuals of hydrophobic and expensive-to-synthesize amino acids as caused by low expression efficiency. However, these findings are consistent with our own estimates of direct peptide effects, as hydrophobic amino acids tend to be order-prone (Linding, et al. 2004; Angyan, et al. 2012), and amino acid size is highly correlated with synthesis cost in *E. coli* (Pearson’s correlation coefficient = 0.85, $P = 2 \times 10^{-6}$, cost for amino acid synthesis in *E. coli* taken from (Akashi and Gojobori 2002)). Indeed, our fitness predictions were significantly correlated with their fitness residuals (fig. 5). The consistency between our results and the findings of Frumkin et al. (2017), who used a different dataset and methodology, provide an external validation of the accuracy of our fitness predictor, even in cases where a random segment makes up only a small portion of a total protein.

Fig. 5. Fitness predictions trained on the random peptides of Neme et al. (2017) also work for short random tags attached to the N-terminus of GFP. Predicted fitness comes from our amino acid composition-only mixed model. "Fitness residuals" of N-terminal tags are from Frumkin et al. (2017), and represent the difference between the fitness of the construct and the expected fitness from expression level. $n = 89$.

Discussion

We found that amino acid composition is the primary predictor of the fitness effects of random peptides expressed in *E. coli*. Amino acids that are small and promote ISD predict high fitness. Amino acids that are good for random peptides in *E. coli* are enriched in young proteins in animals, suggesting that harm is avoided in similar ways across diverse lineages. What is truly astonishing is that simply knowing how many of which amino acids are present in these random

peptides can account for 15% of the variance in fitness among lineages. This indicates both the success of our statistical method for minimizing the noise in our fitness estimates, and that mere amino acid frequencies without amino acid order can be highly informative of peptide properties.

Most studies of random peptides have focused on finding peptides that have specific binding or function (e.g. Kaiser, et al. 1987; Keefe and Szostak 2001; Frulloni, et al. 2009). Some were motivated as proof-of-concept that random peptides can exhibit properties of native proteins, such as folding (Davidson and Sauer 1994; Chiarabelli, et al. 2006; LaBean, et al. 2011) and being soluble (Priyambada, et al. 1996). Others focus on how to increase the percentage of native-like random peptides, e.g. by showing that more hydrophilic random peptide libraries have a higher percentage of stable and soluble peptides (Davidson, et al. 1995). Our work has a different intent, identifying properties that make a peptide less likely to be harmful. Neme et al.'s (2017) experiment was suitable for this purpose because it used a large library of peptides with diverse properties, and selected under relatively permissive conditions, simply for relative growth rates in the presence of peptide expression. In contrast, a study design such as that of Knopp et al. (2019), who selected random peptides that rescue viability in the presence of antibiotics, is less suitable for our purposes because so few peptides, including harm-avoiding peptides, are viable. Neme et al.'s (Neme, et al. 2017) study was also convenient because all peptides were the same length – 65 amino acids with 50 amino acids of random sequence – allowing us to neglect length in our analysis.

Having a higher proportion of random peptides do no harm is expected to increase the success rate of future screens for peptide with specific properties. Nucleotide sequences with

high %GC content tend to encode peptides with more benign fitness effects, suggesting that higher %GC should be used in future random peptide libraries. However, very high GC content will yield low complexity sequences, which our predictor has not been trained on. The marginal fitness effects of each amino acid might be different in this very different context.

While the library used by Neme et al. (2017) was designed to have equal frequencies of each nucleotide in the random region, and thus 50% GC content, the over two million random peptides that had at least one sequencing read had a GC content of ~59% in their random portion. The mean GC content of the peptide clusters we analyzed (see Materials and Methods) was similar, at ~58%, with higher fitness peptides within this group having higher %GC, as discussed in the Results. This might be because many lower GC content sequences were so harmful that lineages that carried them went extinct prior to detection via sequencing. Note that it might also reflect a bias toward GC in sequencing methods (Benjamini and Speed 2012; Choudhari and Grigoriev 2017) – a bias that affects all time points equally and so should not affect our fitness estimates.

Long et al. (2018) proposed that there is direct selection for high GC content, as evidenced in part by a preference for amino acids with G or C at the second position of codons, in excess of that predicted from mutation accumulation experiments. Our findings cannot exclude this hypothesis, but show stronger selection on amino acid composition, selection that is capable of driving increased GC content in coding regions as a correlated trait. In intergenic regions, elevated %GC is likely driven mostly by GC-biased gene conversion. However, if driven in part by selection, it is possibly not just on GC content, but also on peptides from non-coding regions translated by error (Rajon and Masel 2011; Wilson and Masel 2011). Selection on

translation errors is for example strong enough to shape non-coding sequences beyond stop codons in *Saccharomyces cerevisiae* (Kosinski and Masek 2020).

Fitness effects in Neme et al. might not be directly caused by peptide properties alone but instead by the effect of both nucleotide and peptide properties on expression (Knopp and Andersson 2018), with lower expression being less harmful. For example, auto-downregulation at the mRNA level can cause significant difference in expression among peptides, despite identical promoters. However, the properties we find to be predictive, such as ISD and amino acid size, are not *a priori* related to auto-downregulation of mRNA, making the latter an unlikely explanation for our findings.

The fact that IUPred, CamSol, and Tango have predictive power only when amino acid composition is not controlled for suggests that these predictors serve here primarily as summary statistics of amino acid frequencies. Metrics like hydrophobic clustering, which are independent of amino acid frequencies by construction, do not add predictive power.

While driven by amino acid frequencies, our findings still strongly support the hypothesis that peptides with low structural disorder tend to be harmful. Disorder-promoting amino acids may help a peptide remain soluble even if unfolded. Small amino acids also tend to be benign, perhaps because they are hydrophobic enough to promote some amount of folding but flexible enough to avoid too much hydrophobic residue exposure.

Our findings suggest that the easiest way to avoid harm is through disorder and small size, but do not rule out other, more order-prone strategies. Indeed, BCS4, a *de novo* evolved protein in *Saccharomyces cerevisiae*, has a hydrophobic core and is capable of folding (Bungard,

et al. 2017). Vakirlis et al. (2020) found that de novo proteins can emerge as transmembrane proteins, which need to be lipid soluble, presumably requiring different harm-avoidance strategies than peptides that are located in the cytosol.

Plants have different trends in amino acid frequencies as a function of sequence age than animals do, with young genes seeming to prefer readily available amino acids, rather than amino acids that promote ISD (James, et al. 2020). This finding could be because: 1) plants are less susceptible to harm from random peptides, 2) other properties, such as amino acid availability, drive the emergence of *de novo* genes in plants, or 3) the plant data lack the resolution needed to identify the trend. We do not have the ability to differentiate between these three possibilities here.

Nevertheless, our finding of consistency between what is benign in *E. coli* and what is benign in animals suggests a deep concordance in what makes a peptide harmful between two apparently disparate branches of life. The forces that drive protein birth therefore appear to share a key similarity between bacteria and Animalia. To modify Monod's famous quote, what is harmful in *E. coli* is also harmful in elephants, but not necessarily in eucalyptus.

Acknowledgements

This work was supported by the John Templeton Foundation (39667, 60814) and the National Institutes of Health (GM-104040, T32GM-008659, T32GM-084905). We thank Rafik Neme and Diethard Tautz for sharing their data with us and for graciously answering all our questions regarding their analyses, Dvir Schirman and Tzachi Pilpel for sharing their data with us, Joe

Watkins for helpful discussions about our likelihood estimation procedure, and Catherine Weibel for driving the GC content analysis forward.

Works cited

- Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* 99:3695-3700.
- Alba MM, Castresana J. 2007. On homology searches by protein Blast and the characterization of the age of genes. *BMC Evolutionary Biology* 7:53.
- Angyan AF, Perczel A, Gaspari Z. 2012. Estimating intrinsic structural preferences of *de novo* emerging random-sequence proteins: Is aggregation the main bottleneck? *FEBS Letters* 586:2468-2472.
- Basile W, Sachenkova O, Light S, Elofsson A. 2017. High GC content causes orphan proteins to be intrinsically disordered. *PLoS computational biology* 13:e1005375.
- Bates D, Maechler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1-48.
- Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research* 40:e72.
- Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, Masel J, Wysocki VH, Cordes MHJ. 2017. Foldability of a natural *de novo* evolved protein. *Structure* 25:1687-1696.

577 Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charlotiaux B,
578 Hidalgo CA, Barbette J, Santhanam B. 2012. Proto-genes and de novo gene birth. *Nature*
579 487:370-374.

580 Chen SCC, Chuang TJ, Li WH. 2011. The Relationships Among MicroRNA Regulation, Intrinsically
581 Disordered Regions, and Other Indicators of Protein Evolutionary Rate. *Molecular*
582 *Biology and Evolution* 28:2513-2520.

583 Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, Polticelli F, Ottone T, Papa E,
584 Luisi PL. 2006. Investigation of de novo totally random biosequences Part II: On the
585 folding frequency in a totally random library of de novo proteins obtained by phage
586 display. *Chemistry & Biodiversity* 3:840-859.

587 Chiti F, Dobson CM. 2017. Protein misfolding, amyloid formation, and human disease: A
588 summary of progress over the last decade. *Annual Review of Biochemistry* 86:27-68.

589 Choudhari S, Grigoriev A. 2017. Phylogenetic heatmaps highlight composition biases in
590 sequenced reads. *Microorganisms* 5:4.

591 Davidson AR, Lumb KJ, Sauer RT. 1995. Cooperatively folded proteins in random sequence
592 libraries. *Nature Structural Biology* 2:856-864.

593 Davidson AR, Sauer RT. 1994. Folded proteins occur frequently in libraries of random amino-
594 acid sequences. *Proceedings of the National Academy of Sciences of the United States*
595 *of America* 91:2146-2150.

596 Dosztányi Z, Csizmok V, Tompa P, Simon I. 2005. The pairwise energy content estimated from
597 amino acid composition discriminates between folded and intrinsically unstructured
598 proteins. *Journal of Molecular Biology* 347:827-839.

599 Dubrey S, Ackermann E, Gillmore J. 2015. The transthyretin amyloidoses: advances in therapy.
600 Postgraduate Medical Journal 91:439-448.

601 Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-
602 dependent and mutational effects on the aggregation of peptides and proteins. Nature
603 biotechnology 22:1302-1306.

604 Foy SG, Wilson BA, Bertram J, Cordes MHJ, Masel J. 2019. A shift in aggregation avoidance
605 strategy marks a long-term direction to protein evolution. Genetics 211:1345-1355.

606 Frulloni L, Lunardi C, Simone R, Dolcino M, Scattolini C, Falconi M, Benini L, Vantini I, Corrocher
607 R, Puccetti A. 2009. Identification of a Novel Antibody Associated with Autoimmune
608 Pancreatitis. New England Journal of Medicine 361:2135-2142.

609 Frumkin I, Schirman D, Rotman A, Li F, Zahavi L, Mordret E, Asraf O, Wu S, Levy SF, Pilpel Y.
610 2017. Gene Architectures that Minimize Cost of Gene Expression. Molecular Cell 65:142-
611 153.

612 Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in
613 bacterial genes. Science 342:475-479.

614 Heames B, Schmitz J, Bornberg-Bauer E. 2020. A continuum of evolving de novo genes drives
615 protein-coding novelty in Drosophila. Journal of Molecular Evolution 38:382-398.

616 Irback A, Peterson C, Potthast F. 1996. Evidence for nonrandom hydrophobicity structures in
617 protein chains. Proceedings of the National Academy of Sciences of the United States of
618 America 93:9533-9538.

619 Irback A, Sandelin E. 2000. On hydrophobicity correlations in protein chains. Biophysical Journal
620 79:2252-2258.

621 Jacobson DR, Pastore RD, Yaghoubian R, Kane I, Gallo G, Buck FS, Buxbaum JN. 1997. Variant-
622 sequence transthyretin (isoleucine 122) in late-onset cardiac amyloidosis in black
623 Americans. *New England Journal of Medicine* 336:466-473.

624 James JE, Willis SM, Nelson PG, Weibel C, Kosinski LJ, Masel J. 2020. Universal and taxon-
625 specific trends in protein sequences as a function of age.
626 *bioRxiv:2020.2003.2026.010728*.

627 Kaiser CA, Preuss D, Grisafi P, Botstein D. 1987. Many random sequences functionally replace
628 the secretion signal sequence of yeast invertase. *Science* 235:312-317.

629 Keefe AD, Szostak JW. 2001. Functional proteins from a random-sequence library. *Nature*
630 410:715-718.

631 Knopp M, Andersson DI. 2018. No beneficial fitness effects of random peptides. *Nature Ecology*
632 & *Evolution* 2:1046-1047.

633 Knopp M, Gudmundsdottir JS, Nilsson T, Konig F, Warsi O, Rajer F, Adelroth P, Andersson DI.
634 2019. De novo emergence of peptides that confer antibiotic resistance. *Mbio*
635 10:e00837-00819.

636 Kosinski LJ, Masel J. 2020. Readthrough errors purge cryptic sequences, facilitating the birth of
637 coding sequence. *Molecular Biology and Evolution* 37:1761–1774.

638 Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein
639 topology with a hidden Markov model: Application to complete genomes. *Journal of*
640 *Molecular Biology* 305:567-580.

641 LaBean TH, Butt TR, Kauffman SA, Schultes EA. 2011. Protein folding absent selection. *Genes*
642 2:608-626.

643 Larsson SC, Markus HS. 2017. Branched-chain amino acids and Alzheimer's disease: a
644 Mendelian randomization analysis. *Scientific Reports* 7:13604.

645 Levy ED, De S, Teichmann SA. 2012. Cellular crowding imposes global constraints on the
646 chemistry and evolution of proteomes. *Proceedings of the National Academy of*
647 *Sciences of the United States of America* 109:20461-20466.

648 Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. 2015. Quantitative
649 evolutionary dynamics using high-resolution lineage tracking. *Nature* 519:181-186.

650 Li F, Salit ML, Levy SF. 2018. Unbiased fitness estimation of pooled barcode or amplicon
651 sequencing studies. *Cell Systems* 7:521-525.

652 Linding R, Schymkowitz J, Rousseau F, Diella F, Serrano L. 2004. A comparative study of the
653 relationship between protein structure and β -aggregation in globular and intrinsically
654 disordered proteins. *Journal of Molecular Biology* 342:345-353.

655 Liu HX, Zhang RS, Yao XJ, Liu MC, Hu ZD, Fan BT. 2004. Prediction of the isoelectric point of an
656 amino acid based on GA-PLS and SVMs. *Journal of Chemical Information and Computer*
657 *Sciences* 44:161-167.

658 Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss
659 C, Stone C, et al. 2018. Evolutionary determinants of genome-wide nucleotide
660 composition. *Nature Ecology & Evolution* 2:237-240.

661 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for
662 RNA-seq data with DESeq2. *Genome Biology* 15:550.

663 Simpson's Paradox [Internet]. 2016 November 14, 2020]. Available from:
664 <https://plato.stanford.edu/archives/fall2016/entries/paradox-simpson/>

665 Maurer-Stroh S, Debulpaep M, Kuemmerer N, de la Paz ML, Martins IC, Reumers J, Morris KL,
666 Copland A, Serpell L, Serrano L, et al. 2010. Exploring the sequence determinants of
667 amyloid structure using position-specific scoring matrices. *Nature Methods* 7:237-242.

668 McLysaght A, Guerzoni D. 2015. New genes from non-coding sequence: the role of de novo
669 protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions*
670 *of the Royal Society B-Biological Sciences* 370:20140332.

671 Meszaros B, Erdos G, Dosztanyi Z. 2018. IUPred2A: context-dependent prediction of protein
672 disorder as a function of redox state and protein binding. *Nucleic Acids Research*
673 46:W329-W337.

674 Monod J, Jacob F. 1961. General conclusions - teleonomic mechanisms in cellular metabolism,
675 growth, and differentiation. *Cold Spring Harbor Symposia on Quantitative Biology*
676 26:389-401.

677 Moyers BA, Zhang JZ. 2016. Evaluating Phylostratigraphic Evidence for Widespread De Novo
678 Gene Birth in Genome Evolution. *Molecular Biology and Evolution* 33:1245-1256.

679 Moyers BA, Zhang JZ. 2015. Phylostratigraphic Bias Creates Spurious Patterns of Genome
680 Evolution. *Molecular Biology and Evolution* 32:258-267.

681 Neme R, Amador C, Yildirim B, McConnell E, Tautz D. 2017. Random sequences are an abundant
682 source of bioactive RNAs or peptides. *Nature Ecology & Evolution* 1:0127.

683 Prijambada ID, Yomo T, Tanaka F, Kawama T, Yamamoto K, Hasegawa A, Shima Y, Negoro S,
684 Urabe I. 1996. Solubility of artificial proteins with random sequences. *FEBS Letters*
685 382:21-25.

686 R Core Team. 2019. R: A language and environment for statistical computing: R Foundation for
687 Statistical Computing.

688 Rajon E, Masel J. 2011. Evolution of molecular error rates and the consequences for
689 evolvability. *Proceedings of the National Academy of Sciences of the United States of*
690 *America* 108:1082-1087.

691 Rousseau F, Schymkowitz J, Serrano L. 2006. Protein aggregation and amyloidosis: confusion of
692 the kinds? *Current Opinion in Structural Biology* 16:118-126.

693 Sormanni P, Aprile FA, Vendruscolo M. 2015. The CamSol Method of Rational Design of Protein
694 Mutants with Enhanced Solubility. *Journal of Molecular Biology* 427:478-490.

695 Theillet F-X, Kalmar L, Tompa P, Han K-H, Selenko P, Dunker AK, Daughdrill GW, Uversky VN.
696 2013. The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of
697 proline residues in intrinsically disordered proteins. *Intrinsically Disordered Proteins*
698 1:e24360.

699 Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. 2013. Maximum allowed solvent
700 accessibilities of residues in proteins. *PLoS One* 8:e80635.

701 Tretyachenko V, Vymetal J, Bednarova L, Kopecky V, Hofbauerova K, Jindrova H, Hubalek M,
702 Soucek R, Konvalinka J, Vondrasek J, et al. 2017. Random protein sequences can form
703 defined secondary structures and are well-tolerated in vivo. *Scientific Reports* 7:15449.

704 Tsai J, Taylor R, Chothia C, Gerstein M. 1999. The packing density in proteins: Standard radii and
705 volumes. *Journal of Molecular Biology* 290:253-266.

706 Vakirlis N, Acar O, Hsu B, Coelho NC, Van Oss SB, Wacholder A, Medetgul-Ernar K, Bowman RW,
707 Hines CP, Iannotta J, et al. 2020. De novo emergence of adaptive membrane proteins
708 from thymine-rich genomic sequences. *Nature Communications* 11:781.

709 Vakirlis N, Carvunis A-R, McLysaght A. 2020. Synteny-based analyses indicate that sequence
710 divergence is not the main source of orphan genes. *eLife* 9:e53500.

711 Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2018.
712 A Molecular Portrait of De Novo Genes in Yeasts. *Molecular Biology and Evolution*
713 35:631–645.

714 Van Oss SB, Carvunis AR. 2019. *De novo* gene birth. *PLoS Genetics* 15:e1008160.

715 Vecchi G, Sormanni P, Mannini B, Vandelli A, Tartaglia GG, Dobson CM, Hartl FU, Vendruscolo
716 M. 2020. Proteome-wide observation of the phenomenon of life on the edge of
717 solubility. *Proceedings of the National Academy of Sciences of the United States of*
718 *America* 117:1015-1020.

719 Weisman CM, Eddy SR. 2017. Gene Evolution: Getting Something from Nothing. *Current Biology*
720 27:R661-R663.

721 Wickham H. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.

722 Willis S, Masel J. 2018. Gene birth contributes to structural disorder encoded by overlapping
723 genes. *Genetics* 210:303-313.

724 Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by
725 the preadaptation hypothesis of *de novo* gene birth. *Nature Ecology & Evolution* 1:0146.

726 Wilson BA, Masel J. 2011. Putatively noncoding transcripts show extensive association with
727 ribosomes. *Genome Biology and Evolution* 3:1245-1252.