CellPress

Opinion

Does Sequence Conservation Provide Evidence for Biological Function?

Seila Omer,[1] Timothy J. Harlow,[1] and Johann Peter Gogarten[1,2,*]

Finding a signature of purifying selection in a gene is usually interpreted as evidence for the gene providing a function that is targeted by natural selection. This opinion offers a very different hypothesis: purifying selection may be due to removing harmful mutations from the population, that is, the gene and its encoded protein become harmful after a mutation occurred, possibly because the mutated protein interferes with the translation machinery, or because of toxicity of the misfolded protein. Finding a signature of purifying selection should not automatically be considered proof of the gene's selectable function.

Relationship between Sequence Conservation, Purifying Selection, and Biological Function

Natural selection is one of the main drivers of the adaptation and diversification of organisms. Two types of selection are commonly distinguished: positive selection, also known as directional or diversifying selection, acts on mutations that increase organismal fitness; purifying selection acts on mutations that decrease the organismal fitness. Here we challenge the often made assumption that evidence for purifying selection can be equated to the gene under purifying selection making a positive contribution to organismal fitness. We discuss the possibility of mutations leading to toxic or detrimental products that interfere with normal cellular functions. DNA sequence conservation indicates that natural selection operates against the deleterious effects of allelic variants (also known as purifying selection). A popular approach to detect purifying selection in protein-coding DNA sequences is to infer an excess in the rate of synonymous substitutions (dS) relative to the rate of nonsynonymous (dN) substitutions within a set of homologous and very similar protein-coding sequences (Box 1) [1]. It is generally assumed that the rate of amino acid replacements, measured as a low rate of substitutions that lead to changes at the protein sequence level (usually abbreviated as dN), reflects natural selection through the structural and functional constraints imposed on the protein sequence. That is, if a mutation changes an amino acid, and the protein after mutation does not function, or functions less efficiently, the fitness of organisms carrying the mutated gene generally will be lower, and selection will tend to eliminate the gene from the population. A dN/dS ratio estimate significantly smaller than 1 indicates that some nonsynonymous substitutions were removed by natural selection. Because this type of selection removes mutations from the population it is known as purifying selection. In the rare instance that changes in the amino acid sequence prove to be advantageous and are driven to fixation through the increased fitness of the carrier of the mutated alleles, the rate of nonsynonymous substitutions is higher than the rate of synonymous substitutions. For nearly all protein-coding genes, purifying selection is much stronger than positive selection (also known as directional selection), and the latter is usually restricted to individual sequence positions in a multiple sequence alignment. Thus, when present across closely related taxonomic groups, protein-coding DNA sequences appear nearly identical, that is, they are conserved.

Trends

The current accepted definitions of biological function for a gene assume an evolutionary history shaped by natural selection. This assumption is based on the observation that most genes, characterized genotypically and phenotypically, display in their DNA sequence a considerable excess of synonymous substitutions over what would be expected, if the substitution process were random.

Many genes potentially detrimental for bacterial fitness, lacking functionally relevant expression, share a common evolutionary record with their host organisms through vertical descent. Their DNA sequences present the same type of selective footprints found in neighboring functional genes.

The observation that natural selection operates on both functional and putative nonfunctional genes challenges the default connection between sequence conservation and biological function.

[1]Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, CT 06269-3125, USA
[2]Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA

*Correspondence: gogarten@uconn.edu (J.P. Gogarten).

Box 1. dN/dS (a.k.a. Ka/Ks or ω) as a Test for Selection

$$dN = \frac{\text{number of nonsynonymous substitutions}}{\text{number of possible nonsynonymous substitutions}}$$

$$dS = \frac{\text{number of synonymous substitutions}}{\text{number of possible synonymous substitutions}}$$

Nonsynonymous (dN) and synonymous (dS) rates of substitutions are defined as the actual number of synonymous and nonsynonymous substitutions, respectively, relative to the number of all possible synonymous or nonsynonymous substitutions. The ratio of the two rates (dN/dS) measures the strength and direction of natural selection experienced by protein-coding DNA sequences. If mutations have no effect on fitness, the nonsynonymous rate is expected to be the same as the synonymous rate (dN/dS = 1). This neutral rate is commonly used as a null hypothesis against which the presence of selection is tested. If organisms with nonsynonymous mutations in a site are less fit than organisms without the mutation, purifying selection will purge them from lineages so that dN/dS <1. If nonsynonymous mutations are favored by selection, they will be fixed at a higher rate than synonymous mutations resulting in dN/dS >1. The latter is a rare occurrence and is known as positive, directional, or diversifying selection. Averaged over all possible amino acid encoding codons, a random mutation is about three times more likely to be nonsynonymous than synonymous. The redundancy of the genetic code is unevenly distributed over the three codon positions (depicted in Figure I). When considering all codons, a change in the 3rd codon position frequently results in synonymous substitutions, whereas changes in the 2nd codon position always are nonsynonymous. Figure I gives the estimated proportion of possible synonymous and nonsynonymous substitutions for the TTT triplet (UUU codon in the messenger RNA).

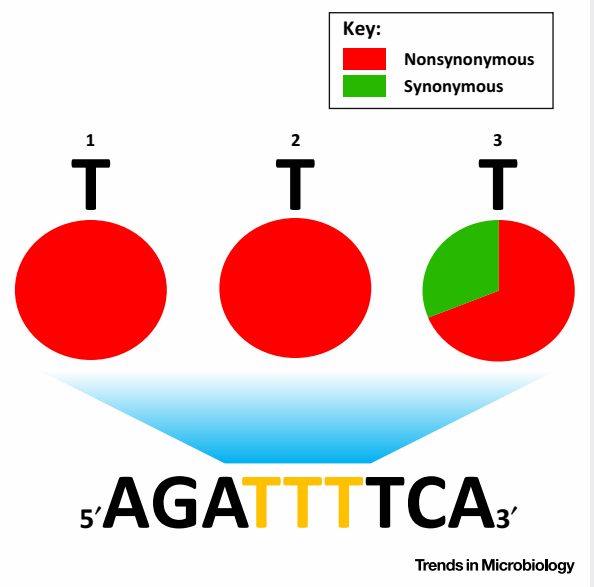

Figure I. Fractions of synonymous and nonsynonymous substitutions for the TTT triplet.

Recent studies have found patterns of purifying selection in bacterial genes with limited distribution for which the positive contribution to the organismal or the gene's fitness was not obvious: ORFans (genes with no recognizable homologs in other genomes) and group-specific genes [2,3], gene transfer agents (GTAs) [4], transposase [5], and prophage genes [6]. Analysis of dN/dS in orthologous genes (i.e., homologs that diverged together with the lineages) spanning the *Escherichia coli* and *Salmonella enterica* clades has found values much lower than 1, interpreted to reflect the selective constraints associated with important, functional roles for cell fitness [3]. A more recent study of group-specific genes within *E. coli* and *Shigella* clades and more widely distributed non-ORFan genes revealed dN values lower than the dS estimates, which led the authors to suggest that most ORFans are, in fact, functional genes [2]. Several studies [7,8] proposed that the observed purifying selection acting on GTA genes in *Rhodobacter capsulatus* and in the genus *Bartonella* argues for functional benefits that GTAs provide to

Table 1. dN/dS Estimates for Genes Not Expressed for Function

Gene/Organism[a]	dN/dS	Refs
ORFans/*Escherichia coli–Salmonella enterica*	0.19 ± 0.030 (average)	[3]
Prophage genes/*E. coli*	0.22 (median)	[6]
orfB IS 6110/Mycobacterium tuberculosis	0.58168	[4]
ORFans/*E. coli, Shigella, Salmonella*	<1[b]	[2]
tn/E. coli	0.1491	NCBI Gene ID: 7152325 + 28 syntenic homologs [c]
ltsu/Lactobacillus casei	0.0722	NCBI Gene ID: 6406903 + 6 syntenic homologs[c]
stsu/L. casei	0.1041	NCBI Gene ID: 6405444 + 6 syntenic homologs[c]
pks1/Burkholderia pseudomallei	0.3129	NCBI Gene ID: 4906294 + 17 syntenic homologs [c]

[a]Abbreviations: *tn*, transposase; *ltsu*, prophage large terminase subunit; *stsu*, prophage small terminase subunit; *pks1*, polyketide synthase (part of the malleilactone cryptic operon).
[b]The authors [2] calculated and reported the difference (dN-dS) not the ratio of dN and dS.
[c]Syntenic homologs were retrieved from Integrated Microbial Genomes (IMG) at Joint Genome Institute (JGI) https://img.jgi.doe.gov/ [36].

the host population or to the GTA genes themselves. More recently, the widespread signature of purifying selection detected in over 300 vertically inherited prophage sequences from *E. coli* and *S. enterica* strains (including structural and regulatory modules) was considered evidence for selection by the host or host population for phage-encoded functions [6]. Table 1 summarizes some of the evidence on the ubiquity of the phenomenon of purifying selection acting on genes without an obvious function beneficial to the gene (in case of selfish genetic elements) [9], the host, or the host population (as an altruistic act benefiting relatives known as kin selection) [10,11].

In this Opinion article, we challenge the default connection between sequence conservation and the functional status of the gene. We argue that low-level expression of a gene may be sufficient to generate a selective footprint on the encoding DNA sequence, that is, a dN/dS value smaller than 1. No additional selection related to the gene product's function may be necessary. In case of dN/dS values not much smaller than 1, assigning biological function to conserved genes must involve additional corroborating pieces of evidence linking the genes to particular traits.

Selection for Function: Other Explanations for Purifying Selection

The *E.coli* genome is estimated to contain upwards of 5% of nonfunctional elements (pseudogenes, defective mobile genetic elements). This dispensable fraction varies in bacterial genomes and experiences a rapid turnover across evolutionary timescales, being prone to deletions as a result of natural selection [12–14]. A recent study finds the energetic cost associated with replication of a DNA segment sufficient to trigger the action of natural selection in a large bacterial population [15]. Despite deletions occurring in nonfunctional genes, some genes are found in closely related strains in a repressed state with no apparent benefit for the organism harboring them, similar to the build-up of items in a car trunk (Figure 1) – these genes might fall victim to deletions in the future. A large number of such genes include mobile genetic elements (phages, transposases, etc.) and operons encoding toxic products with significant effects on bacterial fitness upon expression in certain environments [16]. Under stressful conditions, many dormant prophage genes responsible for phage packaging and lysis become expressed, leading to the demise of the bacterial host. Similar fitness effects are observed in case of transposases. If a transposase is functional, the transposition process via a cut-and-paste mechanism often results in mutagenic effects at the DNA level. Because most mutations caused by transposition in coding regions of the host genome are deleterious to host fitness,

Trends in Microbiology

Figure 1. Car Trunk Analogy for How Purifying Selection Can Act on Bacterial Genes. In many cars the trunk accumulates items most of which are selectively nearly neutral, such as old papers or broken CDs; a few items are useful under certain conditions (e.g., spare tires, tennis racquet, toolbox, map). These items correspond to genes that are neutral or under purifying selection, respectively. Functional genes (the useful items) experience strong selective pressures against mutations that perturb protein stability, lead to a loss of function and overall loss of organismal fitness (purifying selection). Consequently, many mutations that change the amino acid sequence are not kept in populations (analogous with removing a broken tennis racquet, now unusable). However, another reason for the presence of purifying selection is that mutations may lead to a product whose presence is detrimental (e.g., protein toxicity) to the organism. In the trunk of a car analogy, a piece of cheese left in the trunk will rot, producing an obnoxious smell, which will lead to its removal from the trunk. Similarly, a sack of fertilizer might turn into an explosive, prompting fast elimination. In this case, purifying selection occurs because the gene product after the mutation is detrimental, not because the gene before mutation was beneficial.

they will be removed by natural selection. The persistence of mobile genetic elements in bacterial lineages by natural selection raises the question of their biological function either for the benefit of the element itself or for the benefit of the host organism. This question is particularly puzzling for selfish genes that are retained in the same genomic location and were passed on vertically from parent to daughter cells. The debate between kin selection *versus* selfish gene hypotheses as an explanation for apparent altruistic behavior at the organismal level is ongoing in the field of microbial evolution [17]. Several studies involving altruism and cooperation have attempted to explain evolution of certain bacterial traits [18]. By contrast, a theoretical study on DNA secretion in bacteria shows that gene-level selection can be responsible for maintaining a gene promoting gene sharing in bacterial populations [19].

For each individual case of dN/dS values smaller than 1 that is discussed in this article one can invoke complicated scenarios on how selection for function could be acting on the gene in question. For example, a phage may be found in the same location, not because it was vertically inherited, but because the phage has a strong site preference. Even a prophage or transposase that was vertically inherited might have recombined with prophages or recombinases that actually were selected for function as part of their history [20]. Persistence and evolution of

prophage and elements of defective phage, bacteriocins, have been discussed in the context of kin recognition and kin selection, for example under the form of a poison–antidote mechanism [18]. The prophage's function might be to destroy other, less related bacteria after entering the lytic part of its life cycle, thereby creating new ecological niches for the host. Under this scenario the prophage remnant is no longer propagating as a phage, rather the function of lysis would be under group selection, with the cell whose prophage has been activated benefiting other members of the population. In addition, it is impossible to exclude the possibility that the gene in question has acquired a function not yet recognized by researchers, but seen by natural selection; for example, a gene expressed in a lineage for a long time will experience random interactions with other cell constituents, and these may evolve into required interactions through constructive neutral evolution [21], resulting in requirement for protein homeostasis [22] that may be the cause for purifying selection. However, the consistent detection of dN/dS ratios lower than 1 make unknown function or group selection an unlikely explanation, especially, if the gene has not traveled in a lineage for long periods of time.

Expressed Genes May Experience Purifying Selection Regardless of Their Functional Status

Recent studies have advanced the idea that the level of expression may be one of the main determinants of sequence conservation. In support of this assertion, highly expressed genes tend to evolve slower than lowly expressed genes [23,24]. However, a low level of gene expression may not necessarily imply functional relevance. A study analyzing the expression and fitness data for 3247 of the 4467 protein-coding genes in the *Shewanella oneidensis* MR-1 genome has shown that, under 15 laboratory growth conditions tested (stress included), some genes with putative detrimental effects on fitness have significantly higher expression than expected when compared with other genes [16]. These findings support the idea that gene expression regulation via genome-wide and local repression mechanisms in bacteria might not be optimal under some environmental conditions.

Furthermore, the expression levels of proteins in genomes have been shown to be anti-correlated with their evolutionary rate (measured in dN) [24]. Two main hypotheses have been proposed to explain this anti-correlation. First, the protein-misfolding-avoidance hypothesis states that there may be selection pressures acting against error-free and error-induced protein misfolding as a result of the incurring fitness costs. This is supported by the finding of a positive correlation between the expression level and unfolding energy of the polypeptides [23,25]. Second, the protein-misinteraction hypothesis proposes that natural selection also acts against errors in protein–protein interactions. This hypothesis is upheld by studies in yeast and *E. coli*, which suggest a positive correlation between the abundance of proteins and the proportion of charged residues on the surface of the proteins [26,27]. Additionally, differences in sequence conservation between functionally exposed regions of proteins and strictly structure-related regions decrease with increasing level of protein expression [28].

Protein synthesis, folding, and interactions occur in a busy, confined environment [29]. A computational study of an *E. coli* cell modelling the Brownian dynamics of only the 50 most abundant proteins suggests that macromolecular crowding has a considerable effect on protein folding and interactions [30]. Within this small intracellular space, the newly synthesized polypeptides need to fold into their native state, either independently or with the help of molecular chaperones. Because the compartment is very small relative to the size and number of the unfolded polypeptides, aggregation in toxic products is possible [31]. The toxicity of these aggregates is manifested as a 'gain-of-toxic-function' through interactions with many non-cognate partners [32]. In the car trunk analogy, a sack of forgotten fertilizer could become unstable and explode, becoming detrimental for the car's integrity.

Key Figure

Mutations Affecting Protein Structure May Be Subject to Purifying Selection.

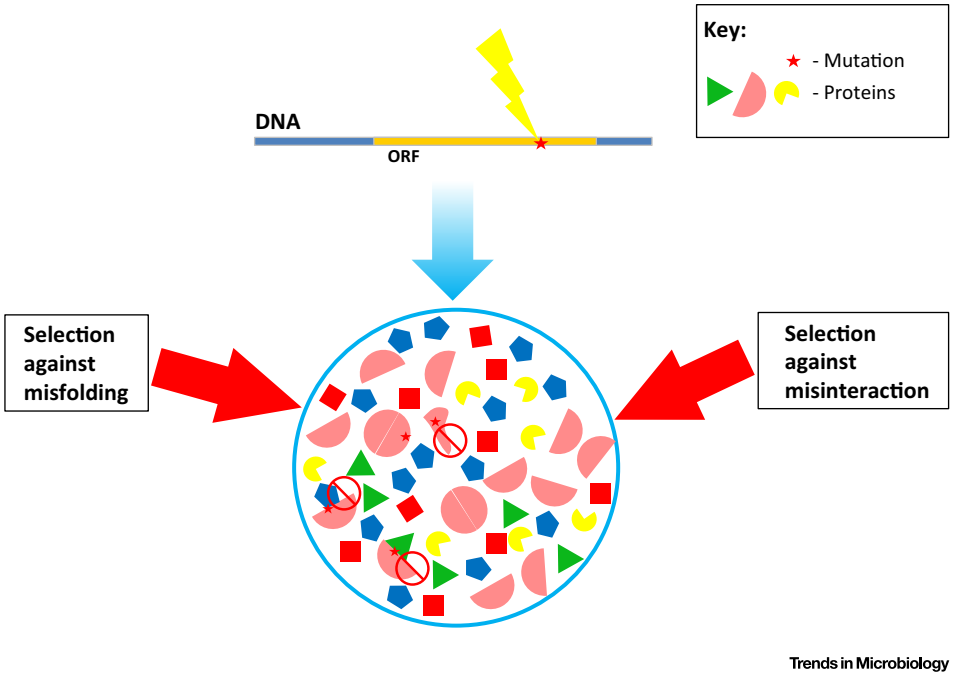

Figure 2. Mutations at the level of DNA may result in changes in the primary structure of proteins. Most changes propagate hierarchically in the newly made macromolecules, to their secondary and tertiary structures, affecting protein stability and interaction with cognate partners. Purifying election acts to remove such mutations, even in cases when protein expression is minimal. Therefore, no selective effects for protein function by the host organism are required to explain signatures of purifying selection in protein-coding DNA sequences.

Mutations can also alter the impact of an encoded gene product by affecting the actual process of protein synthesis. For instance, mutations that create rare codons alter the level of protein expression and ribosome turnover rates by causing ribosome stalling [33]. Furthermore, mutations that alter the secondary structure of an mRNA molecule might impact movement of the ribosomes and translation efficiency [34].

Collectively, these findings constitute arguments for natural selection targeting protein folding and interactions even in the absence of protein function (Figure 2, Key Figure).

We propose that purifying selection signatures detected in genes not expressed for function can result from the potentially deleterious effects of mutations on protein folding and protein interactions. Because prophage structural proteins, transposases, and assembly-line enzymes tend to have multiple interacting partners (other proteins, DNA or RNA), we speculate that their evolution is constrained even when their expression level is low and functionally irrelevant.

Given the mutational target size in bacteria, mutations that inactivate promoters and prevent transcription and translation may be less frequent than mutations that directly affect the open

reading frames (ORFs). Regulatory mutations that prevent transcription can drive rapid inactivation (pseudogenization) of genes not expressed for function. In comparison, mutations in coding regions bring structural and functional constraints on the expressed products, even when the expression level is very low. For example, missense mutations that alter initiation codons or nonsense mutations anywhere within ORFs can result in shorter translation products. These truncated ORFs will still be subject to purifying selection as long as occasional transcriptional and translational events happen.

A comparison between the dN/dS values for the genes we have included in this article and those determined for native genes within the *E. coli– S. enterica* clade (dN/dS = 0.05 ± 0.001) reveals a higher dN rate at similar timescales in case of phenotypically silent genes than for known functional genes [3]. By contrast, a study on *E. coli* prophage genes has advanced the idea that moderately low dN/dS values reflect selection by bacteria for the functions encoded by the phage genes [6]. While we do not dispute that domestication of some phage genes as standalone functional elements can occur, we put forward that, alternatively, this process could be one of the consequences of phage gene coexpression and coexistence with bacterial host proteins within same environment (where interactions are inevitable) rather than resulting from a fitness increase of the bacterial population.

We do not consider the mere fact of providing a template for transcription as providing a function; rather, we use the typical biological definition for function that implies a contribution to the fitness of the organism or to the gene itself [35]. Couched in these terms, our hypothesis is that a gene being transcribed and translated at low levels is sufficient to create a signature of purifying selection; a positive contribution of the gene to the gene's, host's, or group's fitness is not necessary.

Concluding Remarks

There is growing evidence that phenotypically silent genes maintained in bacterial lineages with no apparent functional role, but with possible detrimental effects upon expression, are evolving under purifying selection. Therefore, the mere presence of such selective signatures in the protein-coding DNA sequences should not be used as the only argument to indicate function of the encoded protein. Pseudogenization through mutations of the start codon, frameshift, or nonsense mutations are the likely long-term fate of ORFs that do not contribute towards function. However, along the route towards pseudogenization, the ORF may remain subject to purifying selection, as long as it retains a minimal residual expression. A documented dN/dS <1 is necessary but not sufficient to prove biological function for a gene within an organism. In the context of high-throughput sequencing and genome analysis, it is important to develop clear criteria for determining the functional importance from substitution rates. To prove the positive contribution of a gene to the fitness of the organism (or to the gene itself in case of a selfish genetic element), either dN/dS values much smaller than the ones reported for genes without apparent function, or a clear understanding of the encoded product and data that connect the genotype to the phenotype, may be needed (see Outstanding Questions).

Outstanding Questions

How could we design a selection test using known protein structure data which can differentiate between the effect of mutations on structure and effects on function?

What types of evidence, beyond the evolutionary argument, should be included to assign potential biological functions to genes?

What fraction of the bacterial protein-coding genes is 'junk-in-the-trunk' and what fraction is actually under functional constraints?

References

1. Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736
2. Yu, G. and Stoltzfus, A. (2012) Population diversity of ORFan genes in *Escherichia coli*. *Genome Biol. Evol.* 4, 1176–1187
3. Daubin, V. and Ochman, H. (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* 14, 1036–1042
4. Lang, A.S. *et al.* (2012) Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.* 10, 472–482
5. Thabet, S. *et al.* (2015) Evolutionary trends of the transposase-encoding open reading frames A and B (*orfA* and *orfB*) of the mycobacterial IS6110 insertion sequence. *PLoS ONE* 10, e0130161
6. Bobay, L-M. *et al.* (2014) Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12127–12132
7. Guy, L. *et al.* (2013) A gene transfer agent and a dynamic repertoire of secretion systems hold the keys to the explosive radiation of the emerging pathogen *Bartonella*. *PLoS Genet.* 9, e1003393

8. Lang, A.S. and Beatty, J.T. (2007) Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* 15, 54–62
9. Dawkins, R. (1989) *The Selfish Gene, Oxford.*
10. Hamilton, W.D.D. (1964) The genetical evolution of social behaviour. I. *J. Theor. Biol.* 7, 1–16
11. Hamilton, W.D. (1964) The genetical evolution of social behaviour. II. *J. Theor. Biol.* 7, 17–52
12. Ochman, H. and Davalos, L.M. (2006) The nature and dynamics of bacterial genomes. *Science* 311, 1730–1733
13. Kuo, C.H. and Ochman, H. (2010) The extinction dynamics of bacterial pseudogenes. *PLoS Genet.* 6, e1001050
14. Koskiniemi, S. *et al.* (2012) Selection-driven gene loss in bacteria. *PLoS Genet.* 8, e1002787
15. Lynch, M. and Marinov, G.K. (2015) The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci.* 112, 201514974
16. Price, M.N. *et al.* (2013) Indirect and suboptimal control of gene expression is widespread in bacteria. *Mol. Syst. Biol.* 9, 660
17. Olendzenski, L. and Gogarten, J.P. (2009) Evolution of genes and organisms: the tree/web of life in light of horizontal gene transfer. *Ann. N.Y. Acad. Sci.* 1178, 137–145
18. Strassmann, J.E. *et al.* (2011) Kin discrimination and cooperation in microbes. *Annu. Rev. Microbiol.* 65, 349–367
19. Draghi, J.a. and Turner, P.E. (2006) DNA secretion and gene-level selection in bacteria. *Microbiol. Read. Engl.* 152, 2683–2688
20. Castillo-Ramírez, S. *et al.* (2011) The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog.* 7, e1002129
21. Gray, M.W. *et al.* (2010) Irremediable complexity? *Science* 330, 920–921
22. Bershtein, S. *et al.* (2015) Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in bacteria. *PLoS Genet.* 11, e1005612
23. Drummond, D.A. *et al.* (2005) Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14338–14343
24. Pál, C. *et al.* (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158, 927–931
25. Geiler-Samerotte, K.a. *et al.* (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 108, 680–685
26. Yang, J-R. *et al.* (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl. Acad. Sci. U.S.A.* 109, E831–E840
27. Plata, G. *et al.* (2010) The rate of the molecular clock and the cost of gratuitous protein synthesis. *Genome Biol.* 11, R98
28. Eames, M. and Kortemme, T. (2007) Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. *Struct. Lond. Engl.* 15, 1442–1451
29. Zhou, H-X. *et al.* (2008) Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* 37, 375–397
30. McGuffee, S.R. and Elcock, A.H. (2010) Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.* 6, e1000694
31. Hartl, F.U. *et al.* (2011) Molecular chaperones in protein folding and proteostasis. *Nature* 475, 324–332
32. Oikawa, T. *et al.* (2016) α-Synuclein fibrils exhibit gain-of-toxic-function, promoting tau aggregation and inhibiting microtubule assembly. *J. Biol. Chem.* 291, 15046–15056
33. Keiler, K.C. (2015) Mechanisms of ribosome rescue in bacteria. *Nat. Rev. Microbiol.* 13, 285–297
34. Wachter, A. (2014) Gene regulation by structured mRNA elements. *Trends Genet.* 30, 172–181
35. Doolittle, W.F. *et al.* (2014) Distinguishing between "function" and "effect" in genome biology. *Genome Biol. Evol.* 6, 1234–1237
36. Markowitz, V.M. *et al.* (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* 42, D560–567