Published in final edited form as:

Nat Rev Genet. 2010 July ; 11(7): 487–498. doi:10.1038/nrg2810.

Constraints and plasticity in genome and molecular-phenome evolution

Eugene V. Koonin and Yuri I. Wolf

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.

Abstract

Multiple constraints variously affect different parts of the genomes of diverse life forms. The selective pressures that shape the evolution of viral, archaeal, bacterial and eukaryotic genomes differ markedly, even among relatively closely related animal and bacterial lineages; by contrast, constraints affecting protein evolution seem to be more universal. The constraints that shape the evolution of genomes and phenomes are complemented by the plasticity and robustness of genome architecture, expression and regulation. Taken together, these findings are starting to reveal complex networks of evolutionary processes that must be integrated to attain a new synthesis of evolutionary biology.

Evolution is variously constrained on all levels of biological organization, from genome sequence to genome architecture, gene expression, molecular interactions and organismal phenotypes^{1,2}. Constraints on the rates and paths of evolution can be divided into genomic constraints, which are manifested at the level of the genome sequence and architecture, and phenomic constraints, which affect phenotypic characteristics. The vast amounts of diverse data recently generated by comparative genomics and systems biology provide previously inconceivable insights into the patterns and processes of genome and phenome evolution²⁻⁵.

Comparative genomics has the potential to measure the strength of constraints on different classes of sites in genomes and to elucidate the biological nature of these constraints. Genome comparisons also help to address higher-level questions, including the degree to which constraints act on gene repertoires, genome architecture and the evolution rate itself. Similarly, the avalanche of systems biology data allows researchers to ask new, qualitative questions, such as how do constraints affect metabolic fluxes and the ‘molecular phenome’ (which includes gene-expression, regulatory and interaction networks)?

Here, we attempt a genome-wide and organism-wide assessment of the constraints that operate at different levels of biological organization and discuss the links between constraints and the robustness and plasticity of biological systems. In the interests of space, we omit important areas that are covered in recent reviews — for example, constraints on development⁶ or metabolic energy expenditure⁷ — and only touch upon issues such as the evolution of regulatory networks⁸.

© 2010 Macmillan Publishers Limited. All rights reserved

Correspondence to E.V.K. koonin@ncbi.nlm.nih.gov.

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Authors' homepage: <http://www.ncbi.nlm.nih.gov/CBBresearch/Koonin>

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

We begin by discussing the constraints that affect sequence evolution and proceed to examine constraints on gene and genome architectures, genome size, gene number and gene repertoires. We then switch to considering constraints on molecular phenotypes, discuss universal models of protein evolution and conclude with a synopsis of the main factors that seem to constrain evolution at different levels.

We demonstrate the diversity of constraints on different classes of sites in genomes and contrast the constrained evolution of functionally important sites, particularly in proteins, with the plasticity of genome architecture and expression. Despite the variance of constraints in and across genomes, the emergence of these patterns might be explained by universal evolutionary patterns and simple stochastic models.

Constraints on sequence evolution

Constraints across a genome sequence

Constraints differ greatly across classes of genomic sequences and across taxa. Constraints on a particular class of sites can be measured only by comparison to another class of sites that are assumed to have evolved neutrally. The choice of an appropriate neutral model is a major problem in molecular evolution studies (BOX 1).

Such comparisons have revealed that, typically, the constraints on sites encoding protein sequences and those of structural RNAs (such as ribosomal RNAs and tRNAs) are stronger than those on non-coding sites, although the distributions of constraint strengths are broad and overlapping^{9,10} (BOX 1). Most protein-coding genes evolve under purifying selection of widely varying strength, and few evolve under positive selection (BOX 1). Genes evolving under continuous, long-term positive selection encode specialized proteins for which rapid change is crucial for function, typically involving an ‘arms race’ between competing agencies, such as hosts and parasites; examples include bacterial surface proteins^{11,12} and proteins involved in mammalian spermatogenesis, sperm competition and sperm-egg interaction^{13,14}. Evolution under positive selection is not unconstrained — constraints on the overall protein structure still apply¹⁵ — but evolution along the available trajectories proceeds rapidly.

The fact that most genes encoding proteins and structural RNAs evolve under purifying selection does not imply uniform constraints across sites. on the contrary, the evolutionary rates of codons (sites) in protein-coding genes — and by implication the strength of constraints on different sites — is well-described by a characteristic distribution in which a small fraction of sites are virtually unconstrained or subject to positive selection, and most sites are subject to constraints that vary within a broad range^{16,17}.

The assumption that synonymous sites in protein-coding genes evolve neutrally is useful for measuring selection at the protein level but is a rough approximation at best. The universal, significant positive correlation between K_a and K_s ¹⁸⁻²¹ (BOX 1) is compatible with the view that the evolution of synonymous sites is also constrained and that the forces shaping the evolution of non-synonymous and synonymous sites are related (see ‘Constraints on protein-coding genes’ below).

Constraints across taxa

The distributions of constraints across genomes are dramatically different in life forms with distinct genome architectures. In particular, there is a pronounced difference in constraint distribution between viruses and prokaryotes (bacteria and archaea), which have ‘wall-to-wall’ genomes that consist mostly of protein-coding and RNA-coding genes, and multicellular eukaryotes, in whose genomes the coding nucleotides are a minority^{22,23} (FIG.

1). The constraints on compact genomes, particularly those of prokaryotes, are much stronger than the constraints on the genomes of multicellular eukaryotes (median Ka/Ks values for prokaryotes and multicellular organisms are typically 0.01-0.1 and 0.1-0.5, respectively). In viruses and prokaryotes, nearly all genomic sites are evolutionarily constrained, with the notable exception of pseudogenes, which are common in some parasitic bacteria, such as *Rickettsia* species or *Mycobacterium leprae*²⁴⁻²⁶. Non-coding regions constitute only 10-15% of the genomes of most free-living prokaryotes, and a considerable fraction of these sequences encompasses regulatory elements that are substantially constrained in their evolution²⁷. The genomes of most viruses are even more compact, with almost all of the genome sequence taken up by protein-coding genes²³. The genome architecture of most unicellular eukaryotes resembles that of prokaryotes, although the fraction of non-coding sequences, of which a large portion is expected to evolve without constraints, is greater in these genomes than it is in prokaryotes.

By contrast, multicellular eukaryotes (plants and especially animals) have intron-rich genomes with long intergenic regions, and a large fraction of these non-coding sequences indeed seem to undergo unconstrained evolution (FIG. 1). The estimated fractions of constrained nucleotides in a genome differ substantially even between animals: in *Drosophila melanogaster*, ~70% of sites in the genome, including ~65% of the non-coding sites, seem to be subject to selection²⁸, whereas in mammals this fraction is estimated at only 5-6%²⁹ or even ~3%³⁰. Note, however, that the absolute numbers of sites subject to selection in these animal genomes of widely different size are quite close. By contrast, the fraction of constrained non-coding sites in *Arabidopsis thaliana* is estimated to be much lower than in *D. melanogaster*^{31,32}, although the two genomes are comparable in size and architecture. Of course, the estimates of the fraction of constrained sites (that is, sites that are 'visible' to selection; see discussion below and BOX 1) are based on simple population-genetic models, and the assumptions of these models might be violated to different extents across evolving lineages, leading to illusory differences. Nevertheless, different evolutionary regimes would seem to be operating even among closely related species.

The selectome, the RNome and 'junk' DNA

The estimate of 3-6% for the fraction of constrained sites in mammalian genomes is remarkable from two opposite standpoints. First, it seems that most of the mammalian genome fits with the much-maligned definition of 'junk'³³. Of course, the functional recruitment of 'junk' sequences, such as mobile elements, is common³⁴ but at any given time most of the mammalian genome evolves without appreciable constraints. Second, as protein-coding sequences comprise only ~1.2% of the genome²⁹, the substantial majority of the sites under selection do not encode amino acids. In particular, the selective pressure on 5' and 3' uTRs of mammalian genomes is comparable to or even stronger than that on synonymous sites in coding regions^{35,36}. An even greater contribution to the mammalian 'selectome' (the total number of sites under selection³⁷) is the rapidly growing complement of non-coding RNA genes, the RNome³⁸. The RNome includes numerous regulatory microRNAs that are subject to a broad range of constraints^{39,40} in addition to many long non-coding RNAs that might function in gene regulation and development and seem to be subject to constraints that are comparable to those on protein-coding genes⁴¹.

The known part of the RNome is the tip of the iceberg, especially considering that transcripts have been detected from nearly all sequences in mammalian genomes^{42,43}. Comparative-genomic analysis reveals numerous conserved sequences (including 'ultraconserved elements'⁴⁴) in introns and intergenic regions of animal and plant genomes^{45,46}, but transcription into a specific functional RNA has been shown only for a few of these^{47,48}. Most of the transcription in organisms with complex genomes, such as mammals, is probably functionally irrelevant, but functional non-coding RNAs might still

outnumber protein-coding genes. The extent of sequence conservation that is unrelated to specific functions of transcripts but rather caused by requirements of expression regulation, chromatin structure and other factors remains an open question.

The current understanding of the constraints on different types of sites across known genomes (FIG. 1) can be summarized as follows: sequences encoding structural RNAs and non-synonymous sites in protein-coding sequences are among the most strongly constrained; synonymous sites, sequences coding for regulatory RNAs, and non-coding regulatory sequences on average evolve under weaker selection but are not free of constraints; and characteristic distributions of constraints crucially depend on genome size and architecture (primarily gene density). Evolutionary regimes may differ widely even in closely related taxa, such as arthropods and vertebrates.

Constraints on gene and genome architectures

Constraints on gene architecture

An aspect of gene architecture that is common to all life forms but is particularly prominent in eukaryotes is the multidomain organization of proteins⁴⁹. Numerous proteins consist of multiple ‘evolutionary domains’, and the multidomain organization of some key proteins is conserved throughout the evolution of cellular life. Generally, however, domain rearrangements form an important resource of plasticity in all evolving lineages, particularly in the case of ‘promiscuous domains’^{50,51}.

A eukaryote-specific feature is the exon-intron organization of protein-coding genes. Intron positions are highly conserved over long evolutionary distances: up to 25-30% of intron positions are located in the exact same positions of orthologous genes between animals and plants^{52,53}. Some animal lineages, in particular vertebrates, evolve under almost complete intron stasis, with minimal intron loss and virtually no gain. By contrast, the evolution of other lineages, such as nematodes and many groups of unicellular eukaryotes, features extensive turnover of introns^{54,55}.

On the whole, the evolution of eukaryotic gene architecture is very diverse, with a highly dynamic evolutionary process in some lineages but much less change in others.

Constraints on genome architecture in prokaryotes

Genome architecture refers to the mapping of genetic elements onto the genome, including gene order, and the clustering and co-regulation of genes with related functions^{2,23,54,56}. Long-range gene order is surprisingly poorly conserved among sequenced archaeal and bacterial genomes, which is in contrast to the recurrence of operons in diverse prokaryotes⁵⁷. Long-range gene orders in prokaryotes diverge roughly in proportion to sequence divergence in protein-coding genes; however, evolution of gene order can be so fast that, in many lineages, no long-range conservation is seen even when sequence divergence is very low⁵⁸. Beyond this general pattern, the rates of gene-order decay differ substantially among prokaryotic lineages⁵⁸. Gene order in many prokaryotes seems to be disrupted largely by inversions centred at the origin of replication, a process that does not seem to be strongly constrained by purifying selection and depends primarily on the activity of the relevant recombination machinery^{59,58}. The activity of mobile elements, in particular insertion sequences, is another important — and also weakly constrained — factor affecting genome rearrangement in prokaryotes⁶⁰.

In contrast to the lack of conservation of long-range gene order, prokaryotic operons combine evolutionary resilience and plasticity, forming overlapping gene arrays that are partially shared by evolutionarily distant organisms^{61,62}. Although some operons, notably

several parts of the ‘superoperon’ that encodes ribosomal proteins, show a pronounced signal of long-term vertical evolution⁶³, the wide distribution of many operons among prokaryotes is attributable to horizontal gene transfer (HGT) under the selfish operon concept⁶⁴.

Despite the lack of conservation of the long-range gene order, some constraints affect the gross architecture of prokaryotic genomes: in particular, the preferential codirectionality of gene transcription with replication might arise from selection to minimize the frequency of collisions between RNA polymerase and replication forks⁶⁵.

Constraints on genome architecture in eukaryotes

Most eukaryotes have no operons, and existing operons are unrelated to prokaryotic operons and seem to have evolved *de novo*⁶⁶. In eukaryotes, the search for gene-order non-randomness — for example, the clustering of genes with connected functions or with similar expression levels and patterns — has led to mixed results^{23,66,67}. Evidence exists not only for advantages of clustering of functionally linked genes thanks to enhanced possibilities for co-regulation⁶⁷ but also for disadvantages of such clustering, conceivably due to transcriptional interference⁶⁸. With some striking exceptions, such as the strict order of animal Hox genes⁶⁹ or the clustering of enzymes in certain metabolic pathways in yeast⁷⁰, gene order in eukaryotes seems to be quasi-random²³. Generally, the evolution of gene order in eukaryotes seems to be dominated by random chromosomal rearrangements⁷¹, and no gene arrays are highly conserved between distantly related forms, such as different animal phyla, let alone between animals and fungi or plants²³.

Thus, the evolution of genome architecture seems to be shaped by the interplay of strong constraints that determine the conservation of operons, weak constraints on other forms of functional clustering and large-scale gene organization, and extensive dynamics of genome rearrangements, duplication and HGT. These dynamics both counteract weak constraints by disrupting gene associations and reinforce the effect of stronger constraints, as in the case of the horizontal spread of selfish operons.

Constraints on gene number and gene repertoires

The number of protein-coding genes in cellular life forms varies within a surprisingly narrow range compared with the genome size, especially considering the difference in biological complexity between prokaryotes and multicellular eukaryotes. Excluding the extremely reduced genomes of some intracellular parasitic bacteria and the huge, polyploid plant genomes, the number of encoded proteins varies from ~500 to ~35,000, less than two orders of magnitude²³. In unicellular organisms, especially in prokaryotes, the number of encoded proteins closely correlates with genome size (gene density is roughly constant, at around one gene per kilobase of DNA), whereas in multicellular organisms, especially animals, the two quantities are decoupled.

Constraints on the lower gene bound

What constrains the number of encoded proteins from below and above? The lower threshold intuitively corresponds to a minimal gene set for cellular life^{72,73}. Minimal gene sets derived from comparative-genomic and experimental approaches converge at 250-350 genes — or more precisely, orthologous gene lineages⁷⁴ — and seem to encode most essential cellular functions^{72,73}. An apparent paradox is that the conserved set of 250-350 orthologous genes can be derived only by comparing small sets of not-too-diverse genomes, as pioneered in a comparison of *Haemophilus influenzae* and *Mycoplasma genitalium*⁷⁵. By contrast, the core set of ubiquitously conserved genes is continuously shrinking with the addition of new sequenced genomes and seems to be limited to approximately 30 genes, all

encoding proteins involved in translation and transcription^{57,76}. The solution to the paradox is non-orthologous gene displacement (NoGD): most of the essential cellular functions can be performed by members of more than one orthologous gene set, which in many cases are completely unrelated^{72,77}. The minimal genetic complement of a cell is not a unique minimal gene set but rather a unique set of indispensable functional niches that can be filled with diverse collections of genes.

Constraints on the upper gene bound and gene repertoire

The maximum number of genes in a cellular life form does not substantially increase despite the rapid growth of the collection of sequenced genomes. Thus, although an upper bound of genetic complexity seems to exist, the nature of this limit remains obscure. One attractive hypothesis is the ‘bureaucratic ceiling of complexity’ (BOX 2).

This hypothesis seems to be particularly plausible given the lack of large expansions in gene number in vertebrates, in which the number of genes and the genome size are decoupled. In these organisms, the principal constraint on the upper limit is probably imposed by the cost of regulation and expression rather than the cost of replication. It is not surprising then that vertebrates evolved elaborate means of increasing proteomic complexity — such as alternative splicing, alternative transcription⁷⁸ and regulatory RNAs — that do not involve inflating the number of protein-coding genes.

Gene duplication and constraints on evolution of paralogous gene families

The formation of paralogous gene families through gene duplication is the main route of innovation, especially in eukaryotes^{79,80}. The size distribution of paralogous families in each genome follows a power-law-like function⁸¹ that is well reproduced by a simple gene birth and death model conditioned on the equilibrium (constant size) of genomes during evolution^{82,83}. This process seems to define a fundamental constraint on gene demography that is coupled to the constraint on the upper bound of the total number of genes.

Gene loss is a major factor of evolution, in agreement with the findings on the small and shrinking cores of conserved genes, NoGD and extensive redundancy. Gene loss is extensive in all lineages, in particular in the evolution of animal taxa: there is a high degree of orthology between vertebrates and primitive animals, such as sea anemones and *Trichoplax adhaerens*, which is in contrast to the much more limited orthologous relationships between vertebrates and arthropods or nematodes^{84,85}. Individual genes show a broad distribution of gene-loss rates^{86,87}; moreover, it seems that the observed evolutionary and phenomic features of genes are compatible with a steady-state model of genome evolution under which the distribution of gene-loss and -gain rates remain effectively constant over extended evolutionary spans⁸⁸. This distribution could reflect another important constraint governing genome evolution.

Thus, the genetic complexity of genomes in all life forms (as reflected by the number of genes) is tightly constrained from the above (that is, constrained with respect to the maximum number of genes), whereas the genome size may be either coupled to the gene number and hence similarly constrained (as in prokaryotes and most unicellular eukaryotes) or decoupled and hence much less strictly constrained (as in multicellular eukaryotes, especially animals). The constraints on gene number and the functional distribution of genes remain to be explored in detail but are likely to be determined by fundamental ‘laws’ of cell functioning. The gene repertoire in all organisms depends on the process of evolution by gene duplication and loss that is well-described by birth and death models.

Constraints on protein-coding genes

Protein-coding genes are generally highly constrained, but the distribution of the rates of evolution among non-synonymous sites in orthologous genes in any pair of compared genomes spans three to four orders of magnitude and is much broader than the rate distribution for synonymous sites (BOX 1). Remarkably, the shapes of the rate distributions for orthologous proteins are highly similar for all studied cellular life forms, from bacteria to archaea to mammals⁸⁸ (FIG. 2). Another universal of genomic and phenomic evolution is the anticorrelation between the rate of evolution of a protein-coding gene and its expression level: highly expressed genes evolve slowly, a dependence that is observed in all model organisms for which expression data are available, although the magnitude of the negative correlation widely differs^{86,89-91}.

The existence of these universals suggests that the primary determinants of protein evolution have more to do with fundamental principles of protein structure and folding than are common across all life than with unique biological functions. It has been proposed that the principal selective factor underlying evolution of proteins is robustness to misfolding, as misfolded proteins can be toxic to the cell^{21,90}. This model is compatible with the preferential use of optimal codons (strong codon bias) in highly expressed and highly conserved protein-coding genes⁹²⁻⁹⁴ and with the aforementioned positive correlation between K_a and K_s . Under the misfolding hypothesis, the evolution of synonymous sites is constrained, at least in part, by the same factors as the evolution of proteins owing to the pressure for the preferential use of optimal codons in highly expressed proteins and in specific sites that are important for protein folding^{90,95}.

A recent analysis of protein evolution produced estimates of evolutionary rates under the assumption that misfolding is the only source of fitness cost⁹⁶. The results reproduce the universal distribution of protein evolutionary rates as well as the dependence between evolutionary rate and expression (FIG. 2), and suggest that the universal rate distribution indeed might be a consequence of fundamental physics of proteins (BOX 3).

Thus, the primary constraints on protein evolution might have more to do with the maintenance of the native folding and intermolecular interactions than with unique protein functions, a view that seems to be supported by a recent large-scale analysis of protein-family evolution¹⁵.

Constraints on molecular phenotypes

Owing to advances in systems biology, it is possible to assess evolutionary variance and constraints by comparing various features of the molecular phenotype — such as gene expression, protein abundance and architecture of interaction networks — among different organisms.

Molecular phenomic variables show a distinct structure of dependences among themselves and with evolutionary variables, such as the rates of sequence evolution and gene loss⁹⁷. The correlations between phenomic variables are typically positive (for example, highly expressed proteins also tend to interact with many other proteins and have many paralogues), whereas the correlations between phenomic and evolutionary variables are generally negative (for example, highly expressed genes on average evolve more slowly than those expressed at a low level). Most of these correlations are statistically significant but relatively weak, so caution is required (experimental biases should be investigated as potential causes⁹⁸⁻¹⁰⁰), but the overall pattern of positive and negative correlations seems to be undeniable^{97,101}. Thus, constraints on the ranges of phenomic variables partly seem to

constrain the evolution of gene sequences, gene repertoires and genome architectures, as shown by the model of protein evolution discussed above.

Constraints on gene expression and protein abundance

Comparative expression analysis of orthologous genes has suggested that gene expression in animals is not strongly constrained during evolution or at least has a major neutral component^{102,103}. However, subsequent analyses revealed signatures of selective constraints that affect gene expression^{104-106,107}. The expression level of orthologous proteins is strongly correlated even among distantly related animals: a correlation coefficient greater than 0.8 was observed for approximately 3,000 orthologous genes from *Caenorhabditis elegans* and *D. melanogaster*¹⁰⁸. Similar results have been recently reported for a broad range of model eukaryotes¹⁰⁹, which is in sharp contrast with the correlation coefficients of 0.2-0.4 that are seen in comparisons of other genomic and molecular phenomic variables⁹⁷.

Notably, protein abundance seems to be constrained to a substantially greater extent than gene expression, and the degree of constraint is even stronger than the rate of DNA-sequence evolution within the same set of orthologous genes^{108,110}.

Constraints on network architecture

Global architectures of protein interaction and gene co-expression networks seem to be universal across life forms, and the network node degree (the number of connections) has a characteristic power-law-like distribution¹¹¹. Local network structures are much less constrained and differ even among closely related organisms^{112,113}. However, a comparison of gene co-expression networks from mutation accumulation lineages of *C. elegans*, in which selection is effectively removed¹⁰⁴, with those from the natural isolate suggests that the local wiring of the co-expression network is constrained by selection, whereas the global properties are not¹¹⁴. Thus, the similarity of global-network properties in widely different organisms might reflect 'neutral' rather than selective constraints — that is, these properties could have evolved through simple, stochastic, non-selective processes, as in birth-and-death models of genome and network evolution^{83,115}. This view is reinforced by the fact that certain evolving but non-biological networks, such as the internet, possess similar global properties¹¹¹.

The study of molecular-phenome evolution is still in its infancy, but the advances of evolutionary systems biology have already revealed unexpectedly strong constraints on some phenomic features.

Constraints on evolutionary trajectories

What would happen if the tape of evolution were rewound?

An intriguing question is the degree to which the course of evolution itself is constrained¹¹⁶. To what extent is the evolutionary process free to explore different trajectories between the given initial and end states? Direct studies of evolutionary trajectories in sequence space are very limited but have already shown that, although historical contingency is central to evolution¹¹⁷, exploration of the sequence space is strongly constrained so that only a minority of theoretically possible trajectories are accessible.

In theory, mutational trajectories in sequence space are fundamentally stochastic¹¹⁸. However, experimental evolution studies indicate that paths of adaptive evolution are constrained by interactions between mutations (epistasis and pleiotropy), although not to the point of becoming deterministic. Experimental evolution of bacterial antibiotic resistance resulting from 5 point mutations in the β -lactamase (*blaI*) gene showed that, of the 120

possible trajectories across the sequence space, 102 were inaccessible, and of the remaining 18, several had negligible probability of realization¹¹⁹. Even stronger constraints were identified in a subsequent study that explored a more complex fitness landscape by simultaneously evolving resistance to two antibiotics¹²⁰.

Systematic studies of bacterial evolution under controlled conditions reveal both parallel emergence of the same mutations under a particular selective pressure and the realization of multiple trajectories^{116,121,122}. For instance, the evolution of the same, rare phenotype — the ability to grow on citrate — proceeded along distinct trajectories in different *Escherichia coli* populations¹²³.

These results complement and reinforce previous observations of the convergent and parallel evolution of proteins that perform the same function. Such evolution is limited but demonstrable, and was first observed in the evolution of lysozymes in ungulates and langur monkeys¹²⁴⁻¹²⁷.

The extent of constraints on an evolutionary trajectory and, conversely, the likelihood of parallel evolution crucially depend on the shape of the fitness landscape: the more rugged the landscape, the stronger the constraints. The shape of the landscape itself is determined by the nature, strength and interactions of the relevant selective factors; furthermore, the landscape evolves with time, making it into a fitness seascape¹²⁸.

Ne as the general gauge of evolutionary constraints

Under population genetics theory, the effectiveness of purifying selection is proportional to the effective population size (N_e) of a given organism, assuming a uniform mutation rate. Only mutations for which $|s| > 1/N_e$, (in which s is the selection coefficient) can be fixed or efficiently eliminated during evolution². Conversely, mutations with $|s| < 1/N_e$ are ‘invisible’ to selection. This simple dependence could be the primary determinant of the constraints that affect different aspects of genome and phenome evolution. Differences in N_e seem to underlie the qualitative difference in the genome architectures of unicellular and multicellular organisms^{22,129}. Substantial genome expansion seems to be attainable only in organisms with small populations and the attendant weak selection, as is the case in plants and animals. In these organisms, the deleterious effect of propagating nonfunctional sequences is often too small to allow the ‘detection’ and elimination of such sequences by purifying selection. Accordingly, evolutionary conservation does not automatically imply that the conserved feature is constrained by purifying selection owing to its functional importance but rather, somewhat paradoxically, might reflect weak purifying selection that is insufficient to eliminate non-adaptive ancestral features.

Gene architecture: case example

The evolution of exon- intron gene structure in eukaryotes is an excellent demonstration of this population-genetic paradigm. Most introns are functionless and weakly deleterious for an organism partly due to their energetic cost. However, a simple estimate based on typical mutation rates shows that the deleterious effect of introns is ‘visible’ to purifying selection only in populations with N_e of 10^7 or greater. This is within the typical N_e range of unicellular eukaryotes, whereas multicellular eukaryotes have smaller populations^{2,22,130}. These differences have a dramatic effect on the evolution of genome architecture in eukaryotes. Unlike the genomes of unicellular forms, which typically contain fewer than one intron per gene, plants and animals possess numerous introns⁵³. Furthermore, conservation of intron positions in orthologous genes of animals and plants (see above) seems to be due to the inefficient elimination of introns in organisms with small N_e and not to constraints on intron positions *per se*. Finally, introns in unicellular eukaryotes are short and have

conserved, optimized splice signals^{131,132}. By contrast, introns in intron-rich genomes are often long and bounded by relatively weak splice signals¹³³, thus providing for the evolution of alternative splicing.

Its importance notwithstanding, N_e determines the course of evolution only on a broad scale. A comparative analysis of the K_a/K_s values among prokaryotic lineages did not detect a negative correlation between selective constraints and genome size⁵⁸, as implied by the straightforward population-genetic perspective¹³⁴. on the contrary, larger genomes tend to evolve under stronger constraints (even when only free-living microbes are analysed), implying that lifestyle could be a crucial determinant of genome evolution (favouring, in particular, gene acquisition through HGT in variable environments) independently of N_e ⁵⁸. Thus, evolution driven solely by population-genetic factors could be an appropriate null hypothesis, but the actual evolutionary trajectories are determined — and constrained — by specific biological contexts (FIG. 3).

Robustness and plasticity of biological systems

The aspects of evolution that are complementary to constraints are the plasticity of genomic and phenomic characteristics, and the robustness of molecular phenotypes¹³⁵. As mentioned above, plasticity is pronounced at many levels. In many organisms, large-scale genome organization seems to be only weakly constrained, so gene order substantially differs even among closely related organisms, especially among prokaryotes^{23,58}. The gene repertoires of many organisms, especially prokaryotes, show plasticity that may even exceed the plasticity of genome architecture, as shown by rapid genome reduction in parasitic bacteria²⁶ and by acquisition of pathogenicity islands that may comprise over 30% of the genome in bacterial pathogens¹³⁶. The plasticity of genome organization and composition is paralleled by the evolutionary flexibility of regulatory networks, and complements the more strongly constrained evolution of individual genes^{137,138}.

Evolutionary plasticity and the strength of evolutionary constraints are tightly linked to the robustness of biological systems. Robustness seems to be an evolved property, as shown by the study of specialized buffering mechanisms, the impairment of which reveals hidden genetic variation and capacitates evolution^{139,140}. Recently, the concept of evolutionary capacitance has been extended to numerous genes with extremely diverse functions; stabilization seems to be a general property of interaction networks, so disruption of almost any highly connected node reduces the robustness of the system and leads to increased variation¹⁴¹. A comprehensive study of capacitor properties of yeast mutants revealed that disruption of any one of approximately 300 yeast genes (about 6% of the total) significantly decreased the robustness of yeast to environmental perturbations¹⁴².

Thus, robustness is likely to be a major, selectable mechanism that dampens evolutionary constraints — particularly those caused by interaction between mutations — and enhances plasticity. However, evidence has also been presented that certain forms of robustness, in particular the robustness of metabolic networks that are dependent on redundant reactions, could be byproducts of evolution driven by other factors, such as regulatory versatility¹⁴³.

Concluding remarks and outlook

The prevailing theme that has emerged from recent advances in evolutionary genomics and systems biology is the plurality of constraints that affect the evolution of different types of sequences in genomes, genome architectures, gene repertoires and molecular phenomes. In addition, it has emerged that there are major differences in evolutionary regimens among taxa. Beyond this diversity, comparative-genomic and molecular-phenomic analyses reveal universal patterns that could be compatible with relatively simple, general models of

evolution. As discussed here, these models are starting to suggest fundamental causes underlying important aspects of evolution, such as the universal constraints on the evolution of proteins and gene repertoires (TABLE 1). It seems appropriate to expand the notion of constraints to include not only selective but also ‘neutral’ constraints that are determined by non-selective, stochastic properties of biological systems and are often amenable to modelling using techniques borrowed from statistical physics^{144,145} (TABLE 1).

Evolutionary trajectories in sequence space seem to be strongly constrained, thus substantially limiting the ‘tinkering potential’ of evolution¹¹⁷. The evolutionary process thus seems to be a compromise “between design and bricolage”¹⁴⁶, the design aspect brought about by constraints and the bricolage stemming from the evolved robustness and the ensuing plasticity of evolving organisms.

Comparative genomics and systems approaches are transforming evolutionary biology into a much more complex but also more precise, quantitative field than it was in the twentieth century. of course, evolutionary biology is only at the beginning of the path from ‘stamp collection’ to physics. Many key quantitative relationships, including some of those discussed in this Review, remain open for reinterpretation in light of new data and theoretical models. Nevertheless, it is our belief that the transition has passed the point of no return. Next-generation sequencing, quantitative proteomics and other new methods, combined with more specific approaches of experimental evolution, should reveal the specific constraints that affect diverse aspects of genome and phenome evolution.

Acknowledgments

The authors thank A. Lobkovsky for providing part of the data used in the figure in Box 3 and T. Senkevich for critical reading of the manuscript. We apologize to the many colleagues whose work is not cited here because of space constraints. The authors’ research is funded by the Intramural Research Program of the US Department of Health and Human Services (National Library of Medicine, US National Institutes of Health).

References

1. Kimura, M. The Neutral Theory of Molecular Evolution. Cambridge Univ. Press; 1983.
2. Lynch, M. The Origins of Genome Architecture. Sinauer Associates; Sunderland, Massachusetts: 2007. [A definitive presentation of the population-genetic perspective on genome evolution, with an emphasis on effective population size as the dominant factor of evolution and a non-adaptive origin of genomic complexity.]
3. Loewe L. A framework for evolutionary systems biology. BMC Syst. Biol. 2009; 3:27. [PubMed: 19239699]
4. Koonin EV, Wolf YI. Evolutionary systems biology: links between gene evolution and function. Curr. Opin. Biotechnol. 2006; 17:481–487. [PubMed: 16962765]
5. Yamada T, Bork P. Evolution of biomolecular networks: lessons from metabolic and protein interactions. Nature Rev. Mol. Cell Biol. 2009; 10:791–803. [PubMed: 19851337]
6. Snell-Rood EC, Van Dyken JD, Cruickshank T, Wade MJ, Moczek AP. Toward a population genetic framework of developmental evolution: the costs, limits, and consequences of phenotypic plasticity. Bioessays. 2010; 32:71–81. [PubMed: 20020499]
7. Palsson B. Metabolic systems biology. FEBS Lett. 2009; 583:3900–3904. [PubMed: 19769971]
8. Erwin DH, Davidson EH. The evolution of hierarchical gene regulatory networks. Nature Rev. Genet. 2009; 10:141–148. [PubMed: 19139764]
9. Shabalina SA, Kondrashov AS. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. Genet. Res. 1999; 74:23–30. [PubMed: 10505405]
10. Margulies EH, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. Genome Res. 2007; 17:760–774. [PubMed: 17567995]

11. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. Genes under positive selection in *Escherichia coli*. *Genome Res.* 2007; 17:1336–1343. [PubMed: 17675366]
12. Muzzi A, Moschioni M, Covacci A, Rappuoli R, Donati C. Pilus operon evolution in *Streptococcus pneumoniae* is driven by positive selection and recombination. *PLoS ONE.* 2008; 3:e3660. [PubMed: 18987753]
13. Nielsen R, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 2005; 3:e170. [PubMed: 15869325]
14. Turner LM, Chuong EB, Hoekstra HE. Comparative analysis of testis protein evolution in rodents. *Genetics.* 2008; 179:2075–2089. [PubMed: 18689890]
15. Worth CL, Gong S, Blundell TL. Structural and functional constraints in the evolution of protein families. *Nature Rev. Mol. Cell Biol.* 2009; 10:709–720. [PubMed: 19756040]
16. Grishin NV, Wolf YI, Koonin EV. From complete genomes to measures of substitution rate variability within and between proteins. *Genome Res.* 2000; 10:991–1000. [PubMed: 10899148]
[An early study that suggests that the evolutionary rates of orthologous genes from diverse life forms follow a universal distribution, and that derives a link between intra-gene and across-gene distributions of evolutionary rates.]
17. Nielsen R. Molecular signatures of natural selection. *Annu. Rev. Genet.* 2005; 39:197–218. [PubMed: 16285858]
18. Ohta T, Ina Y. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and nonsynonymous divergences. *J. Mol. Evol.* 1995; 41:717–720. [PubMed: 8587116]
19. Makalowski W, Boguski MS. Synonymous and nonsynonymous substitution distances are correlated in mouse and rat genes. *J. Mol. Evol.* 1998; 47:119–121. [PubMed: 9694659]
20. Ellegren H. Comparative genomics and the study of evolution by natural selection. *Mol. Ecol.* 2008; 17:4586–4596. [PubMed: 19140982]
21. Drummond DA, Wilke CO. The evolutionary consequences of erroneous protein synthesis. *Nature Rev. Genet.* 2009; 10:715–724. [PubMed: 19763154]
22. Lynch M, Conery JS. The origins of genome complexity. *Science.* 2003; 302:1401–1404. [PubMed: 14631042] [A seminal work that expounds the population-genetic perspective on the evolution of genomic complexity. The authors argue that genomic complexity is driven by weak purifying selection in populations with small N_e ; in such populations, slightly deleterious features, such as gene duplications or introns, cannot be efficiently eliminated. Collected data on N_e and genomic complexity in diverse life forms are shown to be compatible with this perspective, at least as a rough approximation.]
23. Koonin EV. Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* 2009; 41:298–306. [PubMed: 18929678]
24. Harrison PM, Gerstein M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* 2002; 318:1155–1174. [PubMed: 12083509]
25. Monot M, et al. Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nature Genet.* 2009; 41:1282–1289. [PubMed: 19881526]
26. Darby AC, Cho NH, Fuxelius HH, Westberg J, Andersson SG. Intracellular pathogens go extreme: genome evolution in the Rickettsiales. *Trends Genet.* 2007; 23:511–520. [PubMed: 17822801]
27. Molina N, van Nimwegen E. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res.* 2008; 18:148–160. [PubMed: 18032729] [A rigorous method for detecting purifying selection in groups of closely related prokaryotes was applied to the study of intergenic region evolution. Universal patterns of purifying selection were detected, and translation-initiation sites were found to be the elements subject to the strongest selective pressure.]
28. Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the *Drosophila* genome? *PLoS Genet.* 2009; 5:e1000495. [PubMed: 19503600] [A critical review of the evidence indicating that most sites in the fruitfly genome are subject to selection.]
29. Waterston RH, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 2002; 420:520–562. [PubMed: 12466850]
30. Lunter G, Ponting CP, Hein J. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* 2006; 2:e5. [PubMed: 16410828]

31. Wright SI, Andolfatto P. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annu. Rev. Ecol. Syst.* 2008; 39:193–213.
32. Gossmann TI, et al. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol. Biol. Evol.* Mar 18.2010 (doi:10.1093/molbev/msq079).
33. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature.* 1980; 284:601–603. [PubMed: 6245369]
34. Bowen NJ, Jordan IK. Exaptation of protein coding sequences from transposable elements. *Genome Dyn.* 2007; 3:147–162. [PubMed: 18753790]
35. Drake JA, et al. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genet.* 2006; 38:223–227. [PubMed: 16380714]
36. Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ. Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucleic Acids Res.* 2004; 32:1774–1782. [PubMed: 15031317]
37. Proux E, Studer RA, Moretti S, Robinson-Rechavi M. Selectome: a database of positive selection. *Nucleic Acids Res.* 2009; 37:D404–D407. [PubMed: 18957445]
38. Costa FF. Non-coding RNAs: new players in eukaryotic biology. *Gene.* 2005; 357:83–94. [PubMed: 16111837]
39. Shabalina SA, Koonin EV. Origins and evolution of eukaryotic RNA interference. *Trends Ecol. Evol.* 2008; 23:578–587. [PubMed: 18715673]
40. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. *Cell.* 2009; 136:642–655. [PubMed: 19239886]
41. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell.* 2009; 136:629–641. [PubMed: 19239885] [A detailed review of long non-coding (macro) RNAs, a recently discovered class of mammalian genes that comprise a substantial part of the RNome.]
42. Bertone P, et al. Global identification of human transcribed sequences with genome tiling arrays. *Science.* 2004; 306:2242–2246. [PubMed: 15539566]
43. Johnson JM, Edwards S, Shoemaker D, Schadt EE. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet.* 2005; 21:93–102. [PubMed: 15661355]
44. Katzman S, et al. Human genome ultraconserved elements are ultraselected. *Science.* 2007; 317:915. [PubMed: 17702936] [A rigorous demonstration of the exceptionally strong selection that affects ultraconserved elements of mammalian genomes that are located outside protein-coding genes.]
45. Dermitzakis ET, Reymond A, Antonarakis SE. Conserved non-genic sequences — an unexpected feature of mammalian genomes. *Nature Rev. Genet.* 2005; 6:151–157. [PubMed: 15716910]
46. Elgar G. Pan-vertebrate conserved non-coding sequences associated with developmental regulation. *Brief. Funct. Genomic. Proteomic.* 2009; 8:256–265. [PubMed: 19752044]
47. Bejerano G, et al. Ultraconserved elements in the human genome. *Science.* 2004; 304:1321–1325. [PubMed: 15131266]
48. Baira E, Greshock J, Coukos G, Zhang L. Ultraconserved elements: genomics, function and disease. *RNA Biol.* 2008; 5:132–134. [PubMed: 18708752]
49. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. *Cell.* 2000; 101:573–576. [PubMed: 10892642]
50. Wuchty S, Almaas E. Evolutionary cores of domain co-occurrence networks. *BMC Evol. Biol.* 2005; 5:24. [PubMed: 15788102]
51. Basu MK, Carmel L, Rogozin IB, Koonin EV. Evolution of protein domain promiscuity in eukaryotes. *Genome Res.* 2008; 18:449–461. [PubMed: 18230802] [A quantitative comparative analysis of promiscuous domains across eukaryotic lineages, including demonstration of a positive correlation between domain promiscuity and the strength of purifying selection.]
52. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* 2003; 13:1512–1517. [PubMed: 12956953]

53. Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Rev. Genet.* 2006; 7:211–221. [PubMed: 16485020]
54. Roy SW, Penny D. Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. *Mol. Biol. Evol.* 2007; 24:171–181. [PubMed: 17065597]
55. Carmel L, Wolf YI, Rogozin IB, Koonin EV. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.* 2007; 17:1034–1044. [PubMed: 17495008] [A detailed analysis of differential dynamics of intron gain and loss across eukaryotic lineages reveals three distinct modes of evolution characterized by pervasive intron loss, equilibrium and relatively rare intron gain, respectively.]
56. Carmel L, Rogozin IB, Wolf YI, Koonin EV. Patterns of intron gain and conservation in eukaryotic genes. *BMC Evol. Biol.* 2007; 7:192. [PubMed: 17935625]
57. Koonin EV, Wolf YI. Genomics of Bacteria and Archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 2008; 36:6688–6719. [PubMed: 18948295]
58. Novichkov PS, Wolf YI, Dubchak I, Koonin EV. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *J. Bacteriol.* 2009; 191:65–73. [PubMed: 18978059] [This study provides a comparative analysis of selective and neutral evolutionary processes between multiple bacterial and archaeal lineages. The article demonstrates high, variable rates of genome rearrangement and the lack of correlation between genome streamlining and selective constraints on sequence evolution.]
59. Eisen JA, Heidelberg JF, White O, Salzberg SL. Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.* 2000; 1:research0011.1–research0011.9. [PubMed: 11178265]
60. Zhou F, Olman V, Xu Y. Insertion sequences show diverse recent activities in Cyanobacteria and Archaea. *BMC Genomics.* 2008; 9:36. [PubMed: 18218090]
61. Rogozin IB, et al. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res.* 2002; 30:2212–2223. [PubMed: 12000841]
62. Ling X, He X, Xin D. Detecting gene clusters under evolutionary constraint in a large number of genomes. *Bioinformatics.* 2009; 25:571–577. [PubMed: 19158161]
63. Wolf YI, Rogozin IB, Kondrashov AS, Koonin EV. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 2001; 11:356–372. [PubMed: 11230160]
64. Lawrence J. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.* 1999; 9:642–648. [PubMed: 10607610]
65. Rocha EP. The organization of the bacterial genome. *Annu. Rev. Genet.* 2008; 42:211–233. [PubMed: 18605898]
66. Osbourn AE, Field B. Operons. *Cell. Mol. Life Sci.* 2009; 66:3755–3775. [PubMed: 19662496]
67. Hurst LD, Pal C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. *Nature Rev. Genet.* 2004; 5:299–310. [PubMed: 15131653]
68. Liao BY, Zhang J. Coexpression of linked genes in Mammalian genomes is generally disadvantageous. *Mol. Biol. Evol.* 2008; 25:1555–1565. [PubMed: 18440951]
69. Lemons D, McGinnis W. Genomic evolution of Hox gene clusters. *Science.* 2006; 313:1918–1922. [PubMed: 17008523]
70. Wong S, Wolfe KH. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nature Genet.* 2005; 37:777–782. [PubMed: 15951822]
71. Eichler EE, Sankoff D. Structural dynamics of eukaryotic chromosome evolution. *Science.* 2003; 301:793–797. [PubMed: 12907789]
72. Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* 2003; 1:127–136. [PubMed: 15035042] [This article demonstrates the difference between the shrinking set of ubiquitously conserved orthologous genes and the larger minimal set of functional niches. Minimal gene sets are also examined in relation to different prokaryotic lifestyles.]

73. Moya A, et al. Toward minimal bacterial cells: evolution vs. design. *FEMS Microbiol Rev.* 2009; 33:225–235. [PubMed: 19067748] [The latest update on minimal gene sets and the promise of synthetic biology for *de novo* synthesis of custom genomes.]
74. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 2005; 39:309–338. [PubMed: 16285863]
75. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes [see comments]. *Proc. Natl Acad. Sci. USA.* 1996; 93:10268–10273. [PubMed: 8816789]
76. Charlebois RL, Doolittle WF. Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.* 2004; 14:2469–2477. [PubMed: 15574825]
77. Koonin EV, Mushegian AR, Bork P. Non-orthologous gene displacement. *Trends Genet.* 1996; 12:334–336. [PubMed: 8855656]
78. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature.* 2010; 463:457–463. [PubMed: 20110989]
79. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science.* 2000; 290:1151–1155. [PubMed: 11073452]
80. Lespinet O, Wolf YI, Koonin EV, Aravind L. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 2002; 12:1048–1059. [PubMed: 12097341]
81. Huynen MA, van Nimwegen E. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* 1998; 15:583–589. [PubMed: 9580988] [The authors report the discovery that the sizes of paralogous gene families follow a power-law-like distribution. They also present a simple model of gene family evolution.]
82. Karev GP, Wolf YI, Rzhetsky AY, Berezhovskaya FS, Koonin EV. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol. Biol.* 2002; 2:18. [PubMed: 12379152]
83. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. *Nature.* 2002; 420:218–223. [PubMed: 12432406] [A discussion of non-adaptive models of genome evolution — in particular, how patterns of gene birth and death reproduce the observed size distributions of paralogous gene families.]
84. Putnam NH, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science.* 2007; 317:86–94. [PubMed: 17615350]
85. Srivastava M, et al. The *Trichoplax* genome and the nature of placozoans. *Nature.* 2008; 454:955–960. [PubMed: 18719581]
86. Krylov DM, Wolf YI, Rogozin IB, Koonin EV. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 2003; 13:2229–2235. [PubMed: 14525925]
87. Wang X, Grus WE, Zhang J. Gene losses during human origins. *PLoS Biol.* 2006; 4:e52. [PubMed: 16464126]
88. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl Acad. Sci. USA.* 2009; 106:7273–7280. [PubMed: 19351897] [This is the definitive demonstration of the universal character of the approximately log-normal distribution of the evolutionary rate of orthologous genes. The distribution of genes by age also follows a similar pattern. The article presents a simple, non-adaptive model according to which the universal distribution of gene-loss rates is a fundamental feature of genome evolution.]
89. Pal C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. *Genetics.* 2001; 158:927–931. [PubMed: 11430355]
90. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell.* 2008; 134:341–352. [PubMed: 18662548] [A comprehensive analysis of the anticorrelation between evolution rate and expression of protein-coding genes in a variety of model organisms. This is a definitive presentation of the mistranslation-induced misfolding hypothesis of protein evolution.]
91. Pal C, Papp B, Lercher MJ. An integrated view of protein evolution. *Nature Rev. Genet.* 2006; 7:337–348. [PubMed: 16619049]

92. Grosjean H, Fiers W. Preferential codon usage in prokaryotic genes: the optimal codon–anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene*. 1982; 18:199–209. [PubMed: 6751939]
93. Lipman DJ, Wilbur WJ. Interaction of silent and replacement changes in eukaryotic coding sequences. *J. Mol. Evol.* 1984; 21:161–167. [PubMed: 6442990]
94. Hershberg R, Petrov DA. Selection on codon bias. *Annu. Rev. Genet.* 2008; 42:287–299. [PubMed: 18983258]
95. Zhou T, Weems M, Wilke CO. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol. Biol. Evol.* 2009; 26:1571–1580. [PubMed: 19349643]
96. Lobkovsky AE, Wolf YI, Koonin EV. Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc. Natl Acad. Sci. USA*. 2010; 107:2983–2988. [PubMed: 20133769] [The universal distribution of evolutionary rates among orthologues is reproduced under a simple model of protein folding and under the assumption that misfolding is the only source of fitness cost in protein evolution.]
97. Wolf YI, Carmel L, Koonin EV. Unifying measures of gene function and evolution. *Proc. Biol. Sci.* 2006; 273:1507–1515. [PubMed: 16777745] [A systematic analysis of correlations between evolutionary and molecular phenomic variables leads to the idea of ‘gene status’, according to which genes with a high expression level, a large number of physical or regulatory interactions and high values of other phenomic variables evolve slowly and are rarely lost in the course of evolution.]
98. Jordan IK, Wolf YI, Koonin EV. No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* 2003; 3:1. [PubMed: 12515583]
99. Bloom JD, Adami C. Evolutionary rate depends on number of protein–protein interactions independently of gene expression level: response. *BMC Evol. Biol.* 2004; 4:14. [PubMed: 15171796]
100. de Silva E, et al. The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol.* 2006; 4:39. [PubMed: 17081312]
101. Jordan IK, Wolf YI, Koonin EV. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* 2004; 4:22. [PubMed: 15238160]
102. Khaitovich P, et al. A neutral model of transcriptome evolution. *PLoS Biol.* 2004; 2:e132. [PubMed: 15138501]
103. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.* 2004; 21:2058–2070. [PubMed: 15282333]
104. Denver DR, et al. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nature Genet.* 2005; 37:544–548. [PubMed: 15852004]
105. Jordan IK, Marino-Ramirez L, Koonin EV. Evolutionary significance of gene expression divergence. *Gene*. 2005; 345:119–126. [PubMed: 15716085]
106. Liao BY, Zhang J. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.* 2006; 23:530–540. [PubMed: 16280543]
107. Gilad Y, Oshlack A, Rifkin SA. Natural selection on gene expression. *Trends Genet.* 2006; 22:456–461. [PubMed: 16806568]
108. Schrimpf SP, et al. Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol.* 2009; 7:e48. [PubMed: 19260763]
109. Weiss M, Schrimpf S, Hengartner MO, Lercher MJ, von Mering C. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. *Proteomics*. 2010; 10:1297–1306. [PubMed: 20077411] [This work extends the pioneering study reported in reference 108. The authors applied quantitative, highly accurate proteomic methods to reveal that the abundance of orthologous proteins is — unexpectedly — highly correlated among distantly related model organisms.]
110. Wolf YI, Gopich IV, Lipman DJ, Koonin EV. Relative contributions of intrinsic structural-functional constraints and translation rate to the evolution of protein-coding genes. *Genome Biol. Evol.* Mar 17.2010 (doi:10.1093/gbe/evq010).

111. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* 2004; 5:101–113. [PubMed: 14735121]
112. Bergmann S, Ihmels J, Barkai N. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol.* 2004; 2:e9. [PubMed: 14737187]
113. Tsaparas P, Marino-Ramirez L, Bodenreider O, Koonin EV, Jordan IK. Global similarity and local divergence in human and mouse gene co-expression networks. *BMC Biol.* 2006; 6:70.
114. Jordan IK, Katz LS, Denver DR, Streelman JT. Natural selection governs local, but not global, evolutionary gene coexpression networks in *Caenorhabditis elegans*. *BMC Syst. Biol.* 2008; 2:96. [PubMed: 19014554]
115. Lynch M. The evolution of genetic networks by non-adaptive processes. *Nature Rev. Genet.* 2007; 8:803–813. [PubMed: 17878896] [A model of the evolution of biological networks that shows how characteristic network properties could evolve through non-adaptive processes of mutation, drift and recombination.]
116. Kassen R. Toward a general theory of adaptive radiation: insights from microbial experimental evolution. *Ann. N. Y. Acad. Sci.* 2009; 1168:3–22. [PubMed: 19566701]
117. Jacob F. Evolution and tinkering. *Science.* 1977; 196:1161–1166. [PubMed: 860134] [A seminal conceptual analysis emphasizing the importance of contingency in evolution: evolution is construed as a bricolage that makes use of pre-existing states and is fundamentally unpredictable.]
118. Mani GS, Clarke BC. Mutational order: a major stochastic process in evolution. *Proc. R. Soc. Lond. B.* 1990; 240:29–37. [PubMed: 1972992]
119. Weinreich DM, Delaney NF, Depristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science.* 2006; 312:111–114. [PubMed: 16601193] [A key study on the landscape of protein evolution that revealed an unexpected level of constraint on evolutionary trajectories, apparently caused by interactions between mutations (epistasis).]
120. Novais A, et al. Evolutionary trajectories of β -lactamase CTX-M-1 cluster enzymes: predicting antibiotic resistance. *PLoS Pathog.* 2010; 6:e1000735. [PubMed: 20107608]
121. Barrick JE, Lenski RE. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Spring Harb. Symp. Quant. Biol.* Sep 23.2009 (doi:10.1101/sqb.2009.74.018). [A summary of a series of long-term, extensive studies of bacterial populations in controlled experimental conditions. The studies revealed that evolutionary trajectories are affected by an interplay between contingency and constraint.]
122. Stanek MT, Cooper TF, Lenski RE. Identification and dynamics of a beneficial mutation in a long-term evolution experiment with *Escherichia coli*. *BMC Evol. Biol.* 2009; 9:302. [PubMed: 20040094]
123. Blount ZD, Borland CZ, Lenski RE. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA.* 2008; 105:7899–7906. [PubMed: 18524956]
124. Stewart CB, Schilling JW, Wilson AC. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature.* 1987; 330:401–404. [PubMed: 3120013]
125. Yokoyama R, Yokoyama S. Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus*, and human. *Proc. Natl Acad. Sci. USA.* 1990; 87:9315–9318. [PubMed: 2123554]
126. Zhang J. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nature Genet.* 2006; 38:819–823. [PubMed: 16767103]
127. Li Y, Liu Z, Shi P, Zhang J. The hearing gene *Prestin* unites echolocating bats and whales. *Curr. Biol.* 2010; 20:R55–R56. [PubMed: 20129037]
128. Mustonen V, Lassig M. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl Acad. Sci. USA.* 2010; 107:4248–4253. [PubMed: 20145113] [A reformulation of the principles of population genetics analogous to the transition from classic to non-equilibrium thermodynamics. The concept of fitness is replaced by fitness flux, and fitness landscape becomes a time-dependent seascape.]
129. Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl Acad. Sci. USA.* 2007; 104(Suppl. 1):8597–8604. [PubMed: 17494740]

130. Lynch M. The origins of eukaryotic gene structure. *Mol. Biol. Evol.* 2006; 23:450–468. [PubMed: 16280547]
131. Irimia M, Penny D, Roy SW. Coevolution of genomic intron number and splice sites. *Trends Genet.* 2007; 23:321–325. [PubMed: 17442445] [A comparative analysis of splice sites showing that intron-poor organisms possess highly conserved splice sites that adhere to a strict consensus, whereas intron-rich genomes contain weak splice sites. A crucial corollary is that the evolution of alternative splicing is conditioned on relatively inefficient splice sites that are prevalent in organisms with weak selective pressure.]
132. Irimia M, Roy SW. Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet.* 2008; 4:e1000148. [PubMed: 18688272]
133. Irimia M, et al. Complex selection on 5' splice sites in intron-rich organisms. *Genome Res.* 2009; 19:2021–2027. [PubMed: 19745111]
134. Lynch M. Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.* 2006; 60:327–349. [PubMed: 16824010]
135. Wagner A. Robustness, evolvability, and neutrality. *FEBS Lett.* 2005; 579:1772–1778. [PubMed: 15763550]
136. Dobrindt U, et al. Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. *J. Bacteriol.* 2003; 185:1831–1840. [PubMed: 12618447]
137. Lozada-Chavez I, Janga SC, Collado-Vides J. Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.* 2006; 34:3434–3445. [PubMed: 16840530]
138. Kazakov AE, et al. Comparative genomics of regulation of fatty acid and branched-chain amino acid utilization in proteobacteria. *J. Bacteriol.* 2009; 191:52–64. [PubMed: 18820024]
139. Wagner A. Neutralism and selectionism: a network-based reconciliation. *Nature Rev. Genet.* 2008; 9:965–974. [PubMed: 18957969] [A conceptual perspective on (nearly) neutral networks that reconciles the neutralistic and adaptationist paradigms of evolution by showing how initially neutral mutations form the basis for subsequent adaptation.]
140. Masel J, Siegal ML. Robustness: mechanisms and consequences. *Trends Genet.* 2009; 25:395–403. [PubMed: 19717203]
141. Bergman A, Siegal ML. Evolutionary capacitance as a general feature of complex gene networks. *Nature.* 2003; 424:549–552. [PubMed: 12891357]
142. Levy SF, Siegal ML. Network hubs buffer environmental variation in *Saccharomyces cerevisiae*. *PLoS Biol.* 2008; 6:e264. [PubMed: 18986213] [An experimental demonstration of the unexpectedly large number of evolution capacitors among yeast genes, a finding that validates the theoretical predictions of reference 141.]
143. Wang Z, Zhang J. Abundant indispensable redundancies in cellular metabolic networks. *Genome Biol. Evol.* 2009; 2009:23–33. [PubMed: 20333174]
144. Koonin EV. Darwinian evolution in the light of genomics. *Nucleic Acids Res.* 2009; 37:1011–1034. [PubMed: 19213802]
145. Frank SA. The common patterns of nature. *J. Evol. Biol.* 2009; 22:1563–1585. [PubMed: 19538344]
146. Wilkins AS. Between 'design' and 'bricolage': genetic networks, levels of selection, and adaptive evolution. *Proc. Natl Acad. Sci. USA.* 2007; 104(Suppl. 1):8590–8596. [PubMed: 17494754]
147. Resch AM, et al. Widespread positive selection in synonymous sites of mammalian genes. *Mol. Biol. Evol.* 2007; 24:1821–1831. [PubMed: 17522087]
148. Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol. Biol. Evol.* 2010; 27:1226–1234. [PubMed: 20150340]
149. Ellegren H, Smith NG, Webster MT. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* 2003; 13:562–568. [PubMed: 14638315]
150. Charlesworth J, Eyre-Walker A. The McDonald–Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.* 2008; 25:1007–1015. [PubMed: 18195052]

151. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* 2009; 26:2097–2108. [PubMed: 19535738]
152. Hurst LD. The *Ka/Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet.* 2002; 18:486–487. [PubMed: 12175810]
153. van Nimwegen E. Scaling laws in the functional content of genomes. *Trends Genet.* 2003; 19:479–484. [PubMed: 12957540] [A key study that reveals distinct scaling laws for different functional classes of genes and their virtual universality across a broad range of taxa.]
154. Molina N, van Nimwegen E. Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends Genet.* 2009; 25:243–247. [PubMed: 19457568]
155. Maslov S, Krishna S, Pang TY, Sneppen K. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proc. Natl Acad. Sci. USA.* 2009; 106:9743–9748. [PubMed: 19482938] [A simple model of evolution of metabolic networks that explains the universal scaling laws for regulators and enzymes.]
156. Lipman DJ, Wilbur WJ. Modelling neutral and selective evolution of protein folding. *Proc. Biol. Sci.* 1991; 245:7–11. [PubMed: 1682931]
157. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA.* 2005; 102:14338–14343. [PubMed: 16176987]
158. Kramer EB, Farabaugh PJ. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA.* 2007; 13:87–96. [PubMed: 17095544]
159. Whitehead DJ, Wilke CO, Vernazobres D, Bornberg-Bauer E. The look-ahead effect of phenotypic mutations. *Biol. Direct.* 2008; 3:18. [PubMed: 18479505] [A modelling study that demonstrates the possibility of evolutionary capacitation through synergistic interactions between mutations and errors of transcription and translation (phenotypic mutations).]

Robustness

The ability to maintain a phenotype or function in the presence of internal or external perturbations.

Purifying selection

(Also known as negative or stabilizing selection.) Mode of natural selection that eliminates deleterious mutations and preserves the *status quo*; in protein-coding genes, it is manifested as $Ka/Ks \ll 1$.

Box 1 Measuring selection

Over 25 years ago, Motoo Kimura proposed that pseudogenes could be used as a neutral baseline for measuring selection¹. In general, this contention stands²⁴; furthermore, genomics revealed additional sources of (apparently) neutrally evolving sequences, such as introns and intergenic regions in animals^{147,148}. However, different parts of a genome differ in their mutation rate¹⁴⁹, so the neutral model can only be a reliable estimate of the strength of selection or constraints if it is derived from the same gene or genomic region in which selection is being measured. Several such measures have been developed^{17,150,151}.

The common gauge of selection pressure on protein-coding sequences follows from the redundancy and non-random structure of the genetic code, in which the same amino acid typically is encoded by codons that differ only in their third (less commonly first) positions. This measure, Ka/Ks (dN/dS), is the ratio of the per-site number or rate of non-synonymous substitutions to the number or rate of synonymous substitutions^{20,152}. The assumption underpinning the use of Ka/Ks as a measure of selection is that, unlike non-synonymous sites, synonymous sites evolve neutrally, allowing the use of synonymous sites as the baseline for measuring the constraints on protein evolution. As a crude approximation, this assumption holds: for most protein-coding genes from any organism, $Ka/Ks \ll 1$, which indicates that most proteins are subject to purifying selection. The figure shows the distributions of evolutionary rates for non-synonymous and synonymous sites of protein-coding genes in primates (**a**) and the Ka/Ks ratios for three diverse pairs of species (**b**)⁸⁸. The distribution of Ka is much broader than the distribution of Ks , indicating that the constraints affecting proteins are qualitatively different from and much more diverse than those affecting synonymous sites. For unconstrained, neutral evolution, $Ka = Ks$, as is the case for most pseudogenes. For a small subset of protein-coding genes, $Ka/Ks > 1$, which might indicate positive selection.

More accurate and powerful tests for detecting purifying and positive selection on different classes of sites are variations of the McDonald-Kreitman test. This test compares the patterns of substitutions for within-species variation (polymorphisms) with those for between-species divergence, under the assumption that the fraction of non-neutral polymorphisms is negligible¹⁷.

An independent approach for estimating the fraction of constrained sites in a genome is based on the deviations from the expected neutral distribution of insertions and deletions³⁰. It has to be kept in mind that population genetics itself is a field in flux, in large part owing to the advances of comparative genomics. The tests outlined here are based on simplified models of evolution, so more realistic and appropriately complex models may affect the estimates of selection strength in different classes of site¹⁵¹.

Non-synonymous substitutions

Nucleotide substitutions in protein-coding genes that lead to amino acid changes in the encoded protein.

Synonymous substitutions

Nucleotide substitutions in protein-coding genes that occur in synonymous positions of codons and accordingly do not lead to amino acid changes in the encoded protein.

Positive selection

(Also known as directional or Darwinian selection.) Mode of natural selection that increases the frequency of initially rare beneficial alleles in a population; in protein-coding genes, this often leads to $K_a > K_s$.

Ultraconserved elements

Sequences in animal genomes that have retained their identity throughout long evolutionary spans, such as the entire course of vertebrate evolution.

Evolutionary domains

Distinct units of gene/protein evolution that form combinations with varying degrees of evolutionary stability. evolutionary domains may or may not correspond to structural domains (that is, an evolutionary domain could encompass one or more structural domains).

Promiscuous domain

A protein domain that combines with diverse other domains in numerous proteins, providing malleable connections in interaction and regulatory networks and complexes.

Orthologues

Genes that evolved from a single ancestral gene in the last common ancestor of the compared genomes (in contrast to paralogues).

Selfish operon concept

A hypothesis according to which the presence of the same or similar operons in different prokaryotes is due more to the horizontal transfer of operons as distinct units than to selection for co-expression and co-regulation. When a transferred piece of DNA includes an entire operon consisting of genes encoding a complete pathway or functional system, the chances of fixation dramatically increase.

Minimal gene set for cellular life

The minimal set of genes that is sufficient to maintain a functional cell.

Non-orthologous gene displacement

The utilization of unrelated or distantly related (not orthologous) genes for the same function.

Box 2 The bureaucratic ceiling of genomic complexity

The bureaucratic ceiling of genomic complexity is a hypothesis that is proposed to explain the upper gene bound. According to this theory, the number of regulators scales much steeper than linearly with the total number of genes; at a certain point the regulatory network inflation would become unsustainable, curbing the expansion of the gene complement.

Different functional classes of genes scale differently with the total number of genes in a genome. Some variation notwithstanding, in bacteria and archaea (prokaryotes) there seem to be three fundamental exponents that characterize these dependences: 0, 1 and 2 (REFS 57,153). The figure shows the differential scaling of four broad classes of genes with the total number of genes in prokaryotic genomes. Genes for proteins involved in information processing (translation, transcription and replication; yellow) scale with a 0 exponent: the number of these genes reaches a plateau even in the smallest genomes and effectively does not depend on overall genomic complexity. Metabolic enzymes and transport proteins (green) scale roughly proportionally to the total number of genes, whereas regulators (blue) and signal transduction system components (red) scale quadratically.

The characteristic exponents of the three broad functional classes of genes show remarkably little variation across prokaryotic lineages, suggesting that the differential evolutionary dynamics of genes with different functions reflect fundamental ‘laws’ of evolution of cellular organization¹⁵⁴ — that is, distinct, strong constraints on the functional composition of genomes. Eukaryotic genes show similar but less pronounced patterns of power-law gene scaling, with the exponent for the regulatory genes being substantially greater than 1 (although less than 2)¹⁵³.

The deep underlying causes of the superlinear scaling of the regulators are not understood. A simple toolbox model of evolution of prokaryotic metabolic networks seems to be compatible with the quadratic scaling of regulators¹⁵⁵. Regardless of the exact underlying mechanisms, the superlinear scaling of the regulators clearly could determine the upper limit of the growth of the gene number. At some point (which is not easy to identify precisely), the cost of adding extra regulation (‘inflating bureaucracy’) will inevitably become unsustainable, curbing the growth of genetic complexity.

The data in the figure are taken from REF. 57.

Toolbox model of evolution

A model according to which enzymes for utilizing new metabolites, together with their dedicated regulators, are added (primarily by horizontal gene transfer) to a progressively versatile reaction network. Because of the growing complexity of the pre-existing network that provides enzymes for intermediate reactions, the ratio of regulators to regulated genes grows steadily.

Paralogous gene families

Gene families that evolved by duplication.

Box 3 Misfolding-driven protein evolution

In general, the rate of evolution of a protein is determined by the size of the (nearly) neutral sequence network^{139,156}. If the fitness of a particular sequence mostly depends on its robustness to misfolding and on expression level^{90,96}, the size of the nearly neutral network depends on the height and shape of the robustness peak occupied by this sequence and its neighbours in the sequence space. The figure shows a conceptual model of misfolding-driven protein evolution.

The cartoon schematically shows the effective robustness/fitness landscapes, which integrate the robustness of the native sequence and all of its mistranslated variants for two protein families (X and Y) at high and low expression levels. The high fitness/robustness area (green) reflects the size of the nearly neutral network in the sequence space.

Recent analysis of the relationships between protein abundances and evolution rates in orthologous proteins reveals a significant non-random association between intrinsic structural properties of proteins and their translation rate (as approximated by measured abundance)¹¹⁰. These results suggest that highly expressed proteins that were selected for robust folding occupy tall, steep peaks that have small areas of high fitness (that is, the proteins occupy a small nearly neutral network); as such, these proteins evolve slowly. By contrast, proteins with lower robustness occupy lower and wider peaks that have larger areas of high fitness at typical expression levels, and this larger nearly neutral network allows faster evolution (see figure).

The original hypothesis on misfolding-dominated evolution of protein-coding genes held that misfolding was largely induced by mistranslation of the coding sequence^{90,157}. Whether this is the case or whether stochastic misfolding of the native sequence is equally or even more common, mistranslation (somatic mutation), which is relatively frequent (10^{-4} - 10^{-5} per codon¹⁵⁸), is likely to be an important factor affecting protein evolution. Error-prone translation of a protein-coding gene produces a 'cloud' of neighbours in sequence space, resulting in smoothing of a potentially very rugged robustness landscape²¹. The fitness of an extremely robust sequence is decreased owing to the appearance of mistranslated variants; conversely, a mutation leading to a poorly folding protein could be less detrimental because of non-negligible production of robustly folding mistranslated proteins. This effect could open up evolutionary trajectories on the fitness landscape that are inaccessible under perfect fidelity of translation and might lead to the striking phenomenon of evolutionary anticipation¹⁵⁹.

Neutral sequence network

A network of sequences connected by effectively single-step mutation distances (although not necessarily by single replacements), and in which there is a negligible fitness difference between neighbours.

Evolutionary anticipation

(Also known as the look-ahead effect.) A scenario for the evolution of complex traits that require multiple mutations. In this scenario, the fixation of the final, beneficial mutation that leads to the emergence of the complex feature is enabled by a preceding random mutational walk over the neutral sequence network or by phenotypic mutations, such as mistranslation.

Experimental evolution

The evolution of organisms with precisely defined genetic backgrounds and known evolutionary histories under controlled laboratory conditions.

Epistasis

When non-allelic genes interact to produce a joint phenotype that differs from the one that would have been produced if the two genes had acted independently.

Pleiotropy

Describes the multiple functions or mutation consequences of a single gene.

Fitness landscape

A multidimensional surface defining the relationships between the fitness and the genotype spaces.

Fitness seascape

A generalization of the concept of a fitness landscape, in which the dependence of fitness on sequence evolves over time.

Effective population size

The size of an idealized panmictic population whose evolutionary behaviour is equivalent to that of the analysed population.

Pathogenicity islands

Large clusters of genes in bacterial genomes that are typically transferred horizontally and contain pathogenicity determinants.

Figure 1. Approximate distribution of evolutionary constraints across genomes with different architectures

The fractions of different classes of sequences that are subject to constraints of varying strength are shown as rough approximations of the values that are typical of the respective class of genomes. The data are from REFS 2,23,29,58, as discussed in the main text.

Figure 2. The universal distribution of evolutionary rates across orthologous gene sets

The evolutionary rates for five pairs of closely related organisms from different branches of life were calculated as nucleotide distances for the complete sets of orthologous genes⁸⁸. The shapes of the rate distributions are very similar from bacteria to humans. The relative evolution rate for each gene was obtained by dividing its evolution rate by the median rate for the respective pair of organisms. ‘Model’ refers to estimated transition rates in 134 mutationally connected networks for simulated, robustly folding 18-mer protein-like molecules⁹⁶. Original model rates were normalized by their median value and scaled to a standard deviation of 0.25 to match the width of the distributions derived from biological data.

Figure 3. Genomic and phenomic constraints on different levels of biological organization

The degree of constraint and plasticity experienced by genomic and phenomic properties across different levels of biological organization are shown. The arrows should be perceived as indicating the midpoints of the respective intervals, with at least the adjacent intervals overlapping. The relationships between some of the constraints on some classes of sites — for example, synonymous sites and disordered segments in proteins — are ‘educated guesses’ based on current data, and could change with further accumulation of the relevant data and advances in methods for quantifying selective pressures. The scales are rough approximations.

Table 1

Main evolutionary constraints and universal dependences across levels of biological organization

Level of organization	Relevant constraints	Universal dependences/patterns	Putative underlying process/model
Sequence of protein-coding genes	Protein robustness to misfolding depends on translation rate *	<ul style="list-style-type: none"> • Approximately log-normal distribution of evolutionary rates of protein-coding genes • Anticorrelation between sequence evolution rate and expression level (translation rate) • Positive correlation among Ka, Ks and 3' UTR evolution rate 	Protein folding
Sequence of genes for non-protein-coding RNAs	RNA secondary structure and binding sites within loops	Differential conservation of ases in stem and loop structures	RNA folding
Non-coding sequences	<ul style="list-style-type: none"> • Regulatory site structure • Chromatin structure 	<ul style="list-style-type: none"> • Weak constraints (on average) • Unconstrained evolution 	<ul style="list-style-type: none"> • Biased mutational process • Sequence elimination by purifying selection depends on Ne
Gene architecture	<ul style="list-style-type: none"> • Compactness • Versatility 	<ul style="list-style-type: none"> • Multidomain organization of proteins • Promiscuous domains 	<ul style="list-style-type: none"> • Streamlining by purifying selection dependent on Ne • Recombination leading to domain rearrangement
Local gene order	Co-regulation of functionally linked genes (primarily in prokaryotes)	Closely spaced genes in partially conserved operons	'Selfish' operon spread
Long-range gene order	<ul style="list-style-type: none"> • Large-scale chromatin organization • Interaction between expression and replication 	<ul style="list-style-type: none"> • Weak constraints • Unconstrained evolution 	Random chromosome rearrangement, including origin-centred inversion in prokaryotes
Gene number (lower bound)	Minimal set of functions for sustaining a functional cell	Roughly constant size of minimal gene sets obtained with different approaches (250–300 genes)	Genome reduction; non-orthologous gene displacement
Gene number (upper bound) and gene repertoire	<ul style="list-style-type: none"> • Ratio of regulators to regulated genes • Lifestyle-dependent functional demands 	Distinct scaling laws for different functional classes of genes	'Toolbox'-like growth of metabolic networks
Paralogous gene families	Dependence of gene gain and loss rates on family size under the condition of genomic equilibrium	Power law-like distribution of the sizes of paralogous gene families	Birth and death process of gene evolution
Interaction and regulatory networks	Contingency of network evolution on the pre-existing network structure	Power law-like distribution of node degrees	Network evolution by preferential attachment and/or by gene duplication

* The total number of molecules of a given protein made per unit time by translation of all cognate mRNA molecules in the cell. Ne , effective population size.