Classification: Biological Sciences, Evolution

Amino acids that are well tolerated in random peptides in *E. coli* are enriched in young animal but not young plant genes

Luke Kosinski[1], Nathan Aviles[2], Kevin Gomez[3], Joanna Masel[4]

[1] Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721

[2] Graduate Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ 85721

[3] Graduate Interdisciplinary Program in Applied Math, University of Arizona, Tucson, AZ 85721

[4] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721

Corresponding author: Joanna Masel, masel@email.arizona.edu +1 520 626 9888

Abstract

Proteins' great potential for harm, via misfolding and aggregation, does not prevent them from being born *de novo* from non-coding DNA. To investigate how newborn proteins might "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 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.

Keywords: Experimental evolution, evolvability, fitness estimation, preadaptation, de novo gene birth

Significance statement

Recent work has shown that protein-coding genes are sometimes born de novo from junk DNA. This is surprising, because random polypeptides are expected to be harmful. When random polypeptides are expressed in Escherichia coli, we can predict 15% of the variation in their fitness effects from their amino acid content. Which amino acids are favored matches that of young genes in animals (but not plants). This suggests a deep concordance across the tree of life in the forces that govern gene birth. To modify a famous Monod quote, what makes a peptide do no harm in *E. coli* also holds in elephants, but not in eucalyptus.

Introduction

Proteins are the workhorses of the cell, but they are dangerous. The polypeptide backbone is the key structural feature of amyloids, putting all proteins at risk of forming insoluble aggregates (1). The risks of aggregation are exacerbated by the fact that most proteins are expressed at or just beyond their solubility limits (2). Despite these dangers, new protein-coding genes are nevertheless born *de novo* from essentially random sequences (3-5). To be beneficial enough for *de novo* birth, a peptide must first do no harm. Here we ask what properties are predictive of whether a peptide is harmful.

Neme et al. (6) 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 including too little evidence in support of the beneficial nature of the top candidates (7, 8). But these criticisms do not detract from using the dataset to identify statistical predictors of harm vs. relatively benign effect.

Conveniently, some properties, such as intrinsic structural disorder and aggregation propensity, can be predicted computationally from peptide sequences alone. Because insoluble proteins have been implicated in toxicity and disease (1), we hypothesize that peptides with high intrinsic structural disorder (9, 10) are least likely to be deleterious. Random sequences with high predicted disorder are well-tolerated *in vivo* (11). Existing mouse (12) and Drosophila (13) 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 (14). In support of the idea that high intrinsic structural disorder is an accessible way to avoid harm, young animal and fungal domains (15) and genes (15), and novel overprinted viral genes (16) have higher predicted disorder than their older counterparts.

High predicted intrinsic structural disorder in young genes is driven primarily by amino acid composition, with amino acid order playing a more minor role (12). Fortunately, the dataset of Neme et al. (6) is large enough to look at the frequencies of each amino acid as predictors, rather than assume that existing prediction programs such as IUPred (17, 18) or Tango (9, 19, 20) 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 content.

How each amino acid frequency predicts peptide harm may also be useful in resolving disputes about the technique of phylostratigraphy, which assigns ages to genes. Detecting homologs is more difficult for fast-evolving sequences, which may be erroneously scored as young (21-23). Disordered proteins tend to be fast evolving (24), 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 (25) 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 (15) 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.

Results

Estimating the fitness effects of random peptides

Assessing predictors of the fitness effects of random peptides requires those fitness effects to be measured accurately and precisely. Neme et al. (6) tracked lineage frequencies over four days. They categorized a peptide as beneficial or deleterious by comparing the counts of days 2, 3, and 4 to day 1 using DESeq2 (6, 26). Out of the 1082 peptides with at least five sequencing reads, they found that 277 were beneficial, while 436 were deleterious.

We use a custom maximum likelihood framework (see Materials & Methods) to quantitatively estimate "fitness" and its associated confidence interval / weight. "Fitness" here refers to allele frequency changes over an entire cycle of population growth and dilution, rather than per generation. Our method accounts for increasing mean population fitness over the four days. In the process, we find statistical support for increases in frequency (i.e. fitness – 1.96 × standard error > 1) for 475 peptides (Dataset S1). Our methods are both more sensitive and more specific. We exclude 22 of the original 277, at least some of which arose from a bug we located in Neme et al.'s (6) pipeline. 509 peptides had statistical support for decreases in frequency, with support lost for 9 of the original 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 (17, 18), water solubility using CamSol (27), and general aggregation propensity using Tango (9, 19, 20). Fewer than 6% of the random peptides have a predicted transmembrane helix (Dataset S1) from TMHMM (28), 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 (29), that specializes in β aggregates, was in the right direction but did not quite meet statistical significance (P = 0.06).

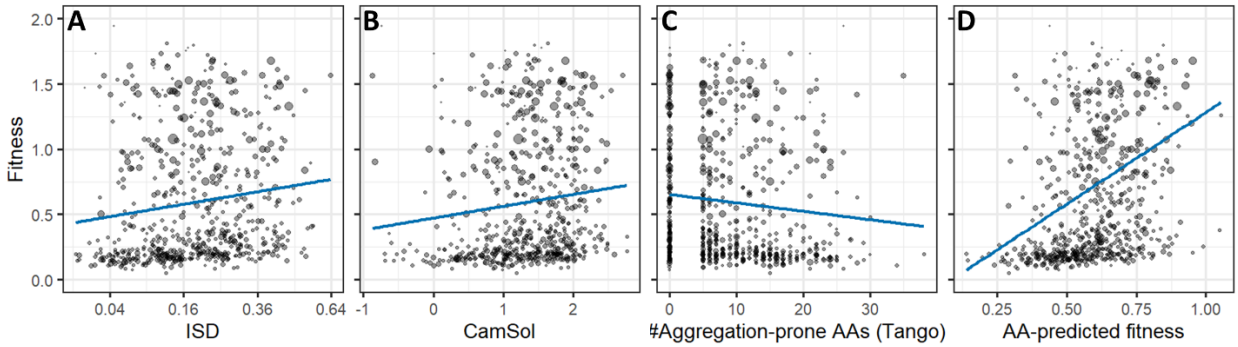

Fig. 1. **Many metrics, on their own, predict peptide fitness effects, but amino acid composition predicts the most.** A) ISD (on a square root scale), B) CamSol, and C) the number of amino acids in Tango-predicted aggregation-prone regions each significantly predict peptide fitness on their own ($P = 7 \times 10^{-4}$, 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. 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), 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 a surprisingly high 0.15.

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 ΔISD = original sequence ISD – scrambled sequence ISD), is higher in younger mouse genes (12). However, this was not the case for the very youngest mouse genes, and including ΔISD 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 (14, 15). This hydrophobic dispersion metric (30, 31), 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).

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 S1, 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 (32) and with a metric of how often an amino acid tends to be found in known disordered versus ordered proteins (33) (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 (34), relative solvent accessibility (35), amino acid cost in *E. coli* (36), and isoelectric point (37) did not provide significant explanatory power on top of disorder propensity and volume (all P > 0.1, likelihood ratio test).

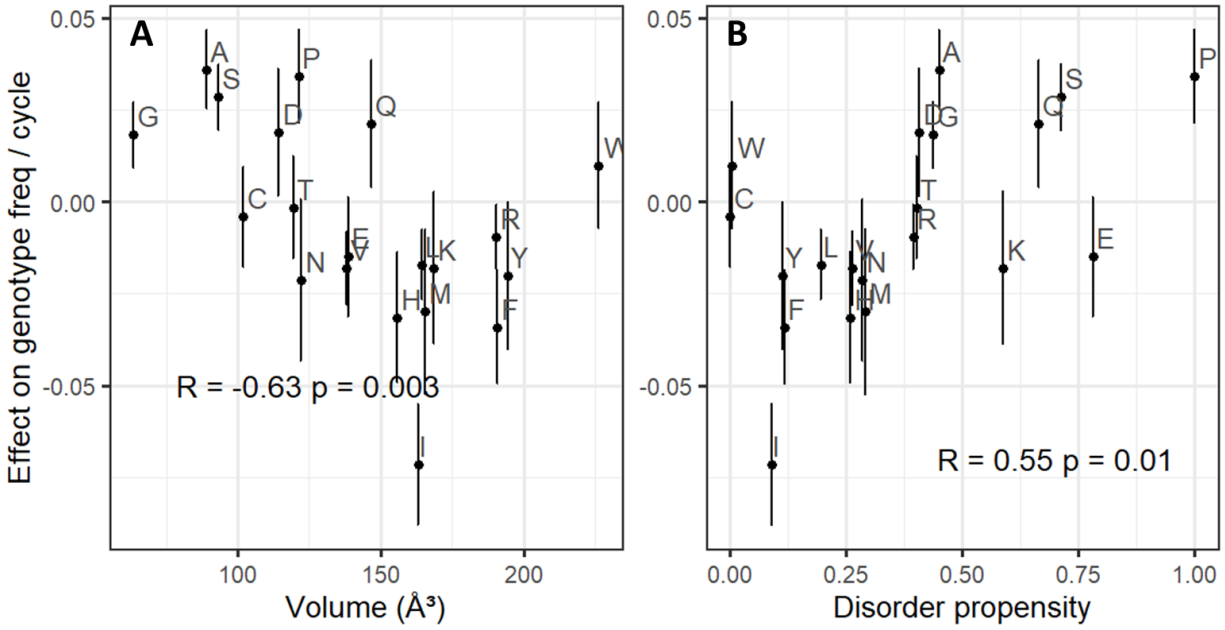

Fig. 2. **Amino acids that are small and are associated with intrinsic structural disorder promote higher fitness**. The y-axis shows each amino acid's marginal effect on fitness, which is the change in fitness when one amino acid of the focal type replaces one randomly chosen amino acid of a different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

Next we tested whether the amino acids that are relatively benign in random peptides expressed under laboratory conditions in *E. coli* are similar to the amino acids that are found in purportedly young protein-coding sequences in natural populations. We took the latter from James et al. (15) who quantified how amino acid frequencies depend on the age of the protein domain, 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).

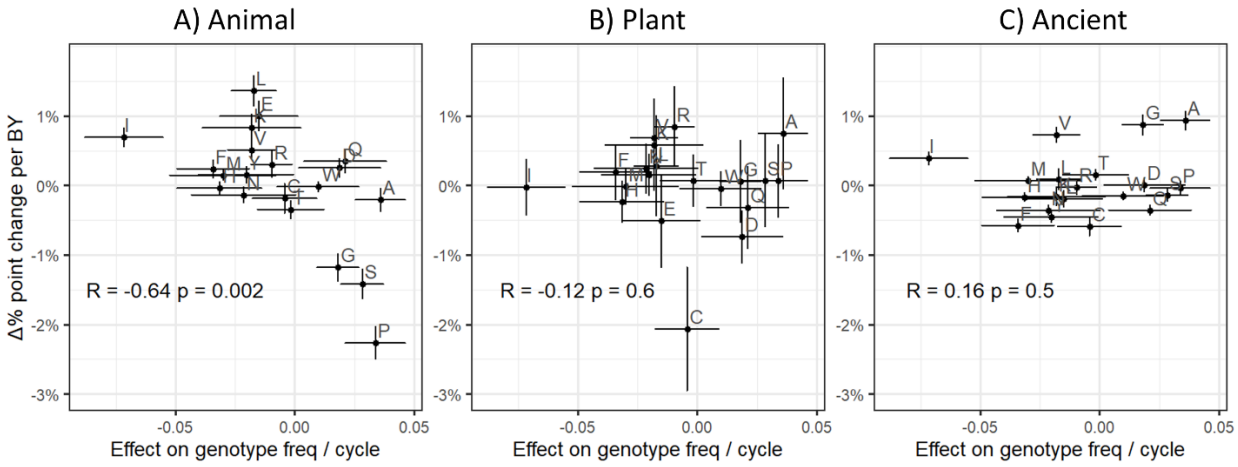

Fig. 3. **Purportedly young animal Pfams are enriched for amino acids that predict high fitness in random peptides.** The y-axis represents the change in the frequency of each amino acid per billion years (BY) for non-transmembrane Pfam domains (15). Change is in number of percentage points, not percentage increase or decrease. For example, a change in glutamic acid content from 5% to 6% is a change of 1 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, rather than for amino acids that promote intrinsic structural disorder (15). We find that intrinsic structural disorder 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 intrinsic structural disorder does shape recent *de novo* gene birth in viruses (16).

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) (21-23). 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. This conclusion is also supported by the fact that these slopes differ among taxa and are not correlated with amino acid changeability (15).

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

Long et al. (38) 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 intrinsic structural disorder (10), 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 lineage 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.

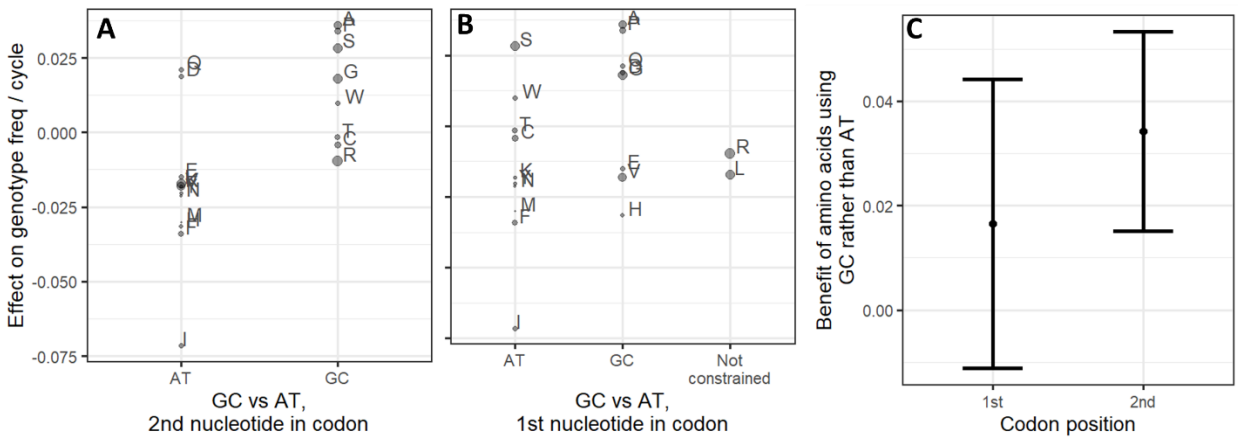

Fig. 4. **Amino acids that are constrained to use Gs and Cs tend to have higher marginal effects on fitness than those constrained to use As and Ts**. The difference is significant for constraints at the second nucleotide position of a codon (A) (P = 0.001, weighted Welch's t-test), but not at the first (B) (P = 0.2). Point area is proportional to weight, which is calculated as 1 / s.e.(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 difference +/- t_{crit} × se), where $t_{crit} \approx 2.1$ 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 to predict the fitness effects of shorter random peptides present as N-terminal tags. Goodman et al. (39) tagged the N-prime end of green fluorescent protein (GFP) with 137 different short random sequences (11 amino acids long), to allow for monitoring of random peptide expression level via fluorescence. Frumkin et al. (40) measured the fitness effects of these random peptide-tagged GFPs in *E. coli* using FitSeq (41). For 89 of them, Frumkin et al. (40) were able to calculate a "fitness residual" based on the deviation from the fitness expected from the level of GFP expression. Our fitness predictions were significantly correlated with their fitness residuals (fig. 5). These findings 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.

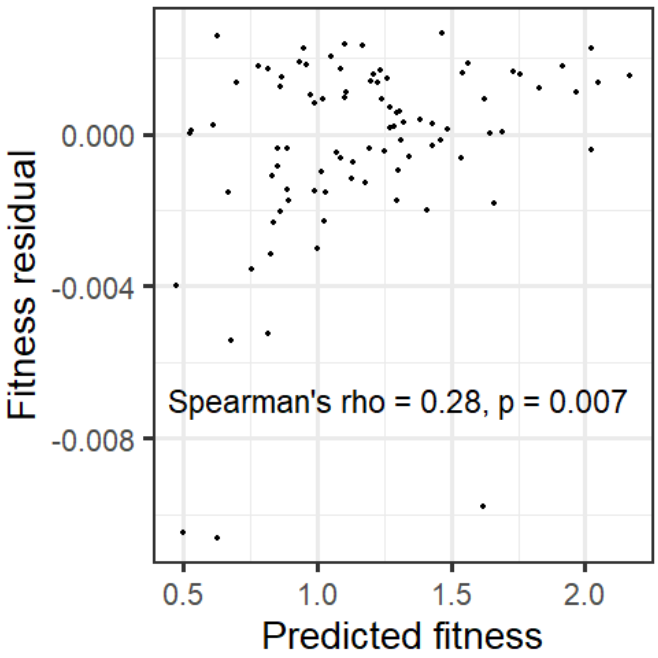

Fig. 5. **Fitness predictions trained on the random peptides of Neme et al. (6) also work for short random tags attached to the N-terminus of GFP**. Predicted fitness comes from our amino acid composition-only mixed model. "Fitness residuals" of N-terminal tags are from Frumkin et al. (40), and represent 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 intrinsic structural disorder 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 a random peptide explained 15% of the variance in fitness.

Most studies of random peptides have focused on finding peptides that have specific binding or function (42, 43, e.g. 44). Some were motivated as proof-of-concept that random peptides can exhibit properties of native proteins, such as folding (45-47) and being soluble (48). 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 (49). Our work has a different intent, identifying properties that make a peptide less likely to be harmful. Neme et al.'s (6) experiment was perfect 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. (50), who selected random peptides that rescue viability in the presence of antibiotics, is less useful for our purposes.

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. (6) 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 suggests that many lower GC content sequences were so harmful that lineages that carried them went extinct prior to detection via sequencing.

Long et al. (38) 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 (51, 52). Selection on translation errors is for example strong enough to shape non-coding sequences beyond stop codons in *Saccharomyces cerevisiae* (53).

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

While driven by amino acid composition, 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.

Plants have different trends in amino acid composition change over time than animals, with young genes seeming to prefer readily available amino acids, rather than amino acids that promote intrinsic structural disorder (15). 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.

Methods

Data retrieval

Neme et al. (6) 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 summed across all 5 replicates to obtain a total number of reads for each polypeptide at each time point. Following Neme et al. (6), we took the 1061 peptides out of over one million that had ≥5 reads across all 5 replicates of experiment 7. 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

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

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

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

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).

348 An existing software package for estimating lineage fitness from sequencing counts is Fit-Seq (41), which
349 captures the amplication of PCR error through a more sophisticated distribution for the number of reads
350 that is derived in the supplementary information of Levy et al. (54). However, Fit-Seq assumes that
351 mean fitness is a simple average of all measured lineages' fitness, requiring all individuals to be tagged
352 and measured. But Neme et al.'s (6) experiment included lineages carrying an empty plasmid, i.e. with
353 the selectable marker but no random peptide. Worse, the proportion of cells with an empty vector can
354 be presumed to increase over time. In the absence of a reliable way to quantify these, we instead
355 consider mean population fitness over time to be a set of independent parameters to be fitted.

356 Clustering non-independent sequences

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

367 Such similar sequences are highly unlikely to arise by chance if the peptides were truly random; $20^{50} \approx$
368 10^{65} peptides are possible, far more than the ~2×10^6 observed. Because we analyze only peptides with
369 at least 5 reads, replicated sequencing error is an unlikely cause. We see the same nearly-identical
370 sequences appearing in every experimental replicate, suggesting either that mutations occurred during
371 Neme et al.'s (6) initial growth phase, or that the "random" peptides synthesized for the experiment are
372 not entirely random. We note that construction of the "random" peptide library involved ligations of a
373 smaller set of "seed" sequences, introducing non-randomness at this stage.

374 Predictors of fitness

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

377 *GC content*

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

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

384 *Disorder*

385 IUPred (17, 18) returns an ISD score between zero and one for each amino acid in a sequence, with
386 higher scores indicating greater intrinsic disorder. To calculate an ISD score for each random peptide, we

387 took the average of the scores for the whole sequence (i.e. including non-random parts). We used a
388 square root transform because it produced a more linear relationship with fitness than no transform.
389 IUPred2 was used for all analyses except ΔISD, which used the original IUPred program – differences
390 between the two are minimal (18).

391 *Aggregation propensity*

392 Tango (9, 19, 20) returns an aggregation score for each amino acid in a sequence. At least five sequential
393 amino acids with a score greater than or equal to five indicates an aggregation-prone region. We scored
394 peptide aggregation propensity as the number of amino acids within regions scored as aggregation-
395 prone, including contributions from non-random regions.

396 *Solubility*

397 CamSol (27) returns a solubility score for each amino acid in a sequence, as well as a simple average of
398 all scores for a sequence, which CamSol calls a "solubility profile." We used the solubility profile of the
399 full sequences, including non-random regions.

400 *Hydrophobic clustering*

401 We used a Perl script developed by Scott Foy ("calculate_hydrophobic_dispersion.pl" at
402 https://github.com/MaselLab/A-shift-in-aggregation-avoidance-strategy-marks-a-long-term-direction-
403 to-protein-evolution/tree/master/script) to calculate Irback's hydrophobicity dispersion metric (30, 31).
404 This metric splits a sequence into blocks of equal length and calculates dispersion as the variance of the
405 hydrophobicity per block by the mean hydrophobicity per block, normalized to account for differences
406 in mean hydrophobicity and length. We used blocks of 6 amino acids, considered I, L, M, F, and V
407 hydrophobic, and scored whole peptide sequences including non-random regions.

408 *Amino acid composition*

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

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

412 Statistics

413 All statistical tests were carried out in R version 3.6.3 (55), with figures generated using "ggplot2" (56).
414 Weighted linear mixed models were implemented using the "lmer" function from the "lme4" package
415 (57), with cluster as a random effect. See Supporting Information for details, including justification of a
416 log-transform for fitness. When R^2 values were needed, we instead averaged peptides within the same
417 cluster into a combined datapoint, allowing us to avoid the use of random effect term, and took R^2
418 values from the base R "lm" function.

419 Code availability

420 All code is available on GitHub at https://github.com/MaselLab/RandomPeptides.

421 Acknowledgements

422 This work was supported by the John Templeton Foundation (39667, 60814) and the National Institutes
423 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

1. Chiti F & Dobson CM (2017) Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* 86:27-68.
2. Vecchi G*, et al.* (2020) Proteome-wide observation of the phenomenon of life on the edge of solubility. *Proc. Natl. Acad. Sci. U.S.A.* 117(2):1015-1020.
3. McLysaght A & Guerzoni D (2015) New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Phil. Trans. R. Soc. B.* 370(1678):20140332.
4. Van Oss SB & Carvunis AR (2019) *De novo* gene birth. *PLoS Genet.* 15(5):e1008160.
5. Vakirlis N, Carvunis A-R, & McLysaght A (2020) Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* 9:e53500.
6. Neme R, Amador C, Yildirim B, McConnell E, & Tautz D (2017) Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* 1:0127.
7. Weisman CM & Eddy SR (2017) Gene Evolution: Getting Something from Nothing. *Current Biology* 27(13):R661-R663.
8. Knopp M & Andersson DI (2018) No beneficial fitness effects of random peptides. *Nat. Ecol. Evol.* 2(7):1046-1047.
9. Linding R, Schymkowitz J, Rousseau F, Diella F, & Serrano L (2004) A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.* 342(1):345-353.
10. Angyan AF, Perczel A, & Gaspari Z (2012) Estimating intrinsic structural preferences of *de novo* emerging random-sequence proteins: Is aggregation the main bottleneck? *FEBS Lett.* 586(16):2468-2472.
11. Tretyachenko V*, et al.* (2017) Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Scientific Reports* 7:15449.
12. Wilson BA, Foy SG, Neme R, & Masel J (2017) Young genes are highly disordered as predicted by the preadaptation hypothesis of *de novo* gene birth. *Nat. Ecol. Evol.* 1(6):0146.
13. Heames B, Schmitz J, & Bornberg-Bauer E (2020) A continuum of evolving de novo genes drives protein-coding novelty in Drosophila. *J. Mol. Evol.* 38:382-398.
14. Foy SG, Wilson BA, Bertram J, Cordes MHJ, & Masel J (2019) A shift in aggregation avoidance strategy marks a long-term direction to protein evolution. *Genetics* 211:1345-1355.
15. James JE*, et al.* (2020) Universal and taxon-specific trends in protein sequences as a function of age. *bioRxiv*:2020.2003.2026.010728.
16. Willis S & Masel J (2018) Gene birth contributes to structural disorder encoded by overlapping genes. *Genetics* 210(1):303-313.
17. Dosztányi Z, Csizmok V, Tompa P, & Simon I (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* 347(4):827-839.
18. Meszaros B, Erdos G, & Dosztanyi Z (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46(W1):W329-W337.

468 19. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, & Serrano L (2004) Prediction of
469 sequence-dependent and mutational effects on the aggregation of peptides and proteins.
470 *Nature biotechnology* 22(10):1302-1306.
471 20. Rousseau F, Schymkowitz J, & Serrano L (2006) Protein aggregation and amyloidosis: confusion
472 of the kinds? *Current opinion in structural biology* 16(1):118-126.
473 21. Alba MM & Castresana J (2007) On homology searches by protein Blast and the characterization
474 of the age of genes. *BMC Evol. Biol.* 7:53.
475 22. Moyers BA & Zhang JZ (2015) Phylostratigraphic Bias Creates Spurious Patterns of Genome
476 Evolution. *Mol. Biol. Evol.* 32(1):258-267.
477 23. Moyers BA & Zhang JZ (2016) Evaluating Phylostratigraphic Evidence for Widespread De Novo
478 Gene Birth in Genome Evolution. *Mol. Biol. Evol.* 33(5):1245-1256.
479 24. Chen SCC, Chuang TJ, & Li WH (2011) The Relationships Among MicroRNA Regulation,
480 Intrinsically Disordered Regions, and Other Indicators of Protein Evolutionary Rate. *Mol. Biol.*
481 *Evol.* 28(9):2513-2520.
482 25. Monod J & Jacob F (1961) General conclusions - teleonomic mechanisms in cellular metabolism,
483 growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.* 26:389-401.
484 26. Love MI, Huber W, & Anders S (2014) Moderated estimation of fold change and dispersion for
485 RNA-seq data with DESeq2. *Genome Biol.* 15(12):550.
486 27. Sormanni P, Aprile FA, & Vendruscolo M (2015) The CamSol Method of Rational Design of
487 Protein Mutants with Enhanced Solubility. *J. Mol. Biol.* 427(2):478-490.
488 28. Krogh A, Larsson B, von Heijne G, & Sonnhammer ELL (2001) Predicting transmembrane protein
489 topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.*
490 305(3):567-580.
491 29. Maurer-Stroh S*, et al.* (2010) Exploring the sequence determinants of amyloid structure using
492 position-specific scoring matrices. *Nature Methods* 7(3):237-242.
493 30. Irback A, Peterson C, & Potthast F (1996) Evidence for nonrandom hydrophobicity structures in
494 protein chains. *Proc. Natl. Acad. Sci. U.S.A.* 93(18):9533-9538.
495 31. Irback A & Sandelin E (2000) On hydrophobicity correlations in protein chains. *Biophysical*
496 *Journal* 79(5):2252-2258.
497 32. Tsai J, Taylor R, Chothia C, & Gerstein M (1999) The packing density in proteins: Standard radii
498 and volumes. *J. Mol. Biol.* 290(1):253-266.
499 33. Theillet F-X*, et al.* (2013) The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance
500 and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disord. Proteins*
501 1(1):e24360.
502 34. Levy ED, De S, & Teichmann SA (2012) Cellular crowding imposes global constraints on the
503 chemistry and evolution of proteomes. *Proc. Natl. Acad. Sci. U.S.A.* 109(50):20461-20466.
504 35. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, & Wilke CO (2013) Maximum allowed solvent
505 accessibilites of residues in proteins. *PLoS One* 8(11):e80635.
506 36. Akashi H & Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes
507 of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U.S.A.* 99(6):3695-3700.
508 37. Liu HX*, et al.* (2004) Prediction of the isoelectric point of an amino acid based on GA-PLS and
509 SVMs. *Journal of Chemical Information and Computer Sciences* 44(1):161-167.
510 38. Long H*, et al.* (2018) Evolutionary determinants of genome-wide nucleotide composition. *Nat.*
511 *Ecol. Evol.* 2(2):237-240.
512 39. Goodman DB, Church GM, & Kosuri S (2013) Causes and effects of N-terminal codon bias in
513 bacterial genes. *Science* 342(6157):475-479.
514 40. Frumkin I*, et al.* (2017) Gene Architectures that Minimize Cost of Gene Expression. *Mol. Cell*
515 65(1):142-153.

41. Li F, Salit ML, & Levy SF (2018) Unbiased fitness estimation of pooled barcode or amplicon sequencing studies. *Cell Syst.* 7(5):521-525.
42. Frulloni L*, et al.* (2009) Identification of a Novel Antibody Associated with Autoimmune Pancreatitis. *N. Engl. J. Med.* 361(22):2135-2142.
43. Keefe AD & Szostak JW (2001) Functional proteins from a random-sequence library. *Nature* 410(6829):715-718.
44. Kaiser CA, Preuss D, Grisafi P, & Botstein D (1987) Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science* 235(4786):312-317.
45. Davidson AR & Sauer RT (1994) Folded proteins occur frequently in libraries of random amino-acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* 91(6):2146-2150.
46. Chiarabelli C*, et al.* (2006) Investigation of de novo totally random biosequences Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display. *Chem. Biodivers.* 3(8):840-859.
47. LaBean TH, Butt TR, Kauffman SA, & Schultes EA (2011) Protein folding absent selection. *Genes* 2(3):608-626.
48. Prijambada ID*, et al.* (1996) Solubility of artificial proteins with random sequences. *FEBS Lett.* 382(1-2):21-25.
49. Davidson AR, Lumb KJ, & Sauer RT (1995) Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.* 2(10):856-864.
50. Knopp M*, et al.* (2019) De novo emergence of peptides that confer antibiotic resistance. *mBio* 10(3):e00837-00819.
51. Wilson BA & Masel J (2011) Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* 3:1245-1252.
52. Rajon E & Masel J (2011) Evolution of molecular error rates and the consequences for evolvability. *Proc. Natl. Acad. Sci. U.S.A.* 108(3):1082-1087.
53. Kosinski LJ & Masel J (2020) Readthrough errors purge cryptic sequences, facilitating the birth of coding sequence. *Mol. Biol. Evol.* msaa046, Published online ahead of print, 26 Feb 2020, https://doi.org/10.1093/molbev/msaa046.
54. Levy SF*, et al.* (2015) Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* 519(7542):181-186.
55. R Core Team (2019) R: A language and environment for statistical computing (R Foundation for Statistical Computing).
56. Wickham H (2016) *ggplot2: Elegant graphics for data analysis* (Springer-Verlag, New York).
57. Bates D, Maechler M, Bolker B, & Walker S (2015) Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1):1-48.