Differences Between the Raw Material and the Products of *de Novo* Gene Birth Can Result from Mutational Biases

Lou Nielly-Thibault[1] and Christian R. Landry
Institut de Biologie Intégrative et des Systèmes, Département de Biologie, and Département de Biochimie, de Microbiologie et de Bio-Informatique, Université Laval, Québec, Quebec G1V 0A6, Canada, and PROTEO, Québec, Quebec G1V 0A6, Canada
ORCID IDs: 0000-0003-3236-651X (L.N.-T.); 0000-0003-3028-6866 (C.R.L.)

ABSTRACT Proteins are among the most important constituents of biological systems. Because all protein-coding genes have a noncoding ancestral form, the properties of noncoding sequences and how they shape the birth of novel proteins may influence the structure and function of all proteins. Differences between the properties of young proteins and random expectations from noncoding sequences have previously been interpreted as the result of natural selection. However, interpreting such deviations requires a yet-unattained understanding of the raw material of *de novo* gene birth and its relation to novel functional proteins. We mathematically show that the average properties and selective filtering of the "junk" polypeptides of which this raw material is composed are not the only factors influencing the properties of novel functional proteins. We find that in some biological scenarios, they also depend on the variance of the properties of junk polypeptides and their correlation with the rate of allelic turnover, which may itself depend on mutational biases. This suggests for instance that any property of polypeptides that accelerates their exploration of the sequence space could be overrepresented in novel functional proteins, even if it has a limited effect on adaptive value. To exemplify the use of our general theoretical results, we build a simple model that predicts the mean length and mean intrinsic disorder of novel functional proteins from the genomic GC content and a single evolutionary parameter. This work provides a theoretical framework that can guide the prediction and interpretation of results when studying the *de novo* emergence of protein-coding genes.

KEYWORDS *de novo* gene birth; novel proteins; random sequences; neutral evolution; GC content

THEORETICAL and empirical studies of how species acquire new proteins have described several mechanisms with distinct effects on genomes. Most of these mechanisms, such as gene duplication (Innan and Kondrashov 2010), horizontal gene transfer (Soucy *et al.* 2015), and gene fusion (Di Roberto and Peisajovich 2014), produce novelty by tweaking and rearranging preexisting gene sequences. In contrast, the mechanism of *de novo* gene birth consists of the emergence of new genes from reading frames that were ancestrally noncoding (McLysaght and Hurst 2016). This mechanism includes emergence from noncoding DNA, but also from alternative, noncoding reading frames in coding DNA (Keese and Gibbs 1992). Although *de novo* gene birth was once thought to be highly improbable (Jacob 1977), lineage-specific genes and proteins are observed in a variety of eukaryotes (McLysaght and Guerzoni 2015), bacteria (Neuhaus *et al.* 2016), and endosymbiotic organelles (Breton *et al.* 2011), which suggests that the contribution of *de novo* gene birth to the evolution of proteomes and cellular systems is not negligible. The biological activities of these novel sequences are often hard to infer since they lack well-studied homologs, but some of them have nevertheless been shown to play important and even vital biological roles (Heinen *et al.* 2009; Chen *et al.* 2010; Reinhardt *et al.* 2013). Since *de novo* gene birth is the only source of novel protein families and thus the only genetic mechanism which can add novel protein elements to cellular networks, it may have significantly influenced the diversity of existing protein structures through the initial emergence of unrelated proteins that gave rise to major protein families (Edwards *et al.* 2013).

doi: https://doi.org/10.1534/genetics.119.302187
Manuscript received January 18, 2019; accepted for publication June 14, 2019; published Early Online June 21, 2019.
Available freely online through the author-supported open access option.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.8304533.
[1]Corresponding author: Département de Biologie, Université Laval, Pavillon Alexandre-Vachon, 1045 Ave. de la Médecine, Québec, QC G1V 0A6, Canada. E-mail: lou.nielly-thibault.1@ulaval.ca

Although many authors agree that conservation by natural selection should be part of the definition of *de novo* genes (Schlötterer 2015; McLysaght and Hurst 2016), the exact moment in the existence of a polypeptide at which *de novo* gene birth occurs has not been agreed upon, which makes "*de novo* gene birth" and related terms confusing in practice. For clarity, we hereinafter avoid these terms. Instead, we define a classification of polypeptides into three types: junk polypeptides (JPs), novel functional polypeptides (novFPs), and derived functional polypeptides (derFPs). Figure 1 illustrates this classification in terms of the underlying evolutionary processes and their implications for the comparison of sequence properties and *cis*-regulatory properties between classes.

A JP is a polypeptide that is encoded by some open reading frame (ORF) but whose beneficial effects, if it has any, has not yet caused the loss of a mutation that modifies its sequence and/or its *cis*-regulatory properties. JPs are a very wide class of polypeptides. They may be encoded by intergenic ORFs, but also by noncoding RNA genes and alternative ORFs in protein-coding genes. A JP may also have any sequence, any *cis*-regulatory properties, and any effect on fitness, with the sole condition that its beneficial effects are either too weak or too recent to have prevented the fixation of mutations at its locus. This definition of JPs is relevant in evolutionary proteomics because they have not been shaped by purifying selection for any activity, although they may be shaped by selection against deleterious effects such as metabolic cost, aggregation, and spurious interactions with other molecules (Figure 1C). As a result, evolutionary models that explain their structural and *cis*-regulatory properties will likely not apply to other polypeptides, and vice versa. The concept of JP is a more precise version of the concept of "spurious" expression producing the raw material of *de novo* gene birth (Wilson and Masel 2011).

Once the beneficial effects of a JP eliminate a mutation which modified its sequence or its *cis*-regulatory properties, it no longer meets the definition of a JP, and yet it is identical to the JP that it recently was. We refer to such a transitory polypeptide as a novFP. We use the term "functionalization" to refer to the transition between a JP and a novFP, which is consistent with the selected-effect definition of biological function (Doolittle *et al.* 2014). Even though each novFP has the same sequence and *cis*-regulation as its last ancestral JP, there may be important statistical differences between JPs and novFPs, since only a select subset of JPs become novFPs. Although the expression of a single JP is presumably unlikely to be strongly beneficial, this barrier to functionalization may be overcome by the "testing" of a large diversity of JPs during evolution. This diversity depends on the number of JPs expressed in the population, but also on their rate of allelic turnover, *i.e.*, the rate of appearance and disappearance of JP-expressing alleles.

Once a novFP undergoes the fixation of at least one nonsynonymous or *cis*-regulatory mutation, its properties are no longer the results of a biased "draw" from the pool of JPs, because they also depend on how the beneficial effects of the polypeptide filter the mutations that modify its sequence and its *cis*-regulation. We use the term derFP for such polypeptides that have changed since their functionalization. As most of the canonical coding genes (ORFs annotated by genome databases) have divergent homologs in multiple species, it is safe to assume that the large majority of the proteins they code for meet the definition of derFPs. Along with point mutations, genetic drift, and natural selection, derFPs are known to evolve through the loss, duplication, fusion, and fission of genes, which may also influence the distributions of their properties (Figure 1C).

This classification of the whole proteome into JPs, novFPs, and derFPs leads to the division of proteomic evolution into four parallel processes (Figure 1B): (1) the allelic turnover of JPs through the evolution of the sequence, transcription, and translation of ORFs without effective purifying selection for their polypeptides' effects; (2) functionalization, which produces novFPs by filtering JPs without modifying their sequence and their *cis*-regulation, but can involve changes in the genetic background or the environment; (3) the subsequent evolution of the pool of novFPs and derFPs through sequence changes and through the loss, duplication, fusion, and fission of genes; and (4) the decay of novFPs and derFPs into JPs through the mitigation of selection, which can be driven by mutations, environmental changes, or an increase in genetic drift. Since most well-studied polypeptides are derFPs, we know very little about the first two processes, either experimentally or theoretically. Studying the allelic turnover of JPs and their functionalization is thus essential for completing our understanding of proteome evolution.

The set of all JPs expressed by a species, which we call the junk proteome, can be seen as a collection of fixed or segregating alleles in the genome. Although the extent of the junk proteome is still unknown, there is evidence of its existence based on diverse experimental approaches: experimental studies have shown that, in a variety of organisms, a large part of intergenic DNA is transcribed into 5′-capped and polyadenylated transcripts (Jensen *et al.* 2013) which can be translated (Ingolia *et al.* 2014; Ruiz-Orera *et al.* 2014). Contrary to canonical genes, these transcripts show signs of suboptimal translation (Guttman *et al.* 2013; Durand *et al.* 2019) and rapid evolution (Neme and Tautz 2016). Additionally, the so-called untranslated regions (UTRs) of canonical transcripts and the alternative reading frames within canonical ORFs are sometimes translated into polypeptides that lack known functions (Vanderperre *et al.* 2013; Ingolia *et al.* 2014; Landry *et al.* 2015; Mouilleron *et al.* 2016) and may thus be JPs. In mice, many translated ORFs in protein-coding genes and long noncoding RNAs were shown to evolve without any detectable selective constraints on the polypeptides that they encode (Ruiz-Orera *et al.* 2018). While it would be tempting to argue that the expression level of JPs should be kept to a minimum because it would represent a cost to the cell, recent studies have shown that this cost may not be high enough to be perceived and eliminated by natural selection in

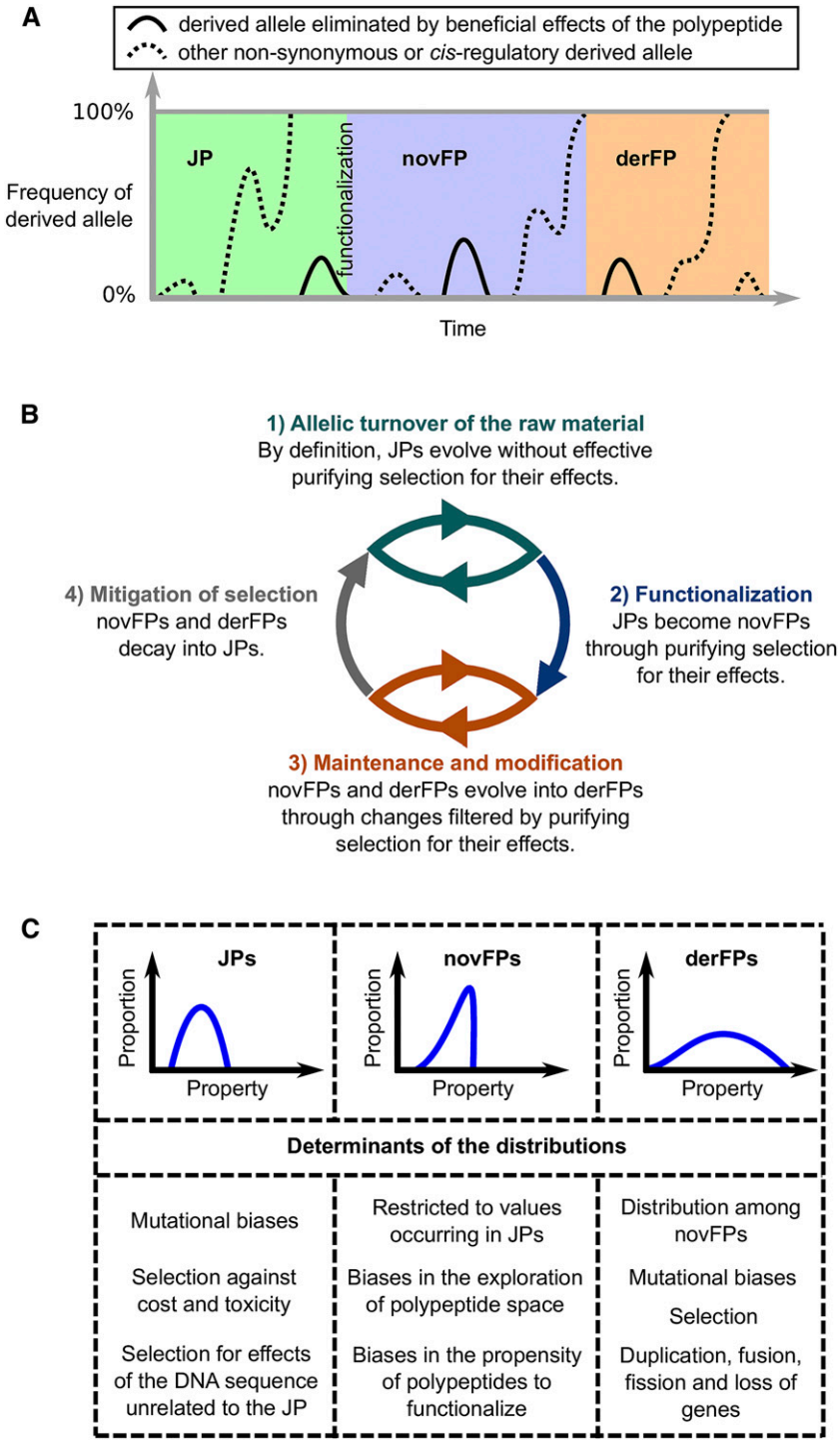

Figure 1 A new classification of polypeptides to clarify the process of *de novo* gene birth. (A) Evolution of a single polypeptide from JP to novFP to derFP. A JP is a polypeptide whose beneficial effects are either nonexistent or have not yet caused the loss of a nonsynonymous or *cis*-regulatory-derived allele of this polypeptide through natural selection. We call such an elimination event the functionalization of the polypeptide. A novFP is the immediate product of functionalization: a polypeptide that is no longer a JP but is identical to its ancestral JP in terms of sequence and *cis*-regulation, while their genetic backgrounds and environments may differ. A derFP is produced when a novFP undergoes the fixation of a nonsynonymous or *cis*-regulatory change. (B) Partitioning of proteome evolution in accordance with the classification of polypeptides. The two loops in the diagram represent the fact that JPs and derFPs can evolve without leaving their respective classes, while a novFP stops being a novFP as soon as it evolves. (C) The general determinants of distributions of polypeptide properties across the three classes of polypeptides. The curves describe hypothetical distributions of an arbitrary property of polypeptides, such as length or ISD. The distribution among novFPs is always restricted to the values that occur in the distribution among JPs, which is a consequence of the fact that functionalization turns a JP into a novFP without modifying it, as can be seen in (A).

many species (Lynch and Marinov 2015). In addition, JPs could be the result of a trade-off between this cost and the need for the transcriptional and translational machineries to be dynamic and reactive for the expression of derFPs, which could decrease the specificity of these machineries (Hausser *et al.* 2019).

The determinants of the properties of novFPs are of particular interest, since they constrain the properties of genes at the very roots of gene families. Several studies have inferred lineage-specific functional polypeptides (*i.e.*, novFPs and young derFPs) and compared them with ancient derFPs, using *in silico* translations of noncoding DNA and with randomly generated polypeptides. Assuming that the young derFPs inferred by these studies are largely similar to novFPs, their results suggest that novFPs typically differ from ancient derFPs by their short length, weak expression (Toll-Riera *et al.* 2009; Neme and Tautz 2013; Schlötterer 2015), peripheral position in cellular networks, and random-sequence-like

secondary structure (Abrusán 2013). It has been proposed that JPs and novFPs may be largely shaped by the genomic GC content through its effects on the properties of ORFs occurring in noncoding DNA sequences (Ángyán *et al.* 2012). Correlations supporting this role of GC content were observed for many quantities computed from the sequences of ORFs encoding inferred novFPs, although the averages of these quantities often depart from random expectations based on GC content (Basile *et al.* 2017). Such discrepancies were previously interpreted as the result of natural selection (Wilson *et al.* 2017), which is in line with the intuition that the probability of functionalization of a beneficial JP increases with its positive effect on fitness. However, several aspects of polypeptide functionalization require clarification before we can confidently interpret the average properties of observed novFPs and their differences from random or noncoding sequences.

In this article, we derive general mathematical results that link the average properties (*e.g.*, average length or average structural disorder) of novFPs to those of JPs. We find that the difference between the mean of a property among novFPs and the corresponding mean among JPs is proportional to the SD of this property among JPs. We also show that such mean discrepancies between JPs and novFPs may not result from natural selection alone, but also from the correlation of polypeptide properties with either the rate of allelic turnover of JPs, their probability of functionalization, or both. To illustrate how our general equations can be used to study polypeptide properties under specific models of the distribution of properties among JPs, we use a GC-content-based random-sequence model of JPs to predict how the genomic GC content and evolutionary parameters interact to determine the mean length and mean intrinsic structural disorder (ISD) of novFPs.

Materials and Methods

An introduction of the mathematical concepts used in the following reasoning (mainly measures and related concepts from measure theory) can be found in the Supplemental Material, in the file "SuppMat_2019-06-20.pdf".

A general model of the link between the properties of JPs and those of novFPs

Let Ω be the space of all possible polypeptides, each characterized by its sequence and its *cis*-regulatory properties. Over a given time period, the time averages of the number of JPs belonging to each possible category of polypeptides (each subset of Ω) can be divided by the time-averaged total number of JPs to form a probability measure P on Ω. In other words, for each subset S of Ω, $P(S)$ is the ratio of the time-averaged number of JPs that belong to S and the time-averaged total number of JPs, which implies that $P(\Omega) = 1$. Similarly, the novFPs that emerge by functionalization in the same period of time form a probability measure P_F on Ω, such that $P_F(S)$ is the proportion of novFPs that belong to S. For any polypeptide property q, *i.e.*, any function that assigns a number to each possible polypeptide in Ω, statistics like the mean and variance of q are defined separately for each probability measure. Hereinafter, we use the subscript F to distinguish between statistics defined for P and those defined for P_F. For example, the mean (expected value) of a property q among JPs will be denoted by $E(q)$, while its mean among novFPs will be denoted by $E_F(q)$.

Because functionalizationunder corresponds to the elimination of a mutation that would otherwise have modified a JP, each novFP is identical to its ancestral JP in terms of sequence and *cis*-regulation. As a result, the probability measures P and P_F have a special relationship: for any subset S of Ω which satisfies $P(S) = 0$, it is also true that $P_F(S) = 0$. In other words, any category of polypeptides that occurs in novFPs necessarily occurred among JPs at some point. Because of this relationship between the two measures (P_F is "absolutely continuous" with respect to P), the Radon–Nikodym theorem for finite measures (Vestrup 2003) implies that there exists a polypeptide property $\hat{r}$ such that, for any subset S of Ω with a well-defined $P(S)$, we have:

$$P_F(S) = \int_S \hat{r}\ dP,$$

where $\int_S \hat{r}\ dP$ is the integral of the function $\hat{r}$ over the set S with respect to the measure P. By dividing each side of the equation by $P(S)$, we get:

$$\frac{P_F(S)}{P(S)} = \frac{1}{P(S)} \int_S \hat{r}\ dP.$$

The right side of this equation is the definition of $E(x|S)$, the conditional average of a variable x knowing an event S (Çinlar 2011), with $x = \hat{r}$ in this specific case. Although we interpret S as a class of polypeptides rather than as an "event" and $\hat{r}$ as a polypeptide property rather than as a "random variable," these terms refer to the same mathematical objects, so that the average of $\hat{r}$ among JPs that belong to S is given by:

$$E(\hat{r}|S) = \frac{1}{P(S)} \int_S \hat{r}\ dP = \frac{P_F(S)}{P(S)}.$$

Since this equation holds for any subset S of Ω with $P(S) \neq 0$, the polypeptide property $\hat{r}$ can be interpreted as the factor by which the relative frequency of a polypeptide changes from JPs to novFPs. The average of $\hat{r}$ among all JPs is:

$$E(\hat{r}) = E(\hat{r}|\Omega) = \frac{P_F(\Omega)}{P(\Omega)} = \frac{1}{1} = 1.$$

This fits the intuition according to which the increase in relative frequency of certain polypeptides between JPs and novFPs ($\hat{r} > 1$) should be counterbalanced by a decrease in relative frequency of other polypeptides ($\hat{r} < 1$), since these frequencies are relative.

If we define T as the duration of the time period considered, F as the total number of functionalization events, and J as the time-averaged number of JPs, then $\frac{F}{T} \times P_F(S)$ is the time-averaged rate of functionalization events in the subset S of polypeptide space and $J \times P(S)$ is the time-averaged number of JPs that belong to S. The ratio of these two numbers is:

$$\frac{F \times P_F(S)}{T \times J \times P(S)} = \frac{F}{T \times J} \times E(\hat{r}|S) = E\left(\frac{F}{T \times J} \times \hat{r}|S\right).$$

If we define $r = \frac{F}{T \times J} \times \hat{r}$, this ratio becomes:

$$\frac{F \times P_F(S)}{T \times J \times P(S)} = E(r|S).$$

Therefore, r is a polypeptide property representing the rate at which each region of the space of polypeptides produces novFPs, normalized by the time-averaged number of JPs that belong to that region. The average of this rate among JPs is:

$$E(r) = E\left(\frac{F}{T \times J} \times \hat{r}\right) = \frac{F}{T \times J} \times E(\hat{r}) = \frac{F}{T \times J}.$$

$\hat{r}$ is thus a normalization of r by its own mean:

$$\hat{r} = \frac{T \times J}{F} \times r = \frac{1}{E(r)} \times r.$$

Given a polypeptide property such as length or ISD, representing its mean among novFPs as a function of its mean among JPs would be useful in the study of *de novo* gene birth. The polypeptide properties $\hat{r}$ and r that we just defined can be used to obtain such a representation. Because of the way we defined $\hat{r}$ from the probability measures P and P_F ($\hat{r}$ is the Radon–Nikodym derivative of P_F with respect to P), it follows (Vestrup 2003) that for any polypeptide property q, we have:

$$\int_{\Omega} q \, dP_F = \int_{\Omega} q\hat{r} \, dP.$$

In probability theory, the expected value or average of a random variable (*e.g.*, q or $q\hat{r}$) among a population represented by a probability measure (*e.g.*, P_F or P) is defined as the integral of the variable with respect to the probability measure over the space of all possibilities (*e.g.*, Ω). Therefore, the above equation is equivalent to

$$E_F(q) = E(q\hat{r}),$$

where $E_F(q)$ is the average of the polypeptide property q among novFPs and $E(q\hat{r})$ is the average of the product $q\hat{r}$ among JPs.

Now that we have an expression for the average of an arbitrary property among novFPs, we can obtain an expression for the difference between this average and the average of the same property among JPs. To achieve this, we will only use universal rules from probability theory without making any assumption, such that the results apply to all biological contexts. By applying the property of covariance $E(xy) = E(x)E(y) + cov(x,y)$, we obtain:

$$E_F(q) = E(q)E(\hat{r}) + cov(q,\hat{r}).$$

Since $E(\hat{r}) = 1$, we obtain Equation 1:

$$E_F(q) - E(q) = cov(q,\hat{r}). \quad (1)$$

The covariance in Equation 1 depends on both the variation of q among JPs and its relation with functionalization (as represented by $\hat{r}$). We now seek to distinguish these two factors by modifying Equation 1. By applying the definition of the Pearson correlation coefficient $\rho(x,y) = \frac{cov(x,y)}{\sigma(x)\sigma(y)}$, we obtain:

$$E_F(q) - E(q) = \sigma(q)\sigma(\hat{r})\rho(q,\hat{r}).$$

Since $E(\hat{r}) = 1$, we can divide the right side of the equation by $E(\hat{r})$:

$$E_F(q) - E(q) = \sigma(q) \, \frac{\sigma(\hat{r})}{E(\hat{r})} \, \rho(q,\hat{r}).$$

By the definition of the coefficient of variation $CV(x) = \frac{\sigma(x)}{E(x)}$:

$$E_F(q) - E(q) = \sigma(q)CV(\hat{r})\rho(q,\hat{r}).$$

If we define $\delta = CV(\hat{r})\rho(q,\hat{r})$, we obtain Equation 2:

$$E_F(q) = E(q) + \sigma(q) \times \delta. \quad (2)$$

Because both the coefficient of variation and the correlation coefficient are insensitive to the multiplication of variables by positive constants, we have:

$$\delta = CV(k\hat{r})\rho(q,k\hat{r}),$$

where k can be any positive constant. Since $r = E(r) \times \hat{r}$ and $E(r)$ is a positive constant, we obtain:

$$\delta = CV(r)\rho(q,r).$$

Distinguishing the role of the allelic turnover of JPs under the assumption of their evolutionary equilibrium

According to Equation 2, the parameter δ, which we call the birth bias, is the only determinant of the average properties of novFPs that we cannot yet interpret in terms of the properties of JPs. To do so, we now make the assumption that the junk proteome is at evolutionary equilibrium, *i.e.*, JPs from any category are gained by mutation as often as they are either lost or functionalized. Since the polypeptide property r represents the ratio of the rate of functionalization events at one point in polypeptide space to the time-averaged number of JPs located at this point, it can be understood as the product $r = \lambda f$, where λ is the ratio of a JP's frequency of appearance by mutation to the time-averaged number of loci expressing this exact JP, and f is the probability that such a gain leads to

the functionalization of the polypeptide (the probability of functionalization). Because we assume the evolutionary equilibrium of the junk proteome, for each JP leaving a point in the polypeptide space, another mutant JP appears at this same point. Therefore, λ is also the rate at which a JP exits the junk proteome by either allele loss or functionalization (the inverse of its expected lifetime as a JP). Since λ is both a rate of arrival and a rate of departure of JPs at each point in polypeptide space, it is a rate of allelic turnover: a region of polypeptide space where λ is low will tend to be populated by mostly the same JPs for a long time, while a region where λ is high will have a large proportion of its JPs replaced by new ones in a short time. By combining $r = \lambda f$ with the definition of δ, we obtain Equation 3:

$$\delta = CV(\lambda f)\rho(q,\lambda f). \tag{3}$$

Equation 3 provides an interpretation of δ in terms of the product of the rate of allelic turnover of JPs with their probability of functionalization. However, it does not indicate how the correlation of only one of these two factors with a polypeptide property q could influence the associated value of δ, and thus the mean of q among novFPs. To find an expression of δ that makes this distinction, we transform Equation 3 using general rules of probability theory, which means that the results depend on the same assumption of evolutionary equilibrium as Equation 3. Using the definitions of the Pearson correlation coefficient and the coefficient of variation, we obtain:

$$\delta = \frac{\sigma(\lambda f)}{E(\lambda f)} \times \frac{cov(q,\lambda f)}{\sigma(q)\sigma(\lambda f)} = \frac{cov(q,\lambda f)}{\sigma(q)E(\lambda f)}.$$

Using the identity $E(\lambda f) = E(\lambda)E(f) + cov(\lambda, f)$:

$$\delta = \frac{cov(q,\lambda f)}{\sigma(q)(E(\lambda)E(f) + cov(\lambda, f))} = \frac{cov(q,\lambda f)}{\sigma(q)E(\lambda)E(f) + \sigma(q)cov(\lambda, f)}.$$

By decomposing the numerator as the covariance of a product of random variables according to (Bohrnstedt and Goldberger 1969):

$$\delta = \frac{E(f)cov(q,\lambda) + E(\lambda)cov(q,f) + E(\Delta q \Delta\lambda \Delta f)}{\sigma(q)E(\lambda)E(f) + \sigma(q)cov(\lambda, f)},$$

where $\Delta x = x - E(x)$. By dividing the numerator and the denominator by $\sigma(q)E(\lambda)E(f)$:

$$\delta = \frac{\frac{cov(q,\lambda)}{\sigma(q)E(\lambda)} + \frac{cov(q,f)}{\sigma(q)E(f)} + \frac{E(\Delta q \Delta\lambda \Delta f)}{\sigma(q)E(\lambda)E(f)}}{1 + \frac{cov(\lambda,f)}{E(\lambda)E(f)}}.$$

By taking the factor $\sigma(\lambda)\sigma(f)$ out of the rightmost term of the numerator:

$$\delta = \frac{\frac{cov(q,\lambda)}{\sigma(q)E(\lambda)} + \frac{cov(q,f)}{\sigma(q)E(f)} + \frac{\sigma(\lambda)\sigma(f)}{E(\lambda)E(f)} \times \frac{E(\Delta q \Delta\lambda \Delta f)}{\sigma(q)\sigma(\lambda)\sigma(f)}}{1 + \frac{cov(\lambda,f)}{E(\lambda)E(f)}}.$$

By applying the definition of the Pearson correlation coefficient three times:

$$\delta = \frac{\frac{\sigma(q)\sigma(\lambda)\rho(q,\lambda)}{\sigma(q)E(\lambda)} + \frac{\sigma(q)\sigma(f)\rho(q,f)}{\sigma(q)E(f)} + \frac{\sigma(\lambda)\sigma(f)}{E(\lambda)E(f)} \times \frac{E(\Delta q \Delta\lambda \Delta f)}{\sigma(q)\sigma(\lambda)\sigma(f)}}{1 + \frac{\sigma(\lambda)\sigma(f)\rho(\lambda,f)}{E(\lambda)E(f)}}.$$

By cancelling and rearranging factors within terms:

$$\delta = \frac{\frac{\sigma(\lambda)}{E(\lambda)} \times \rho(q,\lambda) + \frac{\sigma(f)}{E(f)} \times \rho(q,f) + \frac{\sigma(\lambda)\sigma(f)}{E(\lambda)E(f)} \times \frac{E(\Delta q \Delta\lambda \Delta f)}{\sigma(q)\sigma(\lambda)\sigma(f)}}{1 + \frac{\sigma(\lambda)\sigma(f)}{E(\lambda)E(f)} \times \rho(\lambda,f)}.$$

By applying the definition of the coefficient of variation six times:

$$\delta = \frac{CV(\lambda)\rho(q,\lambda) + CV(f)\rho(q,f) + CV(\lambda)CV(f) \times \frac{E(\Delta q \Delta\lambda \Delta f)}{\sigma(q)\sigma(\lambda)\sigma(f)}}{1 + CV(\lambda)CV(f) \times \rho(\lambda,f)}.$$

By the definition of the coskewness of three variables $cosk(x,y,z) = \frac{E(\Delta x \Delta y \Delta z)}{\sigma(x)\sigma(y)\sigma(z)}$, we obtain Equation 4:

$$\delta = \frac{CV(\lambda)\rho(q,\lambda) + CV(f)\rho(q,f) + CV(\lambda)CV(f)cosk(q,\lambda,f)}{1 + CV(\lambda)CV(f) \times \rho(\lambda,f)}. \tag{4}$$

The denominator in Equation 4 is strictly positive since, by the definition of the coefficient of variation and the properties of covariance

$$1 + CV(\lambda)CV(f)\rho(\lambda,f) = 1 + \frac{cov(\lambda,f)}{E(\lambda)E(f)} = \frac{E(\lambda)E(f) + cov(\lambda,f)}{E(\lambda)E(f)} = \frac{E(\lambda f)}{E(\lambda)E(f)},$$

and since both $E(\lambda f)$ and $E(\lambda)E(f)$ are positive.

Defining a random-sequence model of JPs

This section defines a simple model of the sequence of JPs that can be used to predict the mean and SD of the sequence properties of JPs, which can then be used with Equation 2 to predict the effect of the birth bias δ on novFPs. To build such a model of JPs, we made five assumptions about the DNA encoding them: (1) all sites evolve independently, (2) the transition probability matrix is constant across sites, (3) the transition probability matrix is the same on both strands, (4) each site has reached evolutionary equilibrium, and (5) a random subset of ATG codons define ORFs that are translated into JPs. Assumptions 1 and 2 allow us to focus on a single site and generalize our findings to the whole sequence. Assumptions 3 and 4 entail that if two nucleotides are Watson–Crick complements, then a given site is equally likely to

display either of them. As a result, complementary nucleotides are equally frequent within and between strands, and the frequency of each of the four nucleotides is a function of GC content. Since sites are independent, GC content is the only parameter needed to predict probability distributions for the properties of randomly occurring ORFs under this model. Assumption 5 allows us to extend our predictions to the properties of JPs expressed from these ORFs. In summary, this model implies that each JP is encoded by a random ORF appearing in a random DNA sequence with a fixed GC content.

Predicting the mean length of novFPs from the GC content and the birth bias

In the resulting model, predicting the length distribution of JPs is equivalent to predicting the distribution of the number of in-frame sense codons separating each ATG codon from the closest downstream in-frame stop codon. The frequency p_N of each nucleotide N is a function of the GC content, which we denote by $p_{G/C} = p_G + p_C$. The frequencies of the four nucleotides are given by:

$$p_G = p_C = \frac{p_{G/C}}{2}\; p_A = p_T = \frac{1 - p_{G/C}}{2}.$$

Since, in this model, consecutive nonoverlapping DNA 3-mers are statistically independent and have the same probability of being stop codons, the number of sense codons in an ORF follows a geometric distribution with the following probability mass function:

$$Prob(length = n) = (1-p_S)^{n-1} \times p_S,$$

where n is any positive integer and p_S is the probability that a given DNA 3-mer is a stop codon. Under our assumptions, the frequency of a DNA word is equal to the product of the frequencies of the nucleotides of which it is composed. Using this principle to calculate p_S, we get:

$$p_S = p_T p_A p_A + p_T p_A p_G + p_T p_G p_A$$

$$p_S = \left(\frac{1-p_{G/C}}{2}\right)^3 + \left(\frac{1-p_{G/C}}{2}\right)^2\left(\frac{p_{G/C}}{2}\right) + \left(\frac{1-p_{G/C}}{2}\right)^2\left(\frac{p_{G/C}}{2}\right)$$

$$p_S = \left(\frac{1-p_{G/C}}{2}\right)^3 + 2\left(\frac{1-p_{G/C}}{2}\right)^2\left(\frac{p_{G/C}}{2}\right)$$

$$p_S = \left(\frac{1-p_{G/C}}{2}\right)^2\left(\left(\frac{1-p_{G/C}}{2}\right) + 2\left(\frac{p_{G/C}}{2}\right)\right)$$

$$p_S = \left(\frac{1-p_{G/C}}{2}\right)^2\left(\frac{1+p_{G/C}}{2}\right)$$

$$p_S = \frac{1}{8}\left(1-p_{G/C}\right)^2\left(1+p_{G/C}\right).$$

Using standard equations for the mean and SD of a geometric distribution, we obtained the mean and SD of the length of JPs as functions of the frequency of stop codons, which is itself determined by the GC content:

$$E(length) = \frac{1}{p_S} = \frac{8}{\left(1-p_{G/C}\right)^2\left(1+p_{G/C}\right)}$$

$$\sigma(length) = \frac{\sqrt{1-p_S}}{p_S} = \frac{\sqrt{1-\frac{1}{8}\left(1-p_{G/C}\right)^2\left(1+p_{G/C}\right)}}{\frac{1}{8}\left(1-p_{G/C}\right)^2\left(1+p_{G/C}\right)}.$$

Combining these two equations with Equation 2 results in an expression of the average length of novFPs in terms of the GC content and δ.

Predicting the mean ISD of novFPs from the GC content and the birth bias

To predict the mean and SD of the ISD of JPs as functions of the GC content, we randomly generated the sequences of 100,000 JPs for each value of GC content from 20 to 80% with steps of 2.5%. We computed the per-amino-acid "long" and "short" disorder scores using IUPred (Dosztányi *et al.* 2005), averaged the two types of scores separately within each sequence, and computed the mean and SD of the sequence-wide average of each score for each GC content. We then applied Equation 2 to these means and SDs to predict the mean ISD of novFPs as a function of both the GC content and the birth bias of ISD.

Data availability

All the code used in this study has been made available, along with supplemental figures and methods, in a total of five files. The file "Notebook.ipynb" is a Jupyter notebook containing the Python2 and Bash code used to generate Figure 2 and Figure S1. The Bash code in this notebook makes use of the three Python2 scripts "simulate_junk_proteome.py," "iupred_multi.py," and "stats_iupred_multi.py." The file "SuppMat_2019-06-20.pdf" contains Figure S1, its caption, and the Supplemental Methods text section. The complete procedure for simulations can be found on page 7 of the manuscript. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.8304533.

Results

The difference in the average of a property between JPs and novFPs is proportional to the SD of this property among JPs

We consider the evolution of the proteome in a species over a time period that could be, for instance, a single branch on a phylogenetic tree. We compare two distributions on the space of possible polypeptides. The first one, which we call JPs, is the time-averaged distribution where the relative frequency of any group of polypeptides is the ratio of their time-averaged

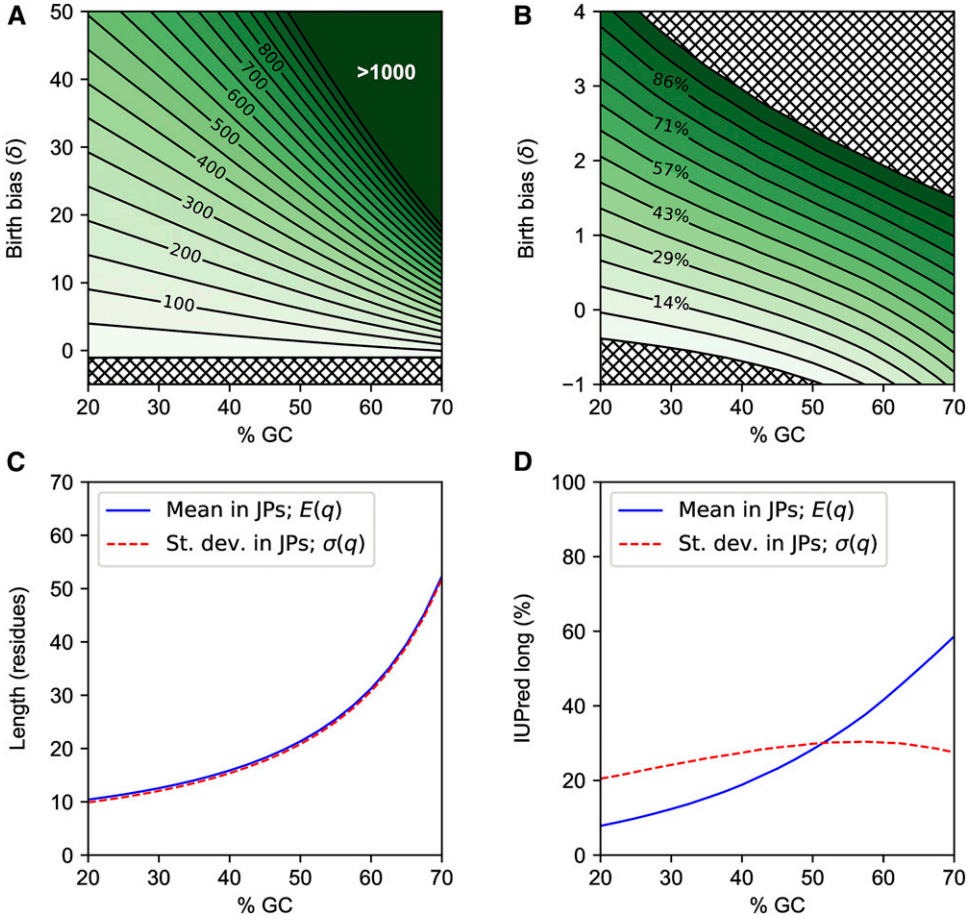

Figure 2 The mean length and mean structural disorder of novFPs are predicted by the birth bias and the genomic GC content under a simple model of the sequences of JPs. (A) Contour plot of the predicted average length of novFPs in amino acid residues. (B) Contour plot of the predicted average of IUPred long disorder (Dosztányi *et al.* 2005) among novFPs. (C) The predicted mean and SD of the length of JPs as functions of the GC content. (D) The predicted mean and SD of IUPred long disorder among JPs as functions of the GC content. Hatched areas indicate impossible scenarios, that is, negative polypeptide lengths and ISD percentages outside the 0–100% interval. Possible values of the birth bias differ between length and ISD because the ranges of these two properties constrain their respective means and SDs differently. The landscapes in A and B can be understood as the results of applying Equation 2 to the curves in C and D, respectively. As a result, the vertical "slice" of a landscape at a given GC content is a straight line whose intercept and slope are respectively the mean and SD associated with this GC content in the corresponding bottom panel. The curve obtained by taking a horizontal slice where there is no birth bias ($\delta = 0$) corresponds to the relation between the mean of the property among JPs, *i.e.*, the solid blue curve in the corresponding bottom panel. Since the vertical distance between contour lines is inversely proportional to the vertical slope of the landscape, it is inversely proportional to the SD of the property among JPs, *i.e.*, the dashed red curve in the corresponding bottom panel. The solid blue curves in C and D are consistent with known effects of an increase in GC content on random polypeptides, namely an increase in their mean length and mean ISD (Basile *et al.* 2017). The fact that contour lines are curved in A and B indicates that the effect of GC content on novFPs is not always proportional to its effect on JPs. As shown in C and D, such inconsistencies of the effect of GC content between JPs and novFPs are due to the fact that the GC content affects both the SDs and the means of the properties of JPs, sometimes in opposite directions. St. dev., standard deviation.

number among JPs to the time-averaged total number of JPs. The second distribution, which we call novFPs, is the distribution of all novFPs that emerge by functionalization during the time period considered. We use the term polypeptide property for a quantity that is determined by the sequence and *cis*-regulation of a polypeptide. Such properties include, for instance, the length of the polypeptide, the prevalence of some amino acid in its sequence, its proportion of disordered residues, and its expression level in a given *trans*-regulatory background. These polypeptide properties have distributions whose summary statistics can be used to compare JPs and novFPs. As illustrated in Figure 1C, the properties of JPs are expected to constrain those of novFPs. The equations shown here specify what the resulting constraints can be, and what can counteract them.

Based on the fact that any novFP must first exist as a JP before functionalizing, we find that the difference in the mean of any polypeptide property q between novFPs and JPs is given by the following equation:

$$E_F(q) - E(q) = cov(q, \hat{r}), \quad (1)$$

where $E_F(q)$ is the expected value (the average) of q among novFPs (the subscript F specifies that a statistic describes novFPs rather than JPs), $E(q)$ is the average of q among JPs, $cov(q, \hat{r})$ is the covariance of q and $\hat{r}$ among JPs, and $\hat{r}$ is the factor by which the relative frequency of a specific polypeptide changes from JPs to novFPs (for instance, because of its fitness effect). Equation 1 is analogous to the Robertson–Price identity from quantitative genetics (Robertson 1966; Price 1970; Lynch and Walsh 1998) which states that during a round of natural selection in a population, the mean of a quantitative phenotypic trait changes by an amount equal to the initial covariance of this trait with relative fitness. This analogy between functionalization and natural selection is due to the fact that they both involve the comparison of distributions between two populations, where the second population (*e.g.*, novFPs or postselection individuals) only features trait values that were present in the first population (*e.g.*, JPs or preselection individuals). This is because both

functionalization and natural selection act on preexisting material without producing novelty on their own.

To better distinguish the effect of the distribution of a polypeptide property among JPs from the effects of other factors, Equation 1 can be transformed into:

$$E_F(q) = E(q) + \sigma(q) \times \delta, \tag{2}$$

$$\text{with } \delta \text{ defined as } CV(\hat{r}) \times \rho(q, \hat{r}),$$

where $\sigma(q)$ is the SD of q among JPs, $CV(\hat{r})$ is its coefficient of variation of $\hat{r}$ among JPs (the ratio of its SD to its mean), and $\rho(q, \hat{r})$ is the Pearson correlation coefficient of q and $\hat{r}$ among JPs. Because of the mathematical properties of the correlation coefficient, the value of δ does not depend on the mean and variance of q among JPs, but rather on its relation with functionalization as symbolized by $\hat{r}$. We call the parameter δ the birth bias of the polypeptide property q. The birth bias is equal to the average difference in q between novFPs and JPs, measured in units of SD of q among JPs. It is thus analogous to the "intensity of selection" in quantitative genetics (Matsumura *et al.* 2012).

Equation 2 has implications for the use of random and noncoding controls in the study of novFPs. Such controls were often used to compute expected means (or other measures of central tendency) for polypeptide properties (Ángyán *et al.* 2012; Abrusán 2013; Basile *et al.* 2017; Wilson *et al.* 2017). According to Equation 2, the SDs of the properties of control sequences could be just as useful as their means for predicting the properties of novFPs, provided that the control is representative of real JPs. Given such a representative control, Equation 2 can be used to estimate the birth bias (δ) of a polypeptide property. Thus, to interpret average differences between JPs and novFPs given a model of JPs, we need to decompose the birth bias into contributions from different evolutionary forces.

Neutral evolutionary forces can cause discrepancies between JPs and novFPs through the rate of allelic turnover of JPs

We sought to further dissect the birth bias (δ in Equation 2) into readily interpretable components. Once we make the additional assumption that JPs are at evolutionary equilibrium—*i.e.*, on average over time, each region of the space of possible polypeptides is entered by mutant JPs as frequently as it loses JPs through allele loss or functionalization—then the birth bias becomes:

$$\delta = CV(\lambda f) \times \rho(q, \lambda f), \tag{3}$$

where λ is the polypeptide-specific rate of allelic turnover and f is the polypeptide-specific probability of functionalization. The rate of allelic turnover λ is the inverse of the expected time from the appearance of a specific JP by mutation to either the loss of its allele or its functionalization. Because of our assumption of evolutionary equilibrium, the landscape of λ across the space of polypeptides measures how fast a single locus can explore a given region of this space. For instance, JPs that evolve slowly (*e.g.*, because of selective constraints on their toxicity and metabolic cost or a low propensity to mutation) will have low values of λ; these polypeptides tend to persist in the population and contribute less to the exploration of polypeptide space than those with a high λ. The polypeptide property f is the probability that the appearance of a given JP by mutation will lead to its functionalization rather than its loss through the fixation of a nonsynonymous or *cis*-regulatory change. In other words, among the events of appearance of a specific JP by mutation, f is the proportion of such events which lead to the functionalization of this JP.

Since the rate λ and the probability f each take a single value in each possible JP, they summarize which genetic backgrounds and environments a JP is likely to encounter at the moment of its appearance and during its existence, and how these factors would determine the longevity of the JP and whether or not it functionalizes. There is no obvious relationship between λ and f : the former is the inverse of the expected time to either one of two events (allele loss or functionalization), while the latter is the probability that one of these events (functionalization) happens before the other (allele loss). However, there are specific scenarios in which λ and f are closely related. For instance, in a case where functionalization would be mainly driven by random environmental and genetic-background changes which are equally likely to favor any JP, the probability that a JP functionalizes before allele loss (f) would be directly proportional to its expected life span ($1/\lambda$) and the product λf would thus be a constant. This would lead to δ being zero for all polypeptide properties (Equation 3) and thus to JPs and novFPs being indistinguishable in terms of their average properties (Equation 2). In fact, they would be indistinguishable in terms of the distributions of properties (not only their averages), since the distribution of a property q is fully determined by the averages of properties of the form q^n, where n is a positive integer. In other words, an inverse proportionality between λ and f would mean that an unbiased sample of JPs become novFPs.

Because of the mathematical properties of the coefficient of variation and the correlation coefficient, the birth bias is insensitive to the scales of q, λ, and f. As a result, each of them can be replaced with a directly proportional quantity without changing the birth bias. This may help to model the birth bias and to estimate it from the observed properties of JPs. For instance, if a model of the evolution of JPs assumed that the allelic turnover rate of a JP is directly proportional to the GC content of its ORF (their ratio is a constant), then λ could be simply replaced by the ORF's GC content in Equation 3.

When the birth bias of a polypeptide property is not zero, its mean differs between JPs and novFPs. While Equation 3 could be used to compute this birth bias given enough data or assumptions about the junk proteome and its evolution, it does not highlight intuitive possible explanations for the existence

of such a bias. To do this, and without adding any assumption to those behind Equation 3, we obtained the following expression of the birth bias using identities from probability theory:

$$\delta = \frac{CV(\lambda) \times \rho(q,\lambda) + CV(f) \times \rho(q,f) + CV(\lambda) \times CV(f) \times cosk(q,\lambda,f)}{1 + CV(\lambda) \times CV(f) \times \rho(\lambda,f)}, \quad (4)$$

where $cosk(q,\lambda,f)$ is the coskewness of the three variables q, λ, and f among JPs.

The second term of the numerator in Equation 4 confirms previous intuitions about the probability of functionalization. All else being equal, an increase in the correlation between the probability of functionalization and a given polypeptide property results in an increase of this property's birth bias and thus of its mean among novFPs (Equation 2). More surprisingly, the first term of the numerator indicates that the same relation exists between the birth bias and the rate of allelic turnover λ. This implies that even if a given polypeptide property does not correlate positively with the probability of functionalization through a positive effect on fitness, the mean of this property can still be different between JPs and novFPs if it is positively correlated with the rate of allelic turnover of JPs. For example, if functionalization were equally likely for JPs of any given length, and if the allelic turnover of long JPs were especially fast because JPs mutate at a frequency proportional to their length, then longer polypeptides would be overrepresented among novFPs relative to JPs. In other words, the frequency of successes (events of functionalization) depends as much on the frequency of trials (the allelic turnover rate) as on the probability of success for a single trial (the probability of functionalization). Consequently, before interpreting observed differences between the average properties of novFPs and those of random sequences as the results of natural selection, we should either show or explicitly assume that these differences are not caused by neutral components of the allelic turnover rate, such as mutational biases resulting from the specific mutation spectrum of the organism under study.

The coskewness that appears in the third term of the numerator in Equation 4 is, roughly speaking, a measure of how any of three variables linearly affects the correlation between the two others (see Supplemental Material, file "SuppMat_2019-06-20.pdf" for a formal explanation). Like the correlation coefficient, coskewness does not depend directly on the means and SDs of variables. Despite the difficulty of its interpretation in the context of Equation 4, it could be estimated from data on the allelic turnover rate and functionalization probability of JPs, or predicted from a model of their evolution.

The denominator of Equation 4 is necessarily positive (proof in Supplemental Material, file "SuppMat_2019-06-20.pdf") and indicates that the overall correlation between the allelic turnover rate and the probability of functionalization negatively affects the magnitude of the birth bias. Interestingly, this term does not involve q, which means that its value is the same for every polypeptide property in a given species. It can be thought of as a measure of the overall tendency of the junk proteome to preferentially explore polypeptides that are likely to functionalize. It constitutes a baseline to which each source of evolutionary bias represented in the numerator must compare favorably to have a strong effect on the average properties of novFPs.

A simple model of the mean length and mean ISD of novFPs as functions of the birth bias and the genomic GC content

Length and secondary structure are properties of polypeptides that can be studied from DNA sequences generated *in silico*, which makes them ideal targets for the modeling of JPs and novFPs. In particular, ISD, which measures a protein's lack of stable tridimensional structure, has been a recurrent topic in previous studies of novel polypeptides (Ángyán *et al.* 2012; Basile *et al.* 2017; Wilson *et al.* 2017). To exemplify the usefulness of our general equations for modeling purposes, we used them to build a model of the mean length and mean ISD of novFPs as functions of the birth bias and the genome-wide GC content.

GC content is known to vary among genomes, typically from 20 to 70% (Long *et al.* 2018). If GC content affects the birth process of novel genes, it could make their properties highly dependent on the species' genomic content. We built a simple model where the sequences of JPs are randomly generated by a single GC content. We used this model to predict the means and SDs of length and ISD among JPs as functions of the GC content. Length predictions were made analytically. For ISD we used the model to simulate 100,000 polypeptide sequences for each of several GC contents from 20 to 70% and we estimated their individual ISD levels using the sequence-wide average of IUPred long disorder (Dosztányi *et al.* 2005). We applied Equation 2 to the resulting means and SDs to compute the expected means of length and ISD among novFPs as functions of their respective birth biases and the GC content (Figure 2).

Figure 2 illustrates the importance of modeling both the SD and the average of a polypeptide property among JPs when trying to predict its average among novFPs. The same birth bias has a larger effect on novFPs when JPs are more diverse, just like phenotypic variation in a population makes

the average of a phenotypic trait more sensitive to natural selection. In the case of polypeptide length, the SD of the length of JPs increases with GC content (Figure 2C), which makes the mean length of novFPs in GC-rich genomes especially sensitive to both neutral and selective sources of birth bias. To a lesser extent, this seems to also be the case for ISD: the mean ISD of novFPs increases less steeply with the birth bias when the GC content is low (Figure 2B), because the SD of ISD among JPs is lower (Figure 2D).

The equations and simulations underlying Figure 2 can be combined with values of genomic GC content, average length, and average ISD from the literature to estimate actual values of the birth bias in real species. We exemplify this process in the case of ISD in the budding yeast *Saccharomyces cerevisiae* and the house mouse *Mus musculus*. In the house mouse, *in silico* predictions suggest that novFPs have higher ISD than potential polypeptides encoded by intergenic DNA, which was interpreted as a result of natural selection in favor of high ISD during *de novo* gene birth (Wilson *et al.* 2017). Other results suggest that this trend may be specific to certain organisms and certain values of genomic GC content (Basile *et al.* 2017). As the average GC content of house mouse DNA is 42% (Elhaik and Graur 2014) and the average IUPred long disorder of its novFPs is close to 55% (Wilson *et al.* 2017), our model predicts that the birth bias of this specific measure of ISD should be >1 in house mouse (Figure 2B), more precisely 1.2. This value being larger than zero is consistent with the conclusion of Wilson *et al.* (2017) that novFPs appear more disordered than the raw material of *de novo* gene birth, assuming that the noncoding control sequences they used are well summarized by a single GC content that is close to 42%. By a similar reasoning, given the 38% GC content observed in yeast DNA (Engel *et al.* 2014) and the 32% average IUPred long disorder of yeast novFPs reported by Wilson *et al.* (2017), the associated birth bias should be between 0 and 1 (Figure 2B), more precisely 0.5. Under our GC-content-based model of JPs, this suggests that given the GC contents of mouse and yeast genomes, the biases of allelic turnover and functionalization in favor of disordered polypeptides are stronger in the mouse than in yeast. As shown by the various terms in Equation 4, many different scenarios could explain this trend in terms of the relation between ISD, the allelic turnover rate of JPs, and their probability of functionalization, which calls for further observation and modeling of these variables.

Discussion

The determinants of the properties of polypeptides resulting from *de novo* gene birth were previously studied empirically by comparing them to random and noncoding sequences (Ángyán *et al.* 2012; Abrusán 2013; Basile *et al.* 2017; Lu *et al.* 2017; Wilson *et al.* 2017), but the field lacked the theoretical tools needed to interpret these comparisons in terms of evolutionary forces. We have defined a classification of polypeptides and their evolutionary history (Figure 1) that clarifies the process of *de novo* gene birth sufficiently to link the properties of its raw material to those of its products through broadly applicable equations. These equations suggest potential roles for both natural selection and neutral forces in biasing the mean properties of novFPs with respect to the properties of the raw material from which they are born. We also showed how a simple GC-content-based model of nonfunctional polypeptides can be combined with our general theoretical framework to infer evolutionary parameters of *de novo* gene birth from the properties of its products, or vice versa.

Various terms have been used to classify polypeptides in studies of *de novo* gene birth, but they do not have exact equivalents in the JP–novFP–derFP classification. It is worth noting that novFPs will be more frequent among polypeptides that are called *de novo* or "novel," although the latter most often correspond to relatively young derFPs (McLysaght and Hurst 2016) and sometimes include JPs (Lu *et al.* 2017). On the other hand, the term "protogene" seems to encompass ORFs encoding JPs, novFPs, and young derFPs (Carvunis *et al.* 2012) since it is associated with viewing *de novo* gene birth as a continuous process.

Our framework can be used to investigate the effects of the relative importance of two types of events that may drive functionalization: (1) "external" changes in the genetic background and the environment, and (2) beneficial mutations at JP-expressing loci. In a hypothetical case where functionalization would only be driven by external changes, each JP would first appear in an unfavorable combination of genetic background and environment, and functionalization would be systematically caused by the occurrence of the "right" external change during the existence of a JP. In such an extreme scenario, the probability of functionalization of any given JP (f) would be the product of its expected life span $(1/\lambda)$ with the rate of external changes that lead to its functionalization. This rate of favorable external changes would thus correspond to the product λf in Equation 3 and would be the only determinant of the birth bias of each polypeptide property. In such a case, modeling differences between JPs and novFPs would only require modeling how the sequence and *cis*-regulation of each JP relate to the rate of external changes that favor this JP, rather than separately modeling both the rate of allelic turnover and the probability of functionalization. This scenario is especially likely if the favorable external changes in question are genetic-background changes that can be conserved by selection along with the JP that they favor. However, if the favorable changes are environmental, they may be reverted before the functionalization of the JPs that they favor. In the more specific case where the rate of favorable external changes would not vary between JPs, the birth bias would be zero for all polypeptide properties and novFPs would be undistinguishable from JPs in all regards, as we mentioned in the *Results* section.

In a hypothetical case where functionalization always follows the appearance of a JP with certain "good" properties that do not depend on the genetic background or the environment, the probability of functionalization would be

essentially binary: favored JPs would functionalize with a probability of one, others with a probability of zero. The product λf in Equation 3 would equal the allelic turnover rate in favored JPs and would equal zero in other JPs. In such a scenario, the properties of nonfavored JPs and their rate of allelic turnover would not have to be modeled, since they would have no effect on the properties of novFPs. This extreme limitation of functionalization by the mutation of individual JPs seems especially likely in species with large effective population sizes and a stable environment. Genome evolution tends to be slower when the effective population size is large (Lanfear *et al.* 2014), which may stabilize the genetic background and thus the selection coefficients of JPs. Moreover, the weakness of genetic drift would allow natural selection to maintain beneficial JPs and eliminate their deleterious modifications, making these JPs likely to functionalize.

Although our framework can be used to develop models for a wide diversity of scenarios, these models are always formulated in terms of the distributions of properties of JPs rather than the evolution of individual JP-expressing loci. Our framework offers no ways to predict, for instance, whether or not a single mutation can modify the rate of allelic turnover of a JP without modifying its probability of functionalization. Such considerations may be useful to understand why the junk proteome reaches a particular state of equilibrium with particular correlations between the properties of JPs. However, once this state of equilibrium is modeled or empirically estimated, our equations can be applied without thinking about the evolution of individual loci.

Implications of the length distribution of JPs have been largely ignored, since studies of *de novo* gene birth usually use random or noncoding controls that are intentionally biased against short ORFs (Ángyán *et al.* 2012; Abrusán 2013; Basile *et al.* 2017; Lu *et al.* 2017; Neme *et al.* 2017; Wilson *et al.* 2017). Such practices may be partly justifiable if the contribution of short JPs to *de novo* gene birth turns out to be negligible because of slow allelic turnover or low probability of functionalization, but this remains to be shown. Even though inferred novFPs tend to be shorter than derFPs, their average length is usually at least 100 residues (Neme and Tautz 2013; Basile *et al.* 2017), which is larger than the expected mean length of JPs for common GC contents (Figure 2C). Thus, if polypeptides that were detected and classified as novel are representative of novFPs, the birth bias of polypeptide length is likely to be positive in many species. As explained in our interpretation of Equation 4, this would not necessarily mean that a long polypeptide is typically more likely to functionalize than a shorter one; it could also have a faster allelic turnover. The fact that derFPs tend to be longer than novFPs is also not conclusive evidence for such a selective advantage of length among JPs, since the evolution of derFPs may be channeled toward long polypeptides that are very different from JPs of the same length. One intuitive alternative explanation for novFPs being longer than JPs is that since the mutation rate of an ORF is proportional to its length, the length of JPs may be positively correlated with their rate of allelic turnover, which would increase the birth bias of polypeptide length as shown in Equation 4. However, the strength of this correlation depends on how much variation in the rate of allelic turnover is independent of ORF length (such as the allelic turnover of promoter sequences), and its contribution to the birth bias also depends on the overall correlation between the rate of allelic turnover and the probability of functionalization, as shown by the denominator of Equation 4. It is therefore currently difficult to tell if this effect is strong enough to fully explain the observed shift in mean length between random ORFs and those expressing novFPs, although this point could be clarified by modeling or estimating the allelic turnover rate and probability of functionalization of JPs.

Despite the ambiguity as to the causes of the apparent length difference between JPs and novFPs, this difference should increase with the SD of the length of JPs, and thus with GC content, unless this effect cancels out with a decrease of δ in GC-rich genomes (Figure 2A). For instance, the mutation spectrum of an organism affects both its genomic GC content and the relation between the sequence of an ORF and its mutation rate, which could lead to an interspecific correlation between GC content and birth bias for various polypeptide properties.

Our positive estimates of δ for the sequence-wide average of IUPred long disorder in mouse and yeast may reflect positive correlations of this polypeptide property with the allelic turnover rate and/or the functionalization probability (see Equation 4). However, IUPred long disorder is an estimator of ISD and we can only assume that the landscape of the actual average proportion of disordered amino acid residues in novFPs is similar to Figure 2B under the GC-content-based model. As a warning against this assumption, the corresponding landscape computed from IUPred short disorder differs in terms of the magnitude of δ because the means and SDs of long and short predictors of ISD among JPs have different relations to the GC content (Figure S1), even though they are both meant to estimate the proportion of disordered residues. Nevertheless, the difference between mouse and yeast in the estimated birth bias for the same measure of ISD suggests that the difference in average ISD between their novFPs is in part driven by a difference in birth bias rather than only differences in the mean and SD of ISD among JPs. Future studies may reveal that some components of the birth bias (such as mutational biases) can be predicted from the GC content, which would make the latter even more useful than Figure 2 suggests for the prediction of interspecific differences in the average properties of novFPs.

When using a model of JPs to infer the effect of birth bias on the properties of novFPs (*e.g.*, Figure 2), the use of Equation 2 is inherently valid since this equation stems from the definitions of JPs and novFPs. However, the means and SDs of the properties of JPs, which are needed to apply Equation 2, are model dependent. Therefore, the predictions that we made in Figure 2 as to the value of δ under a GC-content-based model may not apply to organisms where JPs are not well described

by such a model. For instance, mammalian genomes are known to be organized into compositional domains with various GC contents (Elhaik and Graur 2014). In a study in yeast, candidate *de novo* genes had a significant tendency to be located in GC-enriched regions of the genome (Vakirlis *et al.* 2017). If several different GC contents contribute to a single junk proteome, the means and SDs of the properties of JPs may be different from those expected under our random-sequence model given the average GC content. However, it is possible that such a model will apply to each compositional domain separately, in which case the junk proteome would be readily modeled by drawing values of GC content from an appropriate distribution and using each value to generate a random polypeptide. From there, Equations 1–4 would apply just as they did for the simpler case of a single GC content.

Although the predictions that we made using a random-sequence model of JPs only involve their sequence and structure, *cis*-regulatory aspects of their expression may also be understood as polypeptide properties and analyzed using equations presented here. Transcription and translation levels of JP-encoding ORFs seem especially relevant since, as they approach zero, the probability that a JP functionalizes also goes toward zero and its other properties become irrelevant. Since transcription and translation are controlled by local sequence elements, knowledge of these elements may eventually be combined with a random-sequence model to predict the regulatory properties of JPs, like we did for their length and ISD. Studies of the transcriptional activity of synthetic random DNA in *Escherichia coli* (Yona *et al.* 2018) and yeast (de Boer *et al.* 2018) show that such sequences frequently contain the patterns required for the initiation and regulation of transcription. Factors that are external to intergenic regions also seem to play a role in the expression of JPs and their functionalization, such as bidirectional promoters (Vakirlis *et al.* 2017), translated UTRs, and translated alternative ORFs within canonical ORFs (Vanderperre *et al.* 2013). Understanding the importance of these factors may require more than a simple random-sequence model, but their effects on JPs should be "inherited" by novFPs in accordance with the general equations that we developed.

By empirical and experimental means, the reality of *de novo* gene birth has been made undeniable, and yet the community's quantitative understanding of this process suffers from the scarcity of theoretical contributions to the field. Our results specify how knowledge of the structure, expression, and evolution of the nonfunctional proteome can be used to explain and predict the properties of novFPs. However, much of this knowledge remains to be uncovered by further empirical, experimental, and theoretical investigation.

Acknowledgments

The authors thank C.K. Griswold, H. Martin, A.R. Carvunis, J. Masel, and three anonymous referees for their comments on the manuscript. L.N.-T. is supported by an Alexander Graham Bell Ph.D. fellowship from the Natural Sciences and Engineering Research Council (NSERC). C.R.L. is supported by a Discovery grant from NSERC and a team grant from the Fonds de Recherche du Québec – Nature et Technologies (FRQNT) and holds the Canada Research Chair in Evolutionary Cell and Systems Biology.

Author contributions: L.N.-T. conceived the project with the help of C.R.L. L.N.-T. performed all research, analysis, and interpretation. L.N.-T. wrote the manuscript with feedback from C.R.L.

Literature Cited

Abrusán, G., 2013 Integration of new genes into cellular networks, and their structural maturation. Genetics 195: 1407–1417. https://doi.org/10.1534/genetics.113.152256

Ángyán, A. F., A. Perczel, and Z. Gáspári, 2012 Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? FEBS Lett. 586: 2468–2472. https://doi.org/10.1016/j.febslet.2012.06.007

Basile, W., O. Sachenkova, S. Light, and A. Elofsson, 2017 High GC content causes orphan proteins to be intrinsically disordered. PLoS Comput. Biol. 13: e1005375. https://doi.org/10.1371/journal.pcbi.1005375

Bohrnstedt, G. W., and A. S. Goldberger, 1969 On the exact covariance of products of random variables. J. Am. Stat. Assoc. 64: 1439–1442. https://doi.org/10.1080/01621459.1969.10501069

Breton, S., D. T. Stewart, S. Shepardson, R. J. Trdan, A. E. Bogan *et al.*, 2011 Novel protein genes in animal mtDNA: a new sex determination system in freshwater mussels (Bivalvia: Unionoida)? Mol. Biol. Evol. 28: 1645–1659. https://doi.org/10.1093/molbev/msq345

Carvunis, A. R., T. Rolland, I. Wapinski, M. A. Calderwood, M. A. Yildirim *et al.*, 2012 Proto-genes and de novo gene birth. Nature 487: 370–374. https://doi.org/10.1038/nature11184

Chen, S., Y. E. Zhang, and M. Long, 2010 New genes in Drosophila quickly become essential. Science 330: 1682–1685. https://doi.org/10.1126/science.1196380

Çınlar, E., 2011 Conditioning, pp. 139–170 in *Probability and Stochastics* (Graduate Texts in Mathematics, Vol. 261), edited by S. Axler, and K. A. Ribet. Springer-Verlag, New York. 10.1007/978-0-387-87859-1_4 https://doi.org/10.1007/978-0-387-87859-1_4

de Boer, C. G., E. D. Vaishnav, R. Sadeh, E. L. Abeyta, N. Friedman *et al.*, 2018 Deciphering eukaryotic *cis*-regulatory logic with 100 million random promoters. bioRxiv doi: 10.1101/224907 (Preprint posted September 19, 2018). https://doi.org/10.1101/224907

Di Roberto, R. B., and S. G. Peisajovich, 2014 The role of domain shuffling in the evolution of signaling networks. J. Exp. Zool. B Mol. Dev. Evol. 322: 65–72.

Doolittle, W. F., T. D. Brunet, S. Linquist, and T. R. Gregory, 2014 Distinguishing between "function" and "effect" in genome biology. Genome Biol. Evol. 6: 1234–1237. https://doi.org/10.1093/gbe/evu098

Dosztányi, Z., V. Csizmok, P. Tompa, and I. Simon, 2005 IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21: 3433–3434. https://doi.org/10.1093/bioinformatics/bti541

Durand, E., I. Gagnon-Arsenault, J. Hallin, I. Hatin, A. K. Dubé *et al.*, 2019 Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like characteristics available for de novo gene emergence in wild yeast populations. Genome Res. 29: 932–943. https://doi.org/10.1101/gr.239822.118

Edwards, H., S. Abeln, and C. M. Deane, 2013 Exploring fold space preferences of new-born and ancient protein superfamilies.

PLoS Comput. Biol. 9: e1003325. https://doi.org/10.1371/journal.pcbi.1003325
Elhaik, E., and D. Graur, 2014 A comparative study and a phylogenetic exploration of the compositional architectures of mammalian nuclear genomes. PLoS Comput. Biol. 10: e1003925. https://doi.org/10.1371/journal.pcbi.1003925
Engel, S. R., F. S. Dietrich, D. G. Fisk, G. Binkley, R. Balakrishnan *et al.*, 2014 The reference genome sequence of Saccharomyces cerevisiae: then and now. G3 (Bethesda) 4: 389–398. https://doi.org/10.1534/g3.113.008995
Guttman, M., P. Russell, N. T. Ingolia, J. S. Weissman, and E. S. Lander, 2013 Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell 154: 240–251. https://doi.org/10.1016/j.cell.2013.06.009
Hausser, J., A. Mayo, L. Keren, and U. Alon, 2019 Central dogma rates and the trade-off between precision and economy in gene expression. Nat. Commun. 10: 68. https://doi.org/10.1038/s41467-018-07391-8
Heinen, T. J., F. Staubach, D. Häming, and D. Tautz, 2009 Emergence of a new gene from an intergenic region. Curr. Biol. 19: 1527–1531. https://doi.org/10.1016/j.cub.2009.07.049
Ingolia, N. T., G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. Talhouarne *et al.*, 2014 Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 8: 1365–1379. https://doi.org/10.1016/j.celrep.2014.07.045
Innan, H., and F. Kondrashov, 2010 The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11: 97–108. https://doi.org/10.1038/nrg2689
Jacob, F., 1977 Evolution and tinkering. Science 196: 1161–1166. https://doi.org/10.1126/science.860134
Jensen, T. H., A. Jacquier, and D. Libri, 2013 Dealing with pervasive transcription. Mol. Cell 52: 473–484. https://doi.org/10.1016/j.molcel.2013.10.032
Keese, P. K., and A. Gibbs, 1992 Origins of genes: "big bang" or continuous creation? Proc. Natl. Acad. Sci. USA 89: 9489–9493. https://doi.org/10.1073/pnas.89.20.9489
Landry, C. R., X. F. Zhong, L. Nielly-Thibault, and X. Roucou, 2015 Found in translation: functions and evolution of a recently discovered alternative proteome. Curr. Opin. Struct. Biol. 32: 74–80. https://doi.org/10.1016/j.sbi.2015.02.017
Lanfear, R., H. Kokko, and A. Eyre-Walker, 2014 Population size and the rate of evolution. Trends Ecol. Evol. 29: 33–41. https://doi.org/10.1016/j.tree.2013.09.009
Long, H., W. Sung, S. Kucukyildirim, E. Williams, S. F. Miller *et al.*, 2018 Evolutionary determinants of genome-wide nucleotide composition. Nat. Ecol. Evol. 2: 237–240. https://doi.org/10.1038/s41559-017-0425-y
Lu, T. C., J. Y. Leu, and W. C. Lin, 2017 A comprehensive analysis of transcript-supported de novo genes in Saccharomyces sensu stricto yeasts. Mol. Biol. Evol. 34: 2823–2838. https://doi.org/10.1093/molbev/msx210
Lynch, M., and G. K. Marinov, 2015 The bioenergetic costs of a gene. Proc. Natl. Acad. Sci. USA 112: 15690–15695.
Lynch, M., and B. Walsh, 1998 Covariance, regression, and correlation, pp. 35–50 in *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.
Matsumura, S., R. Arlinghaus, and U. Dieckmann, 2012 Standardizing selection strengths to study selection in the wild: a critical comparison and suggestions for the future. Bioscience 62: 1039–1054. https://doi.org/10.1525/bio.2012.62.12.6
McLysaght, A., and D. Guerzoni, 2015 New genes from non-coding sequence: the role of *de novo* protein-coding genes in eukaryotic evolutionary innovation. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370: 20140332. https://doi.org/10.1098/rstb.2014.0332
McLysaght, A., and L. D. Hurst, 2016 Open questions in the study of *de novo* genes: what, how and why. Nat. Rev. Genet. 17: 567–578. https://doi.org/10.1038/nrg.2016.78
Mouilleron, H., V. Delcourt, and X. Roucou, 2016 Death of a dogma: eukaryotic mRNAs can code for more than one protein. Nucleic Acids Res. 44: 14–23. https://doi.org/10.1093/nar/gkv1218
Neme, R., and D. Tautz, 2013 Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. BMC Genomics 14: 117. https://doi.org/10.1186/1471-2164-14-117
Neme, R., and D. Tautz, 2016 Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to *de novo* gene emergence. eLife 5: e09977. https://doi.org/10.7554/eLife.09977
Neme, R., C. Amador, B. Yildirim, E. McConnell, and D. Tautz, 2017 Random sequences are an abundant source of bioactive RNAs or peptides. Nat. Ecol. Evol. 1: 0217. https://doi.org/10.1038/s41559-017-0127
Neuhaus, K., R. Landstorfer, L. Fellner, S. Simon, A. Schafferhans *et al.*, 2016 Translatomics combined with transcriptomics and proteomics reveals novel functional, recently evolved orphan genes in *Escherichia coli* O157:H7 (EHEC). BMC Genomics 17: 133. https://doi.org/10.1186/s12864-016-2456-1
Price, G. R., 1970 Selection and covariance. Nature 227: 520–521 (erratum: Nature 228: 1011). https://doi.org/10.1038/227520a0
Reinhardt, J. A., B. M. Wanjiru, A. T. Brant, P. Saelao, D. J. Begun *et al.*, 2013 *De novo* ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. PLoS Genet. 9: e1003860. https://doi.org/10.1371/journal.pgen.1003860
Robertson, A., 1966 A mathematical model of the culling process in dairy cattle. Anim. Prod. 8: 108.
Ruiz-Orera, J., X. Messeguer, J. A. Subirana, and M. M. Albà, 2014 Long non-coding RNAs as a source of new peptides. eLife 3: e03523. https://doi.org/10.7554/eLife.03523
Ruiz-Orera, J., P. Verdaguer-Grau, J. L. Villanueva-Cañas, X. Messeguer, and M. M. Albà, 2018 Translation of neutrally evolving peptides provides a basis for *de novo* gene evolution. Nat. Ecol. Evol. 2: 890–896. https://doi.org/10.1038/s41559-018-0506-6
Schlötterer, C., 2015 Genes from scratch – the evolutionary fate of *de novo* genes. Trends Genet. 31: 215–219. https://doi.org/10.1016/j.tig.2015.02.007
Soucy, S. M., J. Huang, and J. P. Gogarten, 2015 Horizontal gene transfer: building the web of life. Nat. Rev. Genet. 16: 472–482. https://doi.org/10.1038/nrg3962
Toll-Riera, M., N. Bosch, N. Bellora, R. Castelo, L. Armengol *et al.*, 2009 Origin of primate orphan genes: a comparative genomics approach. Mol. Biol. Evol. 26: 603–612. https://doi.org/10.1093/molbev/msn281
Vakirlis, N. N., A. S. Hebert, D. A. Opulente, G. Achaz, C. T. Hittinger *et al.*, 2017 A molecular portrait of *de novo* genes in yeasts. Mol. Biol. Evol. 35: 631–645. https://doi.org/10.1093/molbev/msx315
Vanderperre, B., J. F. Lucier, C. Bissonnette, J. Motard, G. Tremblay *et al.*, 2013 Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS One 8: e70698. https://doi.org/10.1371/journal.pone.0070698
Vestrup, E. M., 2003 The radon-nikodym theorem, pp. 367–436 in *The Theory of Measures and Integration*. John Wiley & Sons, Hoboken, NJ.
Wilson, B. A., and J. Masel, 2011 Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol. Evol. 3: 1245–1252. https://doi.org/10.1093/gbe/evr099
Wilson, B. A., S. G. Foy, R. Neme, and J. Masel, 2017 Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nat. Ecol. Evol. 1: 0146. https://doi.org/10.1038/s41559-017-0146
Yona, A. H., E. J. Alm, and J. Gore, 2018 Random sequences rapidly evolve into *de novo* promoters. Nat. Commun. 9: 1530. https://doi.org/10.1038/s41467-018-04026-w

Communicating editor: J. Masel